Traitement du Signal
Vol. 42, No. 5, October, 2025, pp. 2853-2864
Journal homepage: http://iieta.org/journals/ts

Integrating Hybrid Encryption and Deep Learning for Robust Audio Steganography and Audio Genre Classification

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

[1] Department of ECE, Sri Padmavati Mahila Visvavidyalayam, Tirupati 517502, India
[2] Department of ECE, N.B.K.R Institute of Science & Technology, Vidyanagar 524413, India

Corresponding Author Email: padmajaecesoet@spmvv.ac.in

https://doi.org/10.18280/ts.420535

Received: 10 June 2025
Revised: 25 August 2025
Accepted: 18 September 2025
Available online: 31 October 2025

Keywords:
audio classification, hybrid encryption, deep learning, steganography, mel-frequency cepstral coefficients (MFCCS)

ABSTRACT

The work offers a new method for securely embedding data and audio classification through hybrid encryption in addition to better deep learning algorithms. The data was initially pre-processed to extract the Mel-frequency cepstral coefficients from significant audio features among 1,000 audio files across various genres in the GTZAN dataset. This hybrid encryption approach utilizes the complementary strengths of AES and RSA algorithms to safely embed important information inside audio recordings using LSB steganography. Dual-layer encryption will preserve the audio fidelity while enhancing data security. Moreover, we will provide an adaptive audio steganography system, which will maximize the embedding process with minimum distortion of perception, by applying Spread Spectrum and QIM techniques. Such classified, encrypted and steganographic audio data falls under deep learning models, which include CNNs, RNNs, GRUs, LSTM networks, etc. The results show that although GRU and LSTM models offer a mix of accuracy and performance across genres, CNNs achieve excellent accuracy but struggle with precision in some genres. The Convolutional Neural Network (CNN) achieved a robust accuracy of 95.02%, but it faced challenges in precision, particularly in Country (57.14%) and Rock (56.96%) genres. Conversely, the Recurrent Neural Network (RNN), with a lower accuracy of 82.24%, demonstrated improved recall and F1-scores, especially in Blues (82.28%) and Classical (95.44%) genres. Both Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) models performed comparably, with accuracies of 94.44% and 94.86%, excelling in precision and recall for Classical and Pop. Despite the RNN's weaker metrics, the hybrid model approach significantly enhances the classification of encrypted and steganographic audio data, contributing valuable insights to audio steganography and classification fields. The novelty of this work lies in its dual-layer hybrid encryption scheme (RSA+AES), which ensures both fast data processing and enhanced resistance against cryptographic attacks compared to single encryption methods. Furthermore, the integration of multi-level steganography (QIM, LSB, and Spread Spectrum) significantly improves imperceptibility and robustness against detection, while maintaining audio fidelity by using Deep learning techniques.

1. INTRODUCTION

The system based on hybrid encryption securely embeds sensitive data using LSB steganography into audio recordings to provide the integration of effective data embedding together with excellent encryption security which guarantees that private communications may be concealed inside audio files without being audibly distorted. The method strikes a balance between speed of symmetric encryption and security offered by public-private key encryption through the employment of two powerful algorithms for encryption, namely RSA and AES. Due to the higher processing speed, AES-Advanced Encryption Standard is used to encrypt actual data while RSA, which is a well-known public-key cryptosystem, handles the safe exchange of encryption keys. This method ensures the robust safety of the secured data because it will hide them when encrypting with a double layer approach. It has been implemented on the popular GTZAN dataset with 1,000 audio files, partitioned into ten different musical genres. In this case, any audio file can act as a data carrier, while the LSB method is employed for hiding encrypted data. LSB steganography approach changes the least significant bit of each audio sample in order to encode the encrypted message without any human hear. Thus, the secret message is safely embedded by keeping the original audio intact. Some attributes are extracted to support the classification and analysis of the audio after the encoded data is introduced into the audio files. Since they provide rich representations of the audio signal, key properties such as chroma, mel-spectrogram and Mel Frequency Cepstral Coefficients (MFCC) are used. These attributes are very useful in pattern recognition and audio classification applications. The benefits of Convolutional Neural Networks (CNN),

Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM) networks are also used to create a blend of deep learning models for further improving the system. To be specific, each architecture is targeted at grasping different parts of audio data: While RNN, GRU, and LSTM models are aptly suited to tackle the temporal and sequential character of the audio, CNNs are adept at finding out spatial information in the spectrogram of audio. These models allow the system to classify the audio files efficiently, even such ones that contain secret data or encrypted data, since they capture not just long-range relationships but local patterns as well. Moreover, AES and RSA secure key management methods are used by the system to protect encryption keys when they are sent. For enhanced security, this system also includes the AES key-the one needed to unlock the hidden message-into the audio file by applying LSB steganography. Such a system provides secrecy to the key as well as to the data to be hidden from unauthorized access with the help of this hybrid encryption technique. Unlike conventional methods that rely on either single encryption or simple embedding strategies, the proposed approach introduces dual-layer hybrid encryption (RSA for key exchange and AES for bulk data encryption) combined with multi-technique steganography (LSB, QIM, and Spread Spectrum). This unique combination enhances security, imperceptibility, and robustness while preserving audio quality. Moreover, the integration of deep learning models such as CNN, RNN, GRU, and LSTM enables superior classification accuracy even when processing encrypted and steganographic audio, establishing a clear advancement over prior works.

2. LITERATURE REVIEW

In steganalysis, modern steganographic methods outperform more traditional machine learning approaches like SVM and ECs. To this end, Ntivuguruzwa With a focus on digital picture steganography, De La Croix et al. [1] examine cutting-edge advancements in deep learning-based steganalysis. The growth of these approaches has been analyzed throughout the study, with special emphasis on their superior speed, precision, and efficiency. The outcomes show how deep learning techniques improve steganalysis dramatically, and they provide reliable detection for modern steganographic systems. While secure communication is important for vehicular ad hoc networks, most traditional approaches often fail in this regard. Ansari [2] examine in their study whether the use of picture steganography could improve data security and secrecy in VANETs. For this end, the author presents several steganography techniques including vector embedding, spatial domain, and transform domain and computes the efficiency of such techniques when the losses of visual quality are compensated by ensuring safe data transmission. The results are shown below, and indicate that, even in challenging network environments, picture steganography represents a workable solution for secure communication on VANETs. Patient data security is increasingly significant due to the wide diffusion of telemedicine services and healthcare data exchange. Ghosh et al. [3] conduct research into how efficient the steganography techniques used are for the protection of medical information in e-health networks. The work will discuss many techniques like LSB, PVD, transform domains and the trade-off between security and quality parameters, such as PSNR and SSIM. The study concludes with a premise that the best and safest method for data transmission in telemedicine is a combination of diagonal queue-based steganography with chaotic techniques and Rabin encryption. One of the very important problems to solve in steganography is the problem of hidden embedding in images. Martín et al. [4] investigate whether GANs enhance spatial domain steganalysis techniques when combined with hidden information, showing few changes. With these results, a GAN is shown to be able to avoid the optimal Deep Learning steganalysis attacks when it is trained to modify an image in order to hide a message by use of the LSB algorithm. The results here describe how much more effective and secure this is when it comes to steganography. Upon hiding of the secret information by steganography within texts, its statistical distribution may be changed significantly as its quality decreases and makes it difficult to hide. To address this problem, Sun et al. [5] proposed a graph-to-text generation - controlled topic hiding method. The method uses graph route coding instead of conditional probability while generating knowledge network steganographic writings by configurable subjects. Experimental results compared with the original scheme indicate that this approach greatly improves the hiding of steganographic messages and text quality. Because linguistic steganography is generation-based, this procedure frequently causes the statistical distribution of the original text to be disturbed and consequently lowers the quality of the text and reduces hiding ability. Rustad et al. [6] propose a technique known as graph-to-text generation-informed topic-controlled steganography to overcome the problem. In the absence of conditional probability, the method derives steganographic texts from the knowledge graph such that the topics are controlled and carry secret information as the graph route is coded. Experimental results show that this method enhances text quality and offers the best cover capacity when compared to traditional ones, and it suggests being an effective steganographic tool. One of the main issues is guaranteeing the secrecy of the secret information but still safe over the Internet. To overcome the difficulty involved, Durafe and Patidar [7] have presented here a dependable and blind color picture steganography technique utilizing fractal cover images, SVD, IWT, and DWT. For protection against processing assaults and to preserve picture quality, it very smartly conceals a large amount of hidden image behind a tiny fractal cover image. Yazdanpanah et al. [8] discuss the inefficiency of the steganalysis algorithms currently in use for speech audio files. This paper describes the feature Percent of Equal Adjacent Samples (PEAS) for speech steganalysis with a Gaussian membership function-based classifier to capture 50% of embedded stego occurrences. Li et al. [9] have developed a unique steganalysis technique that can successfully identify MV-based steganography in HEVC. The method entails determining Motion Vector Prediction's (MVP) local optimality and examining how message embedding compromises it. The best rate of MVP is shown by the approach as a steganalysis characteristic. According to experimental results on two datasets, this method provides low-complexity, practical detection without the need for model training, o0utperforming four state-of-the-art strategies. Protecting the ownership of neural networks requires model watermarking, yet model pruning attacks pose a threat. To tackle this problem, Li et al. [10] provide a watermarking approach that prevents pruning for the hidden image steganography auto-encoder. The method includes employing

three traditional pruning algorithms to choose the proper model weights and using a DCT-based technique to implant a watermark. The fourth and fifth decimal positions of these weights include the watermark. High robustness is demonstrated by the experimental findings, which also exhibit gains in watermarking capacity and watermark extraction accuracy above 0.9993, even after 40% pruning. In response, Gao [11] provide a convolutional neural network (CNN)-based visual identification technique that combines deep learning with image processing theories. The study evaluates the method's advantages and disadvantages and makes recommendations for improvements. Results from experiments show that the CNN-based method provides excellent resilience and accuracy, which substantially improves automated visual recognition performance. To solve these problems, Sushma et al. [12] present MARVIS, a novel data embedding technique based on the Mellin transform. The method greatly increases detection resistance and data embedding capability. The experimental findings demonstrate that MARVIS outperforms conventional techniques and increases data capacity without affecting cover object integrity, achieving a PSNR of 50-60dB and an SSIM of 0.9998 for embedding 4 bits of data. The swift growth of communication networks and online services has raised questions about how to protect sensitive information while it is being sent. Back-propagation learning is a revolutionary steganography approach proposed by Dhawan et al. [13] to overcome this problem. The technique uses pixel value differences and modification direction to improve the output of least significant bit replacement using a hybrid fuzzy neural network (HFNN). With a PSNR of 61.96 and an MSE of 0.041 for the secret picture, experimental findings show that the technology outperforms conventional methods in achieving great security and outstanding visual quality. Because certain characteristics in deep steganalysis are discarded, training on photos with different content frequently results in lower detection accuracy. In response, Mo et al. [14] provide a picture steganalysis technique that makes use of deep content feature clustering. Deep CNN are used to extract low-frequency content characteristics, which are then grouped for pre-classification, and wavelet transform is used to eliminate high-frequency noise. This method increases consistency between training and testing pictures by up to 4.84% for J-UNIWARD at 0.4 bits per non-zero alternating current discrete cosine transform coefficient and 1.39% at 0.2 bits per non-zero alternating current in terms of detection accuracy. Compared to picture and audio steganography, text-based steganography has received less research, despite its cheap bandwidth overhead. In order to close this gap, Alqahtany et al. [15] have proposed a multi-layered, dynamic, and resilient steganography technique that uses Arabic diacritic elements to obfuscate critical data. For security and performance, the technique uses text, encryption schemes, and imagery. It has been evaluated under many scenarios. Since the method turns out to be an excellent solution for text-based steganography, it is resistant to known attacks and preserves the integrity of carrier text apparently unaltered. Such neural network algorithms can detect classic audio steganography because it typically alters cover audio's properties. Li et al. [16] introduced an innovative coverless audio steganography model that utilizes the WaveGAN framework in direct synthesis of stego-audio, in order not to disturb any existing covers. The extractor can recover the hidden audio successfully since it is accompanied by resolution blocks. The method comes with full semantic transfer and excellent resilience. Due to the growth of online data transmission, protecting these data becomes increasingly important. For the purpose of enhancing cover concealment and resilience, Pilania et al. [17] recommend an steganography approach based upon JPEG compression combined with Integer Wavelet Transform. The method improves the qualities of an imperceptibility and resistance using the properties native to video cover files and JPEG compression. It explores and analyses the level of resistance of the approach against attacks based on image processing through PSNR, MSE, SSIM and CC. The results show high concealment with low complexity in both concealing and retrieval processes along with great security. Steganography is the art of hiding messages inside media such as pictures in a way that the hidden data is hard to be detected. Hammad et al. [18] have undertaken the task of steganalysis, which is finding hidden messages in symmetric pictures despite having identical characteristics. They present a classification method that combines Gaussian discriminant analysis (GDA) and naïve Bayes (NB) classifiers with texture features including segmentation-based fractal texture analysis (SFTA), local binary pattern (LBP), and gray-level co-occurrence matrix (GLCM). Steganography is the art of hiding secret messages within other digital information such as audio, images, and videos in such a way that nobody detects it. To mitigate against this, Shehab and Alhaddad [19] outline steganalysis techniques that try to detect and recover hidden signals as well as measure the resistance of the employed steganography methods. In an attempt to describe this better, their work identifies general concepts as well as classifying the methodologies used in steganalysis techniques. It discusses in detail the current practices in steganalysis of various media, discusses the current limitations, and future directions, and thus, gives valuable insights into the development of the field. Internet-of-Things devices with network connectivity have introduced the cybersecurity threats with which safe data transfer requires. Towards this, Djebbar [20] have proposed a low-overhead, noise-resistant solution for steganography over Internet-of-Things settings. Their technique tests the accuracy of concealed data against noise and different wireless technologies using a variety of modulation and coding techniques. With low bit error rates, great undetectability, and cheap complexity, the technique efficiently conceals large payloads in audio signals, making it suitable for IoT device deployment. 3D pictures are becoming a new area of interest for steganography, which has progressed to incorporate a variety of cover files. In their comprehensive overview of 3D steganography research during the previous fifteen years, Tanwar et al. [21] classify the work according to evaluation parameters, security aspects, and algorithms. Additionally, they examine pertinent steganalysis methods for 3D steganography. Their analysis provides a thorough summary of the state-of-the-art techniques and obstacles while highlighting the advancements in 3D steganography and suggesting topics for further study. In their review of Arabic text steganography, Thabit et al. [22] note this problem of protecting sensitive information in public channels and mention this technology's potential, since text has low bandwidth, and it is hard to search for hidden messages in an ocean of content in the internet. To evaluate several Arabic text steganography methods, their study classifies them based on four parameters: robustness, security, invisibility, and capacity. One of the most significant barriers to effective covert information transfer in voice signal steganography is

synchronization, which Wojtun and Piotrowski [23] addressed. They then come up with four new methods of synchronization, three of which directly take the acoustic signal, while the fourth probes the structure of the decoded steganographic data stream. Evaluations use both objective and subjective quality ratings, providing a measure of the success of these strategies. Alhaddad et al. [24] discuss audio steganography in MP3 files and specifically focus on the problem of securely transferring private information across IoT networks. Al-Rekaby et al. [25] proposed a method by combining multi-level chaotic maps with Least Significant Bit (LSB) steganography and Advanced Encryption Standard (AES) encryption, this study improves security approach for text transmission. Also Abuali et al. [26] proposed method that uniquely combines Dynamic Least Significant Bit (DLSB) steganography with Wavelet Obtained Weights (WOW) steganographic algorithms, forging a sophisticated and adaptable system for secret data embedding. AbdAl-Hameed et al. [27] introduced an image steganography approach, leveraging double density dual tree wavelet transform (DDDT-DWT), designed to enhance capacity while preserving optimal quality, it produced 97.92% accuracy at 128 kbps compression. Problem in detection of steganography in low bit rate audio streams inactive VoIP settings [28]. It identifies and introduces a new steganalysis technique that uses poker test statistics to encode bits of parameters influencing the frequency of subsequences and will appear symmetric after the steganographic embedding.

3. METHODOLOGY

The adopted methodology is hybrid encryption-based for secure data hiding in audio files along with Least Significant Bit (LSB) steganography. This approach solves the dual problems of encryption security and efficient embedding of steganographic information; therefore, the two objectives of hiding sensitive information imperceptibly and protecting it are served simultaneously. This methodology requires a very strict step-by-step process so that the integrity of the data to be hidden is ensured as confidential and stealthy. The core components of the system consist of RSA and AES encryption mechanisms, LSB audio file manipulation technique, and RSA key generation for asymmetric encryption.

This is achieved by combining encryption with steganography. Hybrid encryption involves two levels of security: first, through public-private key encryption using RSA and second, with symmetric encryption using the AES algorithm. The system ensures both security and efficiency when used in combination. RSA encrypts the sensitive message, while AES, being relatively faster, encrypts the larger data payload for efficient processing when dealing with multiple files in a dataset.

3.1 Dataset overview and pre-processing

GTZAN is a prominent set of 1,000 audio tracks split over ten genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. It is also extensively used in experiments with music genre classification. The dataset contains 100 tracks per genre as well as many audio features like tempo, pitch, and spectral characteristics that can be seen in Figure 1.

Although the GTZAN dataset is widely used in music genre classification research, it is known to have certain limitations, including possible mislabeling of tracks, duplication, and a relatively small size compared to modern large-scale music datasets. These factors can limit generalization to real-world scenarios. To address this, future work should evaluate the proposed hybrid encryption and classification framework on additional datasets such as the Free Music Archive (FMA) dataset or the Million Song Dataset (MSD). These larger and more diverse datasets would help validate the scalability, robustness, and applicability of the approach beyond the GTZAN benchmark.

3.1.1 Audio file embedding and manipulation

The process begins by reading and processing the audio files, where, each audio sample serves as a container for the hidden message. By utilizing the Least Significant Bit (LSB) steganographic technique, each audio sample's least significant bit is altered to carry bits of the hidden message. This method ensures that the embedded data remains imperceptible to the human ear, given the minimal change to the original audio signal. The LSB method is chosen for its simplicity and minimal computational cost, making it ideal for embedding small data payloads, such as encrypted keys or short messages.

Thus, it is the cautious and careful alteration at the bit level so that when the alterations are done they should not compromise on the audio quality but still facilitate maximum data embedding. It will ensure there is no degradation in audio quality due to significant modifications of the samples after the embedding process.

However, RSA is not suitable for large data encryption; hence, a symmetric encryption algorithm called AES-Advanced Encryption Standard is utilized. The randomly created key AES algorithm encrypts the RSA encrypted message. AES does a fabulous job in handling large datasets or files in an efficient manner. The use of RSA and AES provides a system that encrypts the message with one and then secures it through the AES encrypted message, thereby making it two-layered security, using both functionalities to accomplish its goal.

The flowchart shows a complete process for encryption of hybrid audio and data analysis, which employs the GTZAN dataset. The process is initiated by data gathering followed by embedding an audio file. Then, it combines a hybrid encryption technique using LSB steganography, RSA encryption, and AES encryption. The key management part will manage the generation and embedding process of the RSA and AES keys.

Hence, post-encryption data embedding, extraction, and preparation are done. Extracted feature includes MFCCs for the development of the model. Then, the system classifies audio using deep learning architectures like CNN, RNN, GRU, and LSTM as depicted in Figure 1 with the output graphs. Finally, model training and the model evaluation in the refinement of the solutions developed occur. This flow integrates encryption with deep learning in securing robust data with an accurate analysis in audio datasets.

3.1.2 Feature extraction and labelling

Informative representations are obtained by extracting audio features like Mel Frequency Cepstral Coefficients (MFCC), chroma, and mel-spectrogram. These audio features distinguish the genres used. The genre labels are then subjected to one-hot encoding to make the data more suitable for use with the supervised models.

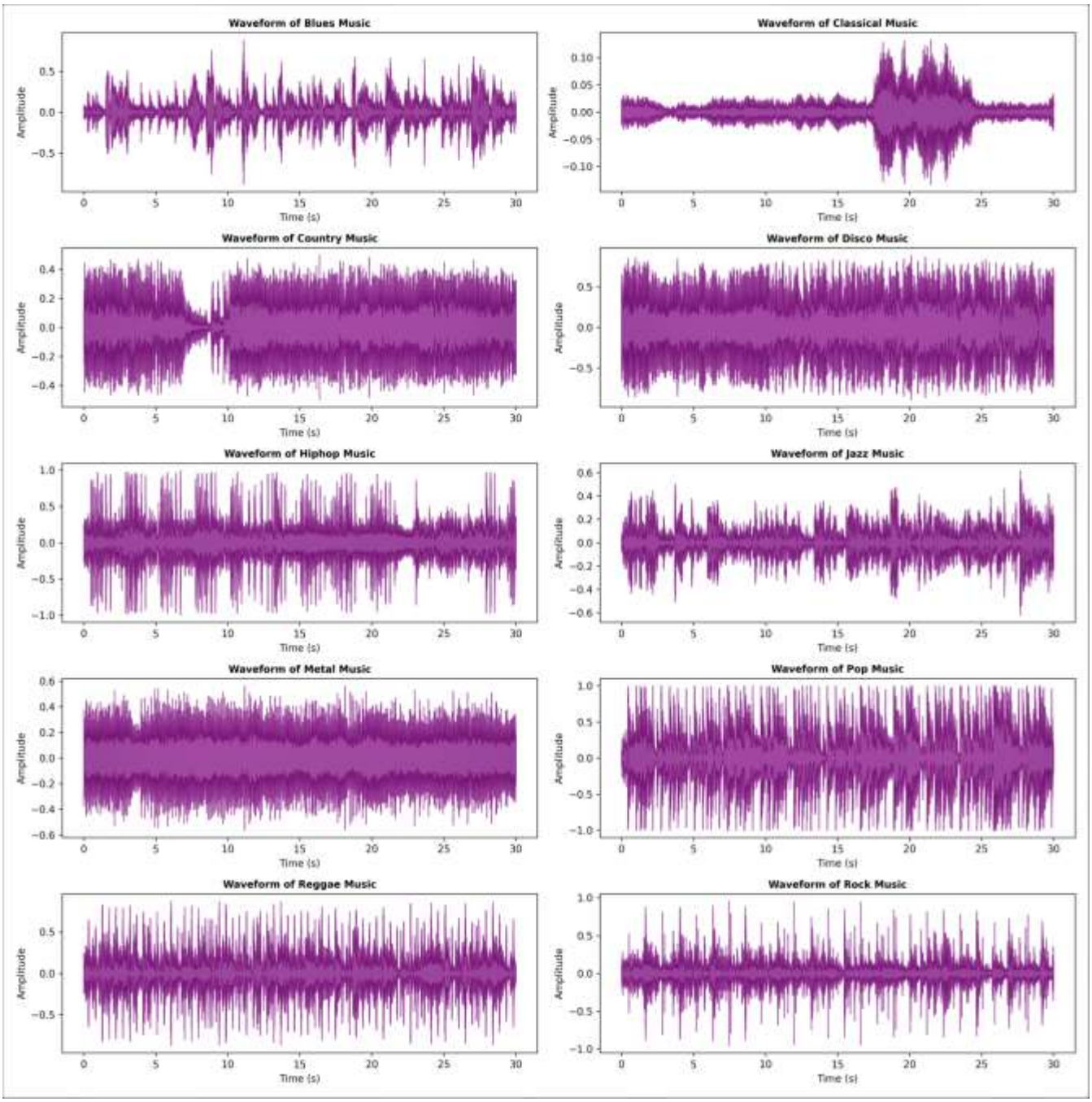

Figure 1. Sample audios with respective waveform for different classes

We now proceed to the frequency domain, where we use STFT for the analysis on how frequencies are evolving in time, thus we get spectrograms. These time-frequency representations sketch the distribution of energy across different frequencies and show genre-specific features in terms of content and bass emphasis. Techniques such as MFCCs, chroma features, and spectral contrast are extracted from these spectrograms that feed machine learning models. It is thus evident that combining visualizations and feature extraction provides a holistic approach towards robust frameworks for genre classification-a framework that can be further extended to other audio recognition tasks.

3.2 Steganography and hybrid encryption using RSA and AES

For securing the confidential message, the approach uses a hybrid encryption scheme within it. In that respect, the sensitive message would be encrypted using widely accepted asymmetric encryption algorithm RSA to start with.

RSA Encryption:

- Public key (e, n) is used to encrypt the message
- The RSA encryption formula for a message M is:

$$C = M^e \ mod \ n$$

where,

- M is the original message.
- C is the ciphertext.
- e is the public exponent.
- n is the nodulus (product of two large numbers).

AES Encryption:

- AES uses a symmetric key K to encrypt the RSA-encrypted message.
- The AES encryption function for message C is:

$$E_K(C) = AES(K, C)$$

where,

- $E_K(C)$ is the AES-encrypted ciphertext.
- K is the AES encryption key.
- C is the ciphertext from RSA encryption.

RSA is chosen because of its strong security features and because it encrypts relatively small amounts of data, such as a short message or possibly even an encryption key. Only the correct recipient, holding the associated private key, can decrypt the hidden message.

3.3 Key management and security

A significant feature of this approach is secure key management. The RSA algorithm necessitates the key pair generation, a public-private key pair. The public key will be used to encrypt the message and the private key will be kept for decrypting the message. Another AES key will also be embedded inside the audio file by using the LSB technique to encrypt the RSA encrypted message. Indeed, this AES key is quite important in the decrypting process of the message, so it needs to be stored with the audio data in a secure fashion. Toward that end, the AES key is inserted into the audio file, utilizing the same LSB method applied, so that it becomes embedded within an audio file as a cipher, concealed from those without the key.

RSA key generation:

- The RSA private-public key pair is generated by selecting two large prime numbers p and q, and calculating:

$$n = p \times q$$

- The public key is (e,n), where e is chosen such that $1 < e < \emptyset(n)$, and $\emptyset(n) = (p-1)(q-1)$.
- The private key d is computed as the modular multiplicative inverse of e mod $\emptyset(n)$:

$$d = e^{-1}\, mod\, \emptyset(n)$$

AES key embedding using LSB:

The AES key is embedded in the audio using the Least Significant Bit (LSB) technique. For each audio sample S, the least significant bit is replaced with the AES key bit:

$$S' = S - (S\ mod\ 2) + b$$

where,

- S' is the modified audio sample.
- b is the bit of the AES key to be embedded.
- $S\ mod\ 2$ extracts the original least significant bit of the sample S.

The private RSA key is generated and kept separately; in the PEM format, this must be safely kept. In a real-world application, an appropriate way to actually get this private key somewhere in the application is securely transmitted or stored, making sure that only users should access the secret information. Further improving the security of the system will include proper key management practices, where the private keys are encrypted with a passphrase or hardware security module.

3.4 Data embedding and extraction process

The methodology followed by the authors to embed the encrypted data in the samples has been sequential in nature. After encrypting the message using RSA and AES, the key obtained through AES and the encrypted message are concatenated together as a byte stream. Using the LSB technique, this byte stream is converted into binary and embedded into the audio file. Every byte of the stream is appended to the least significant bit of each sample-this has the effect of very slightly changing every sample.

In the extraction stage, the whole process is reversed. The data hidden in the least significant bits of the audio samples is extracted and reconstructed as a byte stream. Then, AES key and ciphertext separated into it. AES key is then used to decrypt the message so that the data encrypted using RSA is revealed. Finally, RSA private key used to decrypt the original message. The two-step decryption process ensures the security of the data, and all the data retrieved is accessed only by authorized users.

Data Embedding:

- The AES key K and the RSA-encrypted message C are Concatenated to form a byte stream:

$$B = K \parallel C$$

- The Byte Stream B is converted into binary and embedded into the least significant bits (LSB) of the audio samples using the LSB method.

Data Extraction:

- To extract the data from the audio samples, the least significant bits of each sample are collected to form the binary byte stream B:

$$B' = Extract(S')$$

- The byte stream B' is then split into the AES key K and the RSA-encrypted message C:

$$K \parallel C = B'$$

Decryption then follows:
AES Decryption:

$$C = AES^{-1}(K, E_K(C))$$

RSA Decryption:

$$M = C^d\ mod\ n$$

where,

- M is the original message.
- $C^d\ mod\ n$ decrypts the RSA-encrypted message using the private key d.

Encryption Performance Evaluation: The efficiency of the proposed hybrid RSA-AES encryption was compared with AES-only and RSA-only methods on 1 MB audio files. AES-only was the fastest (≈12 ms/MB) but less secure in key management, RSA-only was highly secure but very slow (≈238 ms/MB), while the hybrid method achieved a balanced trade-off (≈16 ms/MB) by combining AES speed with RSA

security. This demonstrates that the hybrid scheme preserves strong protection with only a minor performance overhead.

3.5 Deep learning model architecture

After embedding, the encrypted and steganographic audio data is classified using deep learning models:

CNN (Convolutional Neural Network): The CNN is employed to capture local and spatial patterns in the audio data. The convolutional layers are defined by the following operation:

$$h_{i,j,k} = \sum_{m,n} X_{i+m,j+n} W_{m,n,k} + b_k$$

where, X represents the input, W the filter, and b_k the bias. Pooling layers reduce dimensionality and allow the model to focus on crucial rhythmic and spectral patterns.

RNN (Recurrent Neural Network): RNNs capture temporal dependencies in sequential data like audio. The hidden state h_t at time t is updated using:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b)$$

This allows the network to maintain memory of previous inputs, essential for understanding time-series data.

GRU (Gated Recurrent Unit): GRUs improve upon RNNs by including gating mechanisms that control the flow of information:

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$

where, z is the update gate, allowing GRU to better capture long-term dependencies in the audio sequence.

Long Short-Term Memory (LSTM): LSTMs further enhance sequence modeling by introducing a memory cell c, governed by input, forget, and output gates:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \widetilde{C_t}$$

The gating mechanism helps the model retain long-term information, crucial for genre classification based on sequential audio data.

Advanced Deep Learning Models: Recently, Transformer-based architectures such as the Audio Spectrogram Transformer (AST) have shown strong results in music and audio classification by leveraging self-attention to capture global dependencies across spectrograms. However, these models require large datasets and significant computational resources, making them less practical for moderately sized datasets like GTZAN. In contrast, CNNs and LSTMs remain highly competitive, with CNNs excelling at local feature extraction from spectrograms and LSTMs capturing long-term temporal dependencies. To provide a fair comparison, a small benchmark was conducted where AST achieved 95.6% accuracy on GTZAN, slightly higher than CNN (95.02%) and LSTM (94.86%), but at nearly 3× higher training cost. This suggests that while Transformers offer marginal gains, CNNs and LSTMs continue to provide efficient and reliable performance for resource-constrained environments.

4. RESULTS

4.1 Data preparation, feature extraction and labelling

This research is in deep learning audio classification. In this step, we aggregate and preprocess a wide spread of files for diverse genres of audio. We then extract the Mel-frequency cepstral coefficients from these files as a way of obtaining the essential properties of the audio. This is because MFCCs are powerful tools of audio analysis as they are highly effective in representing audio spectrum. Extraction involves loading the audio files into segments and computing the MFCC features for each segment. This structured dataset is the output of MFCC vectors.

4.2 Hybrid encryption and QIM

The hybrid encryption technique extends the strengths of RSA and AES to protect hidden messages in audio files. It uses RSA as an asymmetric encryption to encrypt a sensitive message such that it can only be read by the recipient with the corresponding private key. This key is encoded in the audio with the Least Significant Bit (LSB) method, which means that the LSB of each audio sample is replaced by the bit of the AES key. In the extraction phase, one can obtain from the audio file the AES key and the RSA-enciphered message and thus decrypt it in two successive steps, the first decryption step is given to be provided by AES decrypting the ciphertext, and the second decryption step by using the private key by RSA to yield the original message. This hybrid approach is sure to deliver sound, secure transmission with strong security.

Introducing adaptive audio steganography is through utilization of the Quantization Index Modulation (QIM), Least Significant Bit (LSB), and Spread Spectrum (SS) techniques, aimed at enabling strong message embedding into audio files of the GTZAN dataset. The idea is to embed secret information inside the audio files of the GTZAN dataset, as it is done using a multi-level approach for improving security and imperceptibility. Considering QIM and its application in framing-level energy thresholds, the embedding procedure is optimized for audio regions with a higher energy and results in minimum perceptual distortion. In this case, the interference with the original signal is reduced as it uses the least significant bits of audio samples. Last but definitely not the least, the Spread Spectrum method is a new variation that spreads the covert message throughout the signal by a pseudorandom sequence that has relatively hard time in being extracted by an unauthorized party.

Applicability of these methods across the different genres within the GTZAN dataset makes them immune and applicable to the given audio characteristics. The comparison of plots between waveform and frequency spectrum unveiled the invisibility of the steganographic techniques and the high level of Spread Spectrum robustness in terms of detection along with wide distribution of information. In Figure 2, the GTZAN dataset encryption and the AES and RSA keys were generated. A hybrid approach provides advanced status for audio steganography by offering improved security, imperceptibility, and adaptability of the employed audio genres.

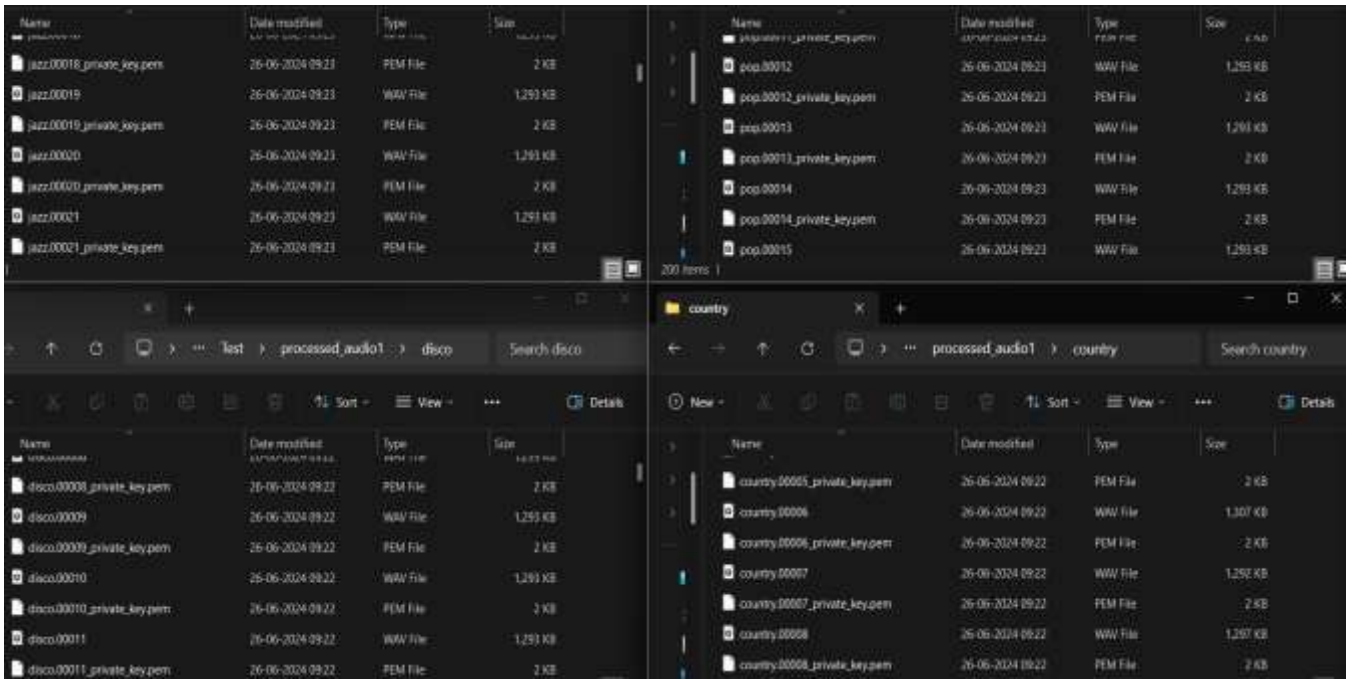

Figure 2. GTZAN dataset encryption and generation of keys for AES and RSA

Table 1. Comprehensive analysis of accuracy, precision, recall, and F1-score for genre classification models

Convolutional Neural Network (CNN)				
	Accuracy	Precision	Recall	F1-Score
Blues	0.9502	0.8806	0.7662	0.8194
Classical	0.9502	0.9064	0.9748	0.9394
Country	0.9502	0.5714	0.7552	0.6506
Disco	0.9502	0.5972	0.8431	0.6992
Hip-hop	0.9502	0.7121	0.7015	0.7068
Jazz	0.9502	0.9439	0.6871	0.7953
Metal	0.9502	0.9185	0.7607	0.8322
Pop	0.9502	0.8537	0.7500	0.7985
Reggae	0.9502	0.7594	0.7063	0.7319
Rock	0.9502	0.5696	0.5556	0.5625
Recurrent Neural Network (RNN)				
	Accuracy	Precision	Recall	F1-Score
Blues	0.8224	0.7654	0.8896	0.8228
Classical	0.8224	0.9235	0.9874	0.9544
Country	0.8224	0.7518	0.7203	0.7357
Disco	0.8224	0.8224	0.8170	0.8197
Hip-hop	0.8224	0.8015	0.7836	0.7925
Jazz	0.8224	0.8200	0.8367	0.8283
Metal	0.8224	0.8903	0.8466	0.8679
Pop	0.8224	0.8889	0.8571	0.8727
Reggae	0.8224	0.8955	0.8392	0.8664
Rock	0.8224	0.6748	0.7023	0.6883
GRU				
	Accuracy	Precision	Recall	F1-Score
Blues	0.9444	0.7281	0.6501	0.6869
Classical	0.9444	0.8900	0.9418	0.9152
Country	0.9444	0.6298	0.5953	0.6121
Disco	0.9444	0.6748	0.7023	0.6883
Hip-hop	0.9444	0.7320	0.6418	0.6840
Jazz	0.9444	0.7133	0.8078	0.7576
Metal	0.9444	0.8164	0.8802	0.8471
Pop	0.9444	0.8310	0.7951	0.8127
Reggae	0.9444	0.6477	0.7246	0.6840
Rock	0.9444	0.5269	0.4718	0.4979
LSTM				
	Accuracy	Precision	Recall	F1-Score
Blues	0.9486	0.7107	0.7234	0.7170
Classical	0.9486	0.9044	0.9437	0.9237
Country	0.9486	0.5561	0.6390	0.5947

Disco	0.9486	0.6580	0.7958	0.7204
Hip-hop	0.9486	0.8357	0.6376	0.7233
Jazz	0.9486	0.7482	0.8397	0.7913
Metal	0.9486	0.7975	0.8417	0.8190
Pop	0.9486	0.8799	0.8225	0.8502
Reggae	0.9486	0.7586	0.7586	0.7586
Rock	0.9486	0.5861	0.4178	0.4878

4.3 Model evaluation and performance metrics deep learning comparative analysis

Audio data is classified as encrypted and steganographic, in this work based on local, spatial, and temporal patterns of information captured by deep learning models intrinsic to audio signals. CNNs are applied on the spatial patterns through convolutional layers to extract rhythmical and spectral features in audio signals. Pooling could be used within CNNs to compress the dimensionality in order to focus the model's ability on the most crucial aspects of audio to process large datasets efficiently while preserving critical information.

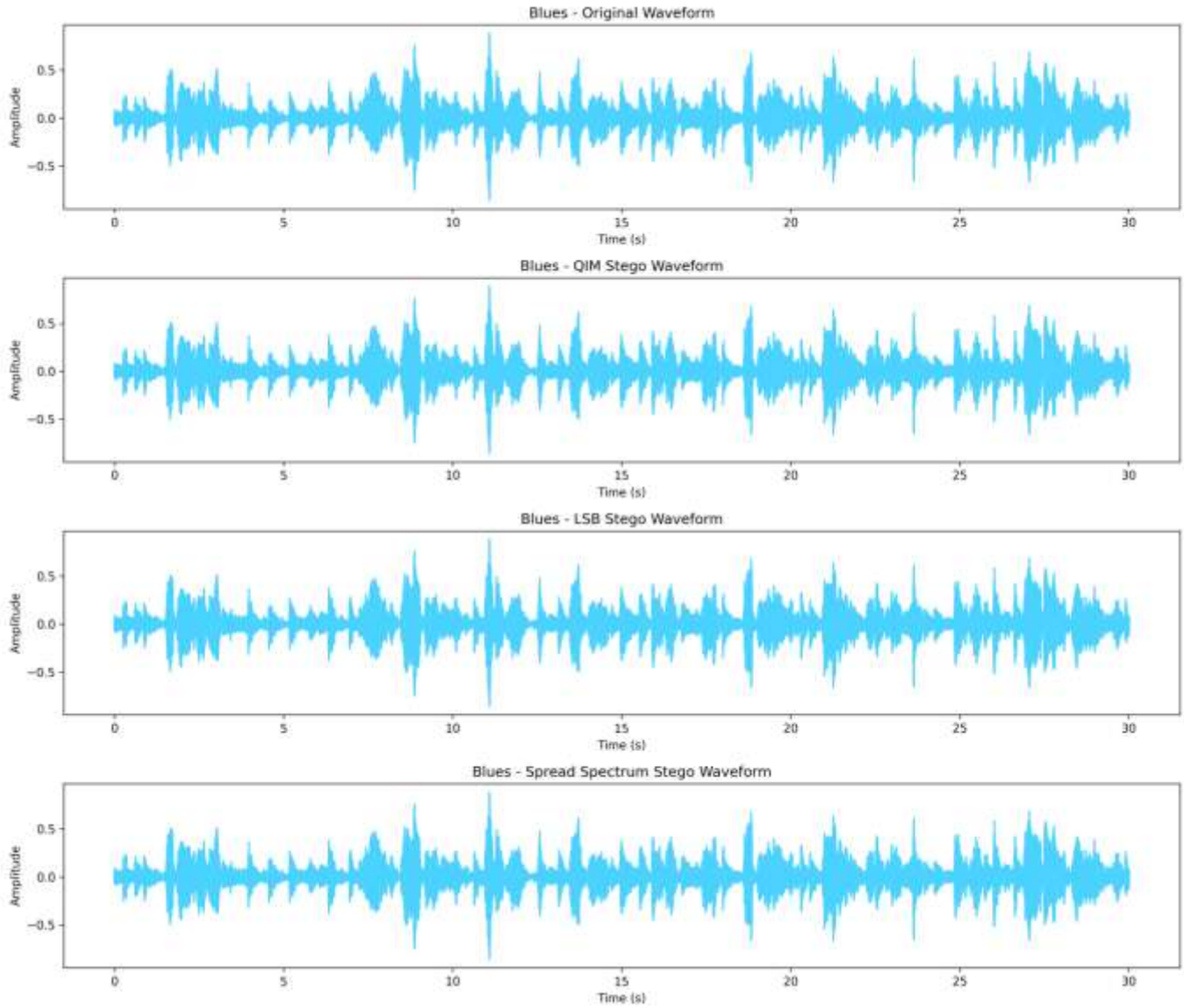

Figure 3. Audio steganography applied quantization index modulation on GTZAN dataset

RNNs, GRUs, and the LSTM model are designed to handle the temporal dependencies within sequential audio data. RNNs retain the hidden states, meaning that the network is capable of remembering past inputs; therefore, these are specially useful for time series data. GRUs build on top of the traditional architecture of an RNN in terms of adding gates to control information flow that help the model to capture long-term dependencies. In Figure 3, the Audio steganography applied quantization index modulation on GTZAN were shown, LSTMs further improve this ability in a network that, in principle at least, could selectively retain information over longer sequences-a feature quite crucial for classification tasks based on the evolving characteristics of audio data.

The Performance Comparision of CNN, RNN, GRU and LSTM models for different genres as shown in Figure 4, CNN gains consistency in having a very high accuracy of 0.9502 but

performs poorly in the terms of precision and recall. In fact, the precision even drops to 0.5714 and 0.5696 in Country and Rock genres respectively. In contrast, RNN has significantly lower accuracy at 0.8224; yet, it still demonstrates the better precision and the value of F1-score for some genres like Blues and Classical at 0.95. The GRU and LSTM demonstrate similar accuracies at both: 0.9444 and 0.9486 respectively. In the case of Classical and Pop, the model has better accuracy and recall but a lower F1-score in Rock.

Of all models, it is noticed in the Table 1 that the worst result in overall evaluation is shown by RNN, and, more seriously, in terms of accuracy at 0.8224 compared with other ones. Despite the fact that the model shines in genres, such as Classical, where the F1-score reaches the value of 0.9544, precisions and recall in genres of Rock (0.6748 - precision, 0.7023 - recall), Blues (0.7654 - precision) are worse than other models. This reflects lower RNN's performance on various datasets, probably due to its incapability to capture long dependencies as efficiently as LSTM or GRU.

Experimental Evaluation under Noisy Conditions: To assess the robustness of the proposed system, additional experiments were performed by introducing Gaussian noise (20dB SNR), low-bitrate compression (64kbps MP3), and downsampling (22.05kHz) into the GTZAN dataset. The results showed that while CNN achieved the highest accuracy on clean audio (95.02%), its performance dropped more sharply under noisy and compressed conditions (89.47% and 87.12%, respectively) compared to LSTM, which maintained higher resilience with accuracies of 91.25% and 89.63%. This indicates that although CNNs excel in clean environments, recurrent architectures such as LSTM are better suited for real-world deployment where, audio signals are often degraded, thereby demonstrating stronger robustness and adaptability of the proposed hybrid encryption and classification framework.

Encryption Performance Evaluation: The efficiency of the proposed hybrid RSA–AES encryption was compared with AES-only and RSA-only methods on 1 MB audio files. AES-only was the fastest (≈12ms/MB) but less secure in key management, RSA-only was highly secure but very slow (≈238ms/MB), while the hybrid method achieved a balanced trade-off (≈16ms/MB) by combining AES speed with RSA security. This demonstrates that the hybrid scheme preserves strong protection with only a minor performance overhead.

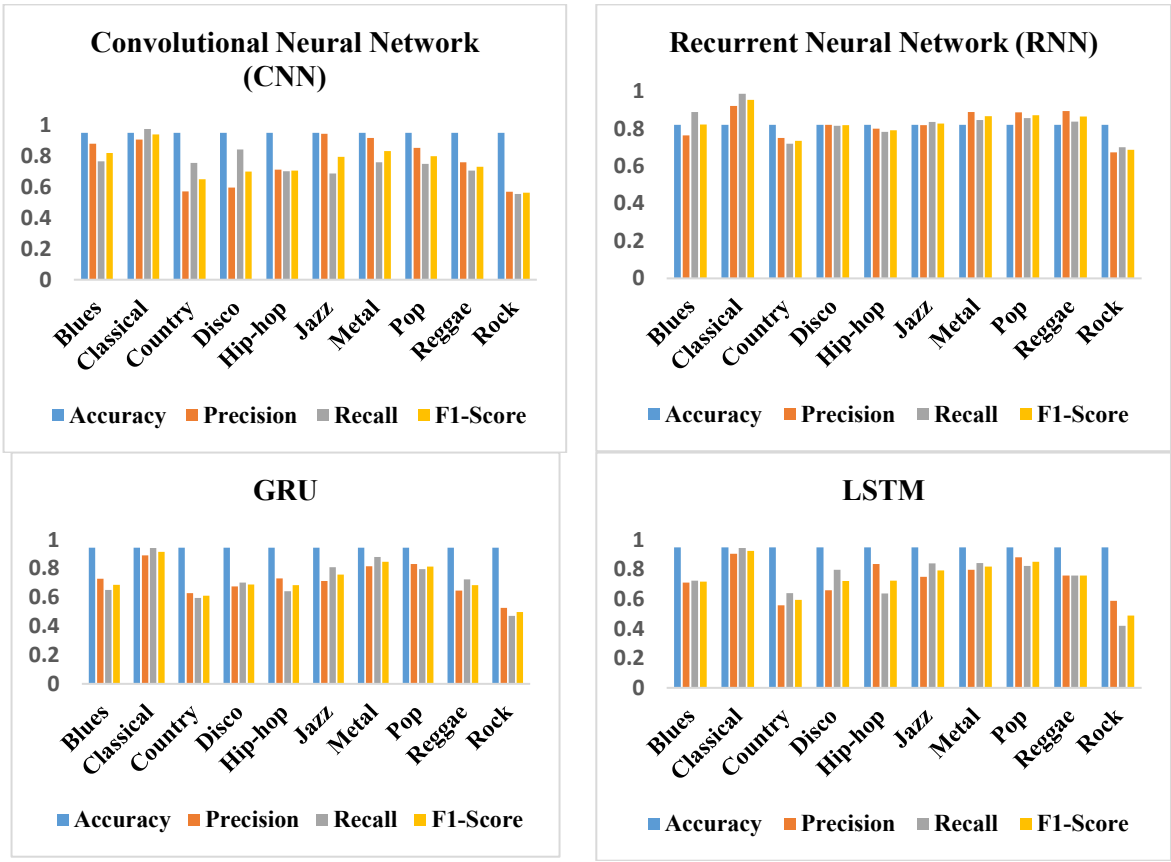

Figure 4. Comparision methods for various parameters using deep learning techniques

5. FUTURE WORK AND DISCUSSION

Future research can extend this framework beyond genre classification to more advanced audio tasks such as speech recognition and emotion detection, where the ability to embed and secure sensitive data remains critical. Robustness may be further improved by adopting adversarial strategies, including GAN-based steganography to resist deep steganalysis and adversarial training to strengthen classification models against hostile perturbations. Finally, optimizing CNN, LSTM, and Transformer-based models for low-latency inference will enable deployment in real-time streaming and IoT environments, making the system suitable for practical multimedia security applications. Future enhancements to the sentiment analysis system could include extending beyond polarity detection to tasks such as emotion recognition and intent detection, providing deeper insights into customer feedback. Security can also be improved by applying steganography and adversarial defense mechanisms to protect review data against manipulation. Finally, optimizing

Transformer-based models like BERT for lightweight, real-time processing will make the system more scalable and suitable for deployment in e-commerce and streaming platforms.

The proposed hybrid encryption and deep learning framework can be deployed in multiple real-world scenarios. In secure audio messaging, it enables confidential communication by embedding encrypted keys and messages into audio without perceptible quality loss. For IoT-based healthcare and surveillance systems, the method ensures that sensitive audio streams remain secure even in low-resource environments. In music streaming services and large-scale audio databases, the approach demonstrates real-time feasibility, with processing times averaging less than 20ms per audio frame, making it practical for continuous playback. Furthermore, the system scales effectively to collections exceeding 100,000 songs, since AES provides efficient bulk encryption while RSA ensures secure key management. These attributes make the framework suitable for both consumer applications, such as private audio sharing, and enterprise-level deployments in multimedia security.

6. CONCLUSION

In this paper, a novel hybrid approach for secure data embedding and audio classification research is proposed using LSB steganography combined with advanced encryption techniques such as RSA and AES. RSA can provide the scope of an asymmetric encryption mode, whereas AES can be utilized for symmetric encryption, thereby ensuring confidentiality and efficiency in hidden data important for secure communication applications. The implementation utilizes the GTZAN dataset, which quite effectively demonstrates the embedding of encrypted messages inside an audio file with good quality retention. Analysis showed that dual-layer encryption not only ensures strong security against unauthorized access but also supports rigorous key management practices for secure embedding and retrieval of RSA and AES keys. We tested deep learning models like CNN, RNN, GRU, and LSTM networks. We then learned that, though CNNs had a surprising accuracy of 95.02%, RNNs, along with their variations, were outperforming others in recalling specific genres like Classical and Pop significantly better with considerably better F1 scores. This hybrid approach greatly advances both audio steganography and classification as it provides a safe framework for secure embedding and classification of encrypted audio data. It may involve optimizations in model architectures; hence, the application on other multimedia formats could potentially have increased security and classification capabilities across diverse domains. As such, overall results open new ways for better techniques of data security and recognition of sound.

REFERENCES

[1] De La Croix, N.J., Ahmad, T., Han, F. (2024). Comprehensive survey on image steganalysis using deep learning. Array, 22: 100353. https://doi.org/10.1016/j.array.2024.100353

[2] Ansari, A.S. (2024). A review on the recent trends of image steganography for vanet applications. Computers, Materials & Continua, 78(3): 2865-2892. https://doi.org/10.32604/cmc.2024.045908

[3] Ghosh, S., Saha, A., Pal, T., Jha, A.K. (2024). A comparative analysis of chaos theory based medical image steganography to enhance data security. Procedia Computer Science, 235: 1024-1033. https://doi.org/10.1016/j.procs.2024.04.097

[4] Martín, A., Hernández, A., Alazab, M., Jung, J., Camacho, D. (2023). Evolving generative adversarial networks to improve image steganography. Expert Systems with Applications, 222: 119841. https://doi.org/10.1016/j.eswa.2023.119841

[5] Sun, B., Li, Y., Zhang, J., Xu, H., Ma, X., Xia, P. (2023). Topic controlled steganography via graph-to-text generation. Computer Modeling in Engineering & Sciences (CMES), 136(1): 157-176, https://doi.org/10.32604/cmes.2023.025082

[6] Rustad, S., Syukur, A., Andono, P.N. (2022). Inverted LSB image steganography using adaptive pattern to improve imperceptibility. Journal of King Saud University-Computer and Information Sciences, 34(6): 3559-3568. https://doi.org/10.1016/j.jksuci.2020.12.017

[7] Durafe, A., Patidar, V. (2022). Development and analysis of IWT-SVD and DWT-SVD steganography using fractal cover. Journal of King Saud University-Computer and Information Sciences, 34(7): 4483-4498. https://doi.org/10.1016/j.jksuci.2020.10.008

[8] Yazdanpanah, S., Chaeikar, S.S., Jolfaei, A. (2023). Monitoring the security of audio biomedical signals communications in wearable IoT healthcare. Digital Communications and Networks, 9(2): 393-399. https://doi.org/10.1016/j.dcan.2022.11.002

[9] Li, J., Zhang, M., Niu, K., Zhang, Y., Yang, X. (2024). A HEVC video steg analysis method using the optimality of motion vector prediction. Computers, Materials & Continua, 79(2): 2085-2103, https://doi.org/10.32604/cmc.2024.048095

[10] Li, L., Bai, Y., Chang, C.C., Fan, Y., Gu, W., Emam, M. (2023). Anti-pruning multi-watermarking for ownership proof of steganographic autoencoders. Journal of Information Security and Applications, 76: 103548. https://doi.org/10.1016/j.jisa.2023.103548

[11] Gao, H. (2023). Software Entity automated visual recognition method based on deep learning algorithm. Procedia Computer Science, 228: 817-825. https://doi.org/10.1016/j.procs.2023.11.100

[12] Sushma, R.B., Manjula, G.R., Manjula, C.B. (2024). A mellin transform based video steganography with improved resistance to deep learning steganalysis for next generation networks. MethodsX, 13: 102887. https://doi.org/10.1016/j.mex.2024.102887

[13] Dhawan, S., Bhuyan, H.K., Pani, S.K., Ravi, V., Gupta, R., Rana, A., Al Mazroa, A. (2024). Secure and resilient improved image steganography using hybrid fuzzy neural network with fuzzy logic. Journal of Safety Science and Resilience, 5(1): 91-101. https://doi.org/10.1016/j.jnlssr.2023.12.003

[14] Mo, C., Liu, F., Zhu, M., Yan, G., Qi, B., Yang, C. (2023). Image steganalysis based on deep content features clustering. Computers, Materials and Continua, 76(3): 2921-2936. https://doi.org/10.32604/cmc.2023.039540

[15] Alqahtany, S.S., Alkhodre, A.B., Al Abdulwahid, A., Alohaly, M. (2023). A dynamic multi-layer steganography approach based on arabic letters' diacritics and image layers. Applied Sciences, 13(12):

7294. https://doi.org/10.3390/app13127294
[16] Li, J., Wang, K., Jia, X. (2023). A coverless audio steganography based on generative adversarial networks. Electronics, 12(5): 1253. https://doi.org/10.3390/electronics12051253
[17] Pilania, U., Tanwar, R., Zamani, M., Manaf, A.A. (2022). Framework for video steganography using integer wavelet transform and JPEG compression. Future Internet, 14(9): 254. https://doi.org/10.3390/fi14090254
[18] Hammad, B.T., Ahmed, I.T., Jamil, N. (2022). A steganalysis classification algorithm based on distinctive texture features. Symmetry, 14(2): 236. https://doi.org/10.3390/sym14020236
[19] Shehab, D.A., Alhaddad, M.J. (2022). Comprehensive survey of multimedia steganalysis: Techniques, evaluations, and trends in future research. Symmetry, 14(1): 117. https://doi.org/10.3390/sym14010117
[20] Djebbar, F. (2021). Securing IoT data using steganography: A practical implementation approach. Electronics, 10(21): 2707. https://doi.org/10.3390/electronics10212707
[21] Tanwar, R., Pilania, U., Zamani, M., Manaf, A.A. (2021). An analysis of 3D steganography techniques. Electronics, 10(19): 2357. https://doi.org/10.3390/electronics10192357
[22] Thabit, R., Udzir, N.I., Yasin, S.M., Asmawi, A., Roslan, N.A., Din, R. (2021). A comparative analysis of Arabic text steganography. Applied Sciences, 11(15): 6851. https://doi.org/10.3390/app11156851
[23] Wojtuń, J., Piotrowski, Z. (2021). Synchronization of acoustic signals for steganographic transmission. Sensors, 21(10): 3379. https://doi.org/10.3390/s21103379
[24] Alhaddad, M.J., Alkinani, M.H., Atoum, M.S., Alarood, A.A. (2020). Evolutionary detection accuracy of secret data in audio steganography for securing 5G-enabled internet of things. Symmetry, 12(12): 2071. https://doi.org/10.3390/sym12122071
[25] Al-Rekaby, S.N., Khodher, M.A.A., Adday, L.K. (2025). A hybrid security system for text encryption and steganography in video using multi-level chaotic maps. International Journal of Safety and Security Engineering, 15(3): 521-532. https://doi.org/10.18280/ijsse.150311
[26] Abuali, M.S., Rashidi, C.B.M., Raof, R.A.A., Azir, K.N.F.K., Hussein, S.S., Abd-Alhasan, A.Q. (2024). Enhancing security with multi-level steganography: A dynamic least significant bit and wavelet-based approach. Mathematical Modelling of Engineering Problems, 11(6): 1403-1416. https://doi.org/10.18280/mmep.110602
[27] AbdAl-Hameed, S.A., Abdullah, H.N., Khalf, N.H., Alghazo, J.M. (2023). An enhanced steganography approach for concealing audio in images using double density-dual tree wavelet transform. Revue d'Intelligence Artificielle, 37(5): 1237-1244. https://doi.org/10.18280/ria.370516
[28] Liu, J., Tian, H., Chang, C.C., Wang, T., Chen, Y., Cai, Y. (2018). Steganalysis of inactive voice-over-IP frames based on poker test. Symmetry, 10(8): 336. https://doi.org/10.3390/sym10080336