An Effective Cell Pathology Image Detection Method Based on Deep Stacked Auto-Encoder Combined with Random Forest

Xin Yan and Lei Wang

1 School of Foreign Languages, Zaozhuang University, Shandong, China
2 Colleges of Information Science and Engineering, Zaozhuang University, Shandong, China
Email: leiwang@ms.xjb.ac.cn; Email: xinyanuzz@gmail.com

Abstract. Cytopathology mainly studies the causes and pathogenesis of disease occurrence, as well as the changes in the physiological functions of cells during the process of disease occurrence, providing a basis for the prevention and diagnosis of diseases. Using computational simulation technology to accurately detect cell pathology images can effectively avoid the influence of subjective factors in diagnostic operation by pathologist, and provide reliable technical means for case diagnosis. In this study, we design an effective model to detect cell pathology images based on deep stacked auto-encoder algorithm. Firstly, the cell pathology images are preprocessed by standardization. Secondly, the processed images are sent to the stacked auto-encoder algorithm to extract theirs hidden feature information. Finally, the random forest algorithm is used to quickly and accurately detect theirs category. We introduce cross-validation in our experiments to test the robustness and stability of the results. In addition, we also compared the model based on a k-nearest neighbor classifier and achieved good results. These excellent experiment results indicated that the proposed method can accurately detect cell pathology images and can be used as a reliable tool for pathologists to diagnose rapidly.

1. Introduction
The recognition and detection of cell pathology images are the focus of biomedical image analysis and the key to computer-aided disease diagnosis. The computer-aided diagnosis system combined with expert experience can provide quantitative characterization and machine identification results, which can offer the objective and quantitative diagnostic basis for pathologists’ clinical review and diagnosis, effectively improve the diagnosis accuracy, and reduce the misdiagnosis caused by the difference of subjective diagnosis experience and fatigue diagnosis [1].

In practical research, however, due to the complexity of cell morphology, cell overlap, aggregation, and imperfect staining in sample making, it is often difficult to realize computer-aided automatic recognition of cell images. In addition, there are many kinds of cells and the differences between various classes are not obvious. Therefore, how to construct an effective calculation model to recognise the characteristics of cell pathology images have become the burning issues to be solved. This issue has attracted the attention of many researchers and proposed constructive solutions [2, 3]. Plissiti et al. [4] designed an image segmentation method for nuclear overlap based on the shape and local features of the nucleus. This algorithm can integrate the shape attributes of the nucleus into the deformation model and segment the contour of the nucleus through finite iterations. Khasawneh et al. [5] proposed the stochastic polygons method to segment glandular structures. In this method, each glandular structure is treated as a polygon consisting of a random number of vertices, where the apex
represents the epithelial nucleus. This method successfully detects and extracts glandular structures in the colon tissue. The study by Beck et al. [6] showed that the features extracted from the image matrix region were significantly associated with the prognosis of breast cancer.

In this study, we propose effective images of cell pathology detection method based on deep stacked auto-encoder method. This method first objectively extracts cell pathology image features using stacked auto-encoder algorithm, and then sends them into random forest classifier for accurate classification. The proposed model implemented the five-fold cross-validation method on warwick-qu data set and achieved good results. To verify the classification ability of the model, we also comparisons with the model based on k-nearest neighbor classifier. These competitive results indicated that the proposed method can accurately detect cell pathology images and can be used as a reliable tool for pathologists to diagnose rapidly.

2. Materials and Methods

2.1. Benchmark Data Set
The data set used in this article is from Hematoxylin and Eosin (H&E) stained slides images, which consists of various histologic grades collected by Khasawneh et al. [5] This benchmark data set contains ground truth annotations which provide by pathologists. The slides are generated by using Zeiss MIRAX MIDI Slide Scanner for 20× magnification digital scanning. The data set contains 165 images extracted from 52 visual fields and saved in BMP format. These fields are divided into five categories by expert pathologist: healthy, adenomatous, moderately differentiated, moderately-to-poorly differentiated and poorly differentiated (well-differentiated adenocarcinoma is not observed). In addition, the benchmark data set provides a handmarked ground truth for glandular structures in each field of visual. Ultimately, the data set consists of 42 healthy, 32 adenomatous, 47 moderately differentiated, 20 moderately-to-poorly differentiated and 24 poorly differentiated images, respectively containing 670, 298, 287, 135 and 195 glandular structures present in them. In size terms, there are 575×430 and 775×520 resolutions in data set. In order to make the algorithm easy to process, we added white edges to the small resolution image to form the same size. There are 14 such images. The pictures of these five types of images are shown in Figure 1.

(a) 
(b) 
(c) 
(d) 
(e) 

Figure 1. The warwick-qu data set contains of the cell pathology images, in which (a)-(e) represents healthy, adenomatous, moderately differentiated, moderately-to-poorly differentiated and poorly differentiated categories, respectively

2.2. Stacked Auto-Encoder
Stacked auto-encoder (SAE) can learn a variety of expressions of raw data layer by layer, and each layer is based on the expression features of the previous layer, extracting more abstract and more suitable for complex features, and then doing some classification tasks [7-9]. Auto-encoder (AE) is the basic component of SAE, which can be considered as a view to restore its original input system, as shown in the figure 2.
Figure 2. Auto-Encoder Structure

AE consists of encoder and decoder, which essentially transforms the input signal. The encoder transforms the input signal $S$ into the encoded signal $R$, while the decoder transforms the encoded signal $R$ into the output signal $V$. The goal of AE is to make the output $V$ reproduce the input $S$ as much as possible. Assuming a training set $S$, the input is transformed by the encoder into a hidden layer representation by mapping $h_c$ as follows:

$$R = h_c(S) = G_c(W_1^TS + b_1)$$

(1)

where $G_c(\cdot)$ is the activation function, $b_1$ is the bias and $W_1^T$ is the weight. Then the hidden layer representation $R$ is transformed by the decoder to output by mapping $G_d$.

$$V = h_d(R) = G_d(W_2^TR + b_2)$$

(2)

where $G_d(\cdot)$ is the activation function, $b_2$ is the bias and $W_2^T$ is the weight. Finally, AE minimizes the loss function $D(S,V)$ by using the back-propagation program.

$$D(S,V) = D_r(S,V) + 0.5\sigma(\|W_1\|_2 + \|W_2\|_2^2)$$

(3)

where $D_r(S,V)$ is the reconstruction error and $\sigma$ is the weight decay cost.

The SAE can be formed by superimposing multiple AEs, which takes the output of the next layer as the input of the previous layer, and extracts the features of the original image layer by layer in this way until it reaches the highest level, thus obtaining the high-level of abstract features. Figure 3 shows this structure. In the experiment, we optimized the SAE algorithm and set the final parameters as follows: the number of layers is 2, and the activation function, learning rate and input zero masked fraction of the first and second layers are sigm, 0.05 and 0.5, respectively.

Figure 3. Structure of Stacked Auto-Encoders

2.3. Random Forest

Random forest (RF) is a classification algorithm consisting of multiple decision trees, which determines the final output category by synthesizing the categories of the single trees [10]. RF classifier has the advantages of high classification accuracy, not easy to over-fit, strong anti-noise ability, fast training speed and easy to achieve parallelization. Its construction process is as follows:

(1) The original training set is $S$, and k new self-service sample sets are randomly extracted by bootstrap method with playback, from which k classification trees are constructed;

(2) Assuming that there are $m$ variables, randomly select $n$ ($n < m$) variables from each node of each tree, then select one of these variables with the strongest classification ability, and determine the threshold of each classification variable by checking each classification point;

(3) Without any pruning, each decision tree is allowed to grow to its maximum.
(4) Random forest consists of these generated multiple classification trees, and random forest classifier identify and classify new data.

Combine all the results of decision tree classifier, and use voting to determine the final classification result. Figure 4 shows the schematic diagram of the proposed model.

![Cell Image Pathology Dataset](image1)

![Stacked Auto-Encoders](image2)

![Random Forest](image3)

**Figure 4. The Proposed Model Schematic Diagram**

3. Results and Discussion

3.1. Model Evaluation Criteria

In this study, we choose general evaluation criteria to verify the performance of our proposed model [11, 12], including accuracy (Accu.), false precision (FPre.) and specificity (Spe.). They can be expressed by the following formulas:

\[
\text{Accu.} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)
\]

\[
\text{FPre.} = \frac{TN}{TN+FP} \quad (5)
\]

\[
\text{Spe.} = \frac{TN}{TN+FN} \quad (6)
\]

where TP (true positive) indicates the number of correct classifications of positive samples, TN (true negative) indicates the number of correct classifications of negative samples, FP (false positive) indicates the number of incorrect classifications of positive samples and FN (false negative) indicates the number of incorrect classifications of negative samples.

3.2. Model performance

In order to verify the performance of our model, we used the five-fold cross-validation method to perform the experiment. We first divided the data set into five equal subsets, using four of them as the training set and the remaining one as the test set. Then the experiments are performed using different subset as test set in turn until all subsets are used as test set once and only once. Finally, we take the average values of the experiments as the experimental results. As the cell pathology images can be divided into 5 types, we have experimented with them separately, and table 1 summarized the specific cross-validation results.
Table 1. The five-fold cross-validation experimental results value (%) were yield on the warwick-qu data set

| Category            | Test set | Accu. | Spe. | FPre. |
|---------------------|----------|-------|------|-------|
| healthy             | 1        | 69.70 | 100.00 | 69.70 |
|                     | 2        | 66.67 | 95.65 | 68.75 |
|                     | 3        | 87.88 | 90.00 | 96.43 |
|                     | 4        | 72.73 | 95.65 | 73.33 |
|                     | 5        | 72.73 | 100.00 | 72.73 |
| **Average**         | **73.94**| **96.26** | **76.19** |       |
| adenomatous         | 1        | 90.91 | 100.00 | 90.91 |
|                     | 2        | 87.88 | 96.55 | 90.32 |
|                     | 3        | 72.73 | 100.00 | 72.73 |
|                     | 4        | 63.64 | 100.00 | 63.64 |
|                     | 5        | 84.85 | 96.55 | 87.50 |
| **Average**         | **80.00**| **98.62** | **81.02** |       |
| moderately          | 1        | 66.67 | 95.45 | 67.74 |
| differentiated      | 2        | 72.73 | 95.65 | 73.33 |
|                     | 3        | 66.67 | 91.30 | 70.00 |
|                     | 4        | 72.73 | 91.67 | 75.86 |
|                     | 5        | 78.79 | 96.15 | 80.65 |
| **Average**         | **71.52**| **94.05** | **73.52** |       |
| moderately-to-poorly| 1        | 87.88 | 100.00 | 87.88 |
| differentiated      | 2        | 90.91 | 96.67 | 93.55 |
|                     | 3        | 87.88 | 100.00 | 87.88 |
|                     | 4        | 84.85 | 100.00 | 84.85 |
|                     | 5        | 87.88 | 100.00 | 87.88 |
| **Average**         | **87.88**| **99.33** | **88.41** |       |
| poorly              | 1        | 78.79 | 100.00 | 78.79 |
| differentiated      | 2        | 87.88 | 96.67 | 90.63 |
|                     | 3        | 84.85 | 100.00 | 84.85 |
|                     | 4        | 87.88 | 100.00 | 87.88 |
|                     | 5        | 81.82 | 96.43 | 84.38 |
| **Average**         | **84.24**| **98.62** | **85.30** |       |

As can be seen from table 1, the proposed model yield an accuracy of 73.94% in the healthy category, an accuracy of 80.00% in the adenomatous category, an accuracy of 71.52% in the moderately differentiated category, an accuracy of 87.88% in the moderately-to-poorly differentiated category, and an accuracy of 84.24% in the poorly differentiated category of the data set. In addition, we also plotted the ROC curves generated by the proposed model in these five categories, as shown in figure 5.
Figure 5. ROC curves generated by our model on the warwick-qu data set, in which (a)-(e) represents healthy, adenomatous, moderately differentiated, moderately-to-poorly differentiated and poorly differentiated categories, respectively.

3.3. Comparison of different classifier models
To objectively verify the performance of our model, we compare it with the famous k-nearest neighbor (KNN) classifier. To be fair, the data fed into the classifier are identical, and all of them have been extracted by the deep learning SAE algorithm. The five-fold cross-validation comparison results are summarized in Table 2. As can be seen from the table that our model achieves high scores except for specificity of moderately differentiated category, false precision of adenomatous category, and specificity of healthy category. This shows that the proposed method is an effective model for detecting images of cell cases and can help the pathologists’ clinical diagnosis.

Table 2. Comparison of experimental results between our model and KNN model on the warwick-qu data set

| Category                     | Evaluation Criteria | Proposed Model   | KNN Model      |
|------------------------------|---------------------|------------------|----------------|
| healthy                      | Accu. 73.94         | Spe. 96.26       | FPre. 76.19    |
|                              |                     |                  | 73.33          |
|                              |                     |                  | 97.59          |
|                              |                     |                  | 74.55          |
| adenomatous                  | Accu. 80.00         | Spe. 98.62       | FPre. 81.02    |
|                              |                     |                  | 73.33          |
|                              |                     |                  | 89.86          |
|                              |                     |                  | 81.57          |
| moderately differentiated    | Accu. 71.52         | Spe. 94.05       | FPre. 73.52    |
|                              |                     |                  | 68.48          |
|                              |                     |                  | 94.94          |
|                              |                     |                  | 70.97          |
| moderately-to-poorly         | Accu. 87.88         | Spe. 99.33       | FPre. 88.41    |
| differentiated              |                     |                  | 80.00          |
|                              |                     |                  | 89.63          |
|                              |                     |                  | 87.78          |
| poorly differentiated        | Accu. 84.24         | Spe. 98.62       | FPre. 85.30    |
|                              |                     |                  | 83.64          |
|                              |                     |                  | 97.85          |
|                              |                     |                  | 85.17          |

4. Conclusions
In this paper, we use random forest classifier to detect cell pathology images based on stacked auto-encoder algorithm. Firstly, we standardized the cell pathology images, and then use SAE to extract their hidden features, and finally use random forest to accurately detect their category. We used the five-fold cross-validation method to verify the performance of our model in the experiment. In addition, the proposed model was compared with the model based on the KNN classifier. In these experiments, our proposed model has achieved excellent results. This indicates that our proposed model is a suitable method for detecting images of cell pathology and can be used as an auxiliary tool to provide useful help for pathologists’ clinical diagnosis. In future research, we will use more feature extraction algorithms to extract the features of cell pathology images in anticipation of better detection results.
5. Acknowledgments
This work is supported in by the National Science Foundation of China, under Grant 61702444, in part by the West Light Foundation of The Chinese Academy of Sciences, under Grant 2018-XBQNZ-B-008, in part by the Zaozhuang Science and Technology Development Plan, under Grant 2018GX07, in part by the Zaozhuang University Research Fund Project. The authors would like to thank all anonymous reviewers for their constructive advices.

6. References
[1] Yan X, Yang L, Liu M, Wang Y, Lai M, Chang IC: Gland Instance Segmentation by Deep Multichannel Side Supervision. IEEE Transactions on Biomedical Engineering 2016, PP(99):1-1.
[2] Madabhushi A, Lee G: Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis 2016, 33:S1361841516301141.
[3] Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, Walliander M, Lundin M, Haglund C, Lundin J: Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific Reports 2018, 8(1).
[4] Plissiti ME, Christophoros N: Overlapping cell nuclei segmentation using a spatially adaptive active physical model. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 2012, 21(11):4568-4580.
[5] Khasawneh S, Al-Wahadni A, Lloyd CH: A Stochastic Polygons Model for Glandular Structures in Colon Histology Images. IEEE Transactions on Medical Imaging 2015, 34(11):2366-2378.
[6] Beck AH, Sangoi AR, Samuel L, Marinelli RJ, Nielsen TO, Vijver MJ, Van De, West RB, Matt VDR, Daphne K: Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Science Translational Medicine 2011, 3(108):108ra113.
[7] Wang L, You Z-H, Yan X, Xia S-X, Liu F, Li L-P, Zhang W, Zhou Y: Using Two-dimensional Principal Component Analysis and Rotation Forest for Prediction of Protein-Protein Interactions. Scientific reports 2018, 8(1):12874.
[8] Gehring J, Miao Y, Metze F, Waibel A: Extracting deep bottleneck features using stacked auto-encoders. In: IEEE International Conference on Acoustics, Speech and Signal Processing: 2013; 2013: 3377-3381.
[9] Shin HC, Orton MR, Collins DJ, Doran SJ, Leach MO: Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. IEEE Transactions on Pattern Analysis & Machine Intelligence 2013, 35(8):1930-1943.
[10] Belgiu M, Drăguţ L: Random forest in remote sensing: A review of applications and future directions. Isprs Journal of Photogrammetry & Remote Sensing 2016, 114:24-31.
[11] Wang L, You Z-H, Xia S-X, Liu F, Chen X, Yan X, Zhou Y: Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier. Journal Of Theoretical Biology 2017, 418:105-110.
[12] Gao ZG, Wang L, Xia SX, You ZH, Yan X, Zhou Y: Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins Using Autocovariance Transformation from PSSM. Biomed Research International 2016:8