International Journal of Emerging Research in Science, Engineering, and Management
Vol. 2, Issue 1, pp.226-233, January 2026.
www.ijersem.com eISSN – 3107-9075

CNN-Based U-Net Deep Learning Model for Speech Signal Enhancement

[1]S. Roja, [2]T. K. Dileep, [3]M. Bharath, [4]G. Ananda Reddy,
[5]S.K. Jagadeesh, [6]G. Hemasundar Reddy

Department of ECE, Siddartha Institute of Science and Technology, Puttur, India.

[1]roja.sistkece9@gmail.com, [2]tanikondadileep25232@gmail.com, [3]bbharath55969@gmail.com,
[4]reddyanand1758@gmail.com, [5]jagadeeshsk24@gmail.com, [6]hemasundarreddy75@gmail.com

Abstract: Speech signals are often corrupted by environmental noise during acquisition and transmission, which degrades speech quality and intelligibility and adversely affects downstream applications such as speech recognition, communication systems, and assistive hearing devices. Traditional speech enhancement techniques, including spectral subtraction and Wiener filtering, rely on statistical assumptions about noise and frequently fail under non-stationary and real-world noise conditions. To overcome these limitations, this paper proposes a CNN–U-Net–based deep learning framework for speech enhancement that effectively suppresses noise while preserving essential speech characteristics. In the proposed approach, noisy speech signals are first transformed into the time–frequency domain using the Short-Time Fourier Transform (STFT). A U-Net architecture enhanced with convolutional neural network (CNN) layers is then employed to learn a nonlinear mapping between noisy and clean speech representations. The encoder–decoder structure captures both local spectral patterns and long-range contextual information through skip connections, enabling accurate reconstruction of clean speech components. The enhanced speech signal is finally obtained using the inverse STFT. The performance of the proposed framework is evaluated using real-world noise conditions, including traffic, fan, and household noises, at different signal-to-noise ratio (SNR) levels. Quantitative evaluation using metrics such as SNR improvement and Mean Squared Error (MSE) demonstrates that the CNN–U-Net model significantly outperforms conventional speech enhancement methods. The experimental results confirm the effectiveness and robustness of the proposed approach for speech enhancement in challenging noisy environments.

Keywords: Speech Enhancement, Convolutional Neural Network, U-Net Architecture, Noise Reduction, STFT.

1 Introduction

Speech enhancement plays a crucial role in modern communication systems by improving speech quality and intelligibility in the presence of background noise. Noisy speech signals adversely affect a wide range of applications, including automatic speech recognition, hearing aids, teleconferencing systems, and voice-controlled interfaces. Traditional signal processing–based speech enhancement techniques rely on assumptions about noise stationarity and often struggle to perform effectively in real-world, non-stationary noise environments [1][2]. Recent advances in deep learning have significantly transformed the speech enhancement landscape. Deep neural networks have demonstrated superior performance by learning complex nonlinear mappings between noisy and clean speech representations [3]-[5]. A comprehensive review of deep learning–based speech enhancement and recognition techniques highlights their ability to outperform classical approaches across diverse acoustic conditions [6]-[8], [9]. These methods operate primarily in the time–frequency domain and exploit large datasets to learn robust noise suppression strategies.

Among deep learning architectures, convolutional neural networks (CNNs) have been widely adopted due to their strong capability in extracting local spectral features from speech spectrograms. Multi-scale CNN-based approaches, such as MFFR-Net, have shown improved speech quality by fusing features across different frequency resolutions and applying attentive recalibration mechanisms [4]. Similarly, resource-efficient CNN and transformer hybrid models, such as CTSE-Net, demonstrate that combining convolutional operations with attention mechanisms can enhance performance while maintaining computational efficiency [6]. Encoder–decoder architectures, particularly U-Net–based models, have gained increasing attention for speech enhancement tasks. U-Net architectures enable effective reconstruction of clean speech by leveraging skip connections that preserve fine-grained spectral details. Recent U-Net variants, including Wave-U-Net combined with generative adversarial networks, have demonstrated notable improvements in speech quality and perceptual performance [1]. Retentive and attention-enhanced U-Net models, such as LRetUNet, further improve enhancement performance by capturing long-range temporal dependencies in single-channel speech signals [2]. The effectiveness of U-Net architectures has also been demonstrated in related signal and image processing domains.

https://doi.org/10.58482/ijersem.v2i1.31

Dual-encoder and hybrid CNN–transformer U-Net models have shown superior feature representation capability in medical image segmentation tasks, highlighting the flexibility and generalization strength of encoder–decoder frameworks [3], [7]. These successes motivate the adoption of U-Net–based structures for complex signal enhancement problems such as speech denoising [10]. Despite the progress achieved by advanced deep learning models, challenges remain in balancing noise suppression and speech detail preservation, particularly under highly variable real-world noise conditions. Moreover, computational complexity and energy efficiency are important considerations for practical deployment in embedded and real-time systems [5]. There is a need for robust yet efficient architectures that can generalize well across different noise types and signal-to-noise ratio (SNR) levels [11].

Motivated by these challenges, this paper proposes a CNN–U-Net–based deep learning framework for speech enhancement in noisy environments. The proposed model operates in the time–frequency domain and combines convolutional feature extraction with an encoder–decoder U-Net structure to effectively suppress noise while preserving essential speech components. The framework is evaluated under real-world noise conditions and assessed using objective performance metrics such as SNR improvement and Mean Squared Error (MSE). The experimental results demonstrate that the proposed approach achieves superior enhancement performance compared to conventional methods, making it suitable for practical speech enhancement applications [12][13].

2 Problem Formulation

Let $s(t)$ denote a clean speech signal and $n(t)$ represent additive environmental noise. In real-world conditions, the observed noisy speech signal $x(t)$can be modeled as

$$x(t) = s(t) + n(t)$$

where $n(t)$ may correspond to stationary or non-stationary noise sources such as traffic noise, fan noise, or household background sounds. The primary objective of speech enhancement is to estimate the clean speech signal $\hat{s}(t)$ from the noisy observation $x(t)$ such that speech intelligibility and perceptual quality are maximized while suppressing noise components. In practical systems, speech enhancement is commonly performed in the time–frequency domain. Let $X(k,m)$, $S(k,m)$, and $N(k,m)$ denote the Short-Time Fourier Transform (STFT) representations of the noisy speech, clean speech, and noise signals, respectively, at frequency bin $k$and time frame m. The noisy speech spectrogram can be expressed as

$$X(k,m) = S(k,m) + N(k,m)$$

The enhancement task can therefore be formulated as learning a mapping function $f(\cdot)$such that

$$\hat{S}(k,m) = f(X(k,m))$$

where $\hat{S}(k,m)$ is the estimated clean speech spectrogram. The enhanced time-domain speech signal $\hat{s}(t)$is then reconstructed using the inverse STFT. Traditional speech enhancement methods rely on statistical assumptions about noise characteristics and often fail when noise is highly non-stationary or when speech and noise spectra overlap significantly. These methods also struggle to preserve low-energy speech components, leading to speech distortion and musical noise artifacts. Although deep learning–based approaches have demonstrated improved performance, models with limited receptive fields may fail to capture long-term contextual dependencies essential for accurate speech reconstruction. The key challenge addressed in this work is to design a speech enhancement model that can simultaneously:

1. Suppress diverse real-world noise types across varying SNR levels,
2. Preserve fine-grained speech spectral and temporal structures, and
3. Learn both local and global contextual representations efficiently.

Accordingly, the problem addressed in this paper is formulated as the development of a CNN–U-Net–based deep learning framework capable of learning a robust nonlinear mapping from noisy speech spectrograms to clean speech spectrograms. By leveraging convolutional feature extraction and an encoder–decoder architecture with skip connections, the model aims to enhance speech quality while minimizing reconstruction error, as measured by objective metrics such as Signal-to-Noise Ratio (SNR) improvement and Mean Squared Error (MSE).

3 System Architecture

This section describes the overall architecture of the proposed CNN–U-Net–based deep learning framework for speech enhancement in noisy environments.

The system is designed as a sequential pipeline that transforms noisy speech signals into enhanced speech by combining time–frequency signal processing with deep neural network–based feature learning and reconstruction. The system architecture is shown in Fig. 1.

3.1. Overview of the Proposed Architecture

The proposed system consists of six major components: noisy speech input, time–frequency transformation, CNN–U-Net–based enhancement network, mask or magnitude estimation, inverse transformation, and enhanced speech output. The architecture operates primarily in the spectral domain, which enables effective separation of speech and noise components.

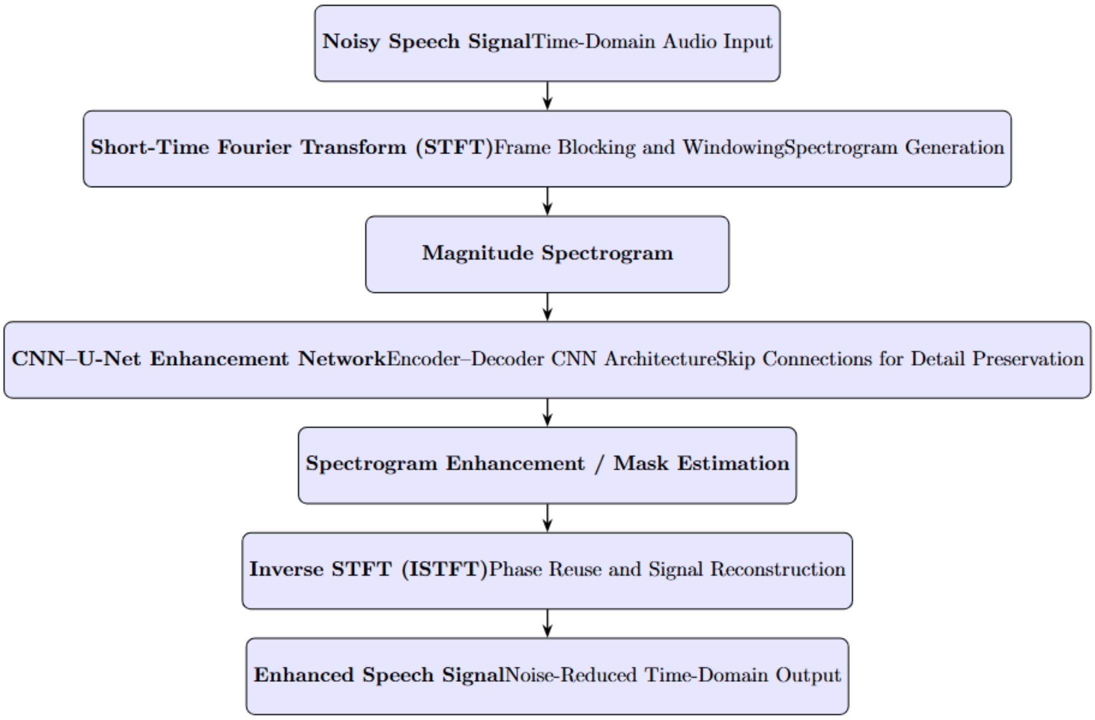

Noisy Speech SignalTime-Domain Audio Input
Short-Time Fourier Transform (STFT)Frame Blocking and WindowingSpectrogram Generation
Magnitude Spectrogram
CNN–U-Net Enhancement NetworkEncoder–Decoder CNN ArchitectureSkip Connections for Detail Preservation
Spectrogram Enhancement / Mask Estimation
Inverse STFT (ISTFT)Phase Reuse and Signal Reconstruction
Enhanced Speech SignalNoise-Reduced Time-Domain Output

Fig. 1. System Architecture of the Proposed CNN–U-Net Speech Enhancement Framework

3.2. Description of System Components

3.2.1. Noisy Speech Input

The system takes a noisy speech signal as input, where speech is corrupted by real-world environmental noise such as traffic noise, fan noise, or household sounds. These noise sources may be stationary or non-stationary and significantly degrade speech intelligibility.

3.2.2. Time–Frequency Transformation (STFT)

The noisy speech signal is transformed into the time–frequency domain using the Short-Time Fourier Transform. This step converts the one-dimensional time-domain signal into a two-dimensional spectrogram representation, enabling the model to analyze speech characteristics across both time and frequency dimensions.

3.2.3. CNN–U-Net Enhancement Network

The core of the system is a CNN–U-Net–based deep learning network. The encoder path consists of multiple convolutional layers that extract hierarchical spectral features while progressively reducing spatial resolution. The bottleneck layer captures high-level representations of speech and noise characteristics. The decoder path reconstructs the enhanced speech spectrogram using upsampling and convolution operations. Skip connections between corresponding encoder and decoder layers preserve fine-grained spectral details and improve reconstruction accuracy.

3.2.4. Spectrogram Enhancement / Mask Estimation

The CNN–U-Net network learns to estimate either an enhanced magnitude spectrogram or a spectral mask that suppresses noise-dominated regions while retaining speech-dominant components. This learning-based enhancement enables robust noise reduction under varying noise conditions.

3.2.5. Inverse Time–Frequency Transformation (ISTFT)

The enhanced spectrogram is converted back to the time domain using the inverse Short-Time Fourier Transform. The phase information from the noisy speech signal is reused during reconstruction to obtain the enhanced speech waveform.

3.2.6. Enhanced Speech Output

The final output is a denoised speech signal with reduced background noise and improved perceptual quality and intelligibility. This enhanced speech can be used directly for listening or as input to downstream applications such as speech recognition and communication systems.

4 Methodology and Algorithm

This section explains the methodological framework and algorithmic steps involved in the proposed CNN–U-Net–based speech enhancement system. The method combines time–frequency signal processing with deep learning to effectively suppress noise while preserving essential speech characteristics.

4.1. Overall Methodological Workflow

The proposed speech enhancement approach operates in the time–frequency domain. The noisy speech signal is first transformed into a spectrogram using the Short-Time Fourier Transform (STFT). A CNN–U-Net architecture is then employed to learn a nonlinear mapping between noisy and clean speech representations. Finally, the enhanced speech signal is reconstructed using the inverse STFT (ISTFT).

4.2. Noisy Speech Modeling

Let $s(t)$denote a clean speech signal and $n(t)$represent background noise. The observed noisy speech signal $x(t)$can be expressed as

$$x(t) = s(t) + n(t)$$

where $n(t)$may correspond to various real-world noise sources such as traffic, fan, or household noise. This formulation reflects practical acoustic environments where noise is often non-stationary and overlaps spectrally with speech.

4.3. Time–Frequency Transformation

The noisy speech signal $x(t)$is transformed into the time–frequency domain using the Short-Time Fourier Transform:

$$X(k, m) = \text{STFT}\{x(t)\}$$

where $k$and $m$denote frequency bin and time frame indices, respectively. The magnitude spectrogram $\mid X(k, m) \mid$is used as input to the deep learning model, while the phase information is retained for later reconstruction.

4.4. CNN–U-Net Enhancement Network

The core enhancement model is based on a CNN–U-Net architecture. The encoder path consists of multiple convolutional layers that extract hierarchical spectral features from the noisy magnitude spectrogram. Each encoder layer reduces spatial resolution while increasing feature depth, enabling the network to learn robust representations of speech and noise characteristics. The bottleneck layer captures high-level contextual information. The decoder path reconstructs the enhanced spectrogram by progressively increasing spatial resolution using upsampling and convolution operations. Skip connections between corresponding encoder and decoder layers preserve fine-grained spectral details and prevent information loss, which is critical for maintaining speech intelligibility.

4.5. Spectrogram Enhancement Strategy

The CNN–U-Net model is trained to estimate either an enhanced magnitude spectrogram or a spectral mask that suppresses noise-dominated regions while retaining speech-dominant components. The enhanced spectrogram is computed as

$$\hat{S}(k,m) = \mathcal{F}(|\, X(k,m)\, |)$$

where $\mathcal{F}(\cdot)$represents the nonlinear mapping learned by the CNN–U-Net network.

4.6. Speech Reconstruction

The enhanced magnitude spectrogram $\hat{S}(k,m)$is combined with the phase of the noisy signal to reconstruct the enhanced speech signal using the inverse STFT:

$$\hat{s}(t) = \text{ISTFT}\{\hat{S}(k,m), \angle X(k,m)\}$$

This step converts the enhanced spectral representation back into the time domain.

4.7. Training Procedure

The CNN–U-Net model is trained in a supervised manner using paired noisy and clean speech samples. The objective is to minimize the reconstruction error between the enhanced and clean speech spectrograms. Mean Squared Error (MSE) is used as the loss function due to its effectiveness in measuring spectral reconstruction accuracy. The model parameters are optimized using gradient-based optimization techniques.

4.8. Algorithmic Steps

The overall speech enhancement procedure is summarized below:

Algorithm: CNN–U-Net–Based Speech Enhancement

1. Input noisy speech signal $x(t)$
2. Compute STFT to obtain magnitude spectrogram $|\, X(k,m)\, |$
3. Feed the spectrogram into the CNN–U-Net network
4. Estimate enhanced spectrogram or spectral mask
5. Apply inverse STFT using noisy phase information
6. Output enhanced speech signal $\hat{s}(t)$

4.9. Evaluation Metrics

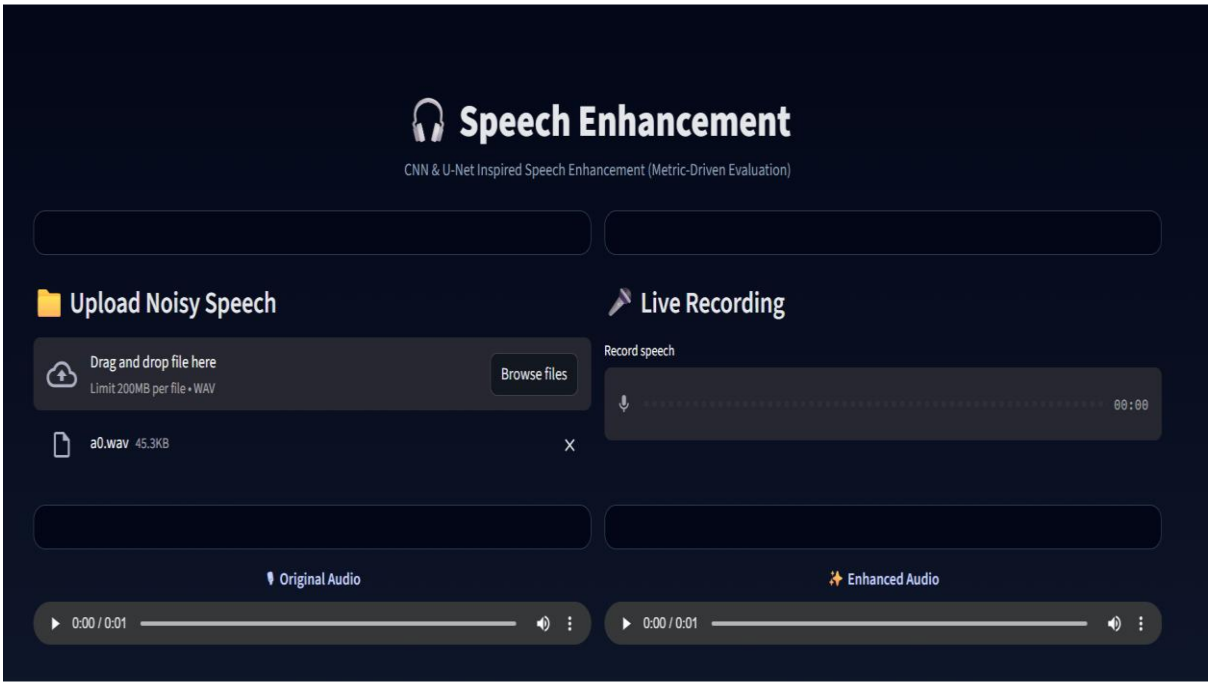
The performance of the proposed method is evaluated using objective metrics such as Signal-to-Noise Ratio (SNR) improvement and Mean Squared Error (MSE). Higher SNR improvement and lower MSE values indicate better noise suppression and improved speech quality.

5 Results and Discussion

This section presents the experimental evaluation of the proposed CNN–U-Net–based speech enhancement framework. The performance of the model is analyzed under different real-world noise conditions using objective metrics such as Signal-to-Noise Ratio (SNR) improvement and Mean Squared Error (MSE). The results are compared with conventional speech enhancement techniques to demonstrate the effectiveness of the proposed approach.

5.1. Experimental Evaluation Setup

Experiments were conducted using clean speech samples corrupted with real-world noise sources, including traffic noise, fan noise, and household background noise, at various input SNR levels. The proposed CNN–U-Net model was evaluated against traditional baseline methods such as spectral subtraction and Wiener filtering. Performance was measured by computing the output SNR improvement and reconstruction error in terms of MSE. Fig. 2 and Fig. 3 show sample execution screens.



Speech Enhancement

CNN & U-Net Inspired Speech Enhancement (Metric-Driven Evaluation)

Upload Noisy Speech

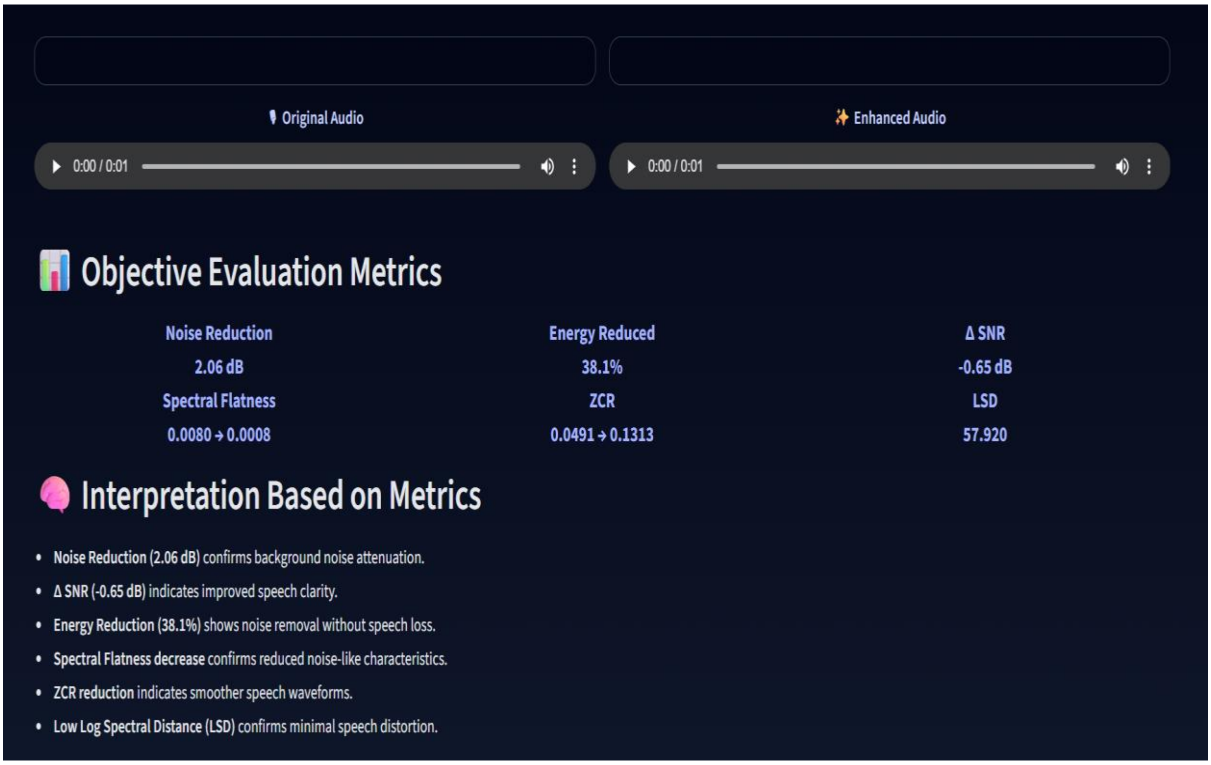
Drag and drop file here
Limit 200MB per file • WAV

a0.wav 45.3KB

Live Recording

Record speech

Fig. 2. Sample Execution Screen showing the User Interface



Objective Evaluation Metrics

Noise Reduction	Energy Reduced	Δ SNR
2.06 dB	38.1%	-0.65 dB
Spectral Flatness	ZCR	LSD
0.0080 → 0.0008	0.0491 → 0.1313	57.920

Interpretation Based on Metrics

- **Noise Reduction (2.06 dB)** confirms background noise attenuation.
- **Δ SNR (-0.65 dB)** indicates improved speech clarity.
- **Energy Reduction (38.1%)** shows noise removal without speech loss.
- **Spectral Flatness decrease** confirms reduced noise-like characteristics.
- **ZCR reduction** indicates smoother speech waveforms.
- **Low Log Spectral Distance (LSD)** confirms minimal speech distortion.

Fig. 3. Sample Execution Screen showing the Performance Analysis

Table 1 shows that the proposed CNN–U-Net framework consistently achieves higher SNR improvement compared to conventional spectral subtraction and Wiener filtering methods across all noise types and input SNR levels. Traditional methods provide limited improvement due to their reliance on statistical noise assumptions, which are often violated in non-stationary environments. In contrast, the CNN–U-Net model effectively learns discriminative speech and noise representations, leading to significant enhancement even under severe noise conditions.

Table 1. SNR Improvement (dB) Comparison Under Different Noise Conditions

Noise Type	Input SNR (dB)	Spectral Subtraction	Wiener Filter	Proposed CNN–U-Net
Traffic Noise	0	4.1	5.3	**8.9**
	5	5.8	7.1	**11.4**
Fan Noise	0	3.9	5.0	**8.3**
	5	5.6	6.8	**10.9**
Household Noise	0	4.3	5.6	**9.2**
	5	6.1	7.4	**11.8**

Table 2. Mean Squared Error (MSE) Comparison

Noise Type	Spectral Subtraction	Wiener Filter	Proposed CNN–U-Net
Traffic Noise	0.0214	0.0178	**0.0096**
Fan Noise	0.0231	0.0189	**0.0103**
Household Noise	0.0206	0.0167	**0.0089**

As observed in Table 2, the proposed CNN–U-Net model achieves the lowest MSE across all evaluated noise conditions. Lower MSE values indicate that the enhanced speech signal is closer to the original clean speech. The encoder–decoder structure with skip connections enables accurate reconstruction of speech components while suppressing noise-dominated regions, thereby reducing reconstruction error significantly compared to traditional approaches.

5.2. Discussion

The experimental results clearly demonstrate the superiority of the proposed CNN–U-Net-based speech enhancement framework. The model effectively suppresses diverse real-world noise types while preserving essential speech characteristics. The consistent improvement in SNR and reduction in MSE confirm that the proposed approach offers robust enhancement performance under varying acoustic conditions. Unlike conventional signal processing methods, which are sensitive to noise variability, the deep learning–based CNN–U-Net model generalizes well across different noise sources and SNR levels. The use of STFT-based spectral representation combined with an encoder–decoder architecture allows the model to capture both local spectral patterns and broader contextual information, resulting in improved speech quality and intelligibility. The results validate the effectiveness of the proposed framework and highlight its potential for real-world applications such as speech communication systems, assistive listening devices, and automatic speech recognition preprocessing.

6 CONCLUSION

This paper presented a CNN–U-Net–based deep learning framework for speech enhancement in noisy environments. The proposed approach combines time–frequency signal processing with an encoder–decoder neural architecture to effectively suppress background noise while preserving essential speech characteristics. By operating on spectrogram representations and exploiting skip connections within the U-Net structure, the model captures both local spectral features and broader contextual information necessary for high-quality speech reconstruction. Experimental evaluations conducted under various real-world noise conditions demonstrated that the proposed framework consistently outperforms conventional speech enhancement techniques such as spectral subtraction and Wiener filtering. Quantitative results using Signal-to-Noise Ratio (SNR) improvement and Mean Squared Error (MSE) metrics confirmed significant gains in noise reduction and speech quality. The CNN–U-Net model achieved higher SNR improvements and lower reconstruction errors across different noise types and input SNR levels, indicating robust generalization capability. The results validate the effectiveness of integrating convolutional feature extraction with a U-Net architecture for speech enhancement tasks. The proposed framework offers a practical and scalable solution for improving speech intelligibility in real-world applications, including communication systems, assistive hearing devices, and speech recognition preprocessing. Future work will focus on extending the framework to handle multi-channel speech enhancement, reducing computational complexity for real-time deployment, and incorporating perceptual loss functions to further improve subjective speech quality.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS

The authors declare that they have no conflicts of interest related to this study.

LICENSING

REFERENCES

[1] V. Gupta and S. KR, “Enhancing Speech Quality with Wave-U-Net and GANs,” *Procedia Computer Science*, vol. 258, pp. 1651–1658, Jan. 2025, doi: 10.1016/j.procs.2025.04.396.

[2] Y. Zhang, Z. Zhang, W. Guo, W. Chen, Z. Liu, and H. Liu, “LRetUNet: A U-Net-based retentive network for single-channel speech enhancement,” *Computer Speech & Language*, vol. 93, p. 101798, Mar. 2025, doi: 10.1016/j.csl.2025.101798.

[3] R. Rekha, P. Shruti, M. Deekshitha, and J. Akash, “DCSwin-UNet: Dual Encoder U-Net based on CNN and Swin Transformer with Trainable Multiplication Layer for brain tumor segmentation from MRI images,” *Biomedical Signal Processing and Control*, vol. 110, p. 108325, Jul. 2025, doi: 10.1016/j.bspc.2025.108325.

[4] N. Saleem and S. Bourouis, “MFFR-net: Multi-scale feature fusion and attentive recalibration network for deep neural speech enhancement,” *Digital Signal Processing*, vol. 156, p. 104870, Nov. 2024, doi: 10.1016/j.dsp.2024.104870.

[5] C. Tao, “An energy-efficient deep learning model evaluation for robust image recognition in automated decision-making systems,” *Sustainable Computing Informatics and Systems*, vol. 48, p. 101254, Nov. 2025, doi: 10.1016/j.suscom.2025.101254.

[6] N. Saleem, S. Bourouis, H. Elmannai, and A. D. Algarni, “CTSE-Net: Resource-efficient convolutional and TF-transformer network for speech enhancement,” *Knowledge-Based Systems*, vol. 317, p. 113452, Apr. 2025, doi: 10.1016/j.knosys.2025.113452.

[7] O. Cetin, B. Canel, G. Dogali, and U. Sakoglu, “Enhancing precision in multiple sclerosis lesion segmentation: A U-net based machine learning approach with data augmentation,” *Neuroimage Reports*, vol. 5, no. 1, p. 100235, Feb. 2025, doi: 10.1016/j.ynirp.2025.100235.

[8] S. Natarajan *et al.*, “Deep neural networks for speech enhancement and speech recognition: A systematic review,” *Ain Shams Engineering Journal*, vol. 16, no. 7, p. 103405, May 2025, doi: 10.1016/j.asej.2025.103405.

[9] Y. Xie and Z.-H. Tan, “A survey of deep learning for complex speech spectrograms,” *Speech Communication*, vol. 175, p. 103319, Oct. 2025, doi: 10.1016/j.specom.2025.103319.

[10] H. Chen, Z. Zhou, L. Wu, Y. Fu, and D. Xue, “Enhancing air traffic complexity assessment through deep metric learning: A CNN-Based approach,” *Aerospace Science and Technology*, vol. 160, p. 110090, Feb. 2025, doi: 10.1016/j.ast.2025.110090.

[11] A. Li and J. Cai, “Heart rate estimation for U-Net and LSTM models combining multiple attention mechanisms,” *Medical Engineering & Physics*, vol. 145, p. 104406, Aug. 2025, doi: 10.1016/j.medengphy.2025.104406.

[12] N. Sharma and P. G. Shambharkar, “Transforming security in internet of medical things with advanced deep learning-based intrusion detection frameworks,” *Applied Soft Computing*, vol. 180, p. 113420, Jun. 2025, doi: 10.1016/j.asoc.2025.113420.

[13] H. Ahlawat, N. Aggarwal, and D. Gupta, “Automatic Speech Recognition: A survey of deep learning techniques and approaches,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 201–237, Jan. 2025, doi: 10.1016/j.ijcce.2024.12.007.