Hindawi
Journal of Healthcare Engineering
Volume 2021, Article ID 6699996, 9 pages
https://doi.org/10.1155/2021/6699996

Research Article

Deep Learning Method for RNA Secondary Structure Prediction with Pseudoknots Based on Large-Scale Data

Bowen Shen,[1,2] Hao Zhang ,[1,2] Cong Li,[1,2] Tianheng Zhao,[1,2] and Yuanning Liu[1,2]

[1]*College of Computer Science and Technology, Jilin University, Changchun, China*
[2]*Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, Changchun, China*

Correspondence should be addressed to Hao Zhang; zhangh@jlu.edu.cn

Received 2 December 2020; Accepted 17 February 2021; Published 25 February 2021

Academic Editor: Saverio Maietta

Traditional machine learning methods are widely used in the field of RNA secondary structure prediction and have achieved good results. However, with the emergence of large-scale data, deep learning methods have more advantages than traditional machine learning methods. As the number of network layers increases in deep learning, there will often be problems such as increased parameters and overfitting. We used two deep learning models, GoogLeNet and TCN, to predict RNA secondary results. And from the perspective of the depth and width of the network, improvements are made based on the neural network model, which can effectively improve the computational efficiency while extracting more feature information. We process the existing real RNA data through experiments, use deep learning models to extract useful features from a large amount of RNA sequence data and structure data, and then predict the extracted features to obtain each base's pairing probability. The characteristics of RNA secondary structure and dynamic programming methods are used to process the base prediction results, and the structure with the largest sum of the probability of each base pairing is obtained, and this structure will be used as the optimal RNA secondary structure. We, respectively, evaluated GoogLeNet and TCN models based on 5sRNA, tRNA data, and tmRNA data, and compared them with other standard prediction algorithms. The sensitivity and specificity of the GoogLeNet model on the 5sRNA and tRNA data sets are about 16% higher than the best prediction results in other algorithms. The sensitivity and specificity of the GoogLeNet model on the tmRNA dataset are about 9% higher than the best prediction results in other algorithms. As deep learning algorithms' performance is related to the size of the data set, as the scale of RNA data continues to expand, the prediction accuracy of deep learning methods for RNA secondary structure will continue to improve.

1. Introduction

RNA's primary structure is a single-stranded sequence of bases randomly composed of bases A, C, G, and U in a specific order. The single-stranded structure of RNA forms the secondary structure of RNA through the principle of complementary base pair pairing. The secondary structure is folded in space to form a complete three-dimensional structure and exhibit its unique functions [1–4]. However, the tertiary structure of RNA is complex and challenging to represent accurately. Therefore, at present, the function of RNA is mainly studied through RNA secondary structure [5–7].

The secondary structure of RNA can be divided into three parts, namely, the loop, the unpaired single-stranded free structure spiral region, and the spiral region. The loop can be divided into bulge loop, internal loop, and so on. The helical region refers to collecting these base pairs when all bases in two disjoint, equal-length regions are paired in reverse. Loop refers to a single-stranded structure in which unpaired bases are bounded by paired base pairs when forming a helical region [8–10]. As shown in Figure 1, each part of the RNA secondary structure is vividly described.

Pseudoknot is a particular structure in RNA, which plays an essential role in RNA secondary structure study. The definition of a pseudoknot is as follows: in a specific RNA sequence, if there are four bases at a, b, c, and d ($a < b < c < d$), where a matches c and b matches d, then the structure formed by (a, c) and (b, d) base pairs is called a pseudoknot structure [11–14]. Pseudoknots

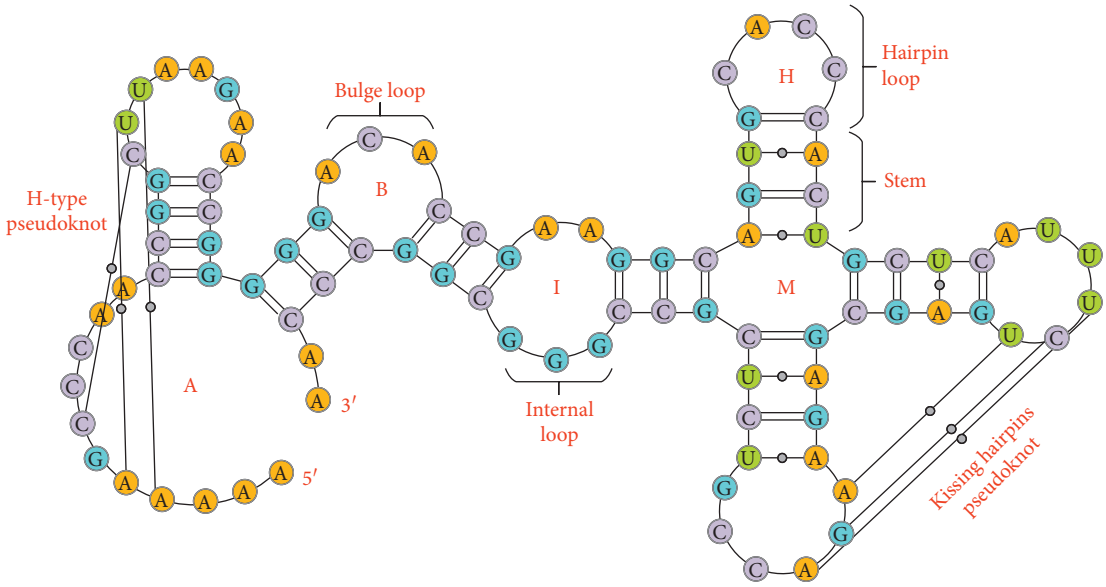

Figure 1: Parts of RNA secondary structure.

can have several different folding topological sorts [15, 16]. In 1990, Pleij proposed 14 types of pseudoknots based on theory. Among them, 4 types are formed by base pairing between free single strands and rings, and 10 types are ring and ring base pairs. Base pairing is between loops. For example, H-shaped pseudoknot is a relatively simple and common structure among all pseudoknot structures [17]. It is formed by complementary base pairing between the unpaired bases of the hairpin loop and the free single-stranded bases outside the stem-loop. This H-type pseudoknot is generally composed of 2 stem regions and 3 loop regions. S1 and S2 represent the stem area, and the loop area is represented by L1, L2, and L3. Among them, S1 and L1 are the stem region and loop close to the 5′ end, S2 and L2 are the stem region and loop close to the 3′ end, respectively, and L3 is the loop connecting different stem regions. In addition to the H-type pseudoknot, there is still a structure in the pseudoknot structure that has been receiving widespread attention. Kissing hairpins pseudoknot: the structure of kissing hairpins pseudoknot is more complicated than that of H-type pseudoknot. Kissing hairpins pseudoknot is formed by complementary pairing of the bases of two hairpin loops [18, 19]. Kissing hairpins pseudoknot is composed of 3 stem regions and 5 loop regions. The stem region is represented by S1, S2, and S3, and the loop region is represented by L1, L2, L3, L4, and L5. The L2 and L4 loop regions' length can be 0, and the other loop regions contain at least one base. Due to the complex structure of kissing hairpins pseudoknot, researchers usually divide a kissing hairpins pseudoknot into two H-shaped pseudoknots in calculations [20]. As shown in Figure 2, the left side of the picture is an H-shaped pseudoknot, and the right side of the picture is a pseudoknot of kissing hairpins.

2. Materials and Methods

2.1. Data Collection and Processing. Researchers in RNA secondary structure research widely use the data in the Mathews lab database, and the data used in the experiment comes from the data set in Mathews lab. The selected RNA structure data set contains 3957 RNA sequences and 10 families in total. Table 1 shows the amount of RNA contained in 10 families.

The experimental data selected three families of 5sRNA, tmRNA, and tRNA from 10 families for research. Among them, 5sRNA and tRNA did not contain pseudoknot structure, while tmRNA contained pseudoknot. The above three families are large and evenly distributed, so they are used as experimental objects. There are similar or identical data in the three families through the analysis of 5sRNA, tmRNA, and tRNA data. Similar data refers to data with the same sequence but different names. These data will affect the accuracy of the experiment, so similar sequences in the data need to be removed. After deredundancy operation, the data of 5sRNA, tmRNA, and tRNA were 1059, 486, and 378, respectively. The original data set uses the CT file format to represent the secondary structure of RNA [21–23]. The CT file contains the sequence information and structure information of the RNA in the data set and contains information that has nothing to do with this experiment. Among them, the sequence formed by the combination of bases "A",

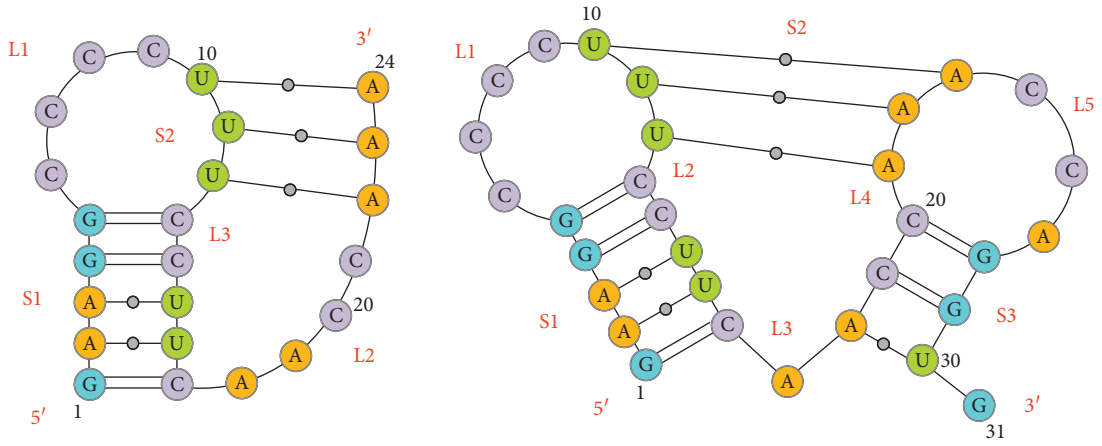

Figure 2: H-type false knot (left) and kissing hairpins false knot (right).

"U", "G", and "C" represents the sequence information of RNA, and ".", "(", ")", "[", "]" are used, "{", "}" dot-bracket notation indicates the structure information of RNA. Therefore, this article needs to extract the available RNA sequence information and structure information. The RNA secondary structure is represented by the CT file. The first line of the CT file contains description information such as the RNA sequence's length and name. The number M indicates the length of a certain RNA sequence, and the string after the number M indicates the name of the RNA. Excluding the first row, each row of the CT file pair includes 6 columns of data: The first and sixth columns indicate the position of the base of the RNA sequence. The second column indicates the sequence of each base of the RNA sequence from the start to the end. The third column indicates the position of the previous base adjacent to a certain position in the RNA sequence. The fourth column indicates the position of the next base adjacent to a certain position in the RNA sequence. The fifth column indicates whether the base in the RNA sequence has a base complementary pairing with the base at that position, where a number other than "0" means that the base at this position has a base with the base at the corresponding position in the first or sixth column base complementary pairing; the number "0" means that the base at this position does not form a base pair with the base at the corresponding position in the first column or the sixth column.

The RNA sequence information is in the second column of the CT file and can be extracted directly. This article uses the dot-bracket notation of seven tags, so when extracting RNA's structure information, compare the numbers in column 1 and column 5 of the CT file with each other. If the data in column 5 is "0", it means no occurrence base complementary pairing; mark it as ".", and if the number in column 5 is not "0" and the number in column 5 is greater than the number in column 1, it means that the number in column 5 is in column 1 for the base at the position. After the base at the corresponding position, use "(" to indicate; on the contrary, use ")" to indicate. Similarly, square brackets and curly brackets are also expressed using corresponding rules [24].

Table 1: The amount of RNA contained in 10 families.

RNA type	Number
5sRNA	1283
16sRNA	110
25sRNA	35
grp1RNA	98
grp2RNA	11
RNasePRNA	454
srpRNA	928
tmRNA	462
tRNA	557
telomeraseRNA	37

If there is an RNA sequence, part of the RNA region is shown in Figure 3. The upper part of the figure indicates that the part of the RNA does not contain false knots. The bottom of the figure indicates that the part of the RNA contains pseudoknots. If three-labeled dot brackets are used, the method indicates that the results are all "(())". Therefore, the traditional dot-bracket method cannot accurately represent the real structure of a pseudoknotted RNA, resulting in the inability to accurately and effectively represent the experiment's classification problem, which will have a great impact on the accuracy of the experiment and subsequent experimental research. In response to the above problems, this paper uses the dot-bracket notation of the combination of ".", "(", ")", "[", "]", "{", "}" to represent the structure information of RNA. Among them, "(", ")" indicate structures without pseudoknot, and "[", "]", "{", "}" indicate structures with pseudoknot. The general idea of the seven-label point bracket notation is to convert the overall problem into individual subproblems for solution and finally integrate the solutions obtained by the subproblems as the solution of the overall problem. The pseudoknot structure in RNA is gradually split into structures without pseudoknots, and finally they are integrated to represent the complete RNA

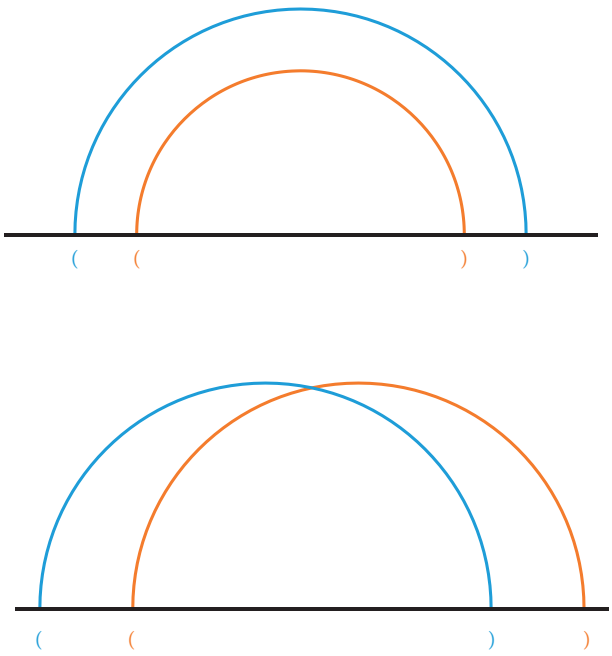

Figure 3: Point bracket notation.

structure [25]. Proceed as follows: in a certain RNA sequence, the base pair that does not have base complementary pairing is marked as ".", and the base pair that has complementary base pairing in the first pair of sequence is marked as "(", ")"; find all disjoint regions. The base pairs are marked as "(", ")" to remove these structures. Then find the first pair of base pairs that have complementary base pairs and the base pairs that do not intersect with them in sequence order and mark them as "[", "]" and so on, and finally get the label representation of the RNA structure, as shown in Figure 4.

2.2. GoogLeNet and TCN. At present, many deep learning methods have been used to predict the secondary structure of RNA. The method of convolutional neural network model to predict RNA secondary structure has got an excellent prediction effect, but there are still some problems. For example, the dot-bracket notation with three tags is used to represent the structural information of RNA. This *f* notation cannot effectively represent the pseudoknot structure in the RNA structure. The dot-bracket notation of seven tags can effectively express the pseudoknot structure. GoogLeNet model training data can get better results.

The innovation and significant advantage of the GoogLeNet method is to start with the structure of the network, increase the depth and width of the network, and improve the computational efficiency of the network. The GoogLeNet method builds a sparse and high computational performance "basic neuron" structure, which is called the Inception network structure. The Inception network structure has gone through multiple versions such as Inception v1, Inception v2, Inception v3, and Inception v4.

The multiple convolution kernels (1×1, 3×3, 5×5) and pooling layer (3×3) in the convolutional neural network are

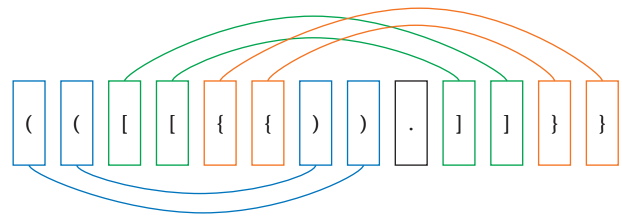

Figure 4: RNA label representation.

placed in the same layer in parallel. The size after convolution and pooling is the same, and each channel is the same. Plus, this design increases the width of the network and increases the network's adaptability to scale. Different convolution kernels can mine every detail feature in the input. In addition, the pooling operation is mainly used to reduce the space size. Therefore, the Inception structure does not need to manually determine whether to add a convolutional layer or a pooling layer. The network can decide by itself whether or not what parameters are needed. However, Native Inception still has big flaws: the number of features that will be output after the three convolutional layers and the pooling layer are spliced is large. If the number of network layers increases, the model will become very complicated and difficult to train and optimize. In addition, the convolution kernel of 5 to 5 will bring about a situation where the amount of calculation is too large and the thickness of the feature map is too large.

In order to solve the above problems, Inception optimized on this basis and proposed a new Inception network structure; that is, 1×1 is added in front of the 3×3 convolution kernel and 5×5 convolution kernel and after max pooling. The convolution kernel not only reduces the dimension but also greatly reduces the number of parameters. For example, if the input of a certain layer is $100 \times 100 \times 128$, one method uses a 5×5 convolutional layer with 256 channels. Another method uses a 1×1 convolutional layer with 32 channels and a 5×5 convolutional layer with 256 outputs. Although the output data obtained by the two methods are the same, the number of parameters of the latter is reduced by about 4 times than that of the former, which greatly improves the calculation efficiency. The number of layers of the GoogLeNet model is 22, of which 9 Inception structures are used. As the number of layers of the network is deepened, the gradient may disappear. In order to avoid the problem of gradient disappearance caused by too many network layers, two auxiliary softmax layers are added to the middle layer of the GoogLeNet network, so that the gradient signal of the network can be backpropagated, which is of great help to the training of the entire GoogLeNet model. The experimental process is shown in Figure 5.

Temporal convolutional network (TCN) is an architecture refined from best practices in convolutional network design. The innovation and advantage of TCN is that its network structure combining causal convolution, dilated convolution, and residual block provides it with flexible receptive fields and network layers. And it can ensure that the key information of the data will not be lost when the model is trained, so as to achieve the ideal prediction effect.

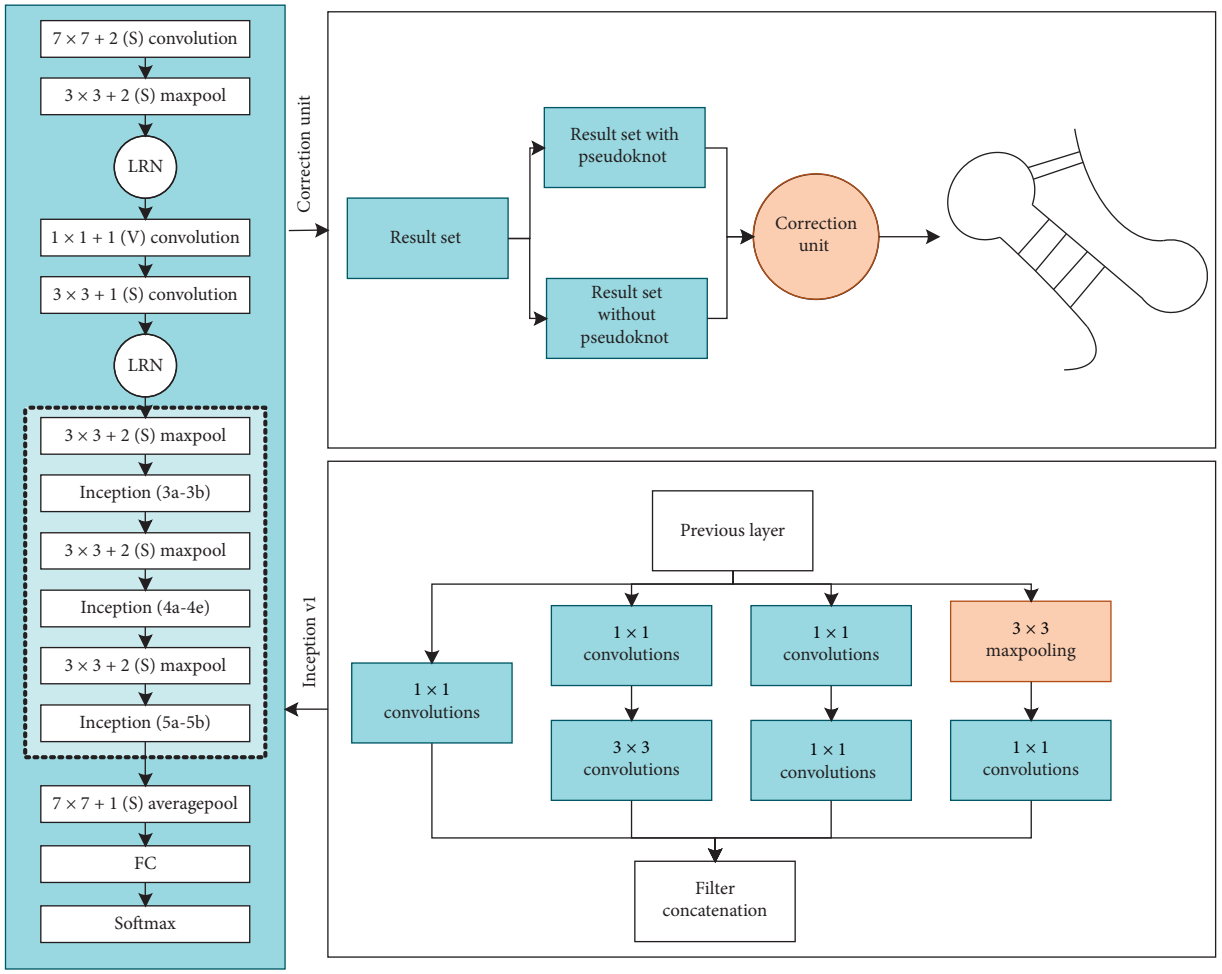

Figure 5: RNA secondary structure prediction of GoogLeNet model.

In addition, TCN also has a more stable gradient and better parallelism, which can improve the efficiency of the network. The RNA secondary structure prediction process of the TCN model is shown in Figure 6. We use the expanded convolutional layer to extract information from the data and add residual links to the network. The dilation convolution method uses interval sampling in the convolution process to obtain a larger receptive field. It not only extracts more data information, but also avoids excessive linear stacking of layers. The residual link allows the network to transmit information in a cross-layer manner and can make the number of layers of the entire network deeper without losing the previous information. A residual block contains two layers of convolution and nonlinear mapping, and WeightNorm and Dropout are added to each layer to regularize the network. The experimental process is shown in Figure 6.

3. Method

The sequence information of the four bases "A", "U", "G", and "C" composing RNA is expressed as a two-dimensional matrix for model training. Create a corresponding two-dimensional matrix for each RNA sequence, and each row of the matrix has a special meaning. For example, the i-th row of the matrix indicates the possibility of base complementary pairing between the base at the i-th position and the base at other positions. In addition, the number of hydrogen bonds between the paired bases determines the size of the weight, where A and U are set to 2, G and C are set to 3, and U and G are swing pairs, set to x. The formula is as follows:

$$P\left(R_i, R_j\right) = \begin{cases} 2, \\ 3, \\ x, x \in (0, 2), \\ 0. \end{cases} \tag{1}$$

RNA secondary structure is also called stem-loop structure. For the bases at positions i and j, it is necessary to consider whether the two are paired and whether they are bases on the stem region. Since the bases closer to the middle of the stem region are more stable than the base pairs on both sides, stability will affect the pairing between bases. Therefore, not only are the bases at positions i and j considered for the RNA sequence, but also the pairing of the bases on the left and right sides of i and j must be taken into consideration. With the help of the concept in locally

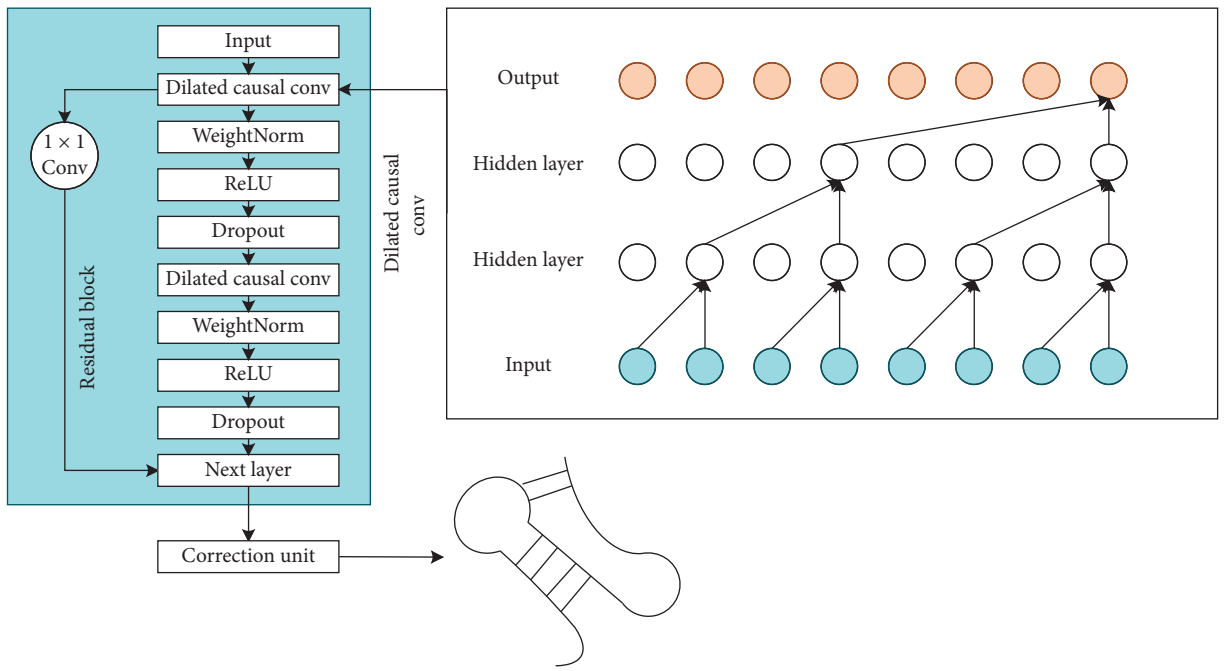

Figure 6: RNA secondary structure prediction of TCN model.

weighted linear regression, a Gaussian function is added to the bases on both sides of i and j, and the closer the base pair to the position i, j, the higher the weight. The farther the base pair is from the i, j position, the lower the weight is [13, 26–28]. For calculating the weight at each position of the matrix, the base pairs at the inner and outer sides of the base pair are the same, as shown in Figure 7.

The RNA sequence of length m is represented as an $m \times m$ two-dimensional matrix. The two-dimensional matrix is split into m matrices by the sliding window. If the size of the sliding window is a, the size of the split matrix is $a \times m$. Therefore, a matrix represents a base of the RNA sequence. The size of the sliding window will affect the accuracy of the experimental model. If the set sliding window is too large, the redundant information contained in the matrix will be extracted. If the sliding window is too small, effective and comprehensive features cannot be extracted. The sliding window size is correlated with the length of the stem region in RNA, which helps to set the size of the sliding window. The input data size of the GoogLeNet network model must be consistent. Therefore, the RNA data needs to be normalized [29, 30].

3.1. Correction Unit. The prediction of RNA secondary structure is a classification problem, but it is not a simple two-classification problem. Regarding the classification problem, although the GoogLeNet method has high accuracy, it cannot avoid its errors. Therefore, it is impossible to ensure whether the prediction results obtained by this method meet the requirements of defining RNA secondary structure. The probabilistic results are corrected by the correction method so that the final result meets the

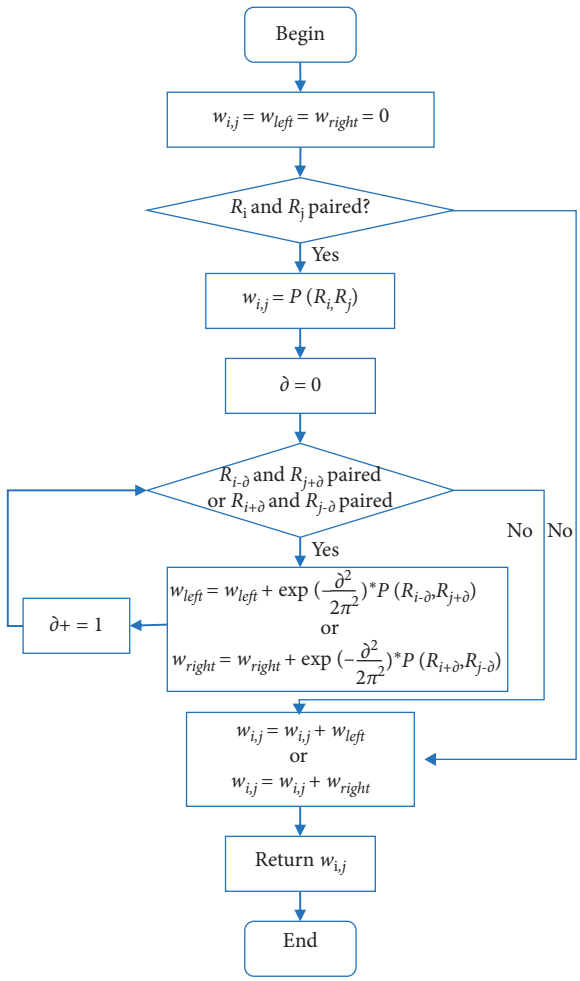

Figure 7: Representation of RNA sequence.

definition of RNA secondary structure. The correction unit is shown in Figure 8.

The pseudoknot structure in RNA macromolecules plays a very important role in its functional expression. If the pseudoknot structure is predicted incorrectly, it will affect the prediction results of the typical stem region. At the same time, correcting the RNA pseudoknot structure and the structure without pseudoknot will affect the final prediction accuracy. Therefore, the prediction results of pseudoknot structure and normal stem area were revised and integrated, respectively.

When the GoogLeNet model predicts RNA secondary structure, Tensorflow's tf.nn.top_k () function returns the maximum target value and index in the output layer. Delete this function, and the probabilities corresponding to the tags ".", "(" and ")" to which each base belongs will be p1, p2, p3. The dot-bracket sequence of RNA secondary structure needs to meet the following conditions. First, the parentheses of the sequence match each other. Second, if there are matched parentheses in the sequence, the bases at the corresponding positions are paired. Finally, after the above two conditions are met, the base probability corresponding to each tag output in the GoogLeNet network is the largest. This paper has made corresponding modifications to the maximum base-pairing algorithm in view of the above situation and proposed the maximum probability and correction algorithm. The formula is as follows:

$$N(i,j) = \max \begin{cases} N(i+1,j) + p_{\text{point}}(R_i), \\ N(i,j-1) + p_{\text{point}}(R_j), \\ N(i+1,j-1) + \delta(R_i,R_j), \\ \max\limits_{i<k<j}[N(i,k) + N(k+1,j)], \end{cases} \tag{2}$$

$$\delta(R_i,R_j) = \begin{pmatrix} p_{\text{left}}(R_i) + p_{\text{right}}(R_j), \\ p_{\text{point}}(R_i) + p_{\text{point}}(R_j). \end{pmatrix}$$

4. Results and Discussion

4.1. Results. We divided the tmRNA data, 5sRNA, and tRNA data into two sets of experimental data. The final experimental results are shown in Figures 9 and 10. As can be seen from the figure, with the result tmRNA data obtained pseudoknot trained about 62%, and the results 5sRNA tRNA data without pseudoknot was approximately 86%. Although the accuracy of the pseudoknot containing the configuration data is lower than the accuracy of the pseudoknot free data structure, this research provides a new breakthrough for the prediction of RNA secondary structure with pseudoknots and lays a solid foundation for subsequent in-depth research.

GoogLeNet model and TCN model are compared with the experimental results of Mfold and RNAfold on the same test data. Table 2 shows the prediction accuracy of GoogLeNet model and TCN model and other algorithms on the tRNA and 5sRNA data sets. It can be seen from this table that the GoogLeNet model and TCN model are significantly better than other algorithms in terms of sensitivity and specificity.

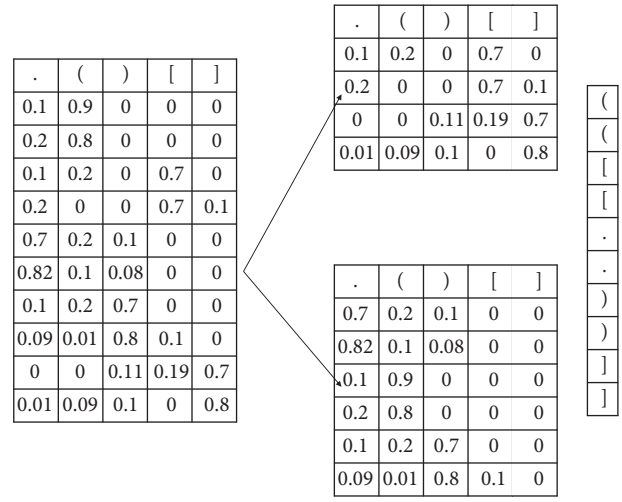

Figure 8: Correction unit.

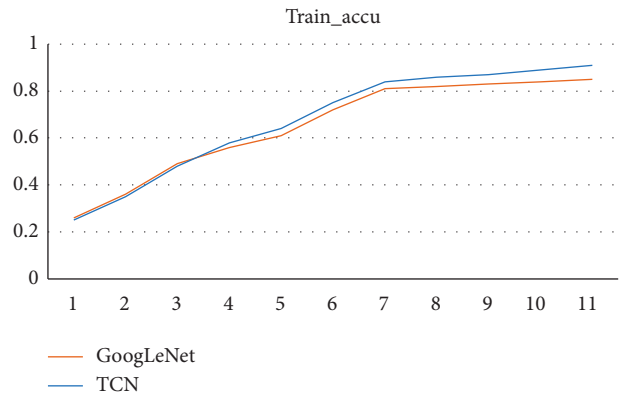

Figure 9: 5sRNA and tRNA.

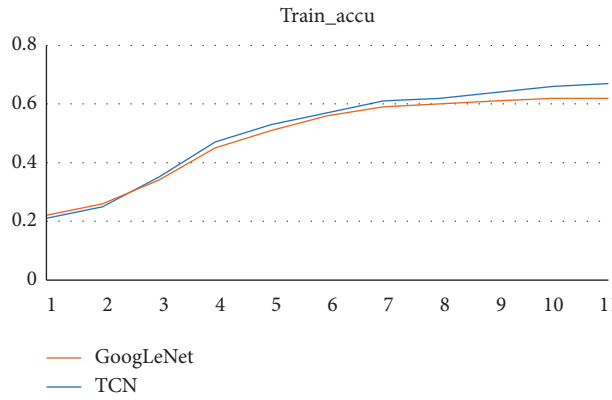

Figure 10: tmRNA.

Table 2: Comparison of prediction accuracy of tRNA and 5sRNA.

Method	tRNA		5sRNA	
	Sensitivity	Specificity	Sensitivity	Specificity
Mfold	0.687	0.685	0.683	0.679
RNAfold	0.698	0.693	0.699	0.699
GoogLeNet	0. 863	0.853	0.861	0.858
TCN	0.925	0.921	0.918	0.919

Table 3: Comparison of prediction accuracy of tmRNA.

Method	tmRNA	
	Sensitivity	Specificity
Mfold	0.537	0.516
RNAfold	0.535	0.512
GoogLeNet	0.629	0.612
TCN	0.685	0.672

Table 3 shows the prediction accuracy of GoogLeNet model and TCN model and other algorithms on the tmRNA data set. It can be seen from this table that when predicting tmRNA data with pseudoknot structure, the prediction accuracy of all algorithms is reduced. However, the prediction accuracy obtained by the GoogLeNet model and TCN model is still in a significant advantage compared with other algorithms, and the prediction accuracy is improved by nearly 9%. It can be seen from Table 3 that although the prediction accuracy of data with pseudoknots is lower than that of data without pseudoknots, it also verifies that the GoogLeNet model and TCN model are sufficient for predicting the structure of pseudoknots.

4.2. Discussion. RNA is an important biological macromolecule in organisms, and it forms the framework of life together with DNA and protein. RNA plays a very important role in the organism, and the key to the role of RNA lies in the structure of RNA in the organism. Traditional experimental methods are expensive and inefficient, and most methods using machine learning cannot effectively predict the secondary structure of RNA with pseudoknots. The seven-tag RNA secondary structure representation accurately represents the pseudoknot structure of RNA. The RNA secondary structure representation algorithm based on the rules of hydrogen bonding between bases can retain the features contained in the RNA secondary structure data. We combined the two methods and innovatively proposed a new data preprocessing method for RNA secondary structure. This method can accurately represent the pseudoknot structure and the structural features in the RNA sequence, which improves the accuracy of model training. At the same time, we applied the GoogLeNet model and TCN model to the field of RNA secondary structure prediction for the first time and achieved remarkable results. However, the training effect of machine learning algorithms depends on the amount of data to a certain extent. The existing data scale cannot fully utilize the advantages of the model. Although our proposed method has achieved significant results, it still lacks accuracy on RNA datasets with false knots. This shows that the method still has a lot of room for improvement.

In the future work, we will do more work to improve the method and enhance the forecasting effect. We will collect more high-throughput data in the RNA structure data set and expand the data set. The data preprocessing method of RNA secondary structure will be more improved to adapt to RNA sequences of different lengths and more complex structures.

5. Conclusions

We applied two deep learning models—GoogLeNet and TCN—to RNA secondary structure prediction with false knots and achieved good results. The results of deep learning models on multiple large-scale RNA sequence data sets are significantly better than traditional prediction methods. This shows that, with the increase of data size, deep learning has significantly improved RNA secondary structure prediction accuracy. Our work proposes two methods for predicting the RNA secondary structure of deep learning models under large-scale data and provides enlightenment and reference significance for the future application of deep learning models under large-scale data.

Data Availability

The data used to support the results of this research are included in GitHub (https://github.com/Demokun/RNA-Secondary-Structure-Database).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to acknowledge the support of the project of Development and Reform Commission of Jilin Province (no. 2019C053-6).

References

[1] P. P. Gardner and R. Giegerich, "A comprehensive comparison of comparative RNA structure prediction approaches," *BMC Bioinformatics*, vol. 5, no. 1, p. 140, 2004.

[2] C. Laing and T. Schlick, "Computational approaches to RNA structure prediction, analysis, and design," *Current Opinion in Structural Biology*, vol. 21, no. 3, pp. 306–318, 2011.

[3] S. Zhang, J. Zhou, H. Hu et al., "A deep learning framework for modeling structural features of RNA-binding protein targets," *Nucleic Acids Research*, vol. 44, no. 4, p. e32, 2016.

[4] R. D. Dowell and S. R Eddy, "Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction," *BMC Bioinformatics*, vol. 5, no. 1, p. 71, 2004.

[5] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *Journal of Molecular Biology*, vol. 288, no. 5, pp. 911–940, 1999.

[6] R. Nussinov and A. B. Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded RNA," *Proceedings of the National Academy of Sciences*, vol. 77, no. 11, pp. 6309–6313, 1980.

[7] Z. Liu, H. Li, and D. Zhu, "A predicting algorithm of RNA secondary structure based on stems," *Kybernetes*, vol. 39, no. 6, pp. 1050–1057, 2010.

[8] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[9] R. B. Lyngso and C. N. S. Pedersen, "RNA Pseudoknot Prediction in Energy-Based Models," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 409–427, 2000.

[10] M. G. Seetin and D. H. Mathews, "RNA structure prediction: an overview of methods," *Methods in Molecular Biology*, vol. 905, pp. 99–122, 2012.

[11] S. H. Bernhart, I. L. Hofacker, S. Will, A. Gruber, and P. F. Stadler, "RNAalifold: improved consensus structure prediction for rna alignments," *BMC Bioinformatics*, vol. 9, no. 1, p. 474, 2008.

[12] J. A. Cruz, M.-F. Blanchet, M. Boniecki et al., "RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction," *RNA*, vol. 18, no. 4, pp. 610–625, 2012.

[13] J. S. Reuter and D. H. Mathews, "RNAstructure: software for rna secondary structure prediction and analysis," *BMC Bioinformatics*, vol. 11, no. 1, p. 129, 2010.

[14] Y. Ding, "Statistical and Bayesian approaches to RNA secondary structure prediction," *RNA*, vol. 12, no. 3, pp. 323–331, 2006.

[15] H. H. Tsang and K. C. Wiese, "SARNA-Predict-pk: predicting RNA secondary structures including pseudoknots," in *Proceedings of the 2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Sun Valley, ID, USA, September 2008.

[16] M. Möhl, R. Salari, S. Will, R. Backofen, and S. C. Sahinalp, "Sparsification of RNA structure prediction including pseudoknots," *Algorithms for Molecular Biology*, vol. 5, no. 1, p. 39, 2010.

[17] E. Y. Jin, J. Qin, and C. M. Reidys, "Combinatorics of RNA structures with pseudoknots," *Bulletin of Mathematical Biology*, vol. 70, no. 1, pp. 45–67, 2008.

[18] C. A. Theimer and D. P. Giedroc, "Equilibrium unfolding pathway of an H-type RNA pseudoknot which promotes programmed −1 ribosomal frameshifting 1 1 edited by D. E. Draper," *Journal of Molecular Biology*, vol. 289, no. 5, pp. 1283–1299, 1999.

[19] E. Rivas and S. R. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots 1 1Edited by I. Tinoco," *Journal of Molecular Biology*, vol. 285, no. 5, pp. 2053–2068, 1999.

[20] X. Huang and H. Ali, "High sensitivity RNA pseudoknot prediction," *Nucleic Acids Research*, vol. 35, no. 2, pp. 656–663, 2007.

[21] M. Zuker, "[20] Computer prediction of RNA structure," *Methods in Enzymology*, vol. 180, pp. 262–288, 1989.

[22] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann, "In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features," *Nature*, vol. 505, no. 7485, pp. 696–700, 2014.

[23] P. F. Fogarty, H. Yamaguchi, A. Wiestner et al., "Late presentation of dyskeratosis congenita as apparently acquired aplastic anaemia due to mutations in telomerase RNA," *The Lancet*, vol. 362, no. 9396, pp. 1628–1630, 2003.

[24] M. Parisien and F. J. N. Major, "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data," *Nature*, vol. 452, no. 7183, pp. 51–55, 2008.

[25] R. M. Dirks and N. A. Pierce, "A partition function algorithm for nucleic acid secondary structure including pseudoknots," *Journal of Computational Chemistry*, vol. 24, no. 13, pp. 1664–1677, 2003.

[26] D. W. Staple and S. E. Butcher, "Pseudoknots: RNA structures with diverse functions," *PLoS Biology*, vol. 3, no. 6, p. e213, 2005.

[27] P. G. Higgs, "RNA secondary structure: physical and computational aspects," *Quarterly Reviews of Biophysics*, vol. 33, no. 3, pp. 199–253, 2000.

[28] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath, "RNAMotif, an RNA secondary structure definition and search algorithm," *Nucleic Acids Research*, vol. 29, no. 22, pp. 4724–4735, 2001.

[29] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster, "Statistics of RNA secondary structures," *Biopolymers*, vol. 33, no. 9, pp. 1389–1404, 1993.

[30] I. Jelinek, J. N. Leonard, G. E. Price et al., "TLR3-Specific double-stranded RNA oligonucleotide adjuvants induce dendritic cell cross-presentation, CTL responses, and antiviral protection," *The Journal of Immunology*, vol. 186, no. 4, pp. 2422–2429, 2011.