Audio steganalysis using multi-scale feature fusion-based attention neural network

Received: 7 September 2023 | Revised: 21 June 2024 | Accepted: 13 July 2024

IET Communications

DOI: 10.1049/cmu2.12806

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Audio steganalysis using multi-scale feature fusion-based attention neural network

Jinghui Peng[1,2] | **Yi Liao**[1] | **Shanyu Tang**[2]

[1]School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou, Guangdong, China

[2]Cybersecurity and Criminology Centre, University of West London, London, UK

Correspondence
Jinghui Peng, Cybersecurity and Criminology Centre, University of West London, St Mary's Road, London W5 5RF, UK.
Email: jinghuipeng@gpnu.edu.cn

Funding information
Education Department of Guangdong Province, Grant/Award Number: 2021KTSCX063; Special topic of basic and applied basic research in Guangzhou, Grant/Award Number: SL2023A04J01043; Guangdong Regional Joint Fund, Grant/Award Number: 2022A1515110693; GPNU Science Foundation, Grant/Award Number: 2021SDKYA 025

Abstract

Deep learning techniques have shown promise in audio steganalysis, which involves detecting the presence of hidden information (steganography) in audio files. However, deep learning models are prone to overfitting, particularly when there is limited data or when the model architecture is too complex relative to the available data for VoIP steganography. To address these issues, new deep-learning approaches need to be explored. In this study, a new convolutional neural network for audio steganalysis, incorporating a multi-scale feature fusion method and an attention mechanism, was devised to enhance the detection of steganographic content in audio signals encoded with G729a. To improve the network's adaptability, a multi-scale parallel multi-branch architecture was employed, allowing characteristic signals to be sampled with varying granularities and adjusting the receptive field effectively. The attention mechanism enables weight learning on the feature information after multi-scale processing, capturing the most relevant information for steganalysis. By combining multiple feature representations using a weighted combination, the deep learning model's performance was significantly enhanced. Through rigorous experimentation, an impressive accuracy rate of 94.55% was achieved in detecting malicious steganography. This outcome demonstrates the efficacy of the proposed neural network, leveraging both the multi-scale feature fusion method and the attention mechanism.

1 | INTRODUCTION

Steganography and cryptography are different types of information security technologies. Cryptography, the use of special codes to keep information safe in computer networks, is a well-known technology that has been widely used in various fields to protect information [1]. On the other hand, steganography is the practice of concealing hidden messages in ordinary text, pictures, audio, video, and other mediums [2, 3]. It allows secret data to be disguised by embedding it into a carrier. Unlike cryptography, which makes the existence of encrypted information obvious, hidden information using steganography 'disappears' in the carrier without subjective consciousness, making its detection challenging. Therefore, the importance of the carrier in steganography is self-evident. Carriers can take the form of images, voice, text, or videos, and they serve as essential means of transmitting information over the Internet. Hidden information conveyed through steganography is not noticeable at first glance. Throughout the process of information transmission, steganography generally does not alter the original attributes of the carrier, enabling the secret data to be sent out secretly, making it difficult to discover its existence. However, if steganography were to be applied to illegal activities, it would pose certain security risks.

To confront the issue of malicious use of steganography in transmitting information, steganalysis technology has come into play. Steganalysis technology is used to detect the presence of secret information in these carriers, which involves a reverse process of steganography. Generally speaking, pattern recognition, machine learning, and other tools can be employed for steganalysis. There are two basic types of steganalysis methods: specific steganalysis and universal steganalysis.

Specific steganalysis involves extracting corresponding proprietary features based on known data embedding methods and analysing and judging based on these proprietary features, which results in strong pertinence. This method has a high degree of

accuracy but an obvious shortcoming, as it can only target a certain type of steganography, limiting its universality.

On the other hand, universal steganalysis, also known as blind steganalysis, does not aim at a specific algorithm or class of algorithms but instead judges by looking for features independent of any particular algorithm. Due to the ever-changing steganographic algorithms, a general analysis of steganography is more suitable and becomes the mainstream research method in order to cope with the seemingly endless variations of steganography.

The method of steganalysis varies depending on the carrier used in steganography. A carrier typically serves as a medium for communication among people. For instance, images and texts are common forms of information transmission. However, information can also be conveyed through sound or video. When information is stored on a computer in formats such as images, text, audio, or video, it can be utilized as the carrier for steganography. As a result, different methods are required to effectively deal with the proprietary characteristics of the information and achieve better results. In the realm of machine learning, convolutional neural networks (CNNs) are frequently employed to process image information, while recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) are generally used to handle audio and text information.

Deep learning is a type of artificial intelligence that uses algorithms (sets of mathematical instructions or rules) based on the way the human brain operates. As a sub-branch of machine learning, it is deeply rooted in the hearts of people because of its excellent performance. A neural network is a computer system or a type of computer program that is designed to mimic the way in which the human brain operates, forming the basis of deep learning. It is inspired by the biological nervous system and serves as a mathematical model that imitates the structure and function of a biological neural network. In the final analysis, deep learning is a mathematical expression, and mathematics provides the theoretical support for deep learning. The process of deep learning is regarded as a journey to find the optimal solution to a mathematical function.

The CNN stands out as a highly effective algorithm in Deep Learning for both regression and classification tasks. Operating as a feed-forward neural network, CNN incorporates convolution calculations and possesses a specific depth structure. Typically, its structure comprises an input layer, convolutional layer, activation layer, pooling layer, and fully connected layer. The convolution operation resembles the use of filters, enabling the activation of pertinent data features while reducing the model's parameter count, thereby decreasing computational complexity. Through the activation function in the activation layer, a non-linear factor is introduced to normalize data and tailor it to the desired function. The pooling layer effectively reduces the feature matrix's size, aiding in processing efficiency. Lastly, the fully connected layer maps the learned distributed representation features to the corresponding sample label space. When utilized judiciously, this function can solve various problems, including the discovery of hidden information. Overall, the deployment of a well-designed CNN can significantly enhance the performance of numerous machine-learning tasks.

The performance of deep learning models in audio steganalysis depends on several factors:

Data availability: Deep learning models require large and diverse datasets for effective training. The availability of sizable and diverse audio steganography datasets directly impacts the performance of these models.

Feature extraction: Designing effective features or representations of audio signals is crucial for the success of deep learning models. Some approaches use spectrograms, mel-frequency cepstral coefficients (MFCCs), or other time-frequency representations as input to deep learning models.

Model architecture: The choice of deep learning architecture plays a crucial role. Various architectures like CNNs, RNNs, long short-term memory (LSTM) networks, or hybrid models have been explored for audio steganalysis.

Steganography techniques: The effectiveness of audio steganalysis models can be influenced by the complexity and sophistication of the steganography techniques used to hide data in audio signals.

The field of deep learning and steganalysis has been evolving rapidly, so new techniques and approaches need to be investigated.

2 | RELATE WORK

In recent years, the field of deep learning has experienced rapid advancements and extensive applications across various industries. A comprehensive literature review reveals that deep learning has emerged as the dominant approach for steganalysis, providing the most reliable methods in recent times. Some deep learning models have been utilised in research within the cybersecurity field [4–13].

The evolution of convolutional neural networks for classification tasks began with the introduction of AlexNet in 2012. Since then, there have been remarkable strides in network architecture, resulting in the proposal of more sophisticated and deeper neural networks like VGG, GoogleNet, and the groundbreaking ResNet. The ResNet architecture effectively addressed challenges related to excessive network depth and gradient disappearance [14], further enhancing the capabilities of deep learning models.

In 2018, Mehdi Boroumand et al. [15] harnessed the power of deep residual networks to construct a steganalysis model, specifically designed for detecting JPEG images. While the experimental results showed only a marginal improvement in detection, this marked a significant step towards leveraging deep learning techniques for steganalysis.

Building upon this foundation, researchers have made continuous efforts to explore various improved and optimized neural network models for steganalysis, yielding promising results. These endeavours have demonstrated the potential of deep learning in detecting and analysing hidden information within digital media.

In the realm of steganalysis for speech signals, notable advancements have been made by various researchers. Initially, Chen et al. [16] introduced an audio steganalysis model called

ChenNet, which relied on a convolutional neural network to identify LSB (least significant bit)-based steganography within the time domain.

Following their work, Lin [17] and others further enhanced the convolutional neural network model by incorporating truncated linear units and residual modules for optimization purposes. The outcomes of their experiments substantiated the effectiveness of these model optimization techniques.

Wang et al. [18] took a different approach and proposed a CNN-based MP3 audio steganalysis method. This method employed MP3 quantization to enhance the discrete cosine transform coefficients, which were then used as input for the neural network. This novel approach demonstrated promising results in detecting steganography in MP3 audio.

In a related endeavour, Ren et al. [19] introduced a steganographic analysis method tailored for MP3 and AAC (Advanced Audio Coding) audio. They utilized a deep residual network (ResNet) and employed audio spectrogram data as the input for the network. This approach proved to be effective in detecting steganography in these audio formats.

Within the domain of VoIP steganography analysis, Yang et al. [20] employed a bi-directional long-term and short-term memory circulation neural network (Bi-LSTM) to effectively capture long-term contextual information within the carrier. Subsequently, they utilized CNN to capture both local and global features, as well as the temporal carrier features.

Furthermore, Yang [21] and his colleagues introduced a lightweight neural network called the fast correlation extraction model (FCEM). This innovative model represents a variant of multi-head attention and exhibits distinct advantages over relatively complex cyclic neural networks like RNN (recurrent neural Network) and CNN in terms of both accuracy and time efficiency.

Steganalysis research has seen significant advancements through the adoption of a multi-scale model structure, inspired by the inception module proposed by Li [22] and others [23]. In this approach, the CNN architecture incorporates convolution kernels of varying widths, interconnected with diverse activations. This novel design, known as the DAM module, has yielded exceptional experimental results, outperforming existing models.

The multi-scale network structures can be categorized into three main types: multi-scale input, multi-scale feature fusion, and multi-scale output. By employing convolution kernels of different sizes, the study generated diverse outputs that were then combined through depth superposition, resulting in new output features. In CNN, the high-level network and the low-level network exhibit distinct perceptual fields, extracting target features through layered abstraction. It is important to note that the characteristics of different scales significantly influence the results of classification tasks, and this model can be optimized through multi-scale convolution kernels.

The inception module, first introduced in [23], lays the foundation for the multi-scale model. Notably, the inception module played a pivotal role in implementing state-of-the-art classification and detection techniques during the ILSVRC 2014 (ImageNet Large-scale Visual Recognition Challenge). One key feature of this architecture is its ability to effectively utilize computing resources within the network.

In a neural network, the attention mechanism plays a crucial role in determining the significance of various sub-projects within a larger project. Its primary task is to prioritize important objectives and enhance the influence of essential sub-tasks, leading to improved results. Commonly used attention mechanisms in convolutional neural networks encompass spatial attention, channel attention, visual attention, among others. For instance, Yang et al. [24] introduced a lightweight neural network named the fast correlation extraction model (FCEM), which relies solely on a variant of attention called multi-head attention to extract features from VoIP frames.

In this study, the network structure incorporates the channel attention mechanism based on SeNet. By assigning weight ratios to different feature channels, the researchers obtained new features that helped optimize the CNN model.

This study introduces a novel convolutional neural network model for audio steganalysis, addressing the lack of research on the effectiveness of a multi-scale feature fusion method and the attention mechanism in this domain. The proposed model incorporates a multi-scale convolution module and integrates the channel attention mechanism. By learning features from audio signals, it aims to distinguish steganographic features from non-steganographic ones. The model's efficacy in audio steganalysis was assessed through various experiments, demonstrating its strong performance.

3 | PROPOSED DEEP LEARNING MODEL FOR AUDIO STEGANALYSIS

A new model for audio steganalysis is illustrated in Figure 1. First, it normalizes the characteristic data of an input network through a normalization layer. Convolution, activation, and normalization operations are then carried out. Next, it enters the multi-scale module to perform sub-multi-scale convolution operations four times. After that, the weight is added from the channel attention module to obtain the new characteristics of the average pooling operation. The features are then flattened on the completely connected layer, and the linear layer is added, followed by activation with ReLu. Dropout is then applied to prevent overfitting. Finally, the classification results are activated with the softmax function.

As shown in Figure 1, BN is the data normalization operation, Conv represents the convolution operation, and the convolution kernel used consists of 16 3 × 3 convolution kernels. The activation functions used in the model are the ReLU and softmax classification functions. MultiScale_Block is the multi-scale convolution module, and Ch_Block is the channel attention module, both of which are described in detail in the next section. Avg_pooling represents the average pooling operation. Dense is a fully connected neural network, and the number of output neural units is 128. The final output result distinguishes steganographic and non-steganographic data.

In the first layer, a feature is normalized with the help of batch normalization (BN), which resets the distribution of the input

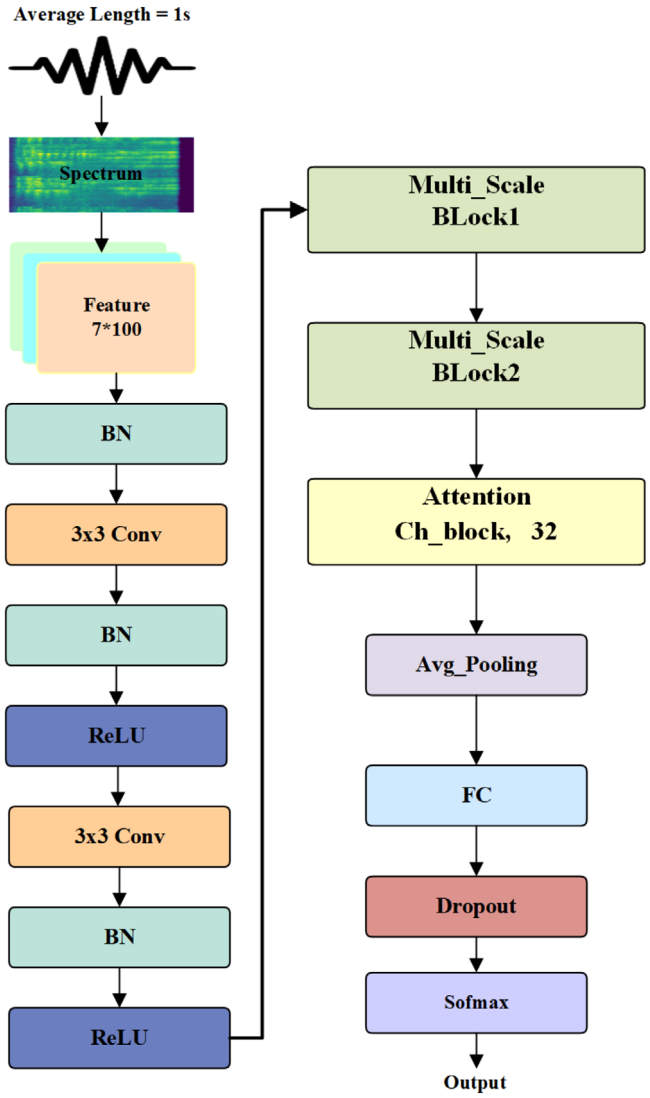

FIGURE 1 Multiple-scale and channel attention-based neural network model.

values for each neuron in every layer of the neural network to a normal distribution with a mean value of 0 and a variance of 1. This one-step operation prevents the gradient of the neural network from exploding and vanishing.

Immediately after entering the convolutional layer, there are 16 3 × 3 convolution kernels. The step size is set to 1, and zero padding is used. The output features have a depth of 16. The convolution operation can be regarded as the perception of local features. Through the perception of local features again and again, an enhanced global feature is then obtained, which improves the expressive ability of features. The BN operation is performed to normalize the features, and the ReLU activation function is used to add non-linear factors and increase the expressive ability of the model. The formula is as follows:

$$f(x) = \max(0, x) \tag{1}$$

The ReLU function is a piecewise linear function. It suppresses negative data by setting them to 0 while leaving positive values unchanged. The ELU function, on the other hand, is a modification of the ReLU function. When the input is negative, it produces a certain output with a specific anti-interference ability. However, there are issues with gradient saturation and exponential calculations.

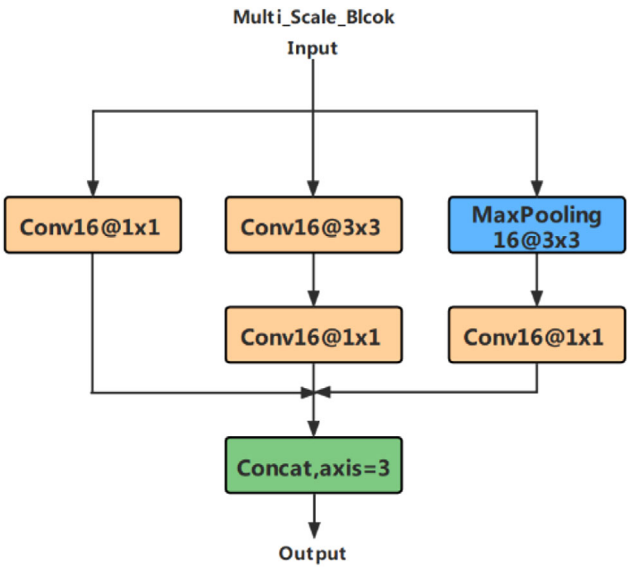

FIGURE 2 Multiple-scale convolution module.

3.1 | Multi-scale module

In the multi-scale module, a multi-scale convolution operation is shown in Figure 2. In Figure 2, Conv represents the convolution operation, 16@1 × 1 denotes the use of 16 1 × 1 convolution kernels, Maxpooling denotes the maximum pooling operation, Concat represents the splicing operation, and axis = 3 indicates the feature depth direction. In the multi-scale module, the input features go through three different paths to achieve the same depth output result.

The first path uses 16 1 × 1 convolution kernels to obtain the output result. The second path goes through 16 3 × 3 convolution kernels, and after the convolution operation, the output result is obtained through a 1 × 1 convolution operation. The third path involves the maximum pooling operation, and the output result is then obtained through a 1 × 1 convolution operation. Finally, the three outputs are stitched by depth to achieve the total output result.

The 1 × 1 convolution kernel in the model can effectively compress data and reduce the amount of computation. It significantly reduces the size of the input layer without compromising the network performance, thus greatly improving the learning efficiency. The width of the network can be increased by using convolution operations of different scales. Although raising the width of the network layer increases the amount of computation and memory overhead to a certain extent, increasing the width of each convolution layer can be more beneficial for expressing local information. For steganalysis, it also provides a

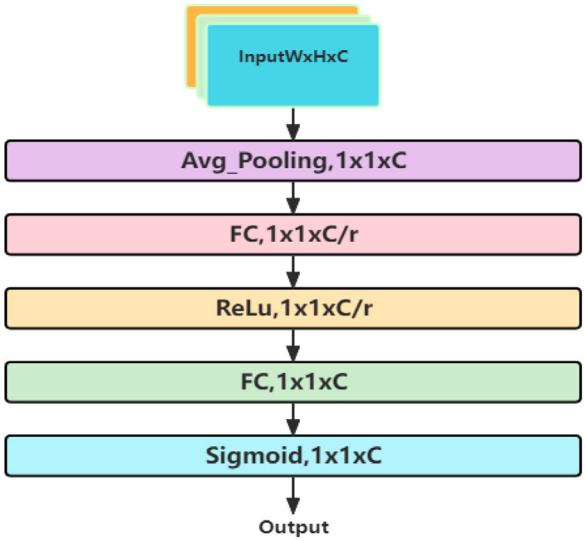

FIGURE 3 Channel attention module.

good solution, extracting as much scale information as possible through the different receptive fields of various branches.

3.2 | Channel attention module

The attention module uses the channel attention mechanism.

Given a feature channel number of c 1, the output channel number is c 2 after a series of operations of convolution, pooling, and full connection. By learning to obtain the weight of each feature channel, the importance of useful channels is strengthened, and the importance of less effective channels is reduced.

In Figure 3, avg_pooling represents average pooling, FC is a fully connected operation, ReLU and Sigmoid are activation functions. W, H, and C represent the width, height, and depth (that is, the number of channels). 'r' represents a certain ratio, and the model uses 16.

The channel attention mechanism can be divided into two steps: Squeeze and Excitation.

Squeeze uses global average pooling to perform feature compression in the spatial dimension and encodes all features in the channel into a global feature that has a global receptive field to some extent. The formula for this operation for a given input feature $H \times W \times C$ is as follows:

$$z = F_{sq}(u_c) = \sum_{i=1}^{H} \sum_{j=1}^{w} u_c(i, j) \quad (2)$$

Formula (1): The formula represents the Squeeze operation, where c represents the depth of the feature channel, H represents the height of the image, and W represents the width of the image.

Excitation employs a fully connected neural network to perform a non-linear transformation on the results of Squeeze. It comprises two fully connected layers: the first layer serves as dimensionality reduction using ReLU activation, while the second layer restores the original dimensionality. Subsequently, it utilizes sigmoid activation to obtain the weight of each feature channel. The formula is as follows:

$$s = F_e(z, w) = \sigma(g, (z, w)) = sigmoid(w_2 ReLu(w_1 z)) \quad (3)$$

Formula (2): W represents the convolution operation, z is the output of Squeeze, ReLu and sigmoid are the activation functions, and the weight value of the output is between 0 and 1.

The excitation result is used as the weight and multiplied by the input features to obtain the final result.

After the two preceding blocks, the neural network enters the pooling layer to reduce feature dimensionality, mitigating overfitting and enhancing the model's fault tolerance. The pooling operations employed are the average pool, which captures global context crucial for effective classification, and the maximum pool, which discards certain feature map information, potentially detrimental to steganalysis.

Subsequently, the learned distributed feature representation undergoes mapping into the sample label space through a fully connected layer. The first layer of this fully connected network consists of 128 neurons, derived from the transformation of collected feature data. To combat overfitting, a Dropout function with a coefficient of 0.5 is employed, preventing excessive reliance on specific connections within the network.

Lastly, the softmax activation function is utilized to produce two classification results, providing the final output of the model.

4 | RESULTS AND DISCUSSION

4.1 | Experimental setup

The model designed in this research was implemented using the Python language based on the Tensor Flow 2.4 deep learning framework and run on a Windows 10 system with an Intel Core i7-12700KF processor. The GPU utilised is the NVIDIA GeForce RTX 4090. The experiment utilised speech samples, comprising 94,599 segments of both steganography and non-steganography audio encoded by the PCM codec. Out of these, 50,000 samples were allocated for training the deep learning model for audio steganalysis, while the remaining 44,599 samples served as the test set.

The audio samples, whether steganographic or non-steganographic, were obtained using the same WAV audio encoding. For creating the steganographic samples, data embedding methods were randomly selected from CNV-QIM [21, 25] and the pitch steganographic method. The data embedding rates varied between 10%, 20%, 30%, and 40%, with each rate being randomly chosen.

To facilitate analysis, each sample was divided into 100 time frames. From each frame, 7 data points of the LPC feature and PD feature in the speech signal were extracted. Consequently,

the input data shape for the model was 7 × 100, wherein the feature data was augmented with a depth value.

By following this process, the deep learning model was trained and evaluated for audio steganalysis.

The QIM steganographic algorithm conceals confidential information by establishing a quantitative mapping relationship between features and information bits. In the context of steganography on linear prediction coding (LPC), it has been observed that the correlation of split vector quantization codewords of the linear prediction coding filter coefficients undergoes changes after QIM steganography. Leveraging this insight, a quantization code word correlation network model was constructed. This model utilizes deep learning to construct steganalysis algorithms, surpassing traditional approaches not only in terms of accuracy but also in real-time performance.

In general, mainstream speech recognition systems commonly rely on two primary feature extraction techniques: mel frequency cepstral coefficients (MFCC) and linear predictive coding (LPC). LPC, specifically, is designed based on the characteristics of the human vocal mechanism and plays a crucial role in speech acoustics. It helps distinguish between different human voices and vowels, mainly by analysing the distribution of resonant peaks in the frequency spectrum. Consequently, the LPC parameters effectively characterize the formant frequency and bandwidth.

In this experiment, a steganographic method was utilized, and it involved modifying the LPC parameters that had undergone steganography. The extracted LPC features were then used as learning parameters for the model. During the pitch steganography process, a small modification determines the value of the pitch period, which allows the model to effectively learn and classify the information encoded in the extracted features.

Overall, the integration of LPC-based steganography and feature extraction proves to be a valuable approach for enhancing speech recognition systems and facilitating efficient information encoding and decoding.

The deep learning framework used in this study was TensorFlow 2.0, and the Adam optimizer was employed during the experiment. Optimal convergence results were achieved when the batch size was set to 64, and the number of epochs was set to 6. The training was conducted on core graphics devices with an i5-1135G7 processor. Throughout the learning process, the model's training set loss consistently decreased, leading to a gradual improvement in accuracy. The evaluation metric used in the experiments was detection accuracy.

In the test set, the loss initially decreased at a slower pace but eventually stabilized. As a result, the accuracy reached an impressive 94.55%.

Figure 4 displays the outcomes of a specific experiment. The left figure illustrates the accuracy of training, with the blue line representing training accuracy and the orange line indicating test accuracy. On the other hand, the right figure showcases the training set's loss (blue line) and the test set's loss (orange line). The figure unequivocally demonstrates the exceptional performance of the model.

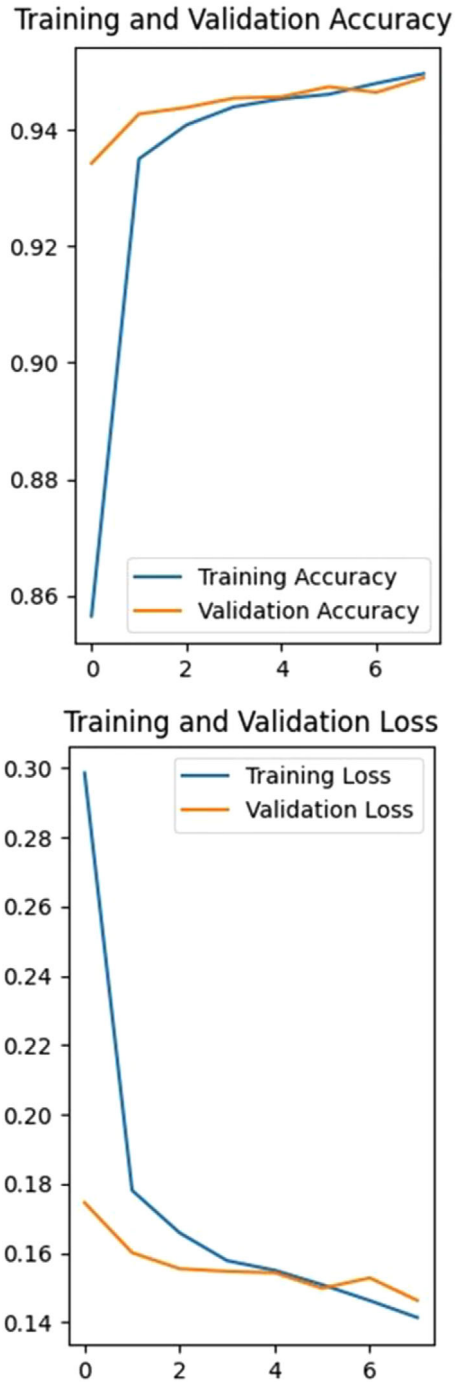

FIGURE 4 Model learning test.

Remarkably, these impressive results were attained with a lightweight network model, highlighting the reliability and precision of the proposed deep learning approach in detecting malicious steganography.

4.2 | Comparisons of different neural network structures

To evaluate the effectiveness of the proposed model for audio steganalysis, a series of experiments were conducted, constantly

TABLE 1 Accuracy of different network structures.

Index	Network description	Accuracy
#1	Full proposed CNN	94.55%
#2	Remove the multi-scale module	70.57%
#3	Remove the channel attention module	93.06%
#4	Use the ELU function	93.11%
#5	Use the Tanh function	93.70%

tuning and testing the model. Several key observations were made during the evaluation:

Impact of multi-scale convolution module: Removing the multi-scale convolution module from the model resulted in a noticeable decrease in accuracy. Hence, it was evident that this module played a crucial role in improving the model's performance, Table 1.

Attention mechanism module: When the attention mechanism module was removed, the model's stability was negatively affected, leading to a reduction of approximately 1% in prediction precision. Hence, the attention mechanism was deemed important for enhancing model stability and prediction accuracy.

Convolution kernel size selection: After testing different convolution kernel sizes and numbers, it was found that 3×3 convolution kernels performed better in the ordinary convolution layer. Consequently, all convolution kernels of this size were utilized in the model's convolution layer.

Multi-scale module with larger convolution kernels: Although using larger convolution kernels in the multi-scale module slightly improved stability and accuracy, it came at the cost of significantly increased computation and reduced training efficiency. As a result, it was determined that the benefits did not outweigh the drawbacks, and the larger kernels were not adopted.

Multi-channel attention mechanism module: The experiment involving the use of the multi-channel attention mechanism module did not yield a significant positive effect. Therefore, this approach was not favoured in the final model.

Activation functions and pooling layers: Different activation functions and pool layers were tested during the evaluation, but no significant advantage was observed in adopting alternative choices.

Based on considerations of training efficiency and accuracy from the experimental results, the proposed model for audio steganalysis was finalized with the following key components: the multi-scale convolution module, the attention mechanism module, and the use of 3×3 convolution kernels in the ordinary convolution layer. These choices were found to provide the best balance between performance and efficiency, making the model well-suited for the task at hand.

4.3 | Comparison with other related methods

In this part, several audio steganalysis methods are compared, and they have similar experimental preconditions.

Experiments were carried out with the CNV-QIM steganographic algorithm [21], setting the embedding rate at 50%. The accuracy of the final model was 92.67% for Chinese audio samples and 92.95% for English audio samples, resulting in significant improvements in the steganalysis rates.

Samples of LSB steganography were tested using the method in [6], achieving a steganalysis accuracy of 88.30%, which improved by 25% and 34%, respectively, compared with typical steganalysis methods using traditional artificial features and classification techniques.

The model proposed by Wang et al. [26] was tested on the EECS steganography algorithm at a bit rate of 128k bps and a relative load of 2 samples, achieving a prediction accuracy of 90.39%. The model of Lin [17] detected steganography with 90.50% accuracy under the condition that the embedding rate of LSB steganography was 1 bps. Lee et al.'s model [27] achieved an impressive accuracy of 89.14% when the embedding rate was set to 0.5 bps in LSB samples. Under the condition that the relative load of the same sample was $w = 4$, the accuracy rate achieved was 80.44% [26].

Additionally, Lin [28] and others made a detailed comparison between English and Chinese audio samples. They found a minor effect of the embedding rate and the sample length on the accuracy rate, and the average accuracy rate of the model reached more than 90%.

Table 2 provides a comprehensive comparison of the accuracy between the proposed model in this study and existing related models. In the table, "-" denotes unknown values, while "Mixed" indicates that the experimental data consist of samples with different embedding rates (0.1, 0.2, 0.3, and 0.5), which were randomly used during the experiment.

According to Table 2, the proposed deep learning model for audio steganalysis achieves an outstanding accuracy of 94.55%, surpassing all other related models [29–31]. This indicates that the neural network, when combined with a multi-scale feature fusion method and attention mechanism, can effectively detect malicious steganography with remarkable precision.

Collectively, these studies have contributed significantly to the advancement of steganalysis in speech signals, showcasing various successful techniques and approaches for detecting hidden information in audio data.

5 | CONCLUSION

This paper has described a lightweight convolutional neural network model that combines multi-scale feature analysis and the channel attention mechanism. This study has shown that using a multi-scale convolution method can enhance feature information for different receptive fields and improve the learning ability of the model for audio steganalysis. Additionally, adding the attention mechanism can help obtain the importance of feature information from different channels and facilitate key learning. Both of these enhancements effectively improve the model's performance.

The findings of this study suggest that to achieve better model performance, the sampling data needs to be increased or a larger convolution kernel needs to be added in the multi-scale

TABLE 2 Accuracy comparisons with related methods.

Method	Sample steganography method	Embedding rate	Sample length	Accuracy
Proposed	CNV-QIM and the pitch steganography	Mixed	1 s	94.55%
Yang [21]	CNV-QIM	50%	1 s	92.95%
Chen [16]	LSB	0.5 bps	–	88.30%
Wang [26]	EECS (The bit rate is 128 kbps and the relative load w is 2)	0.5 bps	10 s	90.39%
Lin [17]	LSB matching	–	1 s	90.50%
Lee [27]	LSB	0.5 bps	1 s	89.14%
Wang [18]	EECS (The bit rate is 128 kbps and the relative load w is 4)	–	10 s	80.44%
Lin [28]	CNV-QIM	Mixed	–	Over 90%

module, such as a 5 × 5 convolution. Considering the efficiency of model learning, the initial module can also be split into 51 × 5 convolution kernels or 23 × 3 convolution kernels.

In terms of the attention mechanism of audio steganalysis, further studies are necessary to determine the effects of different attention mechanisms on the model's performance. For example, a spatial attention mechanism would extract effective information from space to optimize the model. Adding the attention mechanism to the process of extracting feature information from the spectrum would improve the model's performance.

The Transformer's entire network structure is composed solely of an attention mechanism, primarily consisting of self-attention and a feedforward neural network. This architecture has demonstrated significant advantages in the NLP field. However, abandoning the traditional CNN and RNN led to the loss of its ability to capture local features, and simultaneously, it also lost the location information, which is crucial in NLP. Therefore, further research should concentrate on combining CNN or RNN with the characteristics of steganalysis.

AUTHOR CONTRIBUTIONS

Jinghui Peng: Conceptualization; methodology; validation; funding acquisition; writing—original draft; writing—review and editing. **Yi Liao**: Data curation; software; writing-original draft. **Shanyu Tang**: Supervision; validation; writing—review and editing.

ACKNOWLEDGEMENTS

This work was supported in part by the Education Department of Guangdong Province under Grant 2021KTSCX063, Special topic of basic and applied basic research in Guangzhou under Grant SL2023A04J01043, Guangdong Basic and Applied Basic Research Regional Joint Foundation under Grant 2022A1515110693 and GPNU Science Foundation under Grant 2021SDKYA 025.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data will be made available on reasonable request.

ORCID

Jinghui Peng https://orcid.org/0000-0003-0656-5494

REFERENCES

1. Fridrich, J.: Steganography in Digital Media: Principles Algorithms and Applications. pp. 107–129. Cambridge University Press, Cambridge, UK (2014)
2. Peng, J., Jiang, Y., Tang, S., Meziane, F.: Security of streaming media communications with logistic map and self-adaptive detection-based steganography. IEEE Trans. Dependable Secure Comput. 18(4), 1962–1973 (2021)
3. Peng, J., Tang, S.: Covert communication over VoIP streaming media with dynamic key distribution and authentication. IEEE Trans. Ind. Electron. 68(4), 3619–3628 (2021)
4. Zhou, Y., Zhang, Y., Liu, H., Xiong, N., Vasilakos, A.V.: A bare-metal and asymmetric partitioning approach to client virtualization. IEEE Trans. Serv. Comput. 7(1), 40–53 (2012)
5. Xiong, N., Han, W., Vandenberg, A.: Green cloud computing schemes based on networks: A survey. IET Commun. 6(18), 3294–3300 (2012)
6. Zhang, W., Zhu, S., et al.: A novel trust management scheme based on Dempster–Shafer evidence theory for malicious nodes detection in wireless sensor networks. J. Supercomput. 74(4), 1779–1801 (2018)
7. Wan, R., Xiong, N., et al.: Similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks. EURASIA J. Wireless Commun. Networking 2019, 59 (2019)
8. Shen, X., Yi, B., et al.: Deep variational matrix factorization with knowledge embedding for recommendation system. IEEE Trans. Knowl. Data Eng. 33(5), 1906–1918 (2019)
9. Li, X., Zhou, C., Tian, Y.C., Xiong, N., Qin, Y.: Asset-based dynamic impact assessment of cyberattacks for risk analysis in industrial control systems. IEEE Trans. Ind. Inf. 14(2), 608–618 (2017)
10. Li, T., Liu, W., Zeng, Z., Xiong, N.N.: DRLR: A deep-reinforcement-learning-based recruitment scheme for massive data collections in 6G-based IoT networks. IEEE IoT J. 9(16), 14595–14609 (2021)
11. Cheng, J., Yang, Y., Tang, X., Xiong, N., Zhang, Y., Lei, F.: Generative adversarial networks: A literature review. KSII Trans. Internet Inf. Syst. 14(12), 4625–4647 (2020)
12. Xia, Z., Jiang, L., Ma, X., Yang, W., Ji, P., Xiong, N.: A privacy-preserving outsourcing scheme for image local binary pattern in secure industrial internet of things. IEEE Trans. Ind. Inf. 16(1), 629–638 (2019)
13. Guo, J., Liu, A., Ota, K., Dong, M., Deng, X., Xiong, N.: ITCN: An intelligent trust collaboration network system in IoT. IEEE Trans. Network Sci. Eng. 9(1), 203–218 (2021)

14. Qian, Y., Dong, J., Wang, W., et al.: Deep learning for steganalysis via convolutional neural networks. In: Proceedings of the 2015 SPIE—Media Watermarking, Security, and Forensics. SPIE, Bellingham, WA (2015)
15. Boroumand, M., Chen, M., Fridrich, J.: Deep residual network for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. 14(5), 1181–1193 (2019). https://doi.org/10.1109/TIFS.2018.2871749
16. Chen, B., Luo, W., Li, H.: Audio steganalysis with convolutional neural network. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. pp. 85–90. Association for Computing Machinery, New York (2017)
17. Lin, Y., Wang, R., Yan, D., et al.: Audio steganalysis with improved convolutional neural network. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. pp. 210–215. Association for Computing Machinery, New York (2019)
18. Wang, Y., Yi, X., Zhao, X., et al.: RHFCN:: Fully CNN-based steganalysis of MP3 with rich high-pass filtering. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2627–2631. IEEE, Piscataway, NJ (2019)
19. Ren, Y., Liu, D., Xiong, Q., et al.: Spec-resnet: a general audio steganalysis scheme based on deep residual network of spectrogram. arXiv preprint arXiv:1901.06838 (2019)
20. Yang, H., Yang, Z., Huang, Y.: Steganalysis of VoIP streams with cnn-lstm network. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. pp. 204–209. Association for Computing Machinery, New York (2019)
21. Yang, H., Yang, Z.L., Bao, Y.J., et al.: Fcem: A novel fast correlation extract model for real time steganalysis of VOIP stream via multi-head attention. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2822–2826. IEEE, Piscataway, NJ (2020)
22. Li, B., Wei, W., Ferreira, A., et al.: ReST-Net: Diverse activation modules and parallel subnets-based CNN for spatial image steganalysis. IEEE Signal Process Lett. 25(5), 650–654 (2018)
23. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9. IEEE, Piscataway, NJ (2015)
24. Yang, H., Yang, Z.L., Bao, Y.J., et al.: A novel fast correlation extract model for real time steganalysis of VOIP stream via multi-head attention. 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2822–2826. IEEE, Piscataway, NJ (2020)
25. Xiao, B., Huang, Y., Tang, S.: An approach to information hiding in low bit-rate speech stream. In: Proceedings of 2008 IEEE Global Communications Conferences (GLOBECOM 2008). pp. 1940–1944. New Orleans (2008). https://doi.org/10.1109/GLOCOM.2008.ECP.375
26. Wang, Y., Yang, K., Yi, X., et al.: CNN-based steganalysis of MP3 steganography in the entropy code domain. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security. pp. 55–65. Association for Computing Machinery, New York (2018)
27. Lee, D., Oh, T.W., Kim, K.: Deep audio steganalysis in time domain. In: Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security. pp. 11–21. Association for Computing Machinery, New York (2020)
28. Lin, Z., Huang, Y., Wang, J.: Fast steganalysis of VoIP streams using recurrent neural network. IEEE Trans. Inf. Forensics Secur. 13(7), 1854–1868 (2018)
29. Li, S., Jia, Y., Kuo, C.-C.J.: Steganalysis of QIM Steganography in low-bit-rate speech signals. IEEE/ACM Trans. Audio Speech Lang. Process. 25(5), 1011–1022 (2017). https://doi.org/10.1109/TASLP.2017.2676356
30. Li, S., Wang, J., Liu, P., Wei, M., Yan, Q.: Detection of multiple steganography methods in compressed speech based on code element embedding, Bi-LSTM and CNN with attention mechanisms. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 1556–1569 (2021). https://doi.org/10.1109/TASLP.2021.3074752
31. Ahani, S., Ghaemmaghami, S., Wang, Z.J.: A sparse representation-based wavelet domain speech steganography method. IEEE/ACM Trans. Audio Speech Lang. Process. 23(1), 80–91 (2015). https://doi.org/10.1109/TASLP.2014.2372313

How to cite this article: Peng, J., Liao, Y., Tang, S.: Audio steganalysis using multi-scale feature fusion-based attention neural network. IET Commun. 1–9 (2024). https://doi.org/10.1049/cmu2.12806