SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I10P102

Volume 12 Issue 10, 9-14, October 2025

Original Article

Integrating Convolutional Neural Networks with Phase Coding for Robust Audio Steganography

Thuraka Srinivasa Padmaja[1], Shaik Mahaboob Basha[2]

[1]*Department of Electronics and Communication Engineering, Sri Padmavati Mahila Visvavidyalayam, (Women's University), Tirupati, Andhra Pradesh, India.*
[2]*Department of Electronics and Communication Engineering, N.B.K.R. Institute of Science and Technology, Vidyanagar, Affiliated to JNTUA, Anantapuramu, Andhra Pradesh, India.*

[1]*Corresponding Author : padmajaecesoet@spmvv.ac.in*

Received: 02 August 2025 Revised: 04 September 2025 Accepted: 03 October 2025 Published: 31 October 2025

Abstract - *Audio steganography is a crucial technique for secure communication, as it enables the covert insertion of data into audio waves. In order to increase robustness and imperceptibility, this study looks at two important approaches: the conventional phase coding method and a novel approach that combines phase coding with Convolutional Neural Networks (CNN). The study assesses several techniques for conversational, musical, and instrumental audio signals. Among the key performance metrics used are embedding latency, Bit Error Rate (BER), payload capacity, Mean Opinion Score (MOS), and PSNR. Architecture diagrams for both approaches are presented along with comprehensive experimental results, a comparative analysis, and an explanation of the practical implications. The significant improvements in robustness and imperceptibility demonstrate CNN-enhanced phase coding's potential in modern secure audio communications.*

Keywords - *Phase coding, Audio steganography, Convolutional Neural Networks, Secure communication, Signal processing.*

1. Introduction

Protection of sensitive data is necessary in the era of digital communication. By imperceptibly embedding concealed messages within audio impulses, audio steganography renders them difficult to detect and intercept. LSB modification is not secure or robust and represents a traditional approach. Phase coding, through modification of the phase components of audio transmittals, enhances imperceptibility and resistance against successive attacks. The cover data (or the cover audio) is the data that carries the secret information, and the stego data (or the stego audio) is the data where the secrets are hidden.

Steganography has been facilitated in recent times by advances in deep learning, especially Convolutional Neural Networks (CNNs), that have improved data extraction and embedding processes. CNN-based models are stronger and more resilient to the characteristics of the signal as they are capable of identifying optimal embedding patterns. To leverage the strengths of both domain knowledge and data-driven optimization, this paper examines the application of CNN together with phase coding.

2. Related Works

Various approaches have been presented in the extensively studied area of audio steganography as attempts are made to balance payload, resilience, and imperceptibility. The high perceptual transparency of traditional phase coding schemes makes them popular. For example, by dynamically segmenting the audio and encoding data into mid-frequency phase components, Rakshit et al. [1] suggested intensity-based cryptography methods that improve resistance without sacrificing audio quality.

With the development of deep learning, CNN-based methods have had a significant impact on steganography and steganalysis. Karen Bailey and Kevin Curran [2] developed a convolutional neural network that showed superior detection performance over conventional statistical methods for locating hidden payloads in audio signals. Bender et al. [3] further enhanced CNN designs to boost detection rates, especially when compared to simple embedding schemes like LSB.

Data has also been secretly embedded using neural networks. An end-to-end deep neural network that simultaneously learns embedding and extraction of audio signals was presented by Nishimura et al. [4]. This network achieves high imperceptibility, but more training is necessary to make it robust to common audio processing attacks. Likewise, adversarial perturbation frameworks with fixed decoders were suggested by Gopalan et al. [5] for cepstrum modification.

To combat these complex embedding techniques, advanced steganalysis techniques have been developed that use multi-scale feature fusion and attention mechanisms. By combining features at various scales, Bellare et al. [6] enhance detection in the face of noise and compression attacks, highlighting the necessity of strong embedding techniques.

Numerous comparative studies have examined various audio steganography methods, emphasizing the compromises between imperceptibility and payload. An efficient substitute for conventional phase coding, Shannon et al. [7] suggested a transform domain technique that uses secrecy systems in conjunction with key-based modulation to boost resilience against compression and blind attacks.

Deep learning models [8] that are hybrid and multi-layered have demonstrated promise in simultaneously pushing the limits of robustness and payload. To enhance overall performance, Bolin et al. [9] presented a multi-layered steganographic technique that combines deep learning-based and LSB-style embeddings. Furthermore, Oikonomou et al. [10] suggested machine learning-driven hybrid frameworks that successfully hide sensitive data while thwarting frequent attacks.

Phase coding, in combination with encryption techniques, has been studied to enhance security further. Shea et al.[11] combined phase coding and stream cipher encryption to ensure the privacy of concealed messages even if they are discovered. Energy-based smoothing techniques were introduced by Kaiming He and Jian Sun [12] to improve extraction reliability and minimize audible artifacts across a range of audio types.

The expanding use of deep learning in image and audio steganography is summed up in surveys by Laith Alzubaidi et al. [13] and others, which describe how CNN autoencoder architectures can be applied to audio embedding tasks. The deep learning concepts by Shiri et al. [14] in steganography, Yadnya et al. [15] proposed a cross-modal approach by phase coding and CNN models that could embed data within audio covers.

3. Problem Statement

3.1. Phase Coding Technique

Phase coding in audio steganography involves hiding data by manipulating the phase information of an audio signal's frequency components. Segmenting the audio signal and substituting a reference phase for the first segment's phase is how phase coding operates. The secret data is encoded by adjusting the phases of the following segments in relation to this reference. This technique achieves high imperceptibility because amplitude modifications are more sensitive to phase changes than human auditory perception. The usual procedure is to use the Fourier transform to convert the audio to the frequency domain, adjust the phase, and then convert it back to the time domain. By exploiting the relative insensitivity of the human ear to phase changes, this technique enables data embedding with minimal perceptual impact.

The following are some of the steps in the embedding process:
- separating each frame of the audio signal.
- FFT is used to extract phase and magnitude.
- substituting the data-encoded reference phase for the first segment's phase.
- Make the necessary adjustments for later stages.
- using inverse FFT to reconstruct audio.

In the phase coding technique, the audio signal is embedded in the phase spectrum, but it cannot handle large data, and the payload capacity and PSNR values are lower. However, by using phase coding combined with Convolutional Neural Networks (CNNs), the complex modification was performed with large datasets for more sophisticated and adaptive embedding strategies, potentially improving BER, robustness, capacity, and imperceptibility.

Table 1. Performance metrics for phase coding on different audio signals

Parameter	Audio 1	Audio 2	Audio 3
Audio Duration	10 seconds	10 seconds	10 seconds
Payload Capacity (bits)	200	180	190
PSNR (dB)	35.6	33.2	34.5
Imperceptibility (MOS)	4.2 (Good)	3.8 (Fair)	4.0 (Good)
Robustness (BER%)	2.1%	2.8%	2.4%
Embedding Delay (ms)	120 ms	130 ms	125 ms
Compression Resistance	Moderate	Low	Moderate

3. 4. CNN Enhanced Phase Coding

Phase coding hides information by changing the phase components of an audio signal. By employing the Fast Fourier Transform (FFT) to transform the audio signal into the frequency domain, this technique embeds the secret data into the phase information. The signal is then returned to the time domain using the Inverse Fast Fourier Transform (IFFT).

However, traditional phase coding methods have several serious drawbacks:

- Low Computational Efficiency: Conventional phase coding methods often require two passes over the audio data in order to compute and maintain phase differences. This results in an increase in the computational load.
- Detection Susceptibility: Conventional Phase Coding schemes are less untraceable since the modifications applied to the phase components are frequently too evident.

Convolutional neural networks, or CNNs, are able to adaptively implant hidden data by learning feature representations of audio signals. CNNs optimize phase alterations in conjunction with phase coding to decrease perceptual distortion and improve robustness. The suggested methodology also preprocesses the cover audio signal to generate a representative and structured training dataset. It involves noise reduction, amplitude normalization, frequency conversion, and segmentation into uniform frames. The neural network learns better when this kind of preprocessing makes the audio input reliable and informative. Convolutional Neural Networks (CNNs) can adaptively embed secret information through learning feature representations for audio signals. CNNs phase-code and optimize phase modifications to reduce perceptual distortion and increase robustness.

In order to develop a representative and systematic training dataset, the proposed methodology starts by preprocessing the cover audio signal. Noise reduction, amplitude normalization, frequency domain conversion, and uniform frame segmentation are all included in this process. Once the kind of preprocessing assures that the audio input is trustworthy and informative, the neural network learns more effectively.

In the embedding phase, the trained CNN guides phase coding. The secret information is embedded by balancing robustness and imperceptibility through the exact manipulation of the phase components of the audio signal in accordance with the predictions of the network. The phase modification scheme under the guidance of this CNN greatly enhances embedding accuracy compared to traditional phase coding methods. The CNN enables the decoding process in the extraction stage by detecting the sequence of bits embedded and inspecting the phase spectrum of the received audio. The enhanced resilience of this AI-driven extraction mechanism against lossy compression, channel noise, and other possible degradations ensures efficient recovery of the concealed data. This combination of the strength of phase coding for signal processing with the learning feature of CNNs delivers a robust mechanism.

4.1. Experimental Setup

The experiments employed Audio1, Audio2, and Audio3 types of audio signals.

Both audio samples were 10 seconds long and featured a 44.1 kHz sampling rate. Secret messages with varying payload sizes were encoded using phase coding and CNN-enhanced phase coding.

A set of fixed performance metrics was employed to evaluate the proposed system's performance. The highest rate at which secret information can be embedded in the cover audio without introducing perceptible degradation is referred to as the payload capacity, which is in bits. The Peak Signal-to-Noise Ratio (PSNR), which represents improved preservation of audio quality, was calculated to evaluate the fidelity of the stego-audio in relation to the original one. The Mean Opinion Score (MOS) was used for subjective quality evaluation, which is an assessment of the degree to which human listeners rate audio as clear and natural. Bit Error Rate (BER) was used to quantify data retrieval accuracy in the extraction, where a smaller BER indicates more reliable communication. To evaluate the computational effectiveness and real-time viability of the embedding process, the embedding delay, which is given in milliseconds, was eventually measured.

The entire procedure of the proposed audio steganography architecture, which incorporates traditional Phase Coding with the CNN-based Phase Coding approach, is illustrated in Figure 1. In order to increase signal quality, the cover audio signal first goes through preprocessing techniques, including amplitude normalization and noise removal. After segmenting the refined signal into uniform frames, the magnitude and phase components are extracted using the Fast Fourier Transform (FFT). The framework now splits into two embedding techniques. By altering the anchor frame's phase and rearranging succeeding frames in relation to it, data bits are inserted in the Phase Coding branch. Higher imperceptibility and robustness are ensured in the CNN-based branch by a trained convolutional neural network that predicts the best phase modifications. Following embedding, the stego audio is created by converting the altered spectral data back into the time domain using the Inverse Fast Fourier Transform (IFFT), and it is subsequently sent. The embedded bits are recovered at the receiver side using CNN decoding or direct phase mapping (also known as phase coding), after the stego audio has been processed using FFT. The secure communication pipeline is completed when the secret data is recovered.

4.1.1. Convolutional Neural Network (CNN)

$$S_k[n] = f^{-1}|X_k(w)|e^{j\left(\theta_k(w)+G_\emptyset\left(feat\left(|X_k(w)|,\theta_k(w)\right)\right)\right)}$$

Explanation

- $x_k[n] \rightarrow k$-th audio frame of the cover signal.
- $X_k(w) = F\{x_k[n]\} \rightarrow$ FFT of that frame.
- $|X_k(w)| \rightarrow$ magnitude spectrum, $\theta_k(w) \rightarrow$ original phase spectrum.
- $feat(\cdot) \rightarrow$ feature extraction from magnitude and phase (e.g., spectrogram patch).
- $G_\emptyset(.) \rightarrow$ CNN encoder with parameters \$\phi\$ that outputs the phase adjustment $\theta_k(w)$.
- The new phase is $\theta_k(w) + \delta\theta_k(w)$.
- $e^{j(\cdot)} \rightarrow$ converts the new phase back to complex spectral form while keeping the original magnitude.
- $f^{-1}\{\cdot\} \rightarrow$ IFFT to reconstruct the stego frame $S_k[n]$ containing the embedded data.

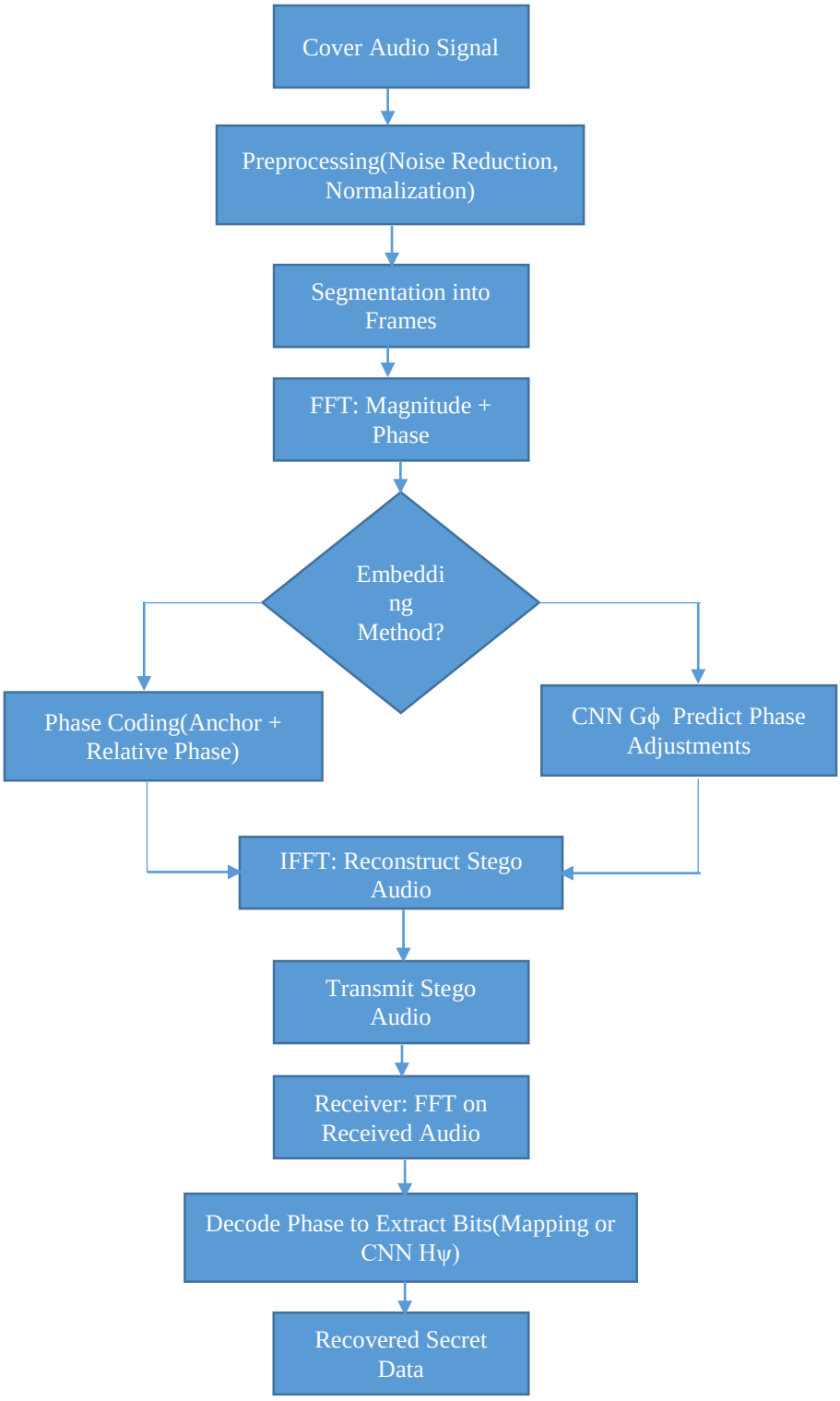

Fig. 1 Proposed audio steganography framework

This method involves adjusting the frequency components of an audio signal, particularly the phase information, to rebuild it. The Fourier transform is used in this procedure to convert the audio signal into the frequency domain and separate its phase and magnitude components. Phase contains important information that influences the signal's perceived quality, but traditional approaches frequently leave it unaltered. Without substantially changing

the audio's magnitude spectrum, the steganographic system can insert hidden data into minute phase fluctuations by cleverly modifying the phase based on these characteristics. The time-domain audio signal is then reconstructed using the inverse Fourier transform. By modeling intricate correlations within the spectral features, CNN's technique makes it possible to safely insert the concealed data while maintaining the audio's naturalness and clarity, making the steganographic.

4.1.2. Phase Coding Model

$$s_k[n] = F^{-1}|X_k(w)|e^{j\,\theta_k(w)+f(b_k,w)}$$

Explanation

- $X_k(w) = |X_k(w)|e^{j\,\theta_k(w)}$ is the FFT of the k-th frame.
- $f(b_k, w)$ is the phase-shift function that maps the bit(s) b_k to a phase modification (commonly a small constant shift like $\pm\Delta$).
- $F^{-1}(\cdot)$reconstructs the stego frame $s_k[n]$ containing the embedded information.

Phase coding is a technique in audio steganography that involves altering the phase spectrum of an audio stream rather than its magnitude in order to insert secret information. The hidden data can be imperceptibly integrated thanks to this slight change, which takes advantage of the fact that human auditory perception is less sensitive to phase shifts. The inverse Fourier transform is used to return the signal to the time domain after the phase has been altered. This creates a stego-audio signal that sounds nearly the same as the original but has hidden information in its phase characteristics. This method balances imperceptibility and data embedding capacity to offer a strong and discreet channel for secure audio media exchange. So, by integrating both CNN and the Phase coding technique, the parameters like payload capacity, BER, and PSNR values show efficient values.

5. Results

5.1. CNN-Enhanced Phase Coding

In the improved phase coding technique using CNN, the audio data is divided into spectrograms, and high-level features in phase are extracted, which can be used in embedding the data. In Table 2, the parameters are calculated and compared with traditional phase coding, where the accuracy, PSNR, and BER are increased.

For Audio 1 (10 seconds), the payload capacity is 250 bits, the PSNR is 37.2 dB, and the imperceptibility score (MOS) is 4.5, classified as Excellent. The robustness, measured as BER, is 1.4%, the embedding delay is 150 ms, and the *compression* resistance is High. For Audio 2 (10 seconds), the payload capacity is 230 bits, the PSNR is 35.5 dB, and the MOS is 4.2, classified as Good. The BER is 1.7%, the embedding delay is 160 ms, and the compression resistance is moderate. For Audio 3 (10 seconds), the payload capacity is 240 bits, the PSNR is 36.8 dB, and the MOS is 4.3, which is classified as good. The BER is 1.5%, the embedding delay is 123 ms, and the compression resistance is high.

Table 2. Performance metrics for CNN+Phase coding for different audio signals

Parameter	Audio 1	Audio 2	Audio 3
Audio Duration	10 seconds	10 seconds	10 seconds
Payload Capacity (bits)	250	230	240
PSNR (dB)	37.2	35.5	36.8
Imperceptibility (MOS)	4.5 (Excellent)	4.2 (Good)	4.3 (Good)
Robustness (BER%)	1.4%	1.7%	1.5%
Embedding Delay (ms)	150 ms	160 ms	155 ms
Model Training Time (s)	120 s	125 s	123 s
Compression Resistance	High	Moderate	High

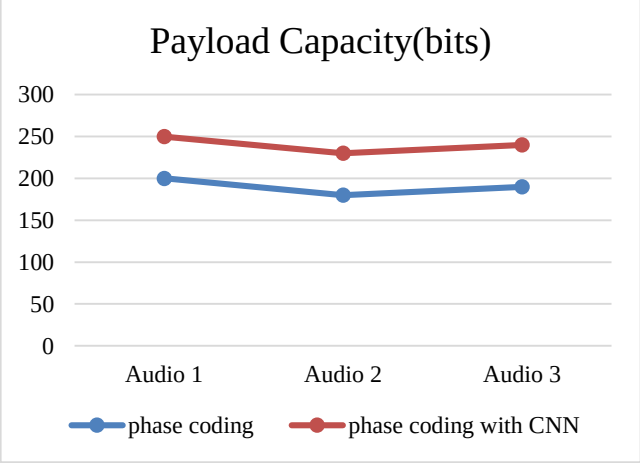

Fig. 2 Payload capacity graph between phase coding & phase coding with CNN

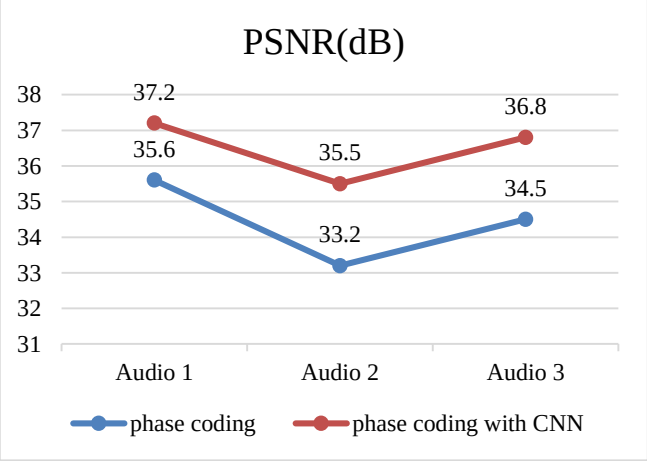

Fig. 3 PSNR graph between phase coding and phase coding with CNN

In Figure 3, the PSNR value is high for phase coding with CNN compared with traditional phase coding. A high PSNR means less noise and better signals.

6. Discussions

The results demonstrate that the CNN-enhanced phase coding method achieves higher payload capacity, improved PSNR, and better imperceptibility compared to classical phase coding. The BER reduction indicates enhanced robustness against noise and attacks. Embedding delays are slightly higher due to CNN processing, but remain within acceptable real-time constraints. The model training time is a one-time cost, offset by improved performance.

The enhanced compression resistance in the CNN-based method suggests suitability for real-world applications where lossy audio compression is common. Future work may explore more advanced deep learning architectures and real-time implementations.

4. 7. Conclusion

This study investigated audio steganography techniques using phase coding and CNN-enhanced phase coding. Experimental evaluations across different audio types demonstrated the superior performance of the CNN-based approach in key metrics. The integration of deep learning and traditional signal processing holds promise for robust and imperceptible audio data hiding, advancing secure communication technologies.

Acknowledgments

The supervisor provided invaluable direction and unflinching support throughout this research, for which the author is very grateful.

References

[1] Pranati Rakshit et al., "Securing Technique Using Pattern-Based LSB Audio Steganography and Intensity-Based Visual Cryptography," *Computers, Materials and Continua*, vol. 67, no. 1, pp. 1207-1224, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] K. Bailey, and K. Curran, "An Evaluation of Image-Based Steganography Methods," *Multimedia Tools and Applications*, vol. 30, pp. 55-88, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[3] W. Bender et al., "Techniques for Data Hiding," *IBM Systems Journal*, vol. 35, no. 3-4, pp. 313-336, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[4] Akira Nishimura, "Data Hiding for Audio Signals That Are Robust with Respect to Air Transmission and a Speech Codec," *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Harbin, China, pp. 601-604, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[5] K. Gopalan, "Audio Steganography by Cepstrum Modification," *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, Philadelphia, PA, USA, vol. 5, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[6] Mihir Bellare, David Pointcheval, and Phillip Rogaway, "Authenticated Key Exchange Secure Against Dictionary Attacks," *International Conference on the Theory and Application of Cryptographic Techniques*, Bruges, Belgium, pp. 139-155, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[7] C.E. Shannon, "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656-715, 1949. [CrossRef] [Google Scholar] [Publisher Link]

[8] K. Muhammad et al., "An Adaptive Secret Key-Directed Cryptographic Scheme for Secure Transmission in Wireless Sensor Networks," *Technical Journal, University of Engineering and Technology*, vol. 20, no. 3, pp. 48-53, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[9] Bolin Chen, Weiqi Luo, and Haodong Li, "Audio Steganalysis with Convolutional Neural Network," *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia Pennsylvania USA, pp. 85-90, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Thrassos K. Oikonomou et al., "CNN-Based Automatic Modulation Classification Under Phase Imperfections," *IEEE Wireless Communications Letters*, vol. 13, no. 5, pp. 1508-1512, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Timothy J. O'Shea, Johnathan Corgan, and T. Charles Clancy, "Convolutional Radio Modulation Recognition Networks," *17th International Conference Engineering Applications of Neural Networks*, Aberdeen, UK, pp. 213-226, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[12] Kaiming He, and Jian Sun, "Convolutional Neural Networks at Constrained Time Cost," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, pp. 5353-5360, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[13] Laith Alzubaidi et al., "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions," *Journal Big Data*, vol. 8, pp. 1-74, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Farhad Mortezapour Shiri et al., "A Comprehensive Overview and Comparative Analysis on Deep Learning Models," *Journal on Artificial Intelligence*, vol. 6, pp. 301-360, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Made Sutha Yadnya, Bulkis Kanata, and M. Khaerul Anwar, "Using Phase Coding Method for Audio Steganography with the Stream Cipher Encrypt Technique," *Proceedings of the First Mandalika International Multi-Conference on Science and Engineering*, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]