Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 7088137, 12 pages
https://doi.org/10.1155/2022/7088137

Research Article

Medical Image Compression Based on Variational Autoencoder

Xuan Liu,[1] **Lu Zhang**,[1] **Zihao Guo**,[1] **Tailin Han**,[1] **Mingchi Ju**,[1] **Bo Xu**,[1] **and Hong Liu**[2]

[1]*College of Electronic Information Engineering, Changchun University of Science and Technology, Changchun, Jilin 130000, China*
[2]*College of Electro-Optical Engineering, Changchun University of Science and Technology, Changchun, Jilin 130000, China*

Correspondence should be addressed to Tailin Han; hantl@cust.edu.cn

Received 6 May 2022; Revised 6 September 2022; Accepted 1 October 2022; Published 2 December 2022

Academic Editor: Muhammad Shahid Farid

With the rapid growth of medical image data, it has become a current research hotspot that how to realize the large amounts of the real-time upload and storage of medical images with limited network bandwidth and storage space. However, currently, medical image compression technology cannot perform joint optimization of rate (the degree of compression) and distortion (reconstruction effect). Therefore, this study proposed a medical image compression algorithm based on a variational autoencoder. This algorithm takes rate and distortion as the common optimization goal and uses the residual network module to directly transmit information, which alleviates the contradiction between improving the degree of compression and optimizing the reconstruction effect. At the same time, the algorithm also reduces image loss in the medical image compression process by adding the residual network. The experimental results show that, compared with the traditional medical image compression algorithm and the deep learning compression algorithm, the algorithm in this study has smaller distortion, better reconstruction effect, and can obtain higher quality medical images at the same compression rate.

1. Introduction

With the advent of the medical and health big data era, medical image data show an "explosive" growth [1]. As a result, the storage and transmission of a large number of medical images become more difficult. So, image data compression becomes one of the important means to solve this problem [2]. At present, medical image compression technology is mainly divided into lossless compression technology and lossy compression technology [3]. In terms of lossless compression of medical images, Fischer et al. [4] used segmented data compression to provide efficient storage and transmission of visualization data. Aldemir et al. [5, 6] employed the interslice correlation between the voxels and used the modern adaptive and context-based reversible methods to compress binary volumetric data, which obtained good compression and storage. Although the images reconstructed by lossless compression technology have no distortion and have a good reconstruction effect, the compression rate is relatively low and lossless compression technology cannot meet the requirements of the compression of the huge amount of medical images currently [7]. The process of lossy compression technology is that the medical image data are compressed by mapping, quantization, coding, and entropy coding [1, 8, 9]. Quantization error caused by quantization coding is the main reason for rate distortion in the restoration process. Therefore, designing a lossy compression algorithm that not only satisfies the lower limit of distortion but also greatly improves the degree of compression has become one of the research hotspots in the field of medical imaging in recent years [10, 11]. At this moment, lossy compression is mainly divided into two categories in the field of medical image compression: image compression technology based on traditional algorithms (nondeep learning algorithms) and image compression technology based on deep learning. Among them, the image compression technology based on traditional algorithms is mainly realized by combining different wavelet transforms and different entropy coding methods [12–19], such as the DCT-based JPEG compression algorithm [12–14], the

DWT-based SPIHT algorithm [15, 16], the Harr Wavelet-based EZW [18], and so on. These methods are relatively mature and have lower complexity, but they rely heavily on manual design. And they need to optimize the three parts of transformation, quantization, and encoding, respectively. According to the rate-distortion theory of Shannon's third theorem, the joint optimization of rate and distortion is difficult. Without further constraints, the optimal quantization problem of high-dimensional space is difficult to solve [20]. Relevant literature shows that the better the compression effect, the more the distortion of the image [21]. At the same time, in the case of a higher compression rate, the image will not only produce serious blocking effects but will also have an effect on the quality of image restoration due to the loss of high-frequency information. Being severely reduced, the reconstructed image will produce a ringing effect on the location of the image with strong edge information [22].

In order to jointly optimize the rate and distortion, scholars put forward further constraints in the training process of the neural network. Much literature has alleviated the contradiction that compression efficiency and reconstruction quality cannot be combined to optimize [23–32]. Balle et al. proposed a variable rate compression algorithm based on LSTM experimental results which show that the proposed algorithm can reconstruct arbitrary input images well under the conditions of given input image quality [23]. However, the input image of the network is limited to a size of 32×32, which indicates that this method is still insufficient in terms of the spatial dependency of the captured image. To solve this problem, Kar et al. [24] proposed a new fully convolutional autoencoder for mammography image compression. Compared with traditional autoencoders, its ability to extract image edge information is stronger and the convergence speed is faster. The reconstructed image quality of this method exceeds that of JPEG. Sushmit et al. [25] proposed an X-ray image compression method based on convolutional recurrent neural network RNN-Conv. The proposed architecture can provide variable compression ratios during deployment, while it requires each network to be trained only once for X-ray images of a specific dimension. This model uses a multistage pooling scheme to learn contextualized features for efficient compression. This is the first evaluation of medical image compression using a deep convolutional RNN. Pareek et al. [26] introduced the IntOPMICM technique, a new image compression scheme that combines GenPSO and VQ, which achieved higher image quality at a given compression ratio. Toderici et al. first proposed an end-to-end nonlinear image compression framework based on deep learning, which extended the concept of transform coding from linear transformation to nonlinear transformation, and it fundamentally improved the qualitative nature of compression artefacts [27]. Compared with linear transformation, nonlinear transformation with higher computing power is more suitable for image statistics and it can simulate the characteristics of data distribution. However, in the case of a lower bit rate, the effect of image reconstruction is bad. In response to this problem, Ballé et al. combined a nonlinear model with a hyperprior model that effectively captures the spatial dependence of the latent representation and tends to produce reconstructed images with more details and a lower bit rate [28]. Compared with the nonlinear network model, it already achieved a better reconstruction effect in the field of natural image compression. The abovementioned analysis proves that this process can not only jointly optimize the network parameters but also can effectively solve the serious distortion problem caused by the traditional compression algorithm while avoiding the complicated algorithm structure design. However, in the field of medical image compression, the compressed details generally cover the information on human lesions, which are of great significance for the detection and diagnosis of diseases; hence, the information loss must be minimized [33]. In response to the above problems, this study proposes a medical image compression algorithm based on a variational autoencoder. By adding the residual network structure, the information before arithmetic coding is directly fused with the information after arithmetic decoding, which reduces information loss. As well as, the problem of gradient dispersion and explosion in the network training process is eased and it can effectively inhibit the degradation of deep neural networks. Thus, the efficiency of network training has been improved. Compared with the introduction of a hyperprior end-to-end compression network, the algorithm in this study showed better rate-distortion performance.

2. Related Theory

The algorithm in this study is a neural network structure combining a variational autoencoder and residual network. In order to obtain the reconstructed image closer to the original image, this study approximates the input samples through a variational autoencoder. Since medical images have more human lesion information than natural images, in this study, a residual network module is introduced to transfer information, thereby directly reducing the loss of information.

2.1. Variational Autoencoder. In image compression technology, model generation is a very ingenious method. It can generate image data similar to the input sample only by inputting a small amount of information. Variational autoencoder is one of the generative models [34].

Variational autoencoder (VAE) [35] is a type of generative model proposed by Huang and Wang. VAE learns the distribution of samples and uses estimated distributions to approximate the true distribution of samples. Then, it estimates the distribution to generate a similar sample to the original sample.

Figure 1 is the VAE neural network frame diagram. The training process of VAE is divided into encoding, sampling, and decoding processes. In the encoding process, a real distribution sample X_k is firstly input, then the unknown posterior distribution through the identification model $q(Z|X)$ that obeys the normal distribution is estimated, and then we can obtain the mean μ_k and standard deviation σ_k of

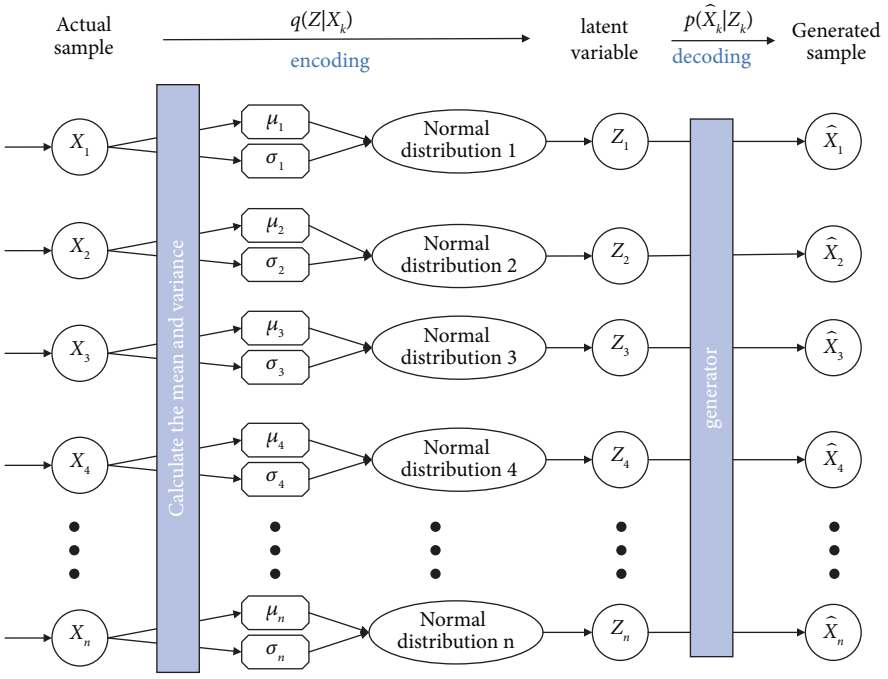

Figure 1: VAE neural network frame diagram.

the hidden variable Z distribution. In the sampling process, VAE generates a random sampling sample corresponding to Z_k through μ_k and σ_k.

In the decoding process, for the latent variable sampling sample Z_k, VAE generates a newly generated sample $\hat{X}_k$ for the hidden variable sampling sample Z_k through the generation model $p(X|Z)$.

The training goal of the VAE network is to reconstruct the image $\hat{X}_k$ infinitely close to the original image X_k. Relative entropy is usually used to measure the distance between two probability distributions. The closer the relative entropy is to zero, the closer the two distributions are, and their target function is

$$D_{KL}(p(X)\|p(\hat{X})) = \int p(X)\frac{p(X)}{p(\hat{X})}\mathrm{d}X. \tag{1}$$

In order to obtain the unknown true posterior distribution $p(Z|X)$ as much as possible, VAE approximates it by introducing a recognition model $q(Z|X)$ and optimizes the objective function through the maximum likelihood method, and then we can obtain the log-likelihood [36] function as follows:

$$\log\, p(X) = D_{KL}(q(Z|X)\|p(Z|X)) + L(X). \tag{2}$$

The relative entropy [37] of the two distributions is as follows:

$$\begin{aligned} D_{KL}(q(Z|X)\|p(Z|X)) &= q(Z|X)\log\int\frac{q(Z|X)}{p(Z|X)}\mathrm{d}Z \\ &= \int q(Z|X)\left(\log\, q(Z|X) - \log\frac{p(Z,X)}{p(X)}\right)dZ \\ &= \int q(Z|X)(\log\, q(Z|X) - \log\, p(Z,X) + \log\, p(X))dZ \\ &= \int q(Z|X)(\log\, q(Z|X) - \log\, p(Z,X))dZ + \log\, p(X) \\ &= E_{Z\sim q(Z|X)}\log\frac{q(Z,X)}{p(Z,X)} + \log\, p(X). \end{aligned} \tag{3}$$

That is, the variational lower bound L(X) of the likelihood function obtained by equations (2) and (3) is

$$\begin{aligned} L(X) &= E_{Z\sim q(Z|X)}\log\frac{p(Z,X)}{q(Z|X)} \\ &= E_{Z\sim q(Z|X)}\log\frac{p(X|Z)p(Z)}{q(Z|X)} \\ &= \int q(Z|X)(\log\ p(Z) - \log\ q(Z|X) + \log\ p(X|Z))dZ \\ &= -\int q(Z|X)\left(\log\frac{q(Z|X)}{p(Z)}\right)dZ + q(Z|X)\log\int p(X|Z)\mathrm{d}Z \\ &= -D_{KL}(q(Z|X)\|p(Z)) + E_{Z\sim q(Z|X)}(\log\ p(X|Z)). \end{aligned} \tag{4}$$

Since the relative entropy is nonnegative, then we can obtain $L(X) \leq \log\ p(X)$. The loss function is the maximum and the variational lower bound is $L(X)$. The closer $L(X)$ and $\log\ p(X)$ are, the smaller the relative entropy gets. Therefore, the loss function can be calculated as

$$\text{Loss} = D_{\text{KL}}(q(Z|X)\|p(Z)) - E_{Z\sim q(Z|X)}(\log\ p(X|Z)). \tag{5}$$

From what has been discussed, the smaller the objective function is, the closer the resulting generated sample is to the input sample.

2.2. Residual Network. The training process of a deep neural network is a process to optimize a feedforward neural network composed of several layers of neurons layer by layer in order to finally obtain an optimal solution. In the training process, the feedforward neural network transmits information forward while transmitting the error term used to update the gradient backwards and then updates the parameters required by neurons at each layer through the updation of the gradient. In other words, the optimization process of the neural network is the process of optimizing the network by optimizing the gradient of the neuron.

Compared with the shallow network, more sample features during the training process can be learned. As the number of network layers increases, the performance of the neural network can be effectively improved. Therefore, most image compression algorithms based on deep learning will have too large network models and deep neural network layers.

However, in the process of backpropagation, the gradient of the current layer is jointly affected by the error term of the current layer and the error term of the next layer. On the one hand, during the backpropagation process, when the error term coefficient of the latter layer is smaller than that of the current layer, it means that the following error term will become smaller and smaller so that the gradient will become smaller and smaller until it disappears. And then the phenomenon that the gradient disappearance caused by failure of updating of parameters required by the network will take place. Nevertheless, when the error coefficient of the latter layer is larger than that of the current one, the gradient will increase exponentially, resulting in gradient explosion. The training coefficient of a deep network cannot guarantee the problem that the gradient cannot be updated easily, which leads to the difficulty of network convergence and poor training effect.

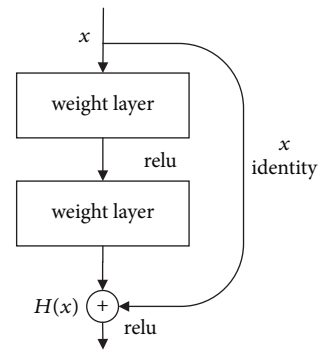

Figure 2: Residual network structure diagram.

On the other hand, the deep network also has the problem of network degradation in the process of forwarding propagation. In the process of neural network training, the network loss function in the early stage increases to saturation due to the gradual increase in the number of network layers and then rapidly converges, which results in network degradation and a higher error rate than a shallow network with the same rounds.

In order to solve the abovementioned problems, a residual network (ResNet) [38] was proposed by Shao et al., and its network structure is shown in Figure 2.

Its equation is expressed as

$$H(x) = F(x) + x. \tag{6}$$

When the error term of the gradient is updated in the backward propagation through the residual network module, the error parameter can be directly propagated to the back layer, and the too-small or too-large error coefficient will not affect the updation of the gradient, thereby alleviating the gradient dispersion or explosion phenomenon; and in the forward propagation of information, the input information can be directly transmitted from the low-level to the high-level via the created identity mapping, which can effectively solve the problem of network degradation. Therefore, the connection of the residual network makes the propagation of parameters in the training process smoother and the training effect is better.

2.3. System Network Architecture. In this study, an image compression algorithm based on a variational autoencoder is proposed for the first time in the research process of medical image compression. The system network flow chart is shown in Figure 3.

The whole flow chart is divided into two parts. The overall structure of the variational autoencoder in the two parts is consistent, one part is the backbone part, and the other part is the scale hyperprior part.

First, in the main part, the main parameter part uses a Gaussian probability model which is inferred from the hyperprior coefficient.

In the encoding portion, the nonlinear transform encoder transforms the sample x into a potential spatial representation y. Then, the discrete values are generated through quantization, and the discrete values are compressed into a binary bit stream in arithmetic coding.

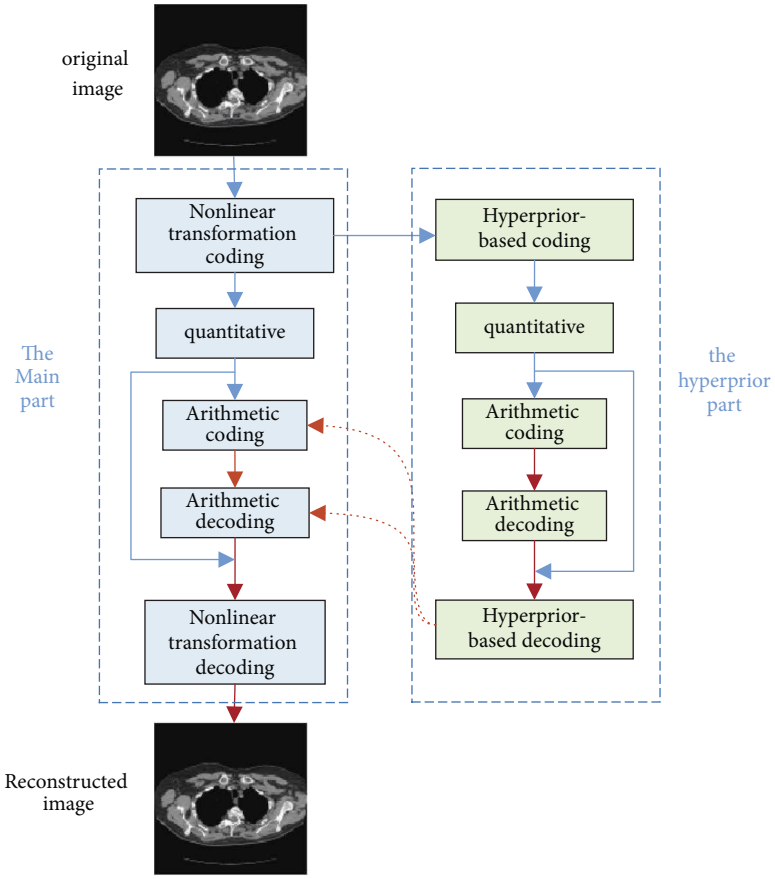

Figure 3: System network flow chart.

In the decoding portion, first, the binary bit stream is decoded through the relative arithmetic decoder, and then the obtained data are fused with the data directly transmitted to the nonlinear transformation decoder through the residual network before compression. Finally, it will be transmitted to the nonlinear transformation decoder for image reconstruction.

Second, in the hyperprior part, the parameters of the hyperprior are encoded by an independent identically distributed model. The parameters obtained from the main network parameter analysis are used to describe the distribution characteristics of the parameters of the main part.

In the encoding portion, the latent space representation data of the original image obtained by the nonlinear encoder are passed into the encoder based on the hyperprior for encoding. Then, a set of random variables to capture spatial dependencies is introduced. The next steps are quantization and arithmetic coding as described in the main part.

Among them, the decoding portion is consistent with the main part and arithmetic decoding. The decoded data are fused with the preencoded data information, and then the data are decoded through a decoder based on a hyperprior to obtain the standard deviation of the distribution of spatially dependent information. This standard deviation is passed to the backbone for estimation to arrive at the correct spatial representation y.

There is a strong spatial correlation between parameters, and the coding performance can be greatly improved by extracting relevant information through additional edge information. The added residual network can also directly transmit information on parameters before data compression, which can effectively preserve the detailed information of medical images.

The neural network structure diagram of this study is shown in Figure 4, where the blue line is the encoding process and the red line is the decoding process. The stride of the encoding part is the step size of down-sampling and the stride of the decoding part is the step size of up-sampling. GDN is generalized normalization, and IGDN is inverse normalization. Q is quantization, and AE and AD are arithmetic encoder and arithmetic decoder, respectively.

In this study, the neural network structure is the posterior superposition of two variational autoencoders. A residual network structure is added to the main part and the hyperprior part, respectively, and the rate-distortion performance is optimized by adjusting the distribution of the latent variables y and z of the bottleneck layer.

From the theoretical knowledge of the variational autoencoder mentioned in Section 2, we can obtain the log-

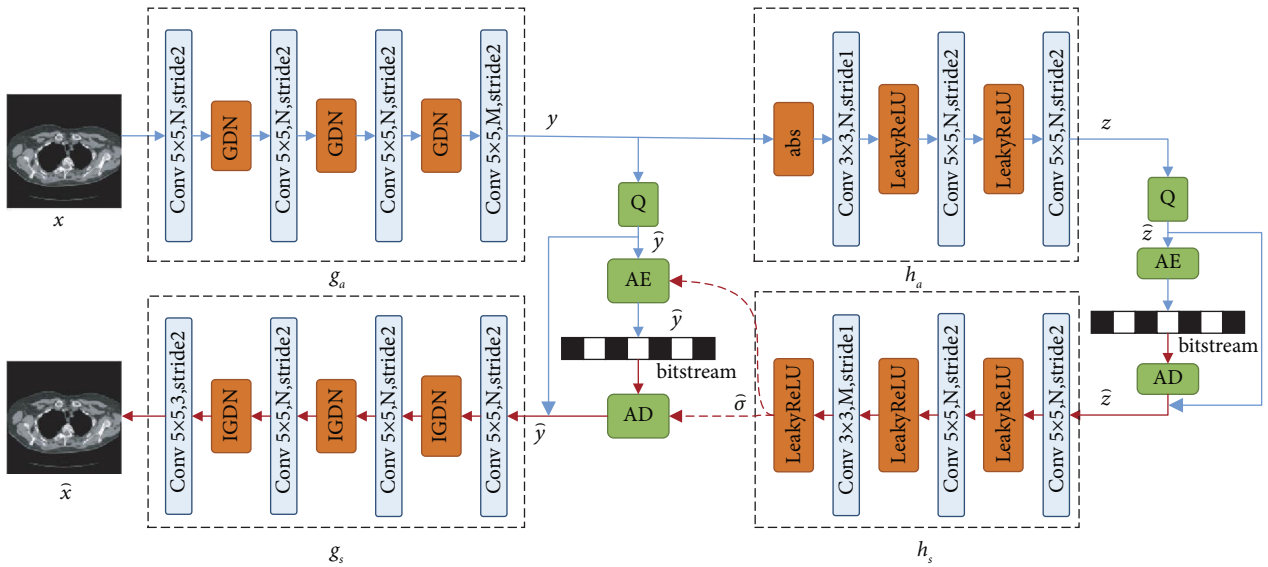

Figure 4: System neural network structure diagram.

likelihood function of the recognition model which can be represented as

$$\log\ p(X) = D_{KL}\left(q \| p_{y,z|x}\right) + L(X). \tag{7}$$

The relative entropy of the distribution is

$$D_{KL}\left(q \| p_{y,z|x}\right) = E_{x\sim p_x} E_{y,z\sim q}\left(\log\ q - \log\ p_{x|y}(x|y) - \log\ p_{y|z}(y|z) - \log\ p_z(z)\right). \tag{8}$$

Then, the loss function can be calculated as

$$\text{Loss} = R + \lambda D, \tag{9}$$

where R and D can be expressed as

$$\begin{aligned} R &= E_{x\sim p_x} E_{y,z\sim q}\left(-\log\ p_{y|z}(y|z) - \log\ p_z(z)\right), \\ D &= E_{x\sim p_x}\left(-\log\ p_{x|y}(x|y)\right). \end{aligned} \tag{10}$$

3. Results

3.1. Experimental Conditions and Main Simulation Parameters

3.1.1. Data Set. In this study, we used 90% medical images (TCGA-LUAD lung cancer CT image dataset in the Tumor Genome Atlas (TCGA) database in the United States) [39, 40] and cropped them to 256 × 256 images for training. For performance evaluation, we averaged the image compression performance of the remaining 10% of the test set to obtain rate-distortion performance.

3.1.2. Lab Environment. All experiments in this study were implemented on the server based on an Nvidia GeForce 2080Ti graphics card, and the network architecture based on TensorFlow 1.15 and the python 3.6 platform is used, and the experiments were carried out on the Linux system.

3.1.3. Main Simulation Parameters. This experiment aims to optimize the two indicators of PSNR and MSE for training, respectively. The main simulation parameters of the experiment are shown in Table 1.

3.2. Evaluation Indicators. In the field of medical image compression, image evaluation criteria can be divided into objective evaluation criteria and subjective evaluation criteria. There are two objective evaluation criteria, the degree of compression evaluation and reconstructed image quality evaluation. Image compression efficiency is evaluated by pixel depth (BPP). The quality of the reconstructed images is evaluated by means of PSNR (peak signal-to-noise ratio) and MS-SSIM (multiscale structural similarity). In this study, the

Table 1: The main simulation parameters of the experiment.

Parameter type and meaning	Value
Loss function	Loss = $BPP + \lambda * MSE$
The value of λ	0.005 0.0075 0.01 0.015 0.02 0.04 0.1
Iterations (times)	1 M
Batch size	8
Patch size (px ∗ px)	256 ∗ 256
Learning rate	0.001 (initial) 0.0001 (after stabilizing)

R-D curve is used to evaluate the reconstruction quality of the image under the same degree of compression.

3.2.1. BPP (Bits per Pixel). BPP is the average number of bits occupied by each pixel. The smaller the BPP, the smaller the amount of information contained in each pixel, that is, a greater amount of compression degree [39]. Its equation can be expressed as

$$BPP = \frac{\text{len}}{\text{pixs}}, \tag{11}$$

where len represents the length of the compressed binary data stream and pixs represents the total number of pixels.

3.2.2. PSNR (Peak Signal-to-Noise Ratio). MSE (mean square error) represents a measure of the distance between the estimated image and the original image. The smaller the MSE is, the closer the estimated image is to the original image [41]. It is expressed as

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[X(i,j) - \hat{X}(i,j)]^2. \tag{12}$$

Among them, X and $\hat{X}$ are, respectively, $M \times N$ original images and reconstructed images.

In the process of lossy compression, compared with the original image, the reconstructed image will have certain information distortion, so scholars often use the image fidelity criterion to measure the quality of the image reconstruction [42]. The peak signal-to-noise ratio can be defined as

$$PSNR = 10\log_{10}\left(\frac{X_{\max}^2}{MSE}\right) = 20\log_{10}\left(\frac{X_{\max}}{\sqrt{MSE}}\right), \tag{13}$$

where $X_{\max}$ is the maximum pixel value of the image. It can be seen from equation (13) that the smaller the distortion after the lossy image reconstruction is, the closer the image quality is to the original image.

3.2.3. MS-SSIM (Multiscale Structural Similarity). Multiscale SSIM is a multiscale and a different resolution image quality assessment SSIM method. The overall SSIM evaluation can be obtained by combining the measurement at different scales using

$$\text{SSIM}(x,y) = [l_M(x,y)]^{aM} \cdot \prod_{j=1}^{M}\left[c_j(x,y)\right]^{\beta j}\left[s_j(x,y)\right]^{\gamma j}, \tag{14}$$

where $c_j(x,y)$ and $s_j(x,y)$ represent the contrast comparison $c(x,y) = 2\sigma_x\sigma_y + C_2/\sigma_x^2 + \sigma_y^2 + C_2$ and structure comparison $s(x,y) = \sigma_{xy} + C_3/\sigma_x\sigma_y + C_3$ at the j-th scale, respectively. $l_M(x,y)$ is the luminance comparison $l(x,y) = 2\mu_x\mu_y + C_1/\mu_x^2 + \mu_y^2 + C_1$ only at scale M. $C_1 = (K_1\ L)^2$, $C_2 = (K_2 L)^2$, and $C_3 = C_2/2$ are small constants.

3.3. Acceptable Compression Ratios for Medical Image Compression. The European Society of Radiology (ESR) [43] published recommendations on acceptable compression ratios for lossy medical image compression, and the acceptable compression ratios for different investigations, body parts, and diseases of medical images based on JPEG compression algorithms are listed. At this compression ratio, the error of medical images based on the JPEG compression algorithm is the maximum error of lossy medical image compression. The error of the compression algorithm proposed in this study should not exceed the maximum error of lossy medical image compression, and the compression ratio is the maximum acceptable compression ratio of the proposed compression algorithm.

3.4. Analysis of Simulation Conclusions

3.4.1. Objective Analysis. The analysis is performed in order to confirm the effectiveness of the residual module in the network training process. Under the same experimental conditions, the relationship between the loss value and the training rounds during the training process is shown in Figure 5.

It can be seen from Figure 5 that the overall trend of the neural network loss value (ordinate) in the training process is smaller than that of the Balle 2018 neural network structure. And compared with Balle 2018, the neural network in this study converges faster, and the downward trend is obvious. After rapidly converging before 400 k, it gradually stabilizes and it has a smaller oscillation range, which alleviates the problem of gradient explosion to a certain extent. Under the same experimental conditions, the training time of the network in this study is shorter than that of Balle 2018. Performance for that the training time of the Balle 2018 network is 31 hours, while the training time of the network in this study is 29 hours.

In order to study the influence of different components of the network on the network performance, we conducted verification experiments on the side information network module and the residual module. In order to verify the effectiveness of the side information network module, we conducted a comparison experiment between BLSour (without the side information network module) and ours; in order to illustrate the effectiveness of the residual module,

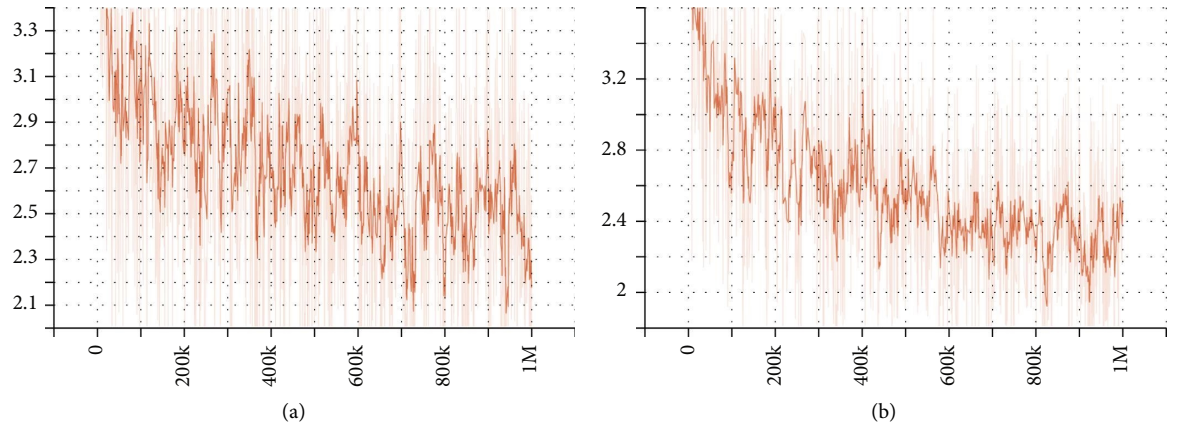

Figure 5: The relationship between the loss function value and the training round during the training process. (a) Ballé 2018 and (b) the proposed model.

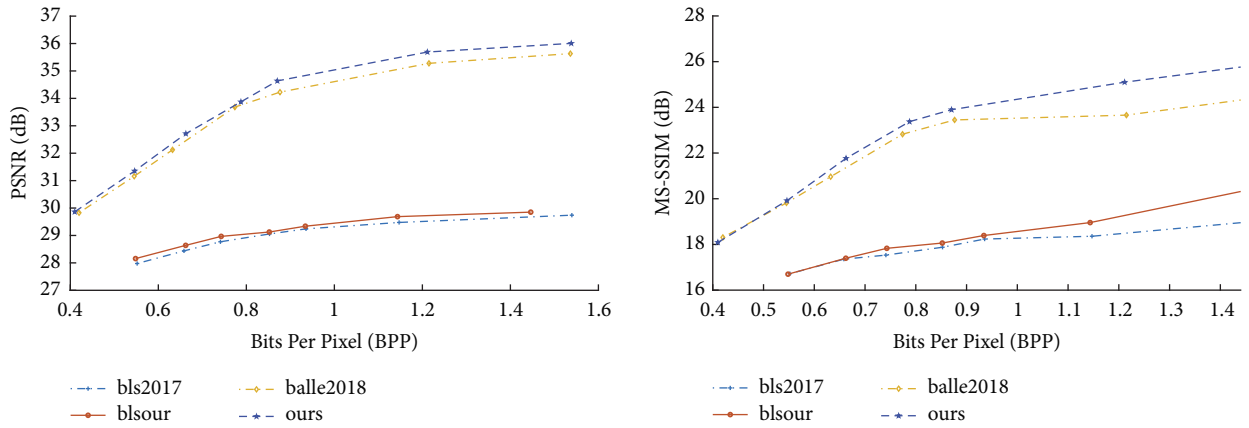

Figure 6: The PSNR-BPP and MS-SSIM-BPP results based on deep-learning compression algorithms.

we conducted a comparison between Balle 2018 (without the residual module) and our experiment.

To illustrate the effectiveness of the proposed network, we use different network structures such as BLS2017, Balle 2018, and BLSour to conduct comparative experiments on medical image data compression. And the difference between BLSour and the algorithm in this study is that it does not include the side information network module. Residual modules are added to the BLSour network and the algorithm in this study, but the BLS2017 and Balle 2018 networks do not add residual modules. To improve the readability of the MS-SSIM, the MS-SSIM was converted to decibels ($-10\log_{10}(1 - MS - SSIM)$). It can be seen from Figure 6 that the PSNR and MS-SSIM values increase with the increase of bpp, which is due to the fact that the larger the compression of the bit depth is, the smaller the compression degree and distortion are. By comparing the results of the BLSour network and BLS2017 network, the results of ours and the Balle2018 network respectively, it can be noticed that the side information network module is a very important part of the network structure, which can effectively extract the detailed information in medical images and achieve better image quality. It can be seen from Figure 6 that the PSNR and MS-SSIM values of BLSour are better than those of the BLS2017 network, and the PSNR and MS-SSIM values of our network are better than those of the Balle 2018 network, which proves that under the same degree of compression, the residual network module added in this study reduces the information loss in the compression process and improves the reconstruction effect.

The comparison results of the compression performance between the method in this study and the traditional algorithm are, respectively, shown in Figure 7. The traditional compression algorithms mainly include SPIHT (set partitioning in hierarchical trees), coiflet (global thresholding of coefficients and Huffman encoding), and JPEG algorithms. Figure 7 shows that under the same degree of compression, the reconstruction effect of the proposed algorithm is significantly better than that of other traditional algorithms for medical image compression. When the compression degree of the JPEG algorithm is small, the reconstruction effect is slightly lower than that of the algorithm in this study, but when the compression degree increases, the distortion of the JPEG gradually increases, and the reconstruction effect is much lower than that of the method proposed in this study. The experimental results show that the method proposed in this study has obvious advantages over traditional methods, and it can learn the detailed information in medical images

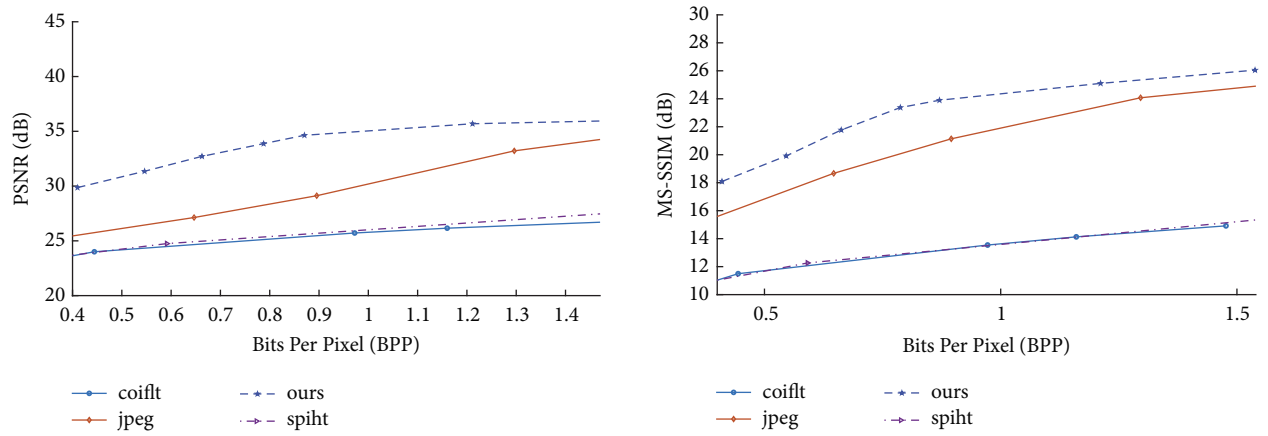

Figure 7: The PSNR-BPP and MS-SSIM-BPP results based on traditional compression algorithms.

Table 2: The main simulation parameters of the experiment.

Parameter type and meaning	Value
Loss function	$\text{Loss} = BPP + \lambda * MSE$
The value of λ	0.005 0.0075 0.1 0.2 0.5 1 8
Iterations (times)	1 M
Batch size	8
Patch size (px ∗ px)	256×256
Learning rate	0.001 (initial) 0.0001 (after stabilizing)

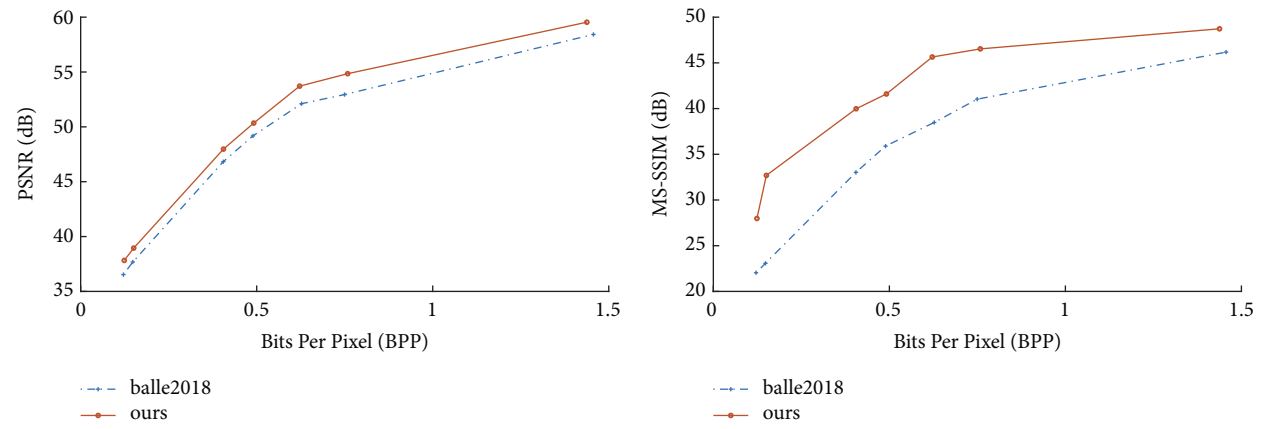

Figure 8: The PSNR-BPP and MS-SSIM-BPP results based on the Chaos-MR dataset.

through training and can obtain high-quality medical images under a high compression ratio.

To confirm the generalizability of the method proposed in this study, the Chaos-MR dataset (dataset of Combined (CT-MR) Healthy Abdominal Organ Segmentation) [44] was used to train and test. It was acquired by 1.5T Philips MRI, producing 12-bit DICOM images with a resolution of 256×256 pixels and includes 120 DICOM datasets from two different MRI sequences that scan the abdomen using different combinations of radio frequency pulses and gradients. The ISD varies from 5.5–9 mm (average 7.84 mm), the xy spacing is 1.36–1.89 mm (average 1.61 mm), and the number of slices ranges from 26 to 50 (average 36). 1594 slices (532 slices per sequence) were used as the training set and 1537 slices were used as the test set. The main simulation parameters of the experiment are shown in Table 2.

Figure 8 shows the PSNR-BPP and MS-SSIM-BPP results of the proposed method and the Balle 2018 network, respectively. It can be seen from Figure 8 that the PSNR and MS-SSIM values of the method proposed in this study are higher than those of Balle 2018. As the BPP gradually increases, the compression performance of the method proposed in this study is also gradually enhanced compared with the Balle 2018 network. The results verify that the

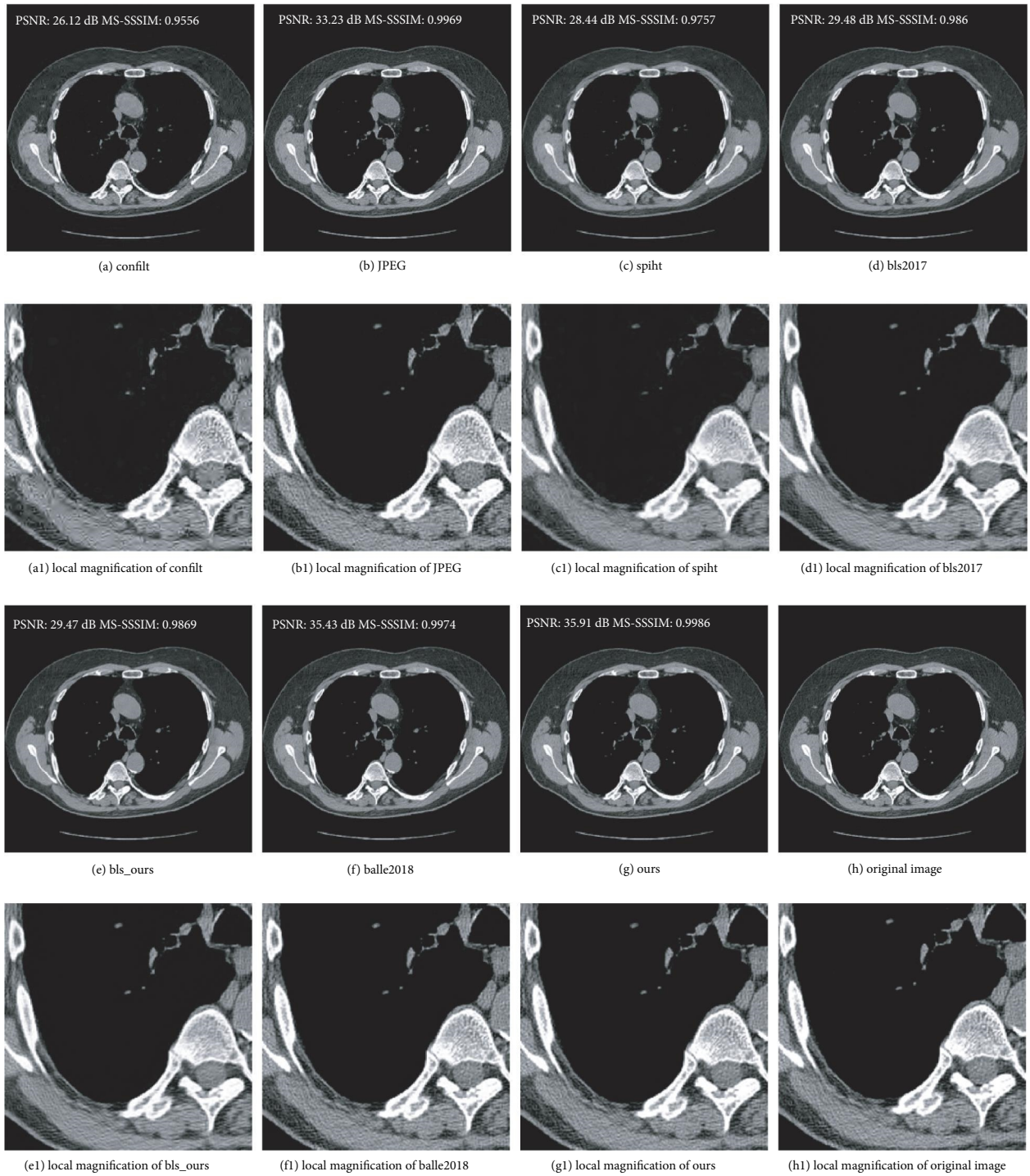

Figure 9: Subjective comparison experiment results. Experimental conditions: $bpp \approx 1$.

proposed method can also obtain compression performance similar to that of the TCGA-LUAD dataset in the Chaos-MR dataset, and it has a good generalization effect.

3.4.2. Subjective Analysis. Figure 9 shows the results of the subjective comparison experiment in the case of $bpp \approx 1$, in which Figures 9(a)–9(g) are reconstructed images using compression algorithms such as coiflet, JPEG, SPIHT, BLS2017, BLSours, Balle 2018, and the proposed algorithm, respectively. And Figure 9(h) is the original image before compression. Figure 9 shows the partial enlargement corresponding to Figures 9(a)–9, respectively. It can be seen from Figures 9(a)–9(g) that under the same compression ratio, the PSNR and MS-SSIM values of the reconstructed image by the deep learning algorithm are higher and closer to the original image. At the same compression ratio, the reconstructed images by the deep

learning algorithms have more detailed information than that of the traditional methods in Figure 9. Subjective comparative experiments demonstrate the effectiveness of the proposed method. The proposed method has higher reconstructed image quality at the same compression ratio.

4. Discussion

There is a problem with increasing data volume in modern medical images. Under the limitation of limited bandwidth and storage space, storing and transmitting more medical image data is a major problem at present. In this study, an end-to-end compression model based on variational autoencoders was introduced into the field of medical image compression. And the structure was similar to the compression algorithm proposed by Balle 2018. The direct transfer of information is carried out by introducing the residual network structure while alleviating the deep network problems during training. Compared with traditional medical image compression algorithms, the algorithm in this study can directly optimize the problem of rate distortion.

However, the method in this study effectively alleviated these problems, but there are still some shortcomings. The network model trained in this study was large, so it had certain requirements for the medical storage hardware equipment. However, the size of the trained neural network model is still very small compared to the size of the medical image data that grows in series. By directly compressing or decompressing the medical image with the trained model, it brings great convenience to the storage and transmission of medical images. In the future, the trained network model can still be compressed to save more storage space.

5. Conclusions

With the sharp increase in the amount of medical image data, under the circumstance of transmission bandwidth, how to effectively compress the image to ensure the effective transmission and storage of the data has become an increasingly challenging task. In this study, a medical image data compression method based on a variational autoencoder is studied. First, the basic theoretical knowledge of variational autoencoder and residual network are introduced and derived. Second, combined with the characteristics of more details of medical images, a network framework with residual network structure is proposed. It proposed to integrate the original data with arithmetic decoded data through the residual network to avoid the loss of a large amount of information. Finally, the experimental results showed that under the same experimental environment and experimental conditions, compared with the compression algorithm introduced by Ballé, the rate-distortion effect of the algorithm in this study is improved by nearly 10% on an average. It can also be seen that the algorithm after adding the residual network structure learning can better reconstruct the original medical image.

Data Availability

In this study, the medical images (TCGA-LUAD lung cancer CT image dataset in the Tumor Genome Atlas database in the United States) can be downloaded from (https://wiki.cancerimagingarchive.net/display/Public/TCGA-LUAD).
CHAOS is a challenge that aims the segmentation of abdominal organs (liver, kidneys, and spleen) from CT and MRI data. It can be downloaded from https://chaos.grand-challenge.org/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project was supported by key research and development projects of Jilin Province Science and and Technology Department (2021021042GX), China.

References

[1] C. T. Selvi, J. Amudha, and R. Sudhakar, "Medical image encryption and compression by adaptive sigma filterized synorr certificateless signcryptive Levenshtein entropy-coding-based deep neural learning," *Multimedia Systems*, vol. 27, no. 6, pp. 1059–1074, 2021.

[2] S. N. Kumar, A. Ahilan, A. K. Haridhas, and J. Sebastian, "Gaussian Hermite polynomial based lossless medical image compression," *Multimedia Systems*, vol. 27, pp. 15–31, 2020.

[3] S. Albahli, T. Nazir, A. Irtaza, and A. Javed, "Recognition and detection of diabetic retinopathy using densenet-65 based faster-RCNN," *Computers, Materials & Continua*, vol. 67, no. 2, pp. 1333–1351, 2021.

[4] F. Fischer, M. A. Selver, S. Gezer, O. Dicle, and W. Hillen, "Systematic parameterization, storage, and representation of volumetric DICOM data," *Journal of Medical and Biological Engineering*, vol. 35, no. 6, pp. 709–723, 2015.

[5] E. Aldemir, G. Tohumoglu, and M. A. Selver, "Binary medical image compression using the volumetric run-length approach," *The Imaging Science Journal*, vol. 67, no. 3, pp. 123–135, 2019.

[6] E. Aldemir, N. S. Gezer, G. Tohumoglu et al., "Reversible 3D compression of segmented medical volumes: usability analysis for teleradiology and storage," *Medical Physics*, vol. 47, no. 4, pp. 1727–1737, 2020.

[7] I. Urbaniak and M. Wolter, "Quality assessment of compressed and resized medical images based on pattern recognition using a convolutional neural network – Science Direct," *Communications in Nonlinear Science and Numerical Simulation*, vol. 95, 2020.

[8] A. Jeromel and B. Zalik, "An efficient lossy cartoon image compression method," *Multimedia Tools and Applications*, vol. 79, no. 1-2, pp. 433–451, 2020.

[9] M. Wu, Z. He, X. Zhao, and S. Zhang, "General generative model-based image compression method using an optimisation encoder," *IET Image Processing*, vol. 14, no. 9, pp. 1750–1758, 2020.

[10] L. Li, V. Muneeswaran, S. Ramkumar, and G. R. Gonzalez, "Metaheuristic FIR filter with game theory based compression technique- A reliable medical image compression technique for online applications," *Pattern Recognition Letters*, vol. 125, pp. 7–12, 2019.

[11] C. Fu, Y. Yi, and F. Luo, "Hyperspectral image compression based on simultaneous sparse representation and general-pixels," *Pattern Recognition Letters*, vol. 116, pp. 65–71, 2018.

[12] L. Zhu, X. Luo, C. Yang, Y. Zhang, and F. Liu, "Invariances of JPEG-quantized DCT coefficients and their application in robust image steganography," *Signal Processing*, vol. 183, no. 2, Article ID 108015, 2021.

[13] J. John, "Discrete cosine transform in JPEG compression," *Computer science*, 2021.

[14] J. Wang, W. Huang, X. Luo, Y. Q. Shi, and S. K. Jha, "Non-aligned double JPEG compression detection based on refined Markov features in QDCT domain," *Journal of Real-Time Image Processing*, vol. 17, no. 1, pp. 7–16, 2020.

[15] B. Khalid, M. Majid, I. F. Nizami, S. M. Anwar, and M. Alnowami, "EEG compression using motion compensated temporal filtering and wavelet based subband coding," *IEEE Access*, vol. 8, pp. 102502–102511, 2020.

[16] M. Tiwari, "STW and SPIHT wavelet compression using MATLAB wavelet tool for color image," *Electrical Engineering and System Science*, 2020.

[17] S. P. Raja, "Wavelet-based image compression encoding techniques— a complete performance analysis," *International Journal of Image and Graphics*, vol. 20, no. 02, Article ID 2050008, 2020.

[18] J. Miya and M. A. Ansari, "Wavelet techniques for medical images performance analysis and observations with EZW and underwater image processing," *Wireless Personal Communications*, vol. 116, pp. 1259–1272, 2020.

[19] T. Brahimi, F. Khelifi, F. Laouir, and A. Kacha, "A new, enhanced EZW image codec with subband classification," *Multimedia Systems*, vol. 28, pp. 1–19, 2021.

[20] E. Stylianou, C. D. Charalambous, and T. Charalambous, "Joint rate distortion function of a tuple of correlated multivariate Gaussian sources with individual fidelity criteria," *Computer Science*, vol. 2, 2021.

[21] H. Yuan, Q. Wang, Q. Liu, J. Huo, and P. Li, "Hybrid distortion-based rate-distortion optimization and rate control for H.265/HEVC," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 2, pp. 97–106, 2021.

[22] K. Padmavathi, C. Asha, and V. K. Maya, "A novel medical image fusion by combining TV-L1 decomposed textures based on adaptive weighting scheme," *Engineering Science and Technology, an International Journal*, vol. 23, no. 1, pp. 225–239, 2020.

[23] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *ICLR*, 2017.

[24] A. Kar, S. P. K. Karri, N. Ghosh, R. Sethuraman, and D. Sheet, "Fully convolutional model for variable bit length and lossy high density compression of mammograms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Piscataway, NJ, USA, October 2018.

[25] A. S. Sushmit, S. U. Zaman, and A. I. Humayun, "X-ray image compression using convolutional recurrent neural networks," in *Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, IEEE, Chicago, IL, USA, May 2019.

[26] P. K. Pareek, C. Sridhar, R. Kalidoss et al., "IntOPMICM: intelligent medical image size reduction model," *Journal of Healthcare Engineering*, vol. 2022, Article ID 5171016, pp. 1–11, 2022.

[27] G. Toderici, S. M. O'Malley, S. J. Hwang et al., "Variable rate image compression with recurrent neural networks," *Computer Science*, 2015.

[28] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *Electrical Engineering and Systems Science*, 2018.

[29] S. K. Shelke, S. K. Sinha, and G. S. Patel, "Study of end to end image processing system including image de-noising, image compression & image security," *Wireless Personal Communications*, vol. 121, pp. 209–220, 2021.

[30] D. Mishra, S. K. Singh, and R. K. Singh, "Lossy medical image compression using residual learning-based dual autoencoder model," in *Proceedings of the 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, India, November 2021.

[31] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Piscataway, NJ, USA, May 2021.

[32] Y. Blau and T. Michaeli, "Rethinking lossy compression: the rate-distortion-perception tradeoff," *Computer Science*, 2019.

[33] S. Bhinge, Q. Long, V. D. Calhoun, and T. Adali, "Adaptive constrained independent vector analysis: an effective solution for analysis of large-scale medical imaging data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1255–1264, 2020.

[34] L. Li, J. Yan, H. Wang, and Y. Jin, "Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1177–1191, 2021.

[35] K. Huang and X. Wang, "ADA-INCVAE: improved data generation using variational autoencoder for imbalanced classification," *Applied Intelligence*, vol. 52, no. 3, pp. 2838–2853, 2021.

[36] R. D. Oeffner, P. V. Afonine, C. Millán et al., "On the application of the expected log-likelihood gain to decision making in molecular replacement," *Acta Crystallographica, Section D, Structural biology*, vol. 74, no. 4, pp. 245–255, 2018.

[37] J. Zhang and E. Sampson, "The mean relative entropy: an invariant measure of estimation error," *The American Statistician*, vol. 75, no. 2, pp. 117–123, 2021.

[38] Y. Shao, J. Lan, Y. Liang, and J. Hu, "Residual networks with multi-attention mechanism for hyperspectral image classification," *Arabian Journal of Geosciences*, vol. 14, no. 4, pp. 252–319, 2021.

[39] B. Albertina, M. Watson, and C. Holback, "Radiology data from the cancer Genome Atlas lung adenocarcinoma [TCGA-LUAD] collection," *The Cancer Imaging Archive*, 2016.

[40] K. Clark, B. Vendt, K. Smith et al., "The cancer imaging archive(TCIA): maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[41] S. Theodoridis, "Mean-square error linear estimation – ScienceDirect," *Machine Learning*, pp. 121–177, Elsevier, Netherlands, 2020.

[42] W. Zhang, H. Zhao, and A. Hu, "Research on rail image preprocessing method based on peak signal-to-noise ratio standard," *Journal of Hunan University of Arts and Science (Natural Science Edition)*, vol. 3, no. 3, 2019.

[43] European Society of Radiology Esr, "Usability of irreversible image compression in radiological imaging. A position study by the European Society of Radiology (ESR)," *Insights into Imaging*, vol. 2, no. 2, pp. 103–115, 2011.

[44] A. E. Kavur, N. S. Gezer, M. Baris et al., "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, Article ID 101950, 2021.