Protein Secondary Structure Prediction Based on Two Dimensional Deep Convolutional Neural Networks

Yihui Liu*, Jinyong Cheng
Institute of Intelligent Information Processing
Qilu University of Technology, China
e-mail: yxl@qlu.edu.cn

Yuming Ma, Yehong Chen
Institute of Intelligent Information Processing
Qilu University of Technology, China

Abstract—**The highest three-state prediction accuracy of protein secondary structure is now at 82-84% without using structure templates, approaching to the theoretical limit 88-90%. Increasingly larger training datasets cover more protein sequences and structures. More powerful deep learning techniques are not only able to deal with the computation load of large data, but also can capture the long-range interactions of protein sequence. In this research, we propose a new approach to design a two dimensional deep convolutional neural networks (2DCNN) with 6 convolutional layers and 5 max-pooling layers. The two dimensional convolutional neural networks keep original amino acid sequence position information based on two dimensional input matrix, and extract features of the sequence interactions better. The performance of our prediction model 2DCNN is 83.09%, 81.74%, 82.41%, 83.56%, 81.16%, and 80.30% for 25PDB, CB513, CASP9, CASP10, CASP11, and CASP12 datasets. Our prediction model achieves better results compared to most state of the art methods. (http://qilubio.qlu.edu.cn/protein)**

Keywords-protein secondary structure prediction; feature extraction; two dimensional deep convolution neural networks (2DCNN).

I. INTRODUCTION

Protein structure prediction is crucial to analyze protein function and applications. Protein secondary structure prediction plays an important role to predict three-dimensional structure. It is widely accepted that amino acid sequence has enough information to decide protein three dimensional structure and it is difficult to predict protein structure using amino acid sequence [1]. The feature extraction of amino acid sequence information is a key step to improve the performance of predicting protein secondary structure.

Protein secondary structure prediction is a fundamental problem and is much more challenging. Many machine learning approaches and statistical approaches are utilized in this research area [2]. The methods using statistical features extracted from protein sequence normally achieve accuracy with no higher than 65% [3, 4]. Position specific scoring matrix (PSSM) based on PSIBLAST reflects evolutionary information and makes most significant improvement in protein secondary structure prediction [5]. Machine learning methods with PSSM profiles improve prediction accuracy to 70-79%, and have a major breakthrough in protein secondary structure prediction, such as support vector machine [6, 7, 8], neural networks [9, 10, 11], K-nearest neighbors [12], etc. Neural networks have been used in protein structure prediction for over a decade. Inspired by the recent success of deep neural networks, several papers use deep learning network to predict protein secondary structure [1, 13, 14]. The paper [1] proposes a deep learning network called DNSS, which uses three separated trained deep networks to predict the results. This deep learning network approach is achieving Q3 accuracy of 80.7% on CASP dataset and 78.8% based on 10 fold cross validation experiments. A supervised generative stochastic network (GSN) is proposed to predict secondary structure using deep hierarchical representations [13]. The performance of 66.4% Q8 accuracy on the CB513 dataset is obtained. Deep convolutional neural fields called DeepCNF [14] are used to predict secondary structure, which are deep learning extension of conditional neural fields (CNF). Experimental results show that DeepCNF obtains 84% Q3 accuracy on the CASP test proteins, and 73.2% Q8 accuracy based on 5 fold cross validation experiments. The highest three-state accuracy without relying on structure templates is now at 82–84%, approaching to the theoretical limit 88-90% [15]. These improvements are based on increasingly larger training datasets of protein sequences and structures, and more powerful deep learning techniques. In our previous work [16, 17], convolutional neural networks (CNN) with two convolutional layers is used to extract the features using a new representation of two dimensional PSSM evolutionary matrix. The extracted CNN features reflect not only the evolutionary information but also the long-range interactions of amino acid sequence, and show the potential ability to tackle with the mutation, insertion, deletion of amino acid sequence.

The previous published papers [1, 13, 14] put PSSM matrix into one dimensional vector, the position information of amino acid sequence does not characterize well. In this research, in order to deal with large dataset, two dimensional convolutional neural networks with 6 convolutional layers and 5 max-pooling layers are designed to predict the three classes of protein secondary structure, The input profile is two dimensional combined PSSM matrix with a binary vector of 20 bits representing the amino acid sequence. The paper is organized as follows. Section 2 describes our proposed method of two dimensional deep convolutional neural networks. Section 3 illustrates the experiments and results. Section 4 is conclusions.

978-1-5090-6352-9/17/$31.00 ©2017 IEEE

II. METHODS

A. Protein Dataset

CulledPDB datasets [18, 19] in PISCES server "provide with the longest list possible of the highest resolution structures that fulfill the sequence identity and structural quality cut-offs". In our research, CulledPDB dataset is selected based on the percentage identity cutoff of 25%, the resolution cutoff of 3 angstroms, and the R-factor cutoff of 0.25. There are 12288 proteins in CulledPDB dataset. About 11000 proteins are used to train the prediction model, and other ones are remaining for test.

Several independent evaluation datasets are collected from the CASP datasets. 122 proteins of CASP9 dataset , 99 proteins of CASP10 dataset, 81 proteins of CASP11 dataset, and 19 proteins of CASP12 dataset are selected according to the availability of crystal structure [20,21,22]. CB513 dataset and 25PDB dataset also are used to evaluate our proposed method. CB513 dataset has 513 protein sequences developed by Cuff and Barton [23]. Any two proteins of CB513 share less than 25% sequence identity with each other. The 25PDB [24] dataset is selected with low sequence similarity of no more than 25%, and has 1673 proteins, consisting of 443 all-a, 443 all-b, 346 a/b and 441 a+b. The number of protein in datasets may be different from with other published paper, because we only use the available data obtained from PSSM program. Table I shows the protein datasets used in our research.

TABLE I. THE PROTEIN DATASETS

Datasets	Protein number
CullPDB	12288
25PDB	1647
CB513	513
CASP 9	122
CASP10	99
CASP11	81
CASP12	19

B. The Input Matrix

Two dimensional PSSM matrix is input to convolutional neural networks. Position specific scoring matrix (PSSM) represents the evolutionary information of acid amino sequence, and can be used as feature vector for predicting secondary structure [5]. PSIBLAST software is used to calculate position specific scoring matrix. In our research, BLOSUM62 evolutionary matrix is adopted to run on nr database using multiple sequence alignments. The PSIBLAST parameters are set with threshold 0.001 and 3 iterations. The resulting PSSM matrix is 20xN matrix, where N is the length of amino acid sequence and 20 is the number of amino acid types. So PSSM matrix has twenty features for each residue in protein sequence, and each feature represents the possibility that the residue mutates into corresponding amino acid type. This mutation matrix is calculated using the algorithm of multiple sequence alignments. We use quadrature encoding to represent the amino acid sequence information of protein. A binary vector of 20 bits is used to indicate the amino acid type. In total there are 40 input features for each residue, 20 features from PSSM and the other 20 features from amino acid types.

A consecutive sliding window of amino acid sequence is used to obtain residue sequence information and predict the secondary structure of the central residue. Figure 1 shows the two dimensional matrix for one slicing window of length 25.

The widely used secondary structure definition DSSP [25] is used to evaluate our proposed method. DSSP has 8 categories of secondary structure: H(alpha-helix), G(3-helix), I(5-helix), E(extended-strand), B(isolated-strand), T(turn), S(bend), and coil ('_'), which are normally reduced into 3 classes. In our research, we map H, G, and I to H; E and B to E; all other states to C, which usually results in lower prediction accuracy than other definitions [26].

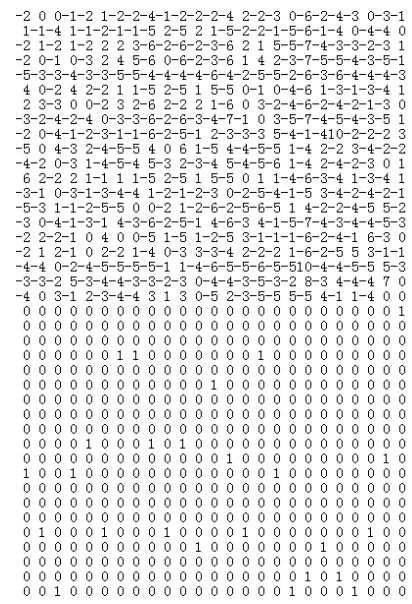

Figure 1. Two dimensional input matrix (40X25) for one slicing window of length 25.

C. 2D Convolutional Neural Network (2DCNN)

Convolutional neural networks [27] have recently drawn attention of researchers due to its three distinguished concepts of local receptive fields, shared weights and spatial sub-sampling. Convolutional neural networks have the ability of shift, scale, and distortion invariance to some extent. For these properties, convolutional neural networks are widely applied into different research fields such as image segmentation [28], crowd density estimation [29], semantic relation classification [30], etc.

It is well known that proteins are genetic diversity. It is necessary that enough protein samples are used to extract the features of the amino acid sequence corresponding to protein secondary structure. The convolutional neural network no doubt is the reasonable choice, because of its capability dealing with large amount of training samples. Compared to the other classifying models, such as support vector machine, one of the best models, which is good at high dimensional features with small dataset, the convolutional neural networks not only capture the features of protein sequence

from large amount samples, but also save the computation time.

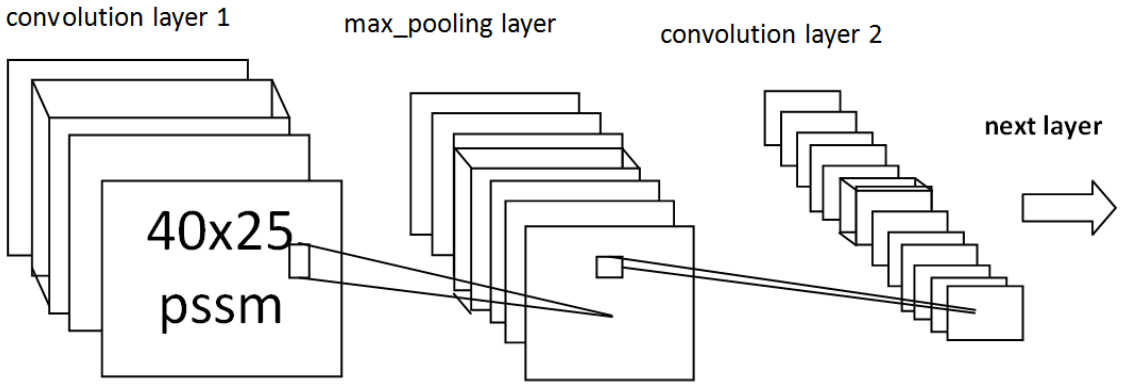

Figure 2. The convolutional neural network for two dimensional input matrix. The first convolutional layer has 720 filters of size 11x11 with stride 1x1 , followed by 4x4 max-pooling layer with stride of 1x1.

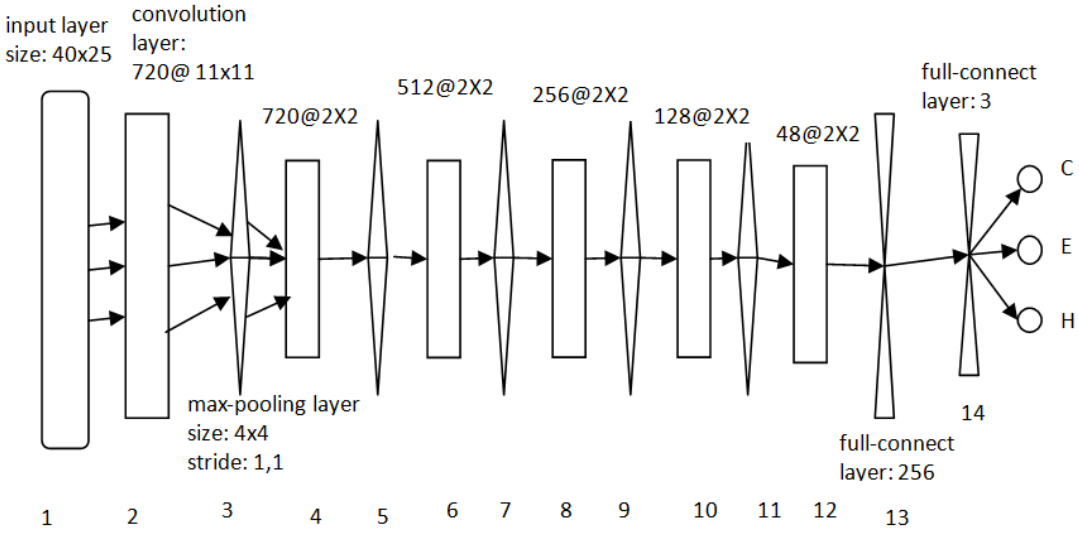

Figure 3. The designed architecture of prediction model based on convolutional neural network. The layers of 2, 4, 6, 8, 10, and 12 position are convolutional layers. The max-pooling layers are at 3,5,7,9, and 11 positions. The 13 and 14 layers are full-connected layers.

In this paper, we use CullPDB dataset with 12288 proteins, 11000 proteins are used to train the model, and 1228 proteins are used for test. Two dimensional PSSM matrix combined with a binary vector is used to extract the features based on deep convolutional neural network, which represents not only the evolutionary information of protein but also amino acid sequence information. Figure 2 shows the convolutional neural network for two dimensional input matrix. Figure 3 shows the designed architecture of prediction model based on 2D convolutional neural networks with 6 convolutional layers.

The convolutional neural network, which has concatenated fourteen layers including one input layer, 6 convolutional layers, 5 max-pooling layers, 2 full-connected layers, and softmax layer, are used to extract features and predict the protein secondary structure. Full connection layers are used to further reduce the feature dimensions. Softmax layer at last classifies three states of protein secondary structure. The first convolutional layer has 720 filters of size 11x11 with stride 1x1, followed by max-pooling layer of 4x4 with stride 1x1. The convolutional layers at 4, 6, 8, 10, and 12 positions have 720, 512, 256, 128, and 48 filters with size 2x2 and stride 1x1, respectively. The max-pooling layers at 5, 7, 9, and 11 positions are doing the max-pooling operation within size 2x2 and stride 1x1, respectively. A first full-connected layer is of 250 units, followed by second full-connected layer of 3 units to further reduce the dimension. The features extracted from last full-connected layer are input to softmax classification model.

III. Experiments and Results

In this research, about nine tenth data of CullPDB is used to train our 2D convolutional neural network, called 2DCNN method. About one tenth data of CullPDB are used to test the results and tune the parameters of convolutional neural network, and 79.86% Q3 is achieved. Then the several

datasets are evaluated using our 2DCNN method. The results are shown in Table II. The performance of 83.09%, 81.74%, 82.41%, 83.56%, 81.16%, and 80.30% for 25PDB,CB513, CASP9, CASP10, CASP11, and CASP12 datasets. We can see that our Q3 results are only lower than DeepCNF method, and better than other state of the art ones based on the papers [14, 1] in Table III and Table IV.

TABLE II. THE RESULTS BASED ON OUR 2DCNN METHOD

Datasets	Q3(%)	SOV(%)
25PDB	83.09	79.70
CB513	81.74	76.09
CASP9	82.41	77.71
CASP10	83.56	78.65
CASP11	81.16	76.68
CASP12	80.30	75.64

TABLE III. THE RESULTS BASED ON THE PAPER [14].

	CB513(%)		CASP10(%)		CASP11(%)	
	Q3	SOV	Q3	SOV	Q3	SOV
SSpro (without template)	78.5	77.2	78.5	75.9	77.6	77.3
SPINE-X	78.9	78.7	80.7	78.7	79.3	79.3
PSIPRED	79.2	81.0	81.2	80.9	80.7	81.4
JPRED	81.7	83.3	81.6	82.4	80.4	82.0
RaptorX-SS8	78.3	79.5	78.9	80.2	79.1	81.1
DeepCNF-SS	82.3	84.8	84.4	85.7	84.7	86.5

SOV results are higher than Q3 results, are different with other published paper. We think that SOV results are not correct, and not used as comparing standards

TABLE IV. THE RESULTS BASED ON THE PAPER [1].

	CASP9		CASP10		Combined	
	Q3	SOV	Q3	SOV	Q3	SOV
DNSS	81.1	74.7	80.2	73.6	80.7	74.2
PSSpred	83.3	72.0	81.0	70.4	82.2	71.3
SSpro	79.6	72.6	78.8	71.9	79.2	72.3
PSIPRED	80.9	69.3	81.2	68.6	81.0	69.0
RaptorX	78.1	74.7	77.9	70.3	78.0	70.3

SOV results are 6-10% lower than Q3 results, and are consistent with our test results.

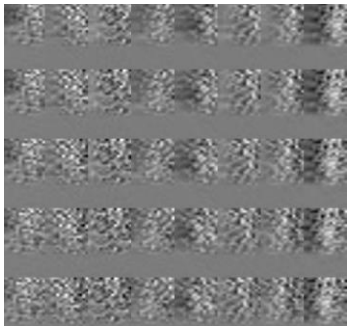

Figure 4. The extracted feature maps from first convolutional layer.

The revolutionary progress of neural networks is the development of deep learning networks. Deep learning networks can build more robust predictors. The convolutional neural networks with 6 convolutional layers can deal with long-range interactions of amino acid sequence better, and tackle with mutation, insertion, deletion of residues better with 5 max-pooling layers. Figure 4 and Figure 5 show the extracted two dimensional feature maps at first and second convolutional layer respectively. We can see that the deeper the convolutional layer is, and more clear the 2D feature maps are corresponding to the protein secondary structure.

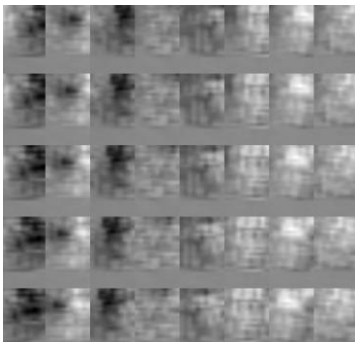

Figure 5. The extracted feature maps from second convolutional layer.

IV. CONCLUSIONS

In this paper a new method is proposed based on two dimensional deep convolutional neural networks (2DCNN) to predict the protein secondary structure. The representation of two dimensional PSSM combined features keeps the original sequence position and extracts sequence interaction of residue better. The deep convolutional neural networks with 6 convolutional layers and 5 max-pooling layers have the ability to tackle with mutation, insertion, deletion of residue in some degree, and reveal the hiding relation between protein sequence and secondary structure. The performance of our 2DCNN prediction model is 83.09%, 81.74%, 82.41%, 83.56%, 81.16%, and 80.30% for 25PDB, CASP9, CASP10, CASP11, and CASP12 datasets. Our prediction model achieves better results compared to most state of the art methods.

ACKNOWLEDGMENT

The work is supported by Shandong Provincial Natural Science Foundation of China under Grant ZR2013FM020, the National Natural Science Foundation of China under Grant 61375013.

REFERENCES

[1] M. Spencer, J. Eickholt, and J. Cheng, "A deep learning network approach to ab initio protein secondary structure," IEEE/ACM Transactions on Computational Biology & Bioinformatics, vol. 12, 2015, pp. 103-112.

[2] J. Cheng, A.N. Tegge, and P. Baldi, "Machine learning methods for protein structure prediction," IEEE Reviews in Biomedical Engineering, vol. 1, 2008, pp. 41–49.

[3] P. Y. Chou, and G. D. Fasman, "Conformational parameters for amino acids in helical, b-sheet, and random coil regions calculated from proteins," Biochemistry, vol. 13, 1974, pp. 211–222.

[4] J. Garnier, J.F. Gibrat, and B. Robson, "GOR method for predicting protein secondary structure from amino acid sequence," Methods Enzymol., vol. 266, 1995, pp. 540–553.

[5] D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” J. Mol. Biol., vol. 292, 1999, pp. 195–202.

[6] C. Chen, Y. Tian, X. Zou, P. Cai, and J. Mo, "Prediction of protein secondary structure content using support vector machine," Talanta. vol. 71, 2007, pp. 2069-2073.

[7] M.N. Islam, S. Iqbal, A.R. Katebi, and M.T. Hoque, "A balanced secondary structure predictor," Journal of Theoretical Biology, vol. 389, 2016, pp. 60-71.

[8] B. Yang, Q. Wu, Z. Ying, and H. Sui, "Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model," Knowledge-Based Systems, vol. 24, pp. 304-313., 2011.

[9] N.P. Bidargaddi, M. Chetty, and J. Kamruzzaman, "Combining segmental semi-Markov models with neural networks for protein secondary structure prediction," Neurocomputing, vol. 72, 2009, pp. 3943-3950.

[10] X. Yao, H. Zhu, and Z. She, "A dynamic Bayesian network approach to protein secondary structure prediction," BMC Bioinformatics, vol. 9, 2008, pp. 49.

[11] S.A. Malekpour, S. Naghizadeh, H. Pezeshk, M. Sadeghi, and C. Eslahchi, "Protein secondary structure prediction using three neural networks and a segmental semi Markov model," Mathematical Biosciences, vol. 217, 2009, pp. 145-150.

[12] Y.T. Tan, B.A. Rosdi, Y.T. Tan, and B.A. Rosdi, "FPGA-based hardware accelerator for the prediction of protein secondary class via fuzzy K-nearest neighbors with Lempel–Ziv complexity based distance measure," Neurocomputing, vol. 148, 2015, pp. 409-419.

[13] J. Zhou and O. Troyanskaya, "Deep supervised and convolutional generative stochastic network for protein secondary structure prediction," Computer Science, 2014, pp. 745-753.

[14] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," Scientific Reports, 2016, DOI: 10.1038/srep18962

[15] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, and Y Zhou, "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?" Brief Bioinform., 2016, doi: 10.1093/bib/bbw129.

[16] Y. Liu, Y. Chen, and J. Cheng, "Feature extraction of protein secondary structure using 2D convolutional neural network," 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI2016), 2016, pp. 1771-1775.

[17] Y. Liu and J. Cheng, "Protein secondary structure prediction based on wavelets and 2D convolutional neural network," 7th International Conference on Computational Systems-Biology and Bioinformatics (CSBio2016), ACM International Conference Proceeding Series, 2016, pp. 53-57.

[18] G. Wang and R.D. Jr, "PISCES: recent improvements to a PDB sequence culling server," Nucleic Acids Research, vol. 33(Web Server issue), 2005, pp. W94–W98.

[19] G. Wang and R.D. Jr, "PISCES: a protein sequence culling server," Bioinformatics, vol. 19, 2003, pp. 1589-1591.

[20] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, “Critical assessment of methods of protein structure prediction (CASP) - round X,” Proteins: Struct., Funct., Bioinformatics, 2012.

[21] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano, “Critical assessment of methods of protein structure prediction (CASP)—round IX,” Proteins: Struct., Funct., Bioinformaticsvol, vol. 79, 2011, pp. 1-5.

[22] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)—round x.," Proteins: Struct. Funct. Bioinform., vol. 82, 2014, pp. 1-6.

[23] J.A. Cuff and G.J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," Proteins: Struct. Funct. Genet., vol. 34, 1999, pp. 508-519.

[24] K.D. Kedarisetti, L.A. Kurgan, and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology," Biochem. Biophys. Res. Commun, vol. 348, 2006, pp. 981-988.

[25] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," Biopolymers, vol. 22, 1983, pp. 2577-2637.

[26] W. Qu, H. Sui, B. Yang, W. Qian, "Improving protein secondary structure prediction using a multi-modal BP method," Computers in Biology & Medicine, vol. 41, 2011, pp. 946-959.

[27] D.H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," J. Physiol. , vol. 160, 1962, pp. 106-154.

[28] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," Neurocomputing, vol. 191, 2016, 214-223.

[29] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," Engineering Applications of Artificial Intelligence, vol. 43, 2015, pp. 81-88.

[30] P. Qin, W. Xu, and J. Guo, "An empirical convolutional neural network approach for semantic relation classification," Neurocomputing, vol. 90, 2016, pp. 1-9.