IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

An advanced automated image analysis model for scoring of ER, PR, HER-2 and Ki-67 in breast carcinoma

Min Feng[1,2],Jie Chen[1],Xuhui Xiang[1],Yang Deng[1],Yanyan Zhou[1],Zhang Zhang[3], Zhongxi Zheng[1], Ji Bao[1] , Hong Bu[1,3*]

1.Laboratory of Pathology, West China Hospital, Sichuan University, Chengdu 610041, China.
2.Department of Pathology, West China Second University Hospital, Sichuan University & key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu 610041, China.
3. Department of Pathology, West China Hospital, Sichuan University, Chengdu 610041, China.

Corresponding author:Hong Bu（E-mail: hongbu@scu.edu.cn）.

ABSTRACT Immunohistochemistry (IHC) plays an important role in evaluating the status of ER, PR, Ki-67 and human epidermal growth factor receptor 2 (HER-2) during diagnosis of breast cancer. Although some existing automated approaches can solve the high time-consumption and inter-/intra-observer variability drawbacks to a certain extent, most of them are can't analyze both nuclear staining and cell membrane staining using the same method. This is attributed to the difference in localization of the positive signal of immunohistochemical staining in different biological markers. The present study proposes a novel automated image analysis model for scoring and grading of ER, PR, Ki-67 and HER-2 immunohistochemical images based on whole tissue sections in breast cancer. The scoring results of the trained model and manual interpretation of ER, PR, Ki-67 and HER-2 were then finally analyzed and compared. Experimental results show that the F1-measure was 0.8450, 0.8533 and 0.7962 for nuclear recognition of Ki-67, ER/PR and HER-2 respectively. For stain grading of Ki-67, ER/PR and HER-2, the F1-measure was 0.9776, 0.8306 and 0.9573 respectively. The scoring consistency of ER/PR, Ki-67 and HER-2 between our model and expert interpretation was 0.9279, 0.9712 and 0.8046 respectively. Our results demonstrate that artificial intelligence technology is a feasible and accurate method for accurate quantitative immunohistochemical analysis that can solve the drawbacks of low repeatability and time consumption brought by manual counting. The main contribution of our proposed model is that it can recognize both nuclear staining and cell membrane staining and grade the staining intensity as a sequential learning task.

INDEX TERMS automatic analysis; breast carcinoma; deep learning; Er; Her-2; immunohistochemistry; Ki-67; Pr.

I. INTRODUCTION

Breast cancer is the most common malignant tumor that harms women's health. Its occurrence even in the younger women has steadily increased in recent years [1]. There were more than 266,000 new cases of breast cancer in women in the United States in 2018. This accounted for 30% of all malignant tumors in women. It also significantly exceeded lung cancer (13%) which came second [1]. In the 2015 Chinese malignant tumor statistics, breast cancer ranked first in women with new malignant tumors (15%). It has become the leading cause of death in women under 45 years old [2]. ER, PR, Ki-67 and human epidermal growth factor receptor 2 (HER-2) proteins are the main biological indicators that guide the diagnosis, molecular classification, treatment plan and prognosis evaluation of breast cancer [3,4]. The expression of these biomarkers is commonly assessed by immunohistochemical (IHC) staining. However, this traditional scoring method is strongly dependent on the expertise and experience of histopathologists, it also has the disadvantages of being time-consuming and non-replicable in practice. Cognizant to this, these common problems remain a challenge for the pathologists to provide an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

accurate scoring for ER, PR, Ki-67 and HER-2 in breast cancer.

Modern artificial intelligence methods such as deep learning supplement pathologists' expertise in ensuring constant diagnostic accuracy. In recent years, increasing models have been developed for ER, PR, Ki-67 and HER-2 assisted computer automated analysis to overcome the major hindrances for evaluating the positive score in lung cancer, pancreatic cancer, gastroesophageal cancer, breast cancer, and other tumors [5-7]. Different deep learning networks and algorithms have been used for detection, segmentation and classification of cell membranes and nuclei from ER, PR, Ki-67 and HER-2 IHC images in breast cancer, this has yielded models such as HER-2Net deep learning network, Gamma mixture model and HscoreNet deep neural network structure among others [8-10]. Although most of those models have achieved good detection results, there are still some problems in practical applications. Saha et al. (2018) proposed a deep neural network structure named HER-2Net, which estimates the expression level of HER-2 by semantic segmentation of cell membranes and nuclei in pathological images of breast cancer tissues based on pixel classification of the entire visual field [8]. But this method does not follow the HER-2 immunohistochemical analysis guidelines which requires to show the membrane staining status of each cell [8]. Some methods for scoring of ER an PR, for example, the HscoreNet deep neural network structure, does not completely segments unevenly colored nuclei of ER and PR. What's more, this model requires labeling of the edges of the entire cell, which is extremely time-consuming [10]. From the reported research we can get, it is almost impossible to analyze both the nuclear staining and the cell membrane staining within the same pipeline with the help of the most existing automated approaches, because the localization of the positive signal of immunohistochemical staining is different in different biological markers [11,12]. That means different models are required to analyze different biomarkers at the same time. In view of the above reasons, an effective model that can solve the problem of accurate immunoassay under different staining modes with the premise of less labeling is needed. Visual saliency detection methods in quantitative analysis of pathological immunity can be applied to many different tasks. Cognizant to this, the fully convolutional networks was chosen as the nuclear detection backbone network in this study, and the DenseNet was used as the backbone network of the intensity classification model to recognize both nuclear and cell membrane staining results. The effectiveness of the pipeline was then verified via immunohistochemical staining of Ki-67, ER, PR and HER-2 in breast cancer tumor.

II. Materials and methods

A. Case selection

500 patients diagnosed with breast cancer who underwent surgery in 2017 or 2018 at West China Hospital of Sichuan University were selected for this study. All patients were pathologically diagnosed with invasive breast cancer by two senior pathologists, excluding those that received chemotherapy, radiotherapy, hormone therapy or immunotherapy before surgery. Collect paraffin samples from all patients for ER, PR, Ki-67 and HER-2 immunohistochemical staining. The dataset included Haematoxylin & Eosin (H&E), ER, PR, Ki-67 and HER-2 stained slides.

B. IMMUNOHISTOCHEMISTRY AND SLIDES COLLATION

For immunohistochemical staining of ER, PR, Ki-67 and HER-2, 4 um sections were freshly cut from the respective representative paraffin blocks and transferred onto slides. The slides were then incubated on a 600C hotplate for 10 minutes. The sections were then deparaffinized and rehydrated using xylene and graded alcohol. They were then stained using ER (clone SP1), PR (clone SP2), HER-2 (clone 4B5) and Ki-67 (clone MIB1) antibodies. Staining patterns were visualized by diaminobezidin (DAB) and counterstained with Mayer's hematoxylin. Appropriate positive and negative controls were included. Slides with objective reasons such as uneven staining and incomplete sections were not included in the study. Finally, only 215 cases with 1075 immunohistochemistry slides were collated.

C. IMAGE ACQUISITION

All the 1075 sections were collated into a complete digital scanning section Whole Slide Images (WSI) using a digital section scanner (Unic PRECICE600) set at a magnification of X40.

D. MANUAL ASSESSMENT

ER/PR positive has been defined as $\geqslant$ 1% labelled invasive tumor cells regardless of the staining intensity based on the 2010 ASCO/CAP (The American Society of Clinical Oncology and the College of American Pathologists) guidelines [13]. For Ki-67, manual counting of the positive tumor cells in the three high power fields (HPFs) and calculation of the average percentage of positive tumor cells was done [14,15]. On the other hand, positive signal localized on the cell membrane were detected for HER-2. HER-2 scoring was done based on the 2018 ASCO/CAP guidelines: HER-2 cases with a score of 0 or 1+ were classified as negative, those with a score of 3+ were classified as positive while those with a score of 2+ were classified as equivocal. The latter were further assessed by fluorescence in-situ hybridization (FISH) to test for gene amplification (A summary of guidelines for HER-2 IHC scoring criteria is shown in Table 1) [16].

III. Machine learning architectures

A. Data preparation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

35 cases of each marker were randomly selected and used as the experimental data. 158 fields of view for Ki-67, 40 fields of view for ER/PR and 47 fields of view for HER-2 under the X40 magnification were selected to train the nuclear detection model. Each field of vision contained positive and negative cells in different ratios. Different staining results for each marker were first classified before labeling. HER-2 staining results were divided into three categories according to the guidelines. Cells with strong, complete and uniform cell membrane staining were defined as category 3. Those with weak to medium intensity and with intact cell membrane staining were defined as category 2, while those with incomplete of no staining and with weak cell membrane staining were defined as category 1. Similarly, ER/PR staining results were divided into four categories: cells with dark brown and uniform cell nuclei staining were defined as category 4, those with medium intensity cell nuclei staining were defined as category 3, those with blue and light blue mixed with light brown cell nuclei staining were defined as category 2 while those with pure blue and light blue cell nuclei staining were defined as category 1. Ki-67 staining results were divided into two categories: cells with brown nuclei or no defined staining intensity were defined as category 2, while those with pure blue and light blue cell nuclei staining were defined as category 1. Different categories with labeled using different colors for each marker (Fig.1)

Table 1 Recommended HER-2 scoring criteria for IHC stained breast cancer tissue slides according to 2018 ASCO/CAP guidelines

Score	Cell Membrane Staining Pattern	Reporting Category
0	No membrane staining or incomplete membrane staining in ⩽10% of invasive tumour cells	Negative
1+	Faint/barely perceptible or weak incomplete membrane stainaing in > 10% of tumour cells	Negative
2+	A weak to moderate complete membrane staining is observed in > 10% of tumour cells or strong complete membrane staining in ⩽10% of tumour cells	Equivocal
3+	A strong (intense and uniform) complete membrane staining is observed in > 10% of invasive tumour cells	Positive

Each square field of view had a side length of 1600 pixels. After the data was marked by the pathologists, it was equally divided into four small images. The top left and bottom right images were used as the training data, the upper right image was used as the validation data while the lower left image was used as the test data. 119 training sets (72 fields for Ki-67, 20 fields for ER/PR and 27 fields for HER-2) of WSI were used to fit the parameters of the model while 58 verification sets (38 fields for Ki-67, 10 fields for ER/PR and 10 fields for HER-2) of WSI were used to tune the model hyperparameters during the training procedures. Another 58 sets (38 fields for Ki-67, 10 fields for ER/PR and 10 fields for HER-2) were selected for testing (Table 2).

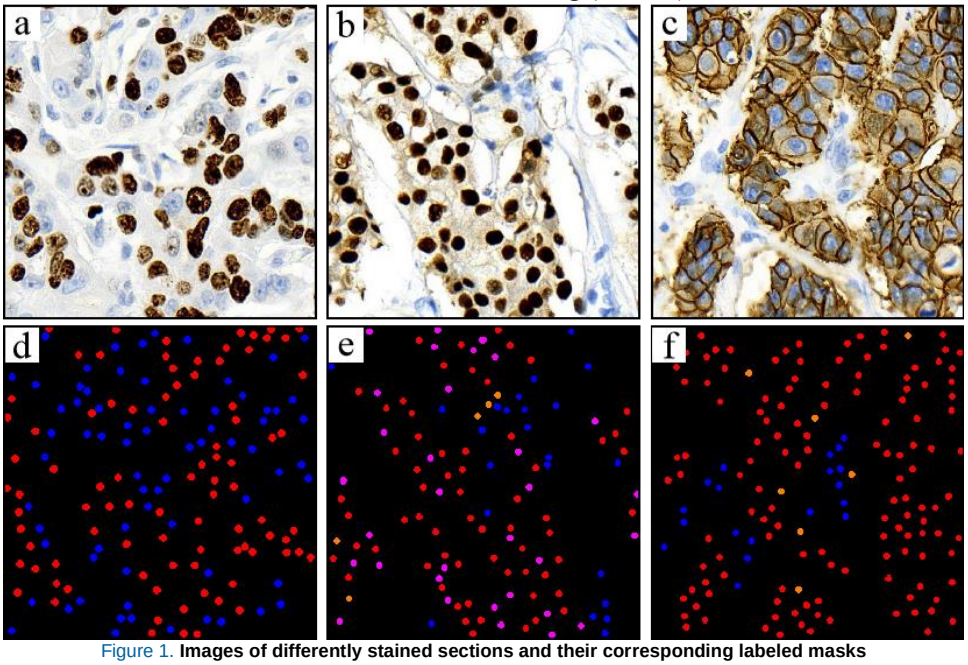


Figure 1. **Images of differently stained sections and their corresponding labeled masks**

Table 2 Selection and establishment of data sets of each marker

	Type	Fields	Resolution	TotalCells	Negative	Positive	+	++	+++
Train	Ki-67	72	800*800	36323	23449	12874	---	---	---
	ER\PR	20	800*800	7670	4612	---	575	645	1838
	HER-2	27	800*800	15126	---	---	4667	7780	2679
Val	Ki-67	38	800*800	20194	12178	8016	---	---	---

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

	ER\PR	10	800*800	4505	2930	---	427	754	394
	HER-2	10	800*800	5736	---	---	2173	2075	1488
	Ki-67	38	800*800	18470	11695	6775	---	---	---
Test	ER\PR	10	800*800	4264	1652	---	504	1593	515
	HER-2	10	800*800	6078	---	---	2277	1448	2353

A bounding box was drawn on 2000 nuclei whose average cell short axis length was found to be 22 pixels after calculations. A 22*22 matrix block that satisfies the gaussian distribution was first generated. A matrix block with a side length of 822 pixels was then generated and a Gaussian distribution small matrix block then inserted into the mask map with the center point coordinates of the cell. However, the newly inserted matrix block interfered with the matrix block inserted later because of cells overlap. Cognizant to this, a maximum pooling process of the block to be inserted and the mask matrix was done prior to insertion. The 800*800 block corresponding to the original image from the mask map was then taken out once the insertion was complete. 0.5 was finally used as the threshold to generate the mask of the equivalent kernel representation (Fig. 2).

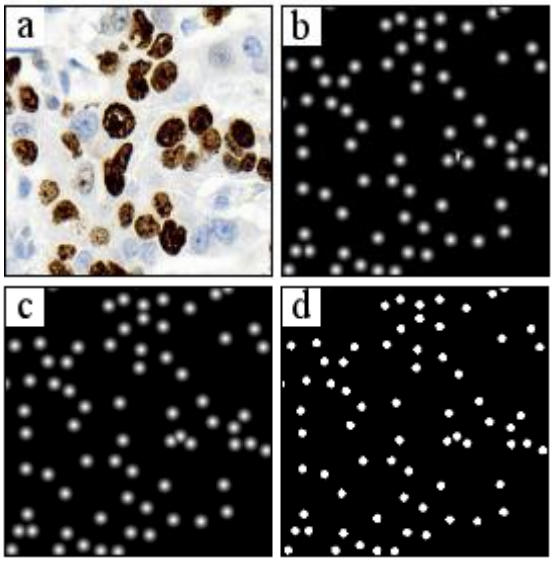


Figure 2. **Ground truth preparation for nuclear center point detection training data**

B. Universal immunohistochemical automatic detection model

Saliency detection is a type of algorithm in computer vision. The goal is to identify more prominent objects from the background. Unlike traditional machine learning algorithms that require manual design of feature extraction operators, models based on deep learning can automatically learn feature expressions in specific tasks [17]. Ki-67, ER and PR were nuclear stained in quantitative immunoassay. Nuclear detection and classification are typical target detection tasks. Since the workload of rectangular box labeling is too large, we define the nuclear detection as a saliency detection task.

Traditional saliency detection algorithms are design feature descriptors that rely heavily on prior knowledge of the data. The problem arise from the non-uniform definite degree of accuracy used in the nuclear and membrane antigen staining. However, this problem does not exist when the deep learning method is introduced. In this method, the kernel of the convolutional neural network is usually set to 3*3. The receptive field accumulated by the multi-layer deep convolutional neural network then extract the cell membrane and cytoplasm staining results to make decisions.

Visual saliency detection methods in quantitative analysis of pathological immunity can be applied to many different tasks. Cognizant to this, the fully convolutional networks was chosen as the backbone network in this study. It was the first model to propose segmentation using full convolutional neural networks.

Fully convolutional networks (FCN) and its variants have lot of successful applications in semantic image segmentation. The basic idea of FCN is to replace dense layers with convolutional layers. This makes the size of the model output to be consistent with the size of the input [18]. The results predicted by the model represent the same shape using ground truth training models of different sizes and shapes. This demonstrate that the FCN has a very strong characterization learning ability. Kainz et al. (2015) found that there are multiple extreme points in the model trained using equivalent kernels to characterize the nucleus. The two real points are easily merged using a post-processing algorithm that combines the multiple extreme points, which may result in missing the detection of the nucleus [19]. The nucleus has a small proportion in the field of view. As such, if a small kernel is used to characterize the nucleus, training the FCN leads to missed detection. At the same time, using larger kernel can cause overlap problems especially in areas with serious atypical nuclei. Both Janowczyk [20] and Xing [21] used the equivalent kernels as Ground Truth in the experiment, but they both took the detection task as a segmentation problem to train the model, that is, each pixel was classified into the foreground and background. While, we regard it as a regression problem, that is, using mean square error (MSE) as a loss function to return the probability that each pixel belongs to the foreground. If solved as a classification problem, the output of the network is N*M*2 and the two values corresponding to each pixel are the probabilities that the pixel belongs to each category. As Fig. 3, a sparse softmax cross entropy with logits is chosen as the loss function of training. When the problem is solved as a regression problem, the output of the model has only one channel. The MSE is thus used as the loss function. As provided in Equation (1), where M and N are the image width and height for each calculation error, j and k represent the pixel position, where $x_{j,\ k}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

represents the value of the ground truth of the pixel, and $x'_{j,k}$ represents the predicted value of the pixel. The convergence rate of regression has been found to be significantly higher than that of classification based on experiments.

$$MSE = \frac{1}{MN}\sum_{j=1}^{M}\sum_{k=1}^{N}\left(x_{j,k} - x'_{j,k}\right)^2 \quad (1)$$

Compared with the Faster R-CNN method of end-to-end regression targets that directly circumscribed rectangles and category classification [22], the regression probability map (PMap) method can only return the probability of the entire image belongs to the target [23]. So, it is necessary to use a post-processing algorithm to calculate the center point of the nucleus.

The probability map predicted using the equivalent nucleus as Ground Truth algorithm have multiple local maxima in the cell region. Moreover, it is not appropriate to use the find peaks algorithm to determine the center point of the nucleus. Cognizant to this, the morphological open operation is used to remove the noise after threshold segmentation of the probability map.

The connected component refers to an image area composed of foreground pixels having the same pixel value and adjacent positions in the image. Connected area analysis refers to finding and marking each connected area in an image and then further calculating the coordinates of the center point and other attributes. The output of the network is accompanied by a large amount of noise. However, a large amount of salt and pepper noise can be removed through the designed post-processing algorithm (Fig. 4).

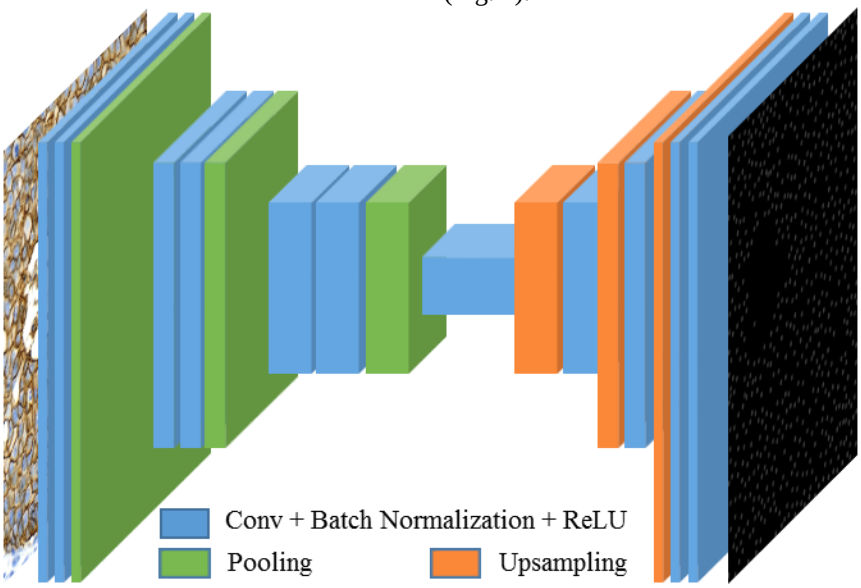


Figure 3. **An illustration of the cell detection network architecture**

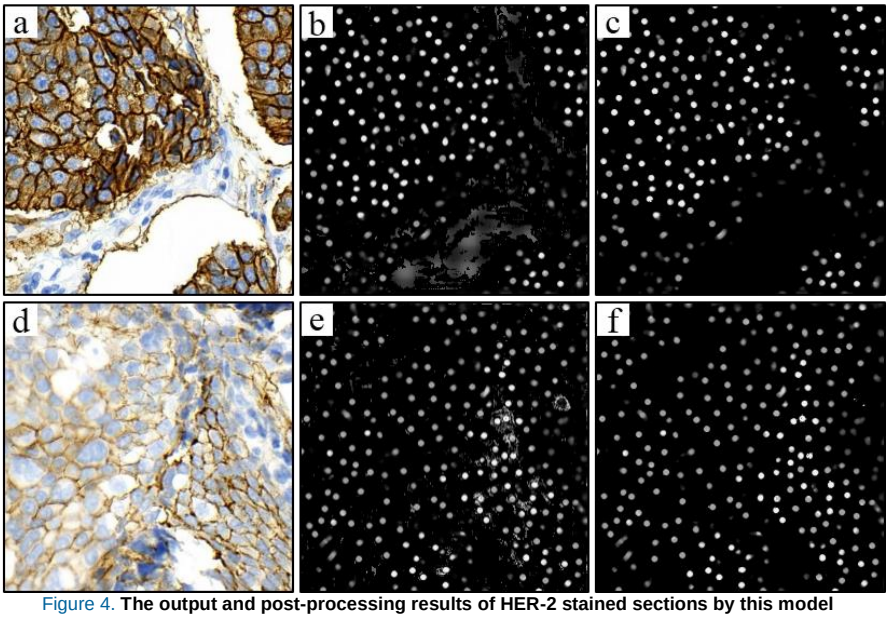


Figure 4. **The output and post-processing results of HER-2 stained sections by this model**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

The intensity of tumor cell immunohistochemical staining is determined by analyzing the Red, Green, Blue (RGB) values of the image. However, this method requires a segmentation algorithm to segment the area where the cell nucleus or cell membrane is located from the image. When non-deep learning methods are used in some staining or in cases of unevenness, segmentation fails resulting to failure of the intensity grading algorithm. On the other hand, when deep learning-based segmentation algorithm is used, the edge data of the segmentation target needs to be labeled for training. However, the labeling method used is time-consuming. Despite this, the classification algorithm based on deep learning can automatically extract the features of the cells to be classified for learning. Moreover, the model can automatically obtain the features of the data to be classified for classification under the training of differently labeled data. Herein, DenseNet was chosen as the backbone network for the classification model [24]. The cells were cropped from the image into small fields of view and then resized to a fixed size based on the detected cell nuclear center points. The fields of view were then fed into the classification model (Fig. 5).

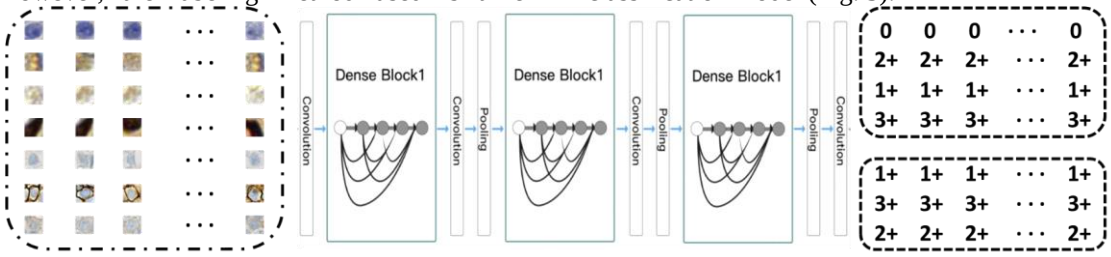


Figure 5. **The input and output of DenseNe network [24]. The input is the patch where each cell is located, the output is the corresponding category, and DenseNet is the backbone network of the hierarchical model.**

The performance of a model is evaluated from the perspective of nuclear detection and positive grading. The positive classification task is defined as a multi-class classification problem. The confusion matrix is first solved and then the model precision and f1 measure calculated. The detection task is then defined as a nuclear center point detection problem. The input of the task is an image while the output are the coordinates of the center point of all the nuclei on the image. The Hungarian algorithm is then used to match each detected nuclear center point with a manual annotation for each image. Recall, precision and f1-measure are then calculated using formula (2), (3) & (4) respectively.

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$f_1 measure = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Where TP refers to all points that match the manual annotation, FP refers to a point that does not match the manual annotation and FN refers to the residual point in the manual annotation that does not match the prediction result.

The model was implemented in python using a system for large-scale machine learning: TensorFlow. It was trained with the Adam algorithm at a learning rate of 0.0001and evaluated on a machine with an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz CPU, and an NVIDIA GTX1080Ti GPU. All the images were processed to RGB channels while the ground truth was given as a set of coordinates of dot annotations (one dot near cell centroid). 512 x 512 x 3 image patches were randomly cropped and used as the training data to prevent over-fitting. Moreover, a threshold of 0.5 was set during the processing of the feature map using threshold segmentation. 5 pixels were used as the structural element size to process noise using median filtering and morphological open operation. The maximum hit distance was set at 11 pixels to calculate the performance of the Hungarian algorithm.

IV. Results analysis

A. Nuclear detection

The prediction results of the model used in nuclear detection are shown in figure 6. The center point (nucleus) of most tumor cells were detected by the model. The F1-scores reached 0.8450, 0.8553 and 0.7962 for the different staining scenes of Ki-67, ER/PR and HER-2 respectively (Table 3). The F1-scores of the nuclear staining were higher compared to those of cell membrane staining. As the training progressed, the model could converge in less than 20 epochs under different staining scenarios of Ki-67, ER/PR and HER-2 (Fig. 7).

Table 3 The prediction results of our model in nuclear detection

	Precision	Recall	F1-score
Ki-67	0.8808	0.8156	0.8450
ER\PR	0.9549	0.7735	0.8533
HER-2	0.8864	0.7259	0.7962

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

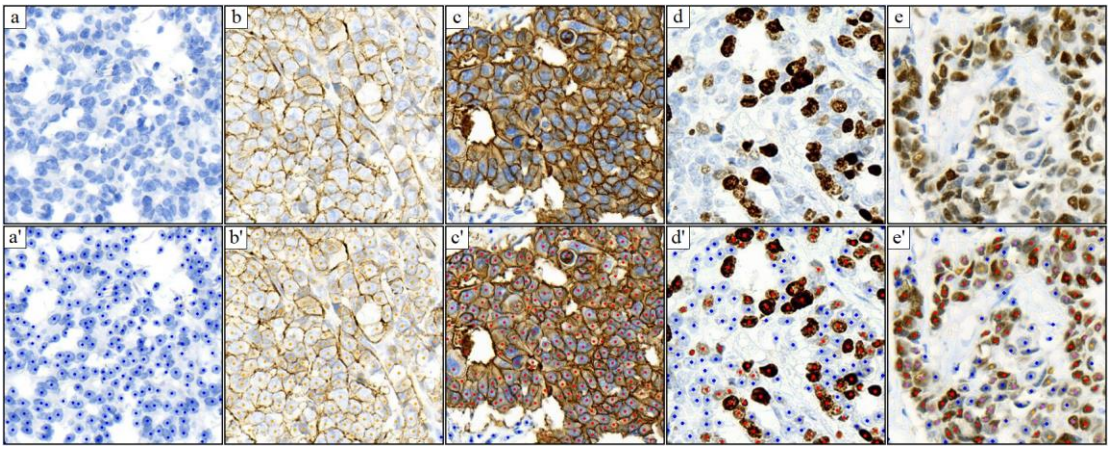


Figure 6. **The nuclear detection and classification results of the proposed approach in the three different scenarios of HER-2, Ki-67 and ER/PR.**

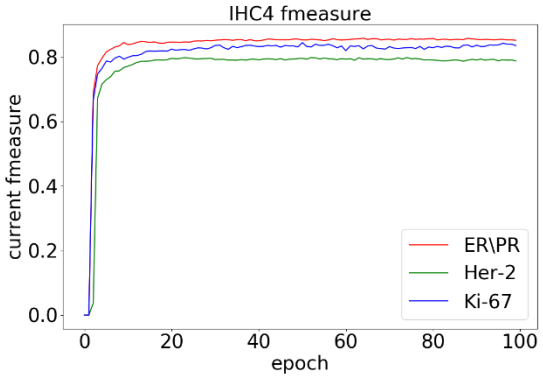


Figure 7. **The performance of nuclear detection and classification results of the proposed approach in the three different scenarios of HER-2, Ki-67 and ER/PR.**

B. Cell staining classification

The classification results after nuclear detection are shown in Fig. 6 (a')-(e'). For HER-2, the blue center points represent a negative result, the orange center points represent for category 2 (scoring 2+), and the red center points represent for category 3 (scoring 3+); For Ki-67, the red center points represent for category 2 (positive result), and the blue center points represent for category 1 (negative result); For ER/PR, the red center points represent for category 4, the purple center points represent for category 3, the orange center points represent for category 2, and the blue center points represent for category 1. The F1-score of Ki-67 negative and positive cell classification was the highest at 0.9776. It was followed by the F1-score of HER-2 1+, 2+ and 3+ three-class classification at 0.9573 while that of the four intensity classifications of ER/PR was the lowest at 0.8306. These differences were attributed to the fuzzy classification of the model between 1+ and 2+ (see the results in Table 4).

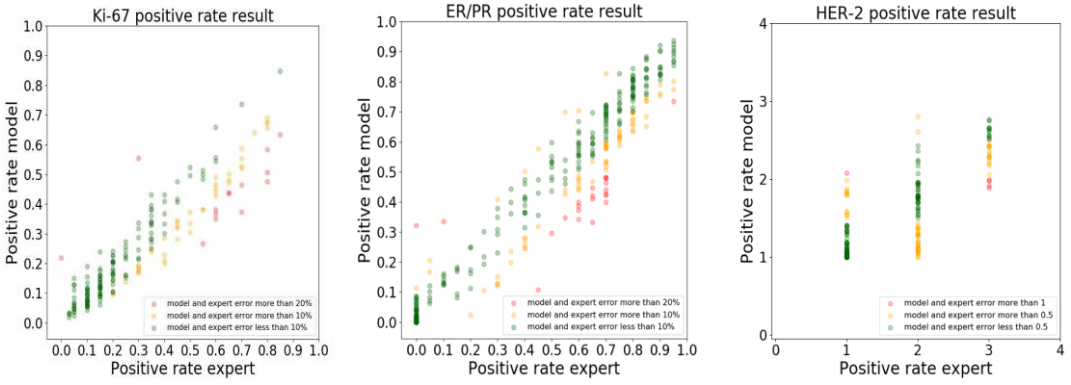
Table 4 The Precision, Recall and F1-score of the classification results after nuclear detection by the model in the three different scenarios of HER-2, ER/PR and Ki-67

Type	Classes	Precision	Recall	F1-score
Ki-67	0	0.9417	0.9700	0.9557
	1	0.9899	0.9800	0.9849
	Weighted avg	0.9779	0.9775	0.9776
ER/PR	0	0.9417	0.9700	0.9557
	1+	0.8690	0.7300	0.7935
	2+	0.7271	0.7140	0.7205
	3+	0.7979	0.9160	0.8529
	Weighted avg	0.8339	0.8325	0.8306
HER-2	1+	0.9898	0.9660	0.9777
	2+	0.9507	0.9260	0.9382
	3+	0.9333	0.9800	0.9561
	Weighted avg	0.9579	0.9573	0.9573

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

C. Correlation between model's automatic scoring and expert analysis in ki-67, er/pr and her-2 immunohistochemical staining scenarios

The correlation between the model's automatic score and the pathologists score under different immunohistochemical staining scenarios was studied to further optimize the model. 200 stained sections of Ki-67, ER/PR and HER-2 were randomly selected for these comparative studies. The correlation coefficients of the model's automatic score and experts score were 0.9279, 0.9712, and 0.8046 for Ki-67, ER/PR and HER-2 respectively. Scatter plots of the model interpretation results and expert interpretation results in different staining scenarios were then plotted to analyze these results more intuitively (Fig.8a-c). The red dots indicated that the absolute errors between the model interpretation and expert interpretation were greater than 10% and 20% respectively. The interpretation of Ki-67 and ER/PR immunohistochemical staining models was positively correlated with that of the experts. However, there were differences in IHC staining models and expert interpretation in HER-2. Expert interpretation only gave three types of results: negative (0/1 +), uncertain (2+) and positive (3+). On the other hand, the model calculated the proportion of various types of cells and then multiplied them by the various scores [1,3]. This gave the output as a continuous result as opposed to definite results. The continuity results were further plotted in a scatter plot (Figure 8c). The green points indicated that the absolute value of error between the expert reading and the model interpretation was less than 0.5 while the yellow and red points indicated that the absolute error value was greater than 0.5 and 1 respectively.



(a) (b) (c)

Figure 8. **Correlation between model's automatic scoring and expert analysis in Ki-67, ER/PR and HER-2 immunohistochemical staining scenarios**

V. Discussion

Automated analysis of pathology images has been in use for more than 20 years [25]. Unlike traditional machine learning algorithms that require manual design of feature extraction operators, deep learning-based models can learn feature expressions automatically in specific tasks [26]. Deep learning-based model was used in a multi-center study of multiple immune markers for breast cancer. This study included 8267 breast cancer patients thus making it the largest research done using this model [27]. In that study, a single tuned algorithm was used to score nuclear (ER, PR), membranous (HER-2, EGFR) and cytoplasmic (CK5/6) markers in tumor cells using the Ariol system based on tissue microarrays (TMA) [27]. For ER, PR and HER-2, the scores were lower than previously reported. This was attributed to the greater variability in TMA preparation in relation to specimen sources compared with whole tissues. Herein, a small large-scale evaluation of the performance of automated image analysis in scoring of breast carcinoma biomarkers based on whole tissue sections was done. This better reflected the real situation of tumor cells. The fully convolutional networks was chosen as the nuclear detection backbone network in this study, and the DenseNet was used as the backbone network of the classification model to recognize both nuclear and cell membrane staining results. The validity of the method was first verified on the Ki-67, ER and PR scenes of nuclear staining and then extended to the HER-2 scene of membrane staining. The saliency detection method could also locate the center point (nucleus) of tumor cells with HER-2 staining. Further post-processing and use of classification models achieved accurate analysis at the cell level. Compared with some existing methods that directly grade fixed-size visual fields or segment cell membranes, this method strictly follows the guidelines for quantitative analysis at the cell level. Moreover, the results of the model were positively correlated (0.8046) with those of expert's interpretation.

Immunohistochemical staining positive signals of ER and PR are located in the nucleus, results of immunohistochemistry are scored based on two important factors: staining intensity and percentage [4, 28]. Positive immunohistochemical staining of Ki-67 is also detected in the nucleus, and it regulates cell cycle as it is expressed at varying levels across G1, S, G2, and M phases. Numerous studies have shown that the combination of Ki-67, ER, PR and other tumor molecular markers provide better prognosis prediction for breast cancer and can guide the application of adjuvant treatment after tumor surgery [15,29,30,31]. Up to nowadays, there are several machine learning approaches have been reported about automated ER/PR and Ki-67

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011294, IEEE Access

image analysis, and most of them with a high concordance between manual scoring and digital image automated analysis [8,14,32,33]. The study by Mungle et al. (2017) characterized segmented ER cells through machine learning employing Markov random fields (MRF)and ANN methods were promising showing an F-measure of 0.9626, which is relatively significant [34]. A centralised evaluation study of 8088 patients from 10 study group also showed that automated image analysis may help to streamline and standardize Ki-67 scoring [35]. However, there are still two drawbacks in most methods of automated image analysis, one is that the edges of the entire tumor cells need to be labeled, and this polygon labeling is extremely time-consuming. The other is that many methods are only suitable for tumor cells with uniformly colored nuclei, ununiform staining images will cause an inaccurate result of a scoring model that depends on segmentation results. In this research, our model uses a post-processing algorithm combined with connected domain analysis based on morphology to accurately locate the position of the cell nucleus. Notably, the subsequent grading algorithm does not rely on the detected contours but directly takes blocks for classification based on the position of the center point of the nucleus. Unlike most of the existing algorithms, it isn't extremely dependent on the segmentation results. It can also be applied to the case of uneven coloring. Detection of nucleus by our model reveals that the F1-measures of Ki-67 and ER/PR are 0.844962 and 0.855311, respectively. We further analyzed 115 groups of slices using this digital nuclear classification algorithm on a digital pathology workstation and compared the automatic analysis results with manual interpretation. Results show that our algorithm is highly adaptable for images with different staining intensities, and the correlation coefficient with the doctor's reading is nearly 95%. The error rate of automatic score and manual score in individual cases exceeds 20%, we postulate that this may be caused by the coexistence of tumor cells and lymphocytes in the interpretation area. The model may interpret some lymphocytes as negative tumor cells resulting in a low positive rate. In our subsequent study, we will further classify negative cells and remove the lymphocytes from the negative cells.

Unlike ER/PR，Ki-67 and other nuclear-biomarkers，HER-2 is a membrane-positive biomarker which regulates cell proliferation and cell growth. In practice, an expert pathologist will report a score between 0 and 3+ and cases scoring 0 or 1+ are classified negative whilst those scoring 3+ are considered as positive. Cases with score 2+ are classified as equivocal and their samples are further examined with FISH to test for gene amplifications [16]. However, in this process, the interpretation of tumor cell membrane staining "integrity" is relatively subjective. Differences in observations from different pathologists, differences in understanding of the testing and in specimen fixation and production will significantly affect the scoring results. The recommendations on HER-2 scoring systems show that up to 20% of the HER-2 IHC results may contain inaccuracies due to variations in the technical rigor and the subjective nature of the scoring [16]. Currently, several software for HER-2 scoring have been developed and some of them were put on the market, such as the Automated Cellular Imaging System III (ACIS III) and the HER-2-CONNECT, with accuracy of up to 82-98% [5,6,36-39]. To our knowledge, most existing HER-2 automatic scoring algorithms, including HER-2Net deep neural network, are mainly based on cell membrane segmentation or small patch classification，which can only compare pixel values, or perform fine-grained image analysis. Automatic analysis of HER-2, that is, applying the nuclear detection method based on saliency detection to the cell membrane staining scene, reveals that the method can detect the center point of each cell, and accurately calculate the membrane staining intensity of each cell using blocks that graded for staining intensity. Although the central point detection algorithm we proposed is applied to the cell membrane scene, the F1-measure is 0.79, which is slightly lower than the nuclear staining scene, it strictly follows the guidelines for quantitative analysis at the cell level and has more than 80% consistency with expert interpretation results. We further analyzed the reason why the F1-measure value is lower in the HER-2 scoring scene, that may be because the HER-2-negatively stained nuclei showed a light blue color, which is similar to the background, resulting in omission of these tumor cells during detection. Comparative analysis for model scores and manual scores of 115 randomly selected HER-2 immunohistochemical staining sections shows that some 2+ cases judged by experts were very close to negative after quantitative analysis by the model, like the results of Koopman's research [36]. The reduction in 2+ cases may be due to that some equivocal cases which judged as 2+ by experts in order to take FISH test were scoring as category 1 by the model. In the next step, we will take FISH test for HER-2 2+ cases, and regard the FISH results as the gold standard, in order to evaluate the model's performance.

VI. Conclusion

In this study, we present a novel deep learning framework to bring an efficient method without labeling each cell outline and perform automatic scoring for nuclear (ER, PR, Ki-67) and membranous (HER-2) markers. Compared with previous algorithms, our method not only greatly reduces the workload of manual labeling，but also provide accurate analysis for nuclear and cell membrane immunohistochemical staining. The method solves the problem of nuclear detection and classification. This highlights the potential of Artificial Intelligence (AI) techniques for examination of IHC slides and accurate quantitation of immune staining. Nevertheless, there are some unsatisfactory points in the results of nuclear detection and staining intensity classification in our method, we still have a lot of work to do in the next research.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018: Cancer Statistics, 2018," CA: A Cancer Journal for Clinicians, vol. 68, no. 1, pp. 7–30, Jan. 2018, doi: 10.3322/caac.21442.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[3] M. Van Bockstal, G. Floris, C. Galant, K. Lambein, and L. Libbrecht, "A plea for appraisal and appreciation of immunohistochemistry in the assessment of prognostic and predictive markers in invasive breast cancer," The Breast, vol. 37, pp. 52–55, Feb. 2018, doi: 10.1016/j.breast.2017.10.012.

[4] M. J. Duffy et al., "Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM)," European Journal of Cancer, vol. 75, pp. 284–298, Apr. 2017, doi: 10.1016/j.ejca.2017.01.017.

[5] J. Jeung, R. Patel, L. Vila, D. Wakefield, and C. Liu, "Quantitation of HER2/ neu Expression in Primary Gastroesophageal Adenocarcinomas Using Conventional Light Microscopy and Quantitative Image Analysis," Archives of Pathology & Laboratory Medicine, vol. 136, no. 6, pp. 610–617, Jun. 2012, doi: 10.5858/arpa.2011-0371-OA.

[6] A. Brügmann et al., "Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains," Breast Cancer Res Treat, vol. 132, no. 1, pp. 41–49, Feb. 2012, doi: 10.1007/s10549-011-1514-2.

[7] M. Jin, R. Roth, V. Gayetsky, N. Niederberger, A. Lehman, and P. E. Wakely, "Grading pancreatic neuroendocrine neoplasms by Ki-67 staining on cytology cell blocks: manual count and digital image analysis of 58 cases," Journal of the American Society of Cytopathology, vol. 5, no. 5, pp. 286–295, Sep. 2016, doi: 10.1016/j.jasc.2016.03.002.

[8] M. Saha and C. Chakraborty, "Her2Net: A Deep Framework for Semantic Segmentation and Classification of Cell Membranes and Nuclei in Breast Cancer Evaluation," IEEE Trans. on Image Process., vol. 27, no. 5, pp. 2189–2200, May 2018, doi: 10.1109/TIP.2018.2795742.

[9] M. Saha, C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee, "An Advanced Deep Learning Approach for Ki-67 Stained Hotspot Detection and Proliferation Rate Scoring for Prognostic Evaluation of Breast Cancer," Sci Rep, vol. 7, no. 1, p. 3213, Dec. 2017, doi: 10.1038/s41598-017-03405-5.

[10] M. Saha, I. Arun, R. Ahmed, S. Chatterjee, and C. Chakraborty, "HscoreNet: A Deep network for estrogen and progesterone scoring using breast IHC images," Pattern Recognition, vol. 102, p. 107200, Jun. 2020, doi: 10.1016/j.patcog.2020.107200.

[11] P. E. Aktan, G. Hatipoğlu, and N. Arıca, "Meme Kanseri Tanısı için HER2 Testine Dayalı Risk Sınıflandırması Risk Classification For Breast Cancer Diagnosis Using HER2 Testing," p. 4.

[12] G. Palacios-Navarro, J. M. Acirón-Pomar, E. Vilchez-Sorribas, and E. G. Zambrano, "Medical image segmentation to estimate HER2 gene status in breast cancer," presented at the PROGRESS IN APPLIED MATHEMATICS IN SCIENCE AND ENGINEERING PROCEEDINGS, Bali, Indonesia, 2016, p. 020026, doi: 10.1063/1.4940274.

[13] C. Liedtke and O. Gluz, "Pathologists' Guideline Recommendations for Immunohistochemical Testing of -Estrogen and Progesterone Receptors in Breast Cancer," p. 3.

[14] T. Koopman, H. J. Buikema, H. Hollema, G. H. de Bock, and B. van der Vegt, "Digital image analysis of Ki67 proliferation index in breast cancer using virtual dual staining on whole tissue sections: clinical validation and inter-platform agreement," Breast Cancer Res Treat, vol. 169, no. 1, pp. 33–42, May 2018, doi: 10.1007/s10549-018-4669-2.

[15] M. K. K. Niazi, C. Senaras, M. Pennell, V. Arole, G. Tozbikian, and M. N. Gurcan, "Relationship between the Ki67 index and its area based approximation in breast cancer," BMC Cancer, vol. 18, no. 1, p. 867, Dec. 2018, doi: 10.1186/s12885-018-4735-5.

[16] A. C. Wolff et al., "Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update," JOURNAL OF CLINICAL ONCOLOGY, p. 20, 2013.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Machine Intell., vol. 20, no. 11, pp. 1254–1259, Nov. 1998, doi: 10.1109/34.730558.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," p. 10.

[19] P. Kainz, M. Urschler, S. Schulter, P. Wohlhart. and V. Lepetit,"You Should Use Regression to Detect Cells， " Presented at 2015 Medical Image Computing and Computer-Assisted Intervention, Munich, Germany. 2015, pp. 276-283.

[20] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," J Pathol Inform, vol. 7, no. 1, p. 29, 2016, doi: 10.4103/2153-3539.186902.

[21] F. Xing, Y. Xie, and L. Yang, "An Automatic Learning-Based Framework for Robust Nucleus Segmentation," IEEE Trans. Med. Imaging, vol. 35, no. 2, pp. 550–566, Feb. 2016, doi: 10.1109/TMI.2015.2481436.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[23] H. Höfener, A. Homeyer, N. Weiss, J. Molin, C. F. Lundström, and H. K. Hahn, "Deep learning nuclei detection: A simple approach can deliver state-of-the-art results," Computerized Medical Imaging and Graphics, vol. 70, pp. 43–52, Dec. 2018, doi: 10.1016/j.compmedimag.2018.08.010.

[24] G. Huang, Z. Liu, and L. van der Maaten, "Densely Connected Convolutional Networks," Presented at 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA. [Online]. Available: https://github.com/liuzhuang13/DenseNet.

[25] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review—Current Status and Future Potential," IEEE Rev. Biomed. Eng., vol. 7, pp. 97–114, 2014, doi: 10.1109/RBME.2013.2295804.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[27] W. J. Howat et al., "Performance of automated scoring of ER, PR, HER2, CK5/6 and EGFR in breast cancer tissue microarrays in the Breast Cancer Association Consortium: Automated scoring of breast tumour TMAs," J Pathol Clin Res, vol. 1, no. 1, pp. 18–32, Jan. 2015, doi: 10.1002/cjp2.3.

[28] S. Tewary, I. Arun, R. Ahmed, S. Chatterjee, and C. Chakraborty, "AutoIHC-scoring : a machine learning framework for automated Allred scoring of molecular expression in ER- and PR-stained breast cancer tissue: AUTOIHC-SCORING," Journal of Microscopy, vol. 268, no. 2, pp. 172–185, Nov. 2017, doi: 10.1111/jmi.12596.

[29] [R. Ohashi, S. Namimatsu, T. Sakatani, Z. Naito, H. Takei, and A. Shimizu, "Prognostic utility of atypical mitoses in patients with breast cancer: A comparative study with Ki67 and phosphohistone H3: OHASHI ET AL.," J Surg Oncol, Aug. 2018, doi: 10.1002/jso.25152.

[30] D. L. Rimm et al., "An international multicenter study to evaluate reproducibility of automated scoring for assessment of Ki67 in breast cancer," Mod Pathol, vol. 32, no. 1, pp. 59–69, Jan. 2019, doi: 10.1038/s41379-018-0109-4.

[31] S. Kurozumi et al., "Power of PgR expression as a prognostic factor for ER-positive/HER2-negative breast cancer patients at intermediate risk classified by the Ki67 labeling index," BMC Cancer, vol. 17, no. 1, p. 354, Dec. 2017, doi: 10.1186/s12885-017-3331-4.

[32] F. Klauschen et al., "Standardized Ki67 Diagnostics Using Automated Scoring--Clinical Validation in the GeparTrio Breast Cancer Study," Clinical Cancer Research, vol. 21, no. 16, pp. 3651–3657, Aug. 2015, doi: 10.1158/1078-0432.CCR-14-1283.

[33] A. L. D. Araújo et al., "The performance of digital microscopy for primary diagnosis in human pathology: a systematic review," Virchows Arch, vol. 474, no. 3, pp. 269–287, Mar. 2019, doi:

10.1007/s00428-018-02519-z.

[34] T. Mungle et al., "MRF-ANN: a machine learning approach for automated ER scoring of breast cancer immunohistochemical images: MRF-ANN," Journal of Microscopy, vol. 267, no. 2, pp. 117–129, Aug. 2017, doi: 10.1111/jmi.12552.

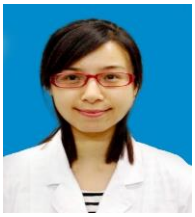
[35] M. Abubakar et al., "Prognostic value of automated KI67 scoring in breast cancer: a centralised evaluation of 8088 patients from 10 study groups," Breast Cancer Res, vol. 18, no. 1, p. 104, Dec. 2016, doi: 10.1186/s13058-016-0765-6.

[36] T. Koopman, H. J. Buikema, H. Hollema, G. H. Bock, and B. Vegt, "What is the added value of digital image analysis of HER 2 immunohistochemistry in breast cancer in clinical practice? A study with multiple platforms," Histopathology, vol. 74, no. 6, pp. 917–924, May 2019, doi: 10.1111/his.13812.

[37] T. Qaiser and N. M. Rajpoot, "Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring," IEEE Trans. Med. Imaging, vol. 38, no. 11, pp. 2620–2631, Nov. 2019, doi: 10.1109/TMI.2019.2907049.

[38] T. Qaiser et al., "HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues," Histopathology, vol. 72, no. 2, pp. 227–238, Jan. 2018, doi: 10.1111/his.13333.

[39] M. E. Vandenberghe, M. L. J. Scott, P. W. Scorer, M. Söderberg, D. Balcerzak, and C. Barker, "Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer," Sci Rep, vol. 7, no. 1,p. 45938, May 2017, doi: 10.1038/srep45938.



Min Feng, graduated from Peking Union Medical College with a master's degree in Pathology in July 2011. From 2018, she began to pursue a PhD in Clinical Medical College of Sichuan University, majoring in Digital Pathology and Artificial Intelligence. In the past decade, the accumulation of professional training and work has provided her the solid basic theoretical knowledge and creative scientific research thinking in the field of gynecological oncology. At present, some meaningful work has been done on breast cancer by artificial intelligence.

Jie Chen was engaged in machine learning and computer vision algorithm research in the data mining laboratory of Jiangxi University of Traditional Chinese Medicine from 2013 to 2017, and obtained a bachelor's degree in 2017. He is engaged in digital pathology and artificial intelligence-related research in the Pathology Laboratory of West China Hospital of Sichuan University.

Xuhui Xiang received a bachelor's degree in mechanical engineering and automation from the Chengdu University, Chengdu, China, in 2013. He obtained the second major of Computer engineering application in university. From 2013 to 2017, he worked as a nuclear power mechanical design engineer at Nuclear Power Institute of China, specializing in computer vision and machine learning.

Yang Deng, graduated from Sichuan University with a master's degree in Software Engineering in 2017, and graduated from Chengdu University of Traditional Chinese Medicine with Bachelor of medicine in 2008. Now, he is working in Sichuan University West China Hospital, specializing in digital pathology (DP) and artificial intelligence (AI), especially the application of AI in breast cancer.

Zhang Zhang , Ph.D. of Pathology and Pathophysiology, Graduated from West China Hospital of Sichuan University in 2015. As a joint-training doctor, she went to the University of Rochester Medical Center in 2012-2013. Now she is working in West China Hospital as a Pathologist, Specializing in breast pathology and molecular pathology. She is a young member of the Professional Committee of Pathology of Sichuan Medical Association, a young member of Chinese Cancer Association Professional Committee of Oncology and Pathology. At present, nearly 20 articles were published as the first author or corresponding author.

YanYan Zhou, Bachelor of Science.Graduated from Chengdu University of Traditional Chinese Medicine in Biotechnology in 2016; She is working at West China Hospital of Sichuan University, engaged in routine pathology and digital pathology. As a research technician, she did a lot of tedious preliminary work in many scientific studies.

Zhongxi Zheng acquired a PhD degree in Pathology Informatics from National Hirosaki University, Japan, and M.S. degree in Pattern Recognition and Artificial Intelligence from China Academy of Sciences. He is working in West China Hospital as a professor, responsible for researches of digital pathology, computer vision and machine learning, especially its applications in diagnosis of breast cancer and cervical cancer. He has published over 30 journal papers and conference proceedings, and received multiple international professional awards so far in the related fields.

Ji Bao is a medical doctor of Sichuan University and a postdoctoral fellow of West China Hospital of Sichuan University. As a joint-training doctor and visiting scientist, he went to the Mayo Clinic Department of Biomedical Engineering in the United States twice in 2010 and 2013. He is currently an associate researcher and master tutor of West China Hospital of Sichuan University. The main research areas are digital pathology and artificial intelligence and bioartificial liver support systems.

Hong Bu is a Ph.D. from West China Medical University and a leader in academic and technology in Sichuan Province. He is the director of the Pathology Laboratory of West China Hospital of Sichuan University, a professor of pathology and a doctoral supervisor. He is the general leader of the "standardized diagnosis standard of tumor pathology" of the National Health and Health Commission, and participated in the diagnosis and treatment of major tumors in China (lung cancer, breast cancer, liver cancer, etc.). He is mainly engaged in breast pathology, molecular pathology diagnosis, digital pathology and artificial intelligence research. He have published more than 100 papers in SCI as the first author and corresponding author, and applied for 19 invention patents, authorized 9 items, and transferred 2 items to Mayo Clinic.