Tumor Type Prediction based on Residual Attention Model

Y H WANG¹, Y TIE¹, L QI¹, F WANG², L WANG³
1 School of Information Engineering, Zhengzhou University, Zhengzhou 450000, Henan, China.
2 The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450000, Henan, China.
3 Center for Dentistry Henan Provincial People’s Hospital, Zhengzhou 450000, Henan, China.

Email: 141851356@qq.com

Abstract. Early detection of tumors is an important part of cancer treatment. In view of the existing algorithms: single data types, low feature extraction efficiency, and low classification network accuracy. A tumor types prediction model based on deep learning is proposed. The network model uses a Variational auto-encoder (VAE) to fuse the RNA expression and DNA methylation data of 32 tumor types, then uses the Hilbert curve to visualize it. Finally fuse data is sent to classification module: embed the attention module in the backbone network ResNet18 framework, convolutional layer instead of fully connected layer. The new sample is used to predict the tumor type. The experimental results show that this network model has excellent performance in tumor type classification and has important guiding significance for the early diagnosis of tumor patients.

Keywords: tumor types prediction; Variational auto-encoder; Data Fusion; Hilbert curve; attention mechanism; full convolution

1. Introduction
With the development of the second-generation high-throughput sequencing technology, more and more biological data is conveniently and quickly available to scientific researchers. Researchers can obtain biological data in a shorter time, and the effectiveness of the data is greatly improved compared to the first-generation sequencing technology. Biological data has general characteristics of high dimensionality, high noise, low samples, and complex relationships between different feature points [1]. For example, the dimensionality of RNA expression usually reaches tens of thousands, while the sample size is relatively small. This phenomenon is called the "dimension disaster" in the field of machine learning [2]. In order to overcome the "dimension disaster", researchers have done a lot of related work. For example: the use of traditional machine learning methods to reduce dimensionality, reduce the impact of "dimensionality disaster" on subsequent research [3-5]. However, due to the close connection between adjacent feature points of biological data, the abnormality of a certain feature point may cause different degrees of body cancer, and general dimensionality reduction algorithms cannot effectively extract its important features [6-8]. In order to detect one-dimensional data more intuitively and reduce the damage to data integrity, Keim first proposed the concept of using Hilbert curve to perform dimensional conversion on stock trends [9]. Later, Simon A used Hilbert curve to map the randomly generated one-dimensional long vector into a two-dimensional matrix, and realized the visualization of the data [10]. In the application of biological data: Chen H used different
algorithms of non-negative matrix factorization (NMF) to perform dimensionality reduction processing on gene expression data respectively, and used the attention mechanism to fuse the processed data, and use the fused data to complete the tumor type. The experiment has achieved good result [11]. Gao Yuanyuan used traditional feature extraction methods: principal component analysis (PCA), kernel principal component analysis (KPCA) and deep learning methods: autoencoder (AE) combined to reduce the dimensionality of gene expression data. The data is more conducive to the processing of subsequent models [12]. Yu X and others used bioinformatics software to analyze the structural characteristics and physicochemical properties of the protein in the transcription and translation of gene sequences[13]. Existing deep learning algorithms often have the following shortcomings in processing biological data. (1): Traditional dimensionality reduction methods cannot effectively extract biological features. (2): The existing algorithms usually only involve one type of data when performing tumor-related analysis, but the internal structure of the organism is complex, and tumor formation is affected by many factors, so a single data cannot effectively explain its cause [7]. (3): Due to the particularity of biological data, general classification models cannot effectively mine its key information. Based on the above shortcomings. This article proposes its own model. Its work is mainly divided into two parts (1) Data preprocessing: First, the data is logarithmic transformed and the missing feature points are removed. The feature points of each sample are reordered according to the chromosome number, so that adjacent feature points have the characteristics of local similarity, so as to give full play to the performance of the classification module in the future. After reordering, a VAE was used to effectively fuse the two data of RNA expression and DNA methylation. The Hilbert curve is used to complete the one-dimensional to two-dimensional mapping of the fused data, and the fused data is mapped from the one-dimensional long vector to the two-dimensional visualization form. (2) Classification module: The generated two-dimensional data is sent to the classification module. Embed the attention module in the backbone network ResNet18 framework. And the idea of fully convolutional structure is integrated into this network, the convolutional layer is used to replace the fully connected layer to optimize the network. The network structure flow chart of this text is shown as in Figure 1. It is mainly composed of three parts: Variational autoencoder, Hilbert curve, and classification module.

![Network structure flow chart](image)

**Figure 1.** The network structure flow chart of this text
2. Method

2.1. Data
In this article, we obtained the RNA expression and DNA methylation data of 19,683 tumor samples from the TCGA database, including 32 tumor types. Combine the two kinds of data as the input of the network.

2.1.1 RNA expression. The original sample form of RNA expression is roughly as shown in Table 1. OR-1 represents the sample number, a, b . . . represents sample feature points.

| OR-1 | a     | b     | c     | d     | ... |
|------|-------|-------|-------|-------|-----|
| OR-1 | 17.5  | 15.00 | 12.33 | 11.29 | ... |

It can be seen from the schematic diagram of the biological center law in Figure 2: that RNA is a bridge in the transmission of genetic information. The translated protein is mainly used for the construction of its own cells. The level of RNA expression can mainly indicate whether cell division is in a vigorous stage at this time. If it is in a state of vigorous division, the RNA expression is higher, otherwise the RNA expression is lower. The level of RNA expression will affect subsequent protein production. Generally, cancer cells have a short growth cycle and vigorous division. They are always in a vigorous state of division during the growth phase, and there will be varying degrees of abnormality in RNA expression. The level of RNA expression largely reflects the difference between different cells [14].

![Figure 2. Schematic diagram of Biological Center Law](image)

2.1.2 DNA methylation. The data format of DNA methylation is similar to that of RNA expression. As shown in Table 1, DNA methylation is one of the earliest DNA modification pathways discovered by scientists. Under the catalysis of methyltransferase, the cytosine of the two nucleotides of CG of DNA is selectively added with methyl groups to form 5-methylcytosine. A large number of studies have confirmed that DNA methylation can cause changes in chromatin structure and DNA stability, thereby controlling the expression of genes in the body. Studies have shown that hypermethylation or hypomethylation of certain genes may lead to the occurrence of different types of tumors [15]. As shown in Figure 3.
2.1.3 Data preprocessing. The main purpose of data preprocessing is to remove noise and offset signals. For RNA expression: remove exons located on the Y chromosome (n=594); exons with 0 expression (n=1904); exons with a deletion rate greater than 10% in all samples (n=248). The number of remaining exons is 58043. For DNA methylation: remove probes that cannot match the human reference genome (n=89512); remove probes that cannot be located on the chromosome (n=2545); remove probes located on the Y chromosome (n=346) and probes with a deletion rate greater than 10% (n=414). The number of remaining probes is 392,761. Since the feature points on the same chromosome are similar in size, the feature points of the two types of data are reordered according to the chromosome number, so that the feature point values of the one-dimensional data have the characteristics of local similarity.

2.2 Variational autoencoder

Variational autoencoder (VAE) is a powerful deep generative model that can learn meaningful data from high-dimensional data [16]. In this paper, VAE is used to reduce the dimensionality of the two data of DNA methylation and RNA expression, extract important features, and effectively merge the two data for the first time. Assuming that the input data is \( x = \{x_1, x_2, x_3, \ldots, x_n\} \). VAE hopes to obtain the distribution of \( x \), so as to obtain the value of \( x \). VAE uses the hidden layer \( z \) to calculate the distribution of \( x \). Its implementation method as shown in the following formula (1):

\[
\int p(x|z)p(z)\,dz \tag{1}
\]

\( p(x|z) \) can be expressed as the process of \( z \) generating \( x \). In the encoding part, the posterior distribution \( p(z|x) \) of \( x \) is obtained. For each sample \( x \), it is considered that there is a posterior distribution \( p(z|x) \), and it is assumed that it obeys the standard normal distribution, which needs to pass the continuous encoder \( \text{Learn} \) to fit it out. In the training process, \( u \) is infinitely close to 0, and \( \sigma^2 \) is infinitely close to 1. In order to make \( p(z|x) \) obey the standard normal distribution, the relative entropy (kl divergence) is used to optimize it, and the relative entropy is shown in formula (2). The smaller the value, the more similar the two probability distributions.

\[
\text{kl}(p(z|x)||\text{N}(0,1)) \tag{2}
\]

The process of minimizing relative entropy can be expressed as formula (3):

\[
L_{kl}=\text{min}(\text{kl}(p(z|x)||\text{N}(0,1))) \tag{3}
\]

Formula (3) can be transformed into formula (4) after calculation:

\[
L_{kl}=\text{min}\frac{1}{2}\sum_{i=1}^{d}(\mu_i^2 +\sigma_i^2 - \log \sigma_i^2 - 1) \tag{4}
\]

In formula (4), \( d \) represents the hidden layer dimension, and \( \mu_i^2 \) and \( \sigma_i^2 \) respectively represent the mean and variance of the normal distribution obeyed by the generated \( p(z|x_i) \) corresponding to the sample \( x_i \). \( p(z|x) \) obeys the normal distribution. \( z \) is sampled from \( p(z|x) \), so \( z \) obeys a normal distribution. In order to sample \( z \), VAE uses heavy parameter techniques to transform \( z \), as shown in the following formula (5):

\[
Z \sim \text{N}(\mu, \sigma^2) \Rightarrow \frac{Z-\mu}{\sigma} \sim (0,1) \tag{5}
\]
Assuming that $\beta$ obeys the standard normal distribution, set $(z-\mu)/\sigma=\beta$ to obtain $z=\sigma\beta+\mu$, and sample $\beta$ to obtain $z$. After obtaining the value of $z$, the input data is reconstructed, and the objective function of the reconstruction process can be expressed by the following formula (6):

$$L_{VAE}=\alpha L_{lk}+E_{VAE}(x,x')^{\gamma}+\text{regularization} \quad (6)$$

$\alpha$ and $\gamma$ are the hyperparameters that control $L_{lk}$ and the regular term, respectively. The optimization process of the variational autoencoder model is realized by minimizing $L_{VAE}$. $E_{VAE}(x,x')$ represents the reconstruction error between $x$ and $x'$, which can be expressed by formula (7) square error or formula (8) cross entropy:

$$E_{AE}(x,x')=|x-x'| \quad (7)$$

$$E_{AE}(x,x')=\sum_{i=1}^{n}(x_i \log x_i'+(1-x_i) \log (1-x_i')) \quad (8)$$

The specific structure of VAE used in this article is shown in Figure 4. DNA methylation data is mainly divided into dimensionality reduction and data fusion. For RNA expression, VAE mainly plays a role of data fusion.

2.3 Hilbert curve

The Hilbert curve was made by the mathematician Hilbert. Its construction method is [9]: When the number of iterations is 1, divide a square into 4 equal small squares, and then start from the small square in the lower left quadrant to the small square in the lower right quadrant, and connect the centers of the small squares with line segments in turn; when the number of iterations is 2, divide each small square into 4 equal smaller squares, and then connect their centers in the above way, and continue this operation infinitely, and finally the curve of the limit situation can fill the entire plane. Experiments show that the adjacent feature points in the one-dimensional form are still in adjacent positions in the two-dimensional form, the pixel values in the two-dimensional picture have local similarities. Figure 5 shows the Hilbert curve with different iteration times, from the upper left to the lower right, the number of iterations is 1-4.
It is experimentally verified that even if a certain point in the two-dimensional image obtained by Hilbert curve transformation has different iteration times, its position in the picture hardly changes. As the iteration times increase, a certain point tends to be a fixed. The location remains unchanged. This means that even if the number of iterations changes, the subsequent classification module does not need to be retrained. We track a point in one dimension and find that its position is almost the same in different iterations. As shown in Figure 6, the position of the point is marked in the image with the number of iterations of 7(a) and 8(b).

![Figure 6. The position of a certain pixel in different iteration times](image)

After the data is fused by VAE, the Hilbert curve is introduced to map the fusion data of 19,683 samples from a one-dimensional long vector to a two-dimensional form, and different pixel values cover the two-dimensional space. The number of iterations is 8, and the resulting picture resolution is 256*256. Each value in one dimension represents a pixel value in two dimensions. Figures 9 and 10 below are two-dimensional example images of esophageal cancer and colon cancer. Label the different tumor types as 1-32. Finally, a random sampling method is used to randomly divide 19,683 sample data into training set and test set.

![Figure 7. example of esophageal cancer](image)

![Figure 8. example of colon cancer](image)

It can be seen intuitively from Figure 7 and Figure 8 that the difference between the two figure is mainly concentrated in the middle right part of the figure.

2.4 Classification module

2.4.1 Residual network. Gao[17] proposed the residual network for applied to facial expression recognition. Experiments show that residual network is easier to optimize and converge faster. The
more common residual network has 18 layers, 34 layers, and 50 layers. After experimental comparison, this article selects ResNet18. Its network structure is shown in Figure 9.

![Figure 9. ResNet18 structure diagram](image)

There are two main designs of residual elements, direct mapping and residual part. As shown in Figure 10, the left side is the direct mapping, and the right side is the residual part. Direct mapping makes the residual possible, and the residual part makes the network deeper. The core of the residual network is direct mapping. Its biggest advantage is to strengthen the transfer of features, but the network parameters and computational complexity have not increased, and the structure can be transplanted to other networks.

![Figure 10. Residual block](image)

In the figure 10, weight is a convolution operation, and addition is a unit addition operation.

2.4.2 Attention module. With the development and application of computers, attention mechanism has become one of the research hotspots in the field of deep learning. The idea is derived from human vision, the human eye quickly scans to focus on the target area that needs to be focused, and then pays more attention to the area to obtain the required detailed information, while suppressing other information, which is human use of limited resources an ability to quickly filter out valuable information from a large amount of information [18]. The attention mechanism in deep learning simulates this process, that is, when the neural network finds the key information of the input data, it
will focus on it in the subsequent prediction stage through learning. The first application of the attention mechanism was Mnih [19].

This article is based on the original ResNet18, the attention mechanism is added. The characteristics of the attention mechanism enable the model built with the attention mechanism to find the key part of the input, especially in the fine-grained image classification, to find the part with discrimination. Suitable for solving weakly supervised learning problems. The residual attention mechanism used in this paper is a mixed-domain attention mechanism, that is, attention mechanism pays attention to the spatial domain and the channel domain at the same time. The residual attention mechanism is mainly composed of a stack of multi-layer residual attention modules. The residual attention module is a stackable attention structure, similar to the residual block in the residual network, and can be well embedded in the network. It mainly contains two branches: mask branch and main branch. In the main branch, two residual blocks are included, and in the mask branch, the maximum pooling layer and the residual block are used in combination. After the two branch processing ends, the two branch features are combined. The final output of the residual attention mechanism is:

\[ P_{i,c}(x) = (1 + B_{i,c}(x)) \times G_{i,c}(x) \]  

(9)

Where: \( G_{i,c}(x) \) is the output of the main branch, \( B_{i,c}(x) \) is the output of the mask branch. The value is in the interval \([0,1]\). \( i \) takes the value at all spatial positions, \( c \in \{1,2,\ldots,c\} \) is the channel index.

Secondly, the convolutional layer is used instead of the fully connected layer to optimize the network. Replacing the fully connected layer with a convolutional layer can reduce the phenomenon of overfitting.

The classification module of this article is shown in Figure 1 (inside the red box): The structure of each combination block is similar, including residual unit and attention module. The difference is that the number of convolutional layers of the residual unit is different, which are 64, 128, and 256 respectively. The feature is extracted from the residual unit first, and the output of the residual unit is used as the input of the attention module. After 3 combination blocks, a residual unit is connected. The residual unit structure here is the same as the residual unit structure in the combined block, except that the number of convolution kernels is different. The number of convolution kernels in each convolution layer of the residual unit is 512, and the feature map size is 1×1. After getting the output, send the output to the dropout layer with a deactivation probability of 0.5. Then classify through a convolutional layer of size 1×1. The final output layer number is 32.

3 Experimental results and analysis

The performance indicators for deep learning generally include Accuracy (A), Precision (P), Recall (R), P-R curve, Balance point, f1-score (F1). [20] This paper uses the ten-fold cross-validation method to calculate the values of P, A, R and F1. Among them:

\[ P = \frac{TP}{TP+FP} \]  

(10)

\[ R = \frac{TP}{TP+FN} \]  

(12)

\[ A = \frac{TP+TN}{All\,Samples} \]  

(11)

\[ F_{1} = \frac{2PR}{P+R} \]  

(13)

Among them, TP, FN, FP, TN represent real cases, false negative cases, false positive cases and true negative cases, respectively. The data is randomly divided into two parts, one part is used for model training, and the other part is used for performance testing. In order to verify the effectiveness of the classification module proposed in this paper, a comparative experiment was done. The experimental results are shown in Table 2 below.

| Methods       | F1   | A    | P    | R    |
|---------------|------|------|------|------|
| LeNet         | 0.8318 | 0.8322 | 0.8498 | 0.8484 |
| AlexNet       | 0.8551 | 0.8565 | 0.8513 | 0.8491 |
| CaffeNet      | 0.8715 | 0.8719 | 0.8800 | 0.8783 |
| AlexNet-Softmax | 0.8745 | 0.8812 | 0.8901 | 0.8792 |
It can be seen from Table 2 that the method proposed in this paper is superior to several commonly used methods in performance. Compared with the ResNet18 method, adding the attention module to the classification network is more excellent in classification performance. Compared with AlexNet-Softmax, the convolutional layer is used to replace the fully connected layer, which not only reduces the consumption of computing resources, but also greatly improves the performance.

In order to verify the excellent performance of VAE in data fusion, we used different data types and different methods to do comparison experiments of classification accuracy. The experimental results are shown in Table 3.

|                | LeNet  | AlexNet | CaffeNet | AlexNet-Softmax | ResNet18 | Method of this article |
|----------------|-------|---------|---------|-----------------|----------|----------------------|
| RNA expression | 0.8215| 0.8254  | 0.8527  | 0.8759          | 0.8902   | 0.9127               |
| DNA methylation| 0.8176| 0.8179  | 0.8596  | 0.8802          | 0.9105   | 0.9304               |
| Fusion data    | 0.8498| 0.8513  | 0.8800  | 0.8901          | 0.9335   | 0.9662               |

It can be seen from Table 3 that the classification accuracy of the method proposed in this article and the current mainstream methods of fusion data is higher than that of a single data type, which shows that VAE retains the data in the process of data fusion and feature extraction. The two types of data affect the formation of tumors from different levels. Moreover, the classification accuracy of our proposed method is higher than that of other methods, whether in single data or in fusion data. In order to verify the excellent performance of the Hilbert curve in the process of dimension transformation, the accuracy of the two transformation strategies was compared in several methods. The results are shown in Table 4 below. Among them, numpy is a toolkit used by python for dimension conversion.

|                  | LeNet  | AlexNet | CaffeNet | AlexNet-Softmax | ResNet18 | Method of this article |
|------------------|-------|---------|---------|-----------------|----------|----------------------|
| numpy            | 0.8027| 0.8125  | 0.8354  | 0.8493          | 0.8921   | 0.9238               |
| Hilbert curve    | 0.8484| 0.8491  | 0.8783  | 0.8901          | 0.9335   | 0.9662               |

It can be seen from Table 4 that the subsequent classification accuracy of the data processed by the Hilbert curve in several methods is greater than that of numpy. It can be seen that the Hilbert curve can largely retain the original data during the dimensional conversion process. The generated two-dimensional data is more beneficial to subsequent network applications.

4 Conclusion
This paper uses VAE for the first time to effectively fuse the two data of DNA methylation and RNA expression, and for the first time uses Hilbert curve to complete the one-dimensional to two-dimensional mapping of the fused data. In the classification module, add attention mechanism and full convolution idea on the basis of ResNet18. The experiment showed excellent performance and provided more accurate guidance for the clinical diagnosis of tumors. This article uses two kinds of data for fusion to improve the classification performance. If it can be combined with the patient’s clinical data such as patient age, gender, the classification performance may be further improved.

5. References:
[1] Ma J J 2019 Research on feature selection strategy of high-dimensional biomedical data set based on intelligent algorithm (Kaifeng: Henan University)
[2] Feng J B 2016 Research on high-dimensional data clustering algorithm based on active learning (Fuzhou: Fuzhou University)
[3] Zhao Z Y 2020 Research on dimensionality reduction optimization of high-dimensional data sets (Neimenggu: Neimenggu University)

[4] Hu S X and Song Z H 2020 Research on big data classification method of complex attributes based on data dimensionality reduction vol 2 (Henan Science and Technology) p 18-20

[5] JUAN L 2020 The curse of dimensionality in inverse problems vol 369 (Journal of Computational and Applied Mathematics) p 571

[6] Shi Y, Dong T 2020 Research progress in the application of circulating tumor DNA in renal cell carcinoma vol 9 (Laboratory Medicine) p 952-956

[7] Zhang Y L and Li S B and Wang Y 2020 Expression of LZTS2 protein in thyroid carcinoma and its relationship with BRAFV600E gene mutation vol 9 (Journal of Clinical and Experimental Pathology) p 1057-1061

[8] Zhu L and Wang Y 2020 Research progress on the relationship between DNA methylation and bladder cancer vol 5 (Southwestern Military Surgeon) p 424-428

[9] Daniel A 1996 Pixel oriented visualization techniques for exploring very large data vol 5 (Journal of Computational and Graphical Statistics) p 58-77

[10] Simon A 2009 Visualising very long data vectors with the Hilbert curve (UK: European Bioinformatics Institute)

[11] Chen H and Gao M and Zhang Y 2019 Attention based multi-NMF deep neural network with multimodality data for breast cancer prognosis model vol 1 (Bio-Med Research International) p 1-11

[12] Gao Yuanyuan 2018 Cancer genome data clustering and cancer immune analysis based on deep learning (Wuhan: School of Electronic Information and Communication Engineering, Huazhong University of Science and Technology)

[13] Yu X and Yu M L 2020 ACE2 protein structure and function bioinformatics analysis vol 42 (Ningxia Medical Journal) p 696-671

[14] Ran H 2019 The heredity of life vol z1 (Inside and Outside the Classroom (Science Fans)) p 64

[15] Wang P and Zhao H L and Shi R F 2020 The role of CDO1 methylation in plasma in the early diagnosis of lung cancer vol 5 (Chinese Journal of Lung Cancer) p 314-320

[16] Zhai Z L and Liang Z M and Zhou W 2019 Overview of variational autoencoder model vol 3 (Computer Engineering and Applications) p 1-9

[17] Gao T 2020 Research on facial expression recognition based on deep residual network vol 28 (Electronic Design Engineering) p 101-104

[18] Kastner S and Ungerleider L G 2000 Mechanisms of Visual Attention in the Human Cortex vol 23 (Annual Review of Neuroscience) p 315-341

[19] Mnih V and Heess N and Graves A 2014 Recurrent Models of Visual Attention (Montreal, Canada: Proceedings of the Conference of Neural Information Processing Systems 2014)

[20] Zhang Y H 2018 The beauty of deep learning (Beijing: China Industry and Information Technology Publishing Group)