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

Cervical Histopathology Image Classification Using Multilayer Hidden Conditional Random Fields and Weakly Supervised Learning

C. LI[1], H. CHEN[1], L. ZHANG[2], N. XU[3], D. XUE[1], Z. HU[1], H. MA[1], H. SUN[2]

[1]Microscopic Image and Medical Image Analysis Group, Northeastern University, Shenyang, China (E-mail: lichen@bmie.neu.edu.cn)
[2]Department of Radiology, Shengjing Hospital, China Medical University, Shenyang, China
[3]Liaoning Shihua University, Fushun, China

Co-first author: H. Chen; Corresponding author: H. Sun (E-mail: sunhongzan@126.com).

This work is supported by the "National Natural Science Foundation of China" (No. 61806047), the "Fundamental Research Funds for the Central Universities" (No. N171903004), and the "Scientific Research Launched Fund of Liaoning Shihua University" (No. 2017XJJ-061).

ABSTRACT In this paper, a novel *Multilayer Hidden Conditional Random Fields* (MHCRFs) based *Cervical Histopathology Image Classification* (CHIC) model is proposed to classify well, moderate and poorly differentiation stages of cervical cancer using a *Weakly Supervised Learning* strategy. First, color, texture and *Deep Learning* features are extracted to represent the histopathological image patches. Then, based on the extracted features, *Artificial Neural Network*, *Support Vector Machine* and *Random Forest* classifiers are designed to calculate the patch-level classification probabilities. Thirdly, effective classifiers are selected to generate unary and binary potentials. Lastly, using the generated potentials, the final image-level classification results are predicted by our MHCRF model, and an overall accuracy around 77.32% is obtained on six practical cervical histopathological image datasets with more than 600 immunohistochemical (IHC) stained samples. Among the six test accuracies, the highest reaches 88%. Furthermore, we also test our MHCRF method with a gastric hematoxylin-eosin (HE) stained histopathological image dataset including 200 images for an extended experiment, and achieve an accuracy of 93%.

INDEX TERMS Cervical Cancer, Conditional Random Fields, Deep Learning, Feature Extraction, Histopathological Image, Weakly Supervised Learning

I. INTRODUCTION

Among females, cervical cancer ranks fourth for both incidence and mortality. In 2018, the number of new cases of cervical cancer is 569847 in worldwide, accounting for 3.2% of all new cancer cases; the number of cervical cancer deaths is 311365, accounting for 3.3% of all cancer deaths [4]. In all of the 185 countries surveyed, the incidence of cervical cancer is the highest among women in 28 countries, and the number of countries with the highest mortality rate reaches 42 [4], [50]. Hence, the study of cervical cancer is very important and attracts a lot of attention from different scientific fields.

In recent years, *Machine Learning* (ML) plays a more and more important role in the computer-aided diagnosis (CAD) of cervical cancer. In terms of *Cervical Histopathology Image Classification* (CHIC), a variety of ML methods are developed and applied to image segmentation, feature extraction and classification tasks. From decision trees to *Support Vector Machines* (SVMs), from classical *Artificial Neural Networks* (ANNs) to complex *Deep Learning* (DL), the ML methods are constantly updated with the development of technology in the CHIC field. However, the existing approaches usually focus on individual characteristics and properties, such as color, shape or texture features, without a strategy to describe the integral information. Therefore, some advanced methods are proposed to integrate these individual existing methods to obtain an even better performance. Especially, because *Conditional Random Fields* (CRFs) can characterize the spatial relationship of images, they are effective and robust methods for analysing the contents of complex images. Whilst, due to medical doctors usually are too busy to support a large number of ground truth images for

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.2019.2924467, IEEE Access

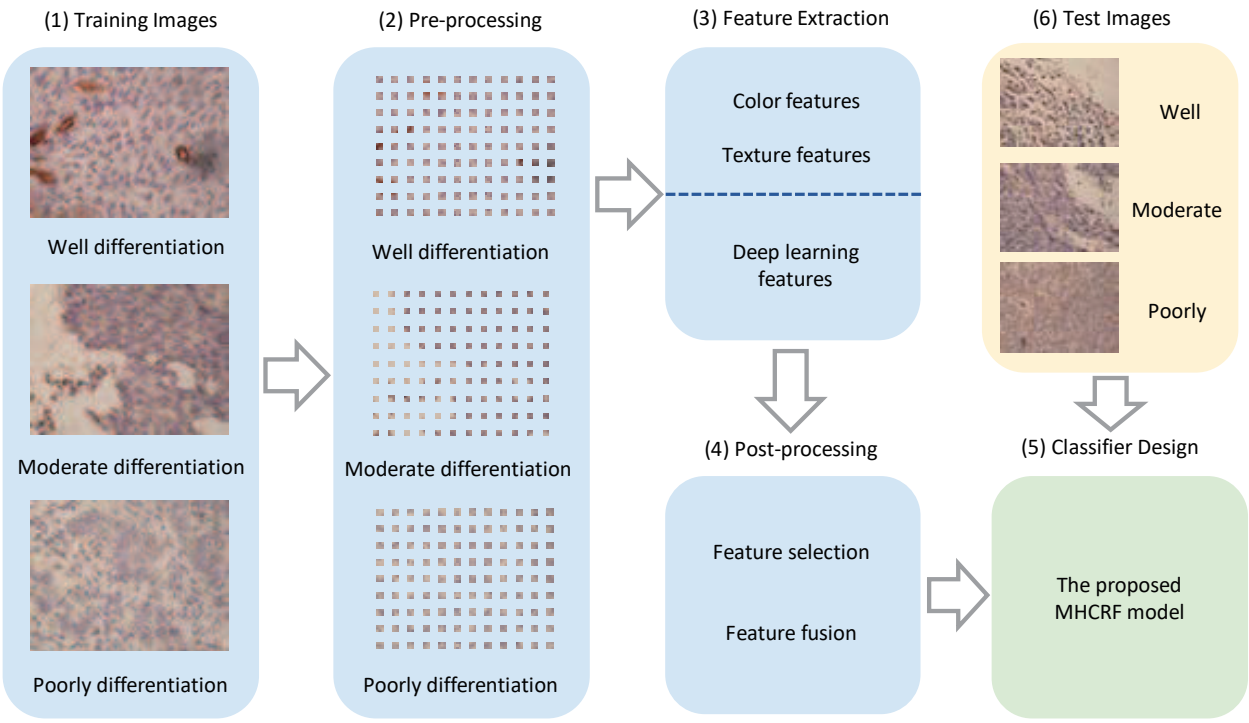

FIGURE 1. The workflow of the proposed weakly supervised MHCRF method. The blue boxes denote the training process, the green box represents the MHCRF model, and the yellow box shows the test process.

a strongly supervised learning process, an effective weakly supervised learning approach is expected. Hence, we propose a weakly supervised *Multilayer Hidden Conditional Random Field* (MHCRF) framework to address the CHIC problem, where the cervical histopathological images are mixed with the complicated nucleus, interstitial and tissue fluid. As far as we know, in the CHIC field, MHCRFs are not used before our work. Furthermore, because cellular visual features in a histopathological image are observed on patch scales, training a patch-level classifier has good performance as training an image-level classifier [20].

The workflow of the proposed MHCRF model is shown in Fig. 1, and the details are introduced in Sec. III:

- Step 1 (Data Input): Input cervical histopathological digital images of training set and validation set to the proposed MHCRF framework for a *Weakly Supervised Learning* process.
- Step 2 (Image Pre-processing): Image meshing is used as a method of image pre-processing to support the next feature extraction step. We unify the image size to 1280×960 pixels, then mesh the image into patches (100×100 pixels).
- Step 3 (Feature Extraction): We extract multiple features from the pre-processed image patches, including color, texture and DL features. Color features: Intensity histograms of R, G, B channels and gray-level images. Texture features: Scale-invariant Feature Transform (SIFT), DAISY, Gray-level Co-occurrence Matrix (GLCM) and Histogram of Oriented Gradient (HOG) features. DL features: Inception-V3 [42] and VGG-16 [40] features.
- Step 4 (Post-processing): In order to obtain a priori probability, we use SVMs, ANNs, and RFs to pre-classify image patches. In the SVMs, 'RBF' and 'linear' kernels are applied. In the ANNs, different hidden layers are compared. In the RFs, different numbers of trees are tested. Finally, we obtain 19 types of classifiers, and they are trained with seven features, resulting in 133 patch-level pre-classification results. Furthermore, we select the top 8% of these 133 feature-classifier combinations to generate our unary and binary potentials.
- Step 5 (Classifier Design): Based on the selected patch-level pre-classification results, we generate unary and binary potentials of the MHCRF, and combine them to calculate the joint probability for the final image-level classification.
- Step 6 (System Evaluation): We input test images to the trained MHCRF framework to evaluate the effectiveness of the proposed method.

The structure of this paper is as follows: Sec. II is the related work about cervical cancer, ML techniques for cervical cancer and applications of CRFs. Sec. III specifies how the MHCRF in this paper is designed. Sec. IV is experiments and results of the proposed method. Finally, Sec. V closes this paper with a brief conclusion.

II. RELATED WORK

This section summarizes the related work of this paper and consists of three subsections. Sec. II-A is about cervical cancer, Sec. II-B is the ML techniques for cervical cancer, and Sec. II-C is the applications of CRFs.

A. CERVICAL CANCER

According to the data of [4], in the worldwide female cancer diseases, the incidence of cervical cancer in 2018 accounts for 6.6% and the mortality rate is 7.5%. Both morbidity and mortality are ranked fourth in the world. Therefore, the prevention and timely diagnosis of cervical cancer are particularly important. The biopsy diagnosis of cervical cancer can be roughly divided into three types. The first is the Pap test [11], which can be used as a screening test, but produces a false negative in up to 50% of cases of cervical cancer [3]. Confirmation of the diagnosis of cervical cancer or precancer requires a biopsy of the cervix. This is often done through colposcopy, a magnified visual inspection of the cervix aided by using a dilute acetic acid (e.g. vinegar) solution to highlight abnormal cells on the surface of the cervix [26]. Further diagnostic and treatment procedures are loop electrical excision procedure [48] and cervical conization [34], in which the inner lining of the cervix is removed to be examined pathologically. These are carried out if the biopsy confirms severe cervical intraepithelial neoplasia [26]. With the rise and development of computer technology, more and more ML related technologies have been applied to the medical field. The diagnosis of histopathological images of cervical cancer is no exception.

B. ML TECHNIQUES IN CERVICAL CANCER RESEARCH

a: *Feature Extraction Methods:*

Scale-invariant feature transform (SIFT) is a classical operator for extracting local features of images [30]. The essence of SIFT algorithm is to find the key points in different scale space and calculate the direction of the key points [31]. The description is characterized by scale invariance.

DAISY [44] is a description operator which can quickly calculate local image features in the face of dense feature extraction. DAISY extends the basic idea of SIFT: Block statistical gradient direction histogram. The difference is that DAISY uses Gauss convolution to aggregate the histograms of gradient direction [45]. Because of the convolution property of Gauss kernels, the gradient graphs with different weights can be obtained only by convoluting the gradient graphs several times when calculating DAISY operators.

The gray-level of the pixel appears repeatedly in the spatial position to form the texture of the image [16]. Gray-level co-occurrence matrix (GLCM) describes the joint distribution of the gray-level of two pixels with some spatial position relationship. GLCM not only reflects the distribution characteristics of brightness, but also reflects the location distribution characteristics between pixels with the same brightness or near brightness [22]. It is a second-order statistical feature of image brightness change.

Histogram of oriented gradient (HOG) feature is a feature descriptor for object detection in computer vision and image processing [9]. The HOG method is based on the computation of normalized local orientation gradient histograms in dense grids [41]. The HOG method can maintain good geometric invariance and optical deformation. At present, Hog features are widely used in image recognition fields.

In image processing, a color histogram is a representation of the distribution of colors in an image [33]. For digital images, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges, that span the image's color space, the set of all possible colors. The color histogram can be built for any kind of color space, although the term is more often used for three-dimensional spaces like RGB or HSV [46]. Like other kinds of histograms, the color histogram is a statistic that can be viewed as an approximation of an underlying continuous distribution of colors values.

In order to describe the properties of uterine cervical cancer histology images, six shape features are extracted, including the average area of triangles, standard deviation of area of the triangles, average edge length, the standard deviation of edge length, average nuclei area, and the ratio of background over nucleus area. The first four features describe the global shape the tissue, and the last two features represent the local shape of single cells. [13]

A survey paper about 'histology image analysis for carcinoma detection and grading' is proposed, where image segmentation, feature extraction and classification approaches for cervix, prostate, breast, and lung cancers are mainly summarized and discussed. This paper refers to 161 related works, including ten papers that focus on the cervical cancer [17].

b: *Classifier Models:*

Support vector machine (SVM) is a type of supervised learning method [8]. The basic model of SVM is to find the best separating hyperplane in the feature space to maximize the interval between positive and negative samples in the training set. The core idea of SVM is to make every effort to maximize the separation between the two categories, so as to make the separation more credible. Moreover, it has good classification and prediction abilities for unknown new samples (called generalization ability in ML) [10]. Currently, SVMs are also applied in handwritten digit string recognition with a new cascade of hybrid principal component analysis network (PCANet) and support vector machine (SVM) classifier called PCA-SVMNet [1].

Artificial neural networks (ANNs) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains [23]. The neural network itself is not an algorithm, but rather a framework for many different ML algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules [43]. Components

of an ANN include neurons, connections, weights, biases, propagation functions and learning rules. Recently, a novel probabilistic process neural network (PPNN) was purposed to classify electrocardiogram signals [12]. In addition, a new content-based medical image retrieval (CBMIR) framework using convolutional neural network (CNN) and hash coding is proposed [6].

Random forest (RF) algorithm is to train multiple decision trees, generate models, and then use multiple decision trees to classify [19]. Random forests have improved the prediction accuracy without increasing the computational complexity [5]. Currently, RFs are also applied in driving assistance systems to recognize the propensity of drivers [28]. Random forests are insensitive to multivariate collinearity. The results are robust to missing data and unbalanced data, and can predict the effects of up to thousands of explanatory variables. Random forests are regarded as one of the most effective classification algorithms at present.

Weakly supervised learning is a branch of the ML strategy, which only annotates an image with the label of its category, but do not give any other annotations. Hence, the weakly supervised learning approach is a suitable solution for big dataset or complex image labelling problems. For example, in our previous work [29], a sparse coding based weakly supervised learning framework is introduced to address a microscopic image classification task.

C. APPLICATIONS OF CONDITIONAL RANDOM FIELDS

Currently, conditional random fields (CRFs) are used for labelling or parsing of sequential data, such as natural language processing or biological sequences [27] and in computer vision [18]. Specifically, CRFs find applications in part-of-speech (POS) tagging, shallow parsing [39], named entity recognition [38], gene finding and peptide critical functional region finding [7], among other tasks, being an alternative to the related hidden Markov models (HMMs). In computer vision, CRFs are often used for object recognition, image classification and image segmentation [37].

In [47], a probabilistic discriminative method is proposed to fuse contextual constraints in functional images based on the CRFs and it is applied to the detection of brain activation from both synthetic and real fMRI data. Experimental results show that the proposed CRF approach effectively integrates contextual constraints within the detection process and robustly detects brain activities from fMRI data.

In [2], a new segmentation method is introduced by combining CRFs with a cost-sensitive framework. The experiment shows that this method further improves its previous cost-sensitive SVM results by incorporating spatial information with the CRFs.

In [32], [35], colposcopy images of cervical cancer neoplasia are used as data in a CRF framework to extract the domain-specific diagnostic features in probabilistic form. They judge and locate precancerous and cancerous areas based on the optical and tissue relationships of different tissues.

With the development of DL, some researchers combine DL methods with CRF frameworks to obtain a even better classification performance. For example, in our previous work [24], a microscopic image classification engine is proposed, which can automatically classify and segment the images using DL features in a strongly supervised CRF model. In contrast to it, we further improve our CRF framework in this paper, which includes more ML techniques and only uses weakly supervised learning (does not need any ground truth images for CRF training, which means that labels of patches are endowed from the labels of corresponding images directly).

III. MULTILAYER HIDDEN CONDITIONAL RANDOM FIELDS

In this section, the basic knowledge of CRFs is first introduced in Sec. III-A. Then, the details of our MHCRF model is proposed in Sec III-B, including unary potential, binary potential and the combination of them.

A. BASIC KNOWLEDGE OF CRFS

Conditional Random Field (CRF) is first proposed in [27]. The definition of a CRF is as follows: Firstly, X is defined as a random variable of the data sequence to be labelled, and Y is a random variable of the corresponding label sequence. Then, let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices (or nodes) of G. V is the array of all sites, corresponding to the nodes of an associated undirected graph $G = (V, E)$, whose edges E model interactions between adjacent sites. So, (X, Y) is a CRF in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $p = (Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbours in G. This means that the CRF is an undirected graphical model whose nodes can be divided into two disjoint sets X and Y, which is the observed variable and the output variable. Then, the model conditional distribution is $p(Y|X)$.

According to the basic theorem of the random fields in [15], the joint distribution on the label sequence Y of a given X has the form as Eq. (1).

$$p_\theta(y|x) \propto \exp\Big(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x)\Big), \quad (1)$$

where x is a data sequence, y is a label sequence, and $y|_S$ is the set of components of y associated with the nodes in sub-graph S.

It can be known from [14], [51] that Eq. (1) can be rewritten as Eq. (2).

$$p(Y|X) = \frac{1}{Z} \prod_C \psi_C(Y_C, X), \quad (2)$$

where $\psi_C(Y_C, X)$ is the potential function on the clique C and $Z = \sum_{XY} P(Y|X)$ is the normalization factor. A

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.2019.2924467, IEEE Access

clique, C, in an undirected graph $G = (V, E)$ is a subset of the vertices, $C \subseteq V$, such that every two distinct nodes are adjacent.

B. THE PROPOSED MHCRF MODEL

1) Structure of the MHCRF Model

Because we focus on the visual features of cellular scales in histopathological images [20], we first design a unary potential to represent the information of cells, and then we design a binary potential to describe the surrounding special relationships among different cells. Hence, based on the basic definition of CRFs introduced in Sec. III-A, our MHCRF is expressed by Eq. (3).

$$p(X|Y) = \frac{1}{Z} \prod_{i \in V} \varphi_i(x_i; Y) \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j; Y), \quad (3)$$

Where

$$Z = \sum_{XY} \prod_{i \in V} \varphi_i(x_i; Y) \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j; Y) \quad (4)$$

is the normalization factor; V is the set of all nodes in the graph G; E is the set of the all edges. The clique potential function consists of two parts (terms): The unary potential function $\varphi_i(x_i, Y)$ is used to measure the probability that a node i is labelled as x_i for a given observation vector Y; the binary potential function $\psi_{ij}(x_i, x_j; Y)$ is used to describe the adjacent nodes i and j in the graph. The spatial context relationship between them is related not only to the tag of node i but also to the tag of its neighbour node j. Finally, we find the largest posterior label $\tilde{X} = \arg\max_X p(X|Y)$ and solve the problem of image classification. The structure of our MHCRF model is shown in Fig. 2.

In Fig. 2, the structure of our MHCRF model is shown:

- Layer 1 shows the real labels x_i of the patches in an image.
- Layer 2 represents the original image patches y_i which correspond one-to-one with the labels x_i in Layer 1.
- Layer 3 denotes seven types of features extracted from each image patch y_i, including RGBGray histogram, SIFT, DAISY, GLCM, HOG, Inception-V3 and VGG-16 features. Especially, there is an additional layer, namely the Layer 3.5, in the binary potential, where the features of the target image patches y_i are obtained by calculating the characteristics of their surrounding image patches according to the layout in Fig. 3.
- In Layer 4, the extracted features in Layer 3 are classified by four categories of classifiers, including Linear-SVMs, RBF-SVMs, ANNs and RFs, to obtain a priori probability for patch-level classification.
- In Layer 5, according to the Gaussian distribution and proportion, the most effective patch-level feature-classifier combinations are selected.
- In Layer 6, the selected patch-level feature-classifier combinations are jointly used for image-level classification.
- In Layer 7, the best image-level classification combinations are further selected to generate the unary or binary potentials.
- Finally, in Layer 8, the generated unary and binary potentials are combined in the proposed MHCRF model and used as a classifier for the CHIC task.

2) Unary Potential

The probability of a label x_i taking a value $c \in \mathbb{L}$ is connected with the unary potential parts $\varphi_i(x_i; Y)$ of the Eq. (3) given the data Y by $\varphi_i(x_i; Y) \propto p(x_i = c|f_i(Y))$ [25], where the image data is expressed as site-wise feature vectors $f_i(Y)$ which may depend on all the data of Y.

We extract seven types of features from each image patch:

- Color features: We extract the intensity histograms of the R, G, B color channels and gray-level images, and link them together to obtain a 1024-dimensional feature vector [46].
- Texture features: Four texture features are extracted, including SIFT [31], DAISY [44], GLCM [22] and HOG [9] features. The vector dimensions of these texture features are 128, 200, 64, and 4356, respectively.
- DL features: One is Inception-V3 feature [42] and another is VGG-16 feature [40], where transfer learning by ImageNet is applied to pre-train the Inception-V3 and VGG-16 networks [21], and the second last layer are fine-tuned with our cervical histopathological images. Finally, the length of the extracted DL feature vectors is set to 1000 dimensions based on our pre-tests.

In order to obtain the label probability, we use a total of 19 classifiers in four categories:

- Linear-SVMs: SVMs with a linear kernel.
- RBF-SVMs: SVMs with a radial based function kernel.
- ANNs: We use the quantization gradient algorithm "trainscg", and the hidden layer uses the six forms from one to six layers, respectively.
- RFs: We use RFs as classifiers, where the number of trees is set to $2^n (n = 1, 2, ..., 11)$.

Based on the features and classifiers mentioned above, we generate our unary potential as follows:

- First, according to the combinations of seven features and 19 classifier types, we build 133 single patch-level classifiers and obtain 133 pre-classification results.
- Then, based on the Gaussian distribution of these 133 results, we select the top 8% feature-classifier combinations of them (about 11). Among these 133 combinations, the number of the top 8% of the DL features is about three, and the number of the top 8% of the handicraft features is about eight. In the handicraft features, color features and texture features have the same numbers, so the most effective four features of each of them are selected, respectively.
- Thirdly, based on the patch-level pre-classification probabilities of the selected 11 feature-classifier combina-

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.2019.2924467, IEEE Access

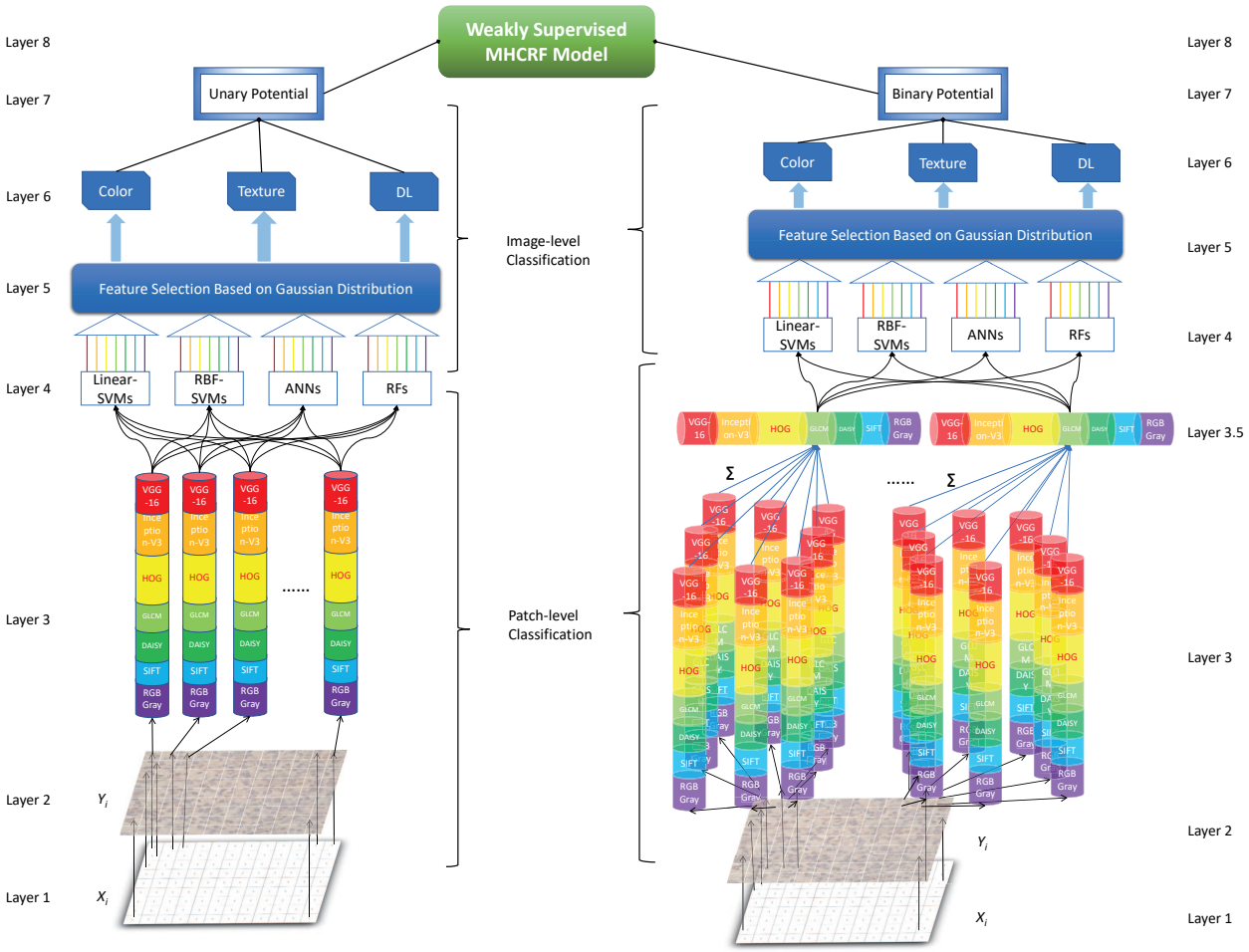

FIGURE 2. The structure of our proposed weakly supervised MHCRF model. The left part denotes the term of the unary potential, and the right part shows the term of the binary potential.

tions, the joint probabilities of all patches are calculated as the image-level classification probabilities.

- Fourthly, we combine the calculated image-level classification probabilities to obtain a second order joint probability, where the number of the combinations is a factorial of 11, i.e. 39916800.
- Fifthly, the top ten from 39916800 combinations are further selected as promising candidates to generate the unary potential in the image-level classification. Here, we combine each two of these ten candidates together, so 100 third order joint probabilities are achieved.
- Finally, the combination with the best classification result is selected from 100 joint probabilities, and it is used as the unary potential in our MHCRF model.

3) Binary Potential

The binary potential term $\psi_{ij}(x_i, x_j; Y)$ of the Eq. (3) shows how probably the pair of adjacent sites i and j is to take the label $(x_i, x_j) = (c, c')$ given the data: $\psi_{ij}(x_i, x_j; Y) = p(x_i = c; x_j = c' | f_i(Y) f_j(Y))$ [25]. Fig. 3 shows the layout of our binary potential. We use this "lattice" layout to characterize the feature vector of each patch by calculating the sum of each patch of eight neighbourhood feature vectors. The other steps are consistent with the unary potential in Sec. III-B2.

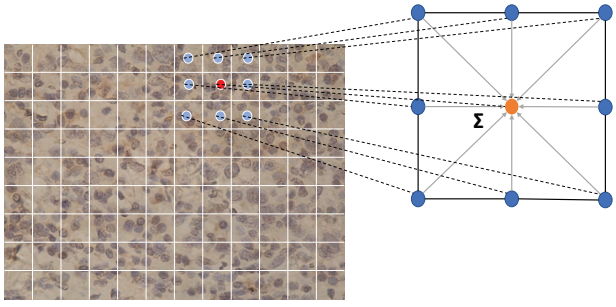

FIGURE 3. "Lattice" layout of the binary potential. "$\sum$" denotes that the sum of the eight neighbourhood feature vectors is used as the feature vector of the target patch (the central patch).

IV. EXPERIMENTS

A. EXPERIMENTAL SETTING

1) Image Dataset

In this paper, we use six immunohistochemical (IHC) stained cervical histopathological image datasets (AQP1, AQP2, HIF1, HIF2, VEGF1 and VEGF2 datasets) to test the effectiveness of our proposed MHCRF model in the CHIC task. In addition, we have a hematoxylin-eosin (HE) stained gastric histopathological image datasets to verify the generalization ability of the MHCRF model. The details of the applied datasets are introduced as follows.

Six Cervical IHC Histopathological Image Datasets:

- Data source: Three practical medical doctors (Doctor A, Dr. B and Dr. C) from Shengjing Hospital of China Medical University provide image samples and give image-level labels for the weakly supervised learning process. Dr. A and Dr. B provide us with the AQP1, HIF1 and VEGF1 datasets. Dr. A and Dr. C provide us with the AQP2, HIF2 and VEGF2 datasets. The image labelling rules are as follows:
  Rule 1: When the doctors find only one differentiation stage in an image, they give this stage as the label to this image;
  Rule 2: When the doctors find multiple differentiation stages in an image, they give the most significant stage as the label to this image;
  Rule 3: In Rule 2, if different differentiation stages have similar distributions in the image, the doctors give the most serious stage as the label to this image.
  Rule 4: When Dr. A has a different judgement with other doctors, we follow the judgement of Dr. B or Dr. C, due to they have a higher-level professional qualification.
- Staining method: IHC Staining, AQP, HIF, VEGF.
- Magnification: $400\times$.
- Microscope: Nikon (Japan).
- Acquisition software: NIS-Elements F 3.2.
- Image size: 1280×960 pixels.
- Image format: "*.tiff" or "*.png".
- Image types:
  Well differentiation: The tumour cells are closer to normal cells, cell heteromorphism is relatively small, cell sizes and morphology are similar;
  Moderate differentiation: Most cancer cells are concentrated in moderately differentiated, the characteristic is between well differentiated and poorly differentiated cervical cancer cells;
  Poorly differentiation: The cell structure is not visible, and the topological structure is disordered [50]. An example of these six datasets is shown in Fig. 4, and the detailed information of it is as follows:

Gastric HE Histopathological Image Dataset:

- Data source: A public dataset of gastric histopathological images is additionally tested in our paper [49]. Especially, due to the original dataset is not balanced on normal or abnormal classes, we randomly select 100 images from each class to build a sub-dataset in our work.
- Staining method: HE Staining.
- Magnification: $20\times$.
- Image size: 2048×2048 pixels. Furthermore, to use the original images in our system, we resize them into 1280×960 pixels.
- Image format: "*.tiff" or "*.png".
- Image types:
  Normal: No cancerous cells appeared in the section.;
  Abnormal: Cancerous cells appear in this section. An example of the gastric HE datasets is shown in Fig. 5, and the detailed information of it is as follows:

2) Training, Validation and Test Data Setting

We randomly divide all the datasets into training, validation and test sets at a ratio of 1:1:2. The specific data settings for these six cervical IHC datasets are shown in TABLE 1. And the data settings for the gastric HE dataset is shown in TABLE 2.

TABLE 1. Experimental data settings (Cervical IHC Datasets). The first column shows the names of datasets. The second column shows the differentiation stages. The third to the fifth columns show usage of data, respectively. The last column is the number of well images, the number of Moderate images, the number of Poorly images, and the total number of images in the dataset for each dataset.

Dataset	Usage / Stage	Train	Validation	Test	Sum
AQP1	Well	9	9	17	35
	Moderate	9	9	17	35
	Poorly	9	8	16	33
	Sum	27	26	50	103
AQP2	Well	9	8	16	33
	Moderate	9	9	17	35
	Poorly	7	7	13	27
	Sum	25	24	46	95
HIF1	Well	8	7	15	30
	Moderate	9	8	17	34
	Poorly	8	8	16	32
	Sum	25	23	48	96
HIF2	Well	10	10	19	39
	Moderate	10	9	19	38
	Poorly	9	8	17	34
	Sum	29	27	55	111
VEGF1	Well	9	9	17	35
	Moderate	8	8	16	32
	Poorly	10	9	18	37
	Sum	27	26	51	104
VEGF2	Well	8	7	14	29
	Moderate	9	8	16	33
	Poorly	10	10	19	39
	Sum	27	25	49	101

B. EVALUATION OF UNARY AND BINARY POTENTIALS

1) Evaluation of Unary Potential

First, in order to select effective feature-classifier combinations to general our unary potential, we compare the 133 accuracies of single classifiers on all the six cervical IHC validation sets in the patch-level, and the classification result of AQP1 validation set is shown in TABLE 3 as an example.

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.2019.2924467, IEEE Access

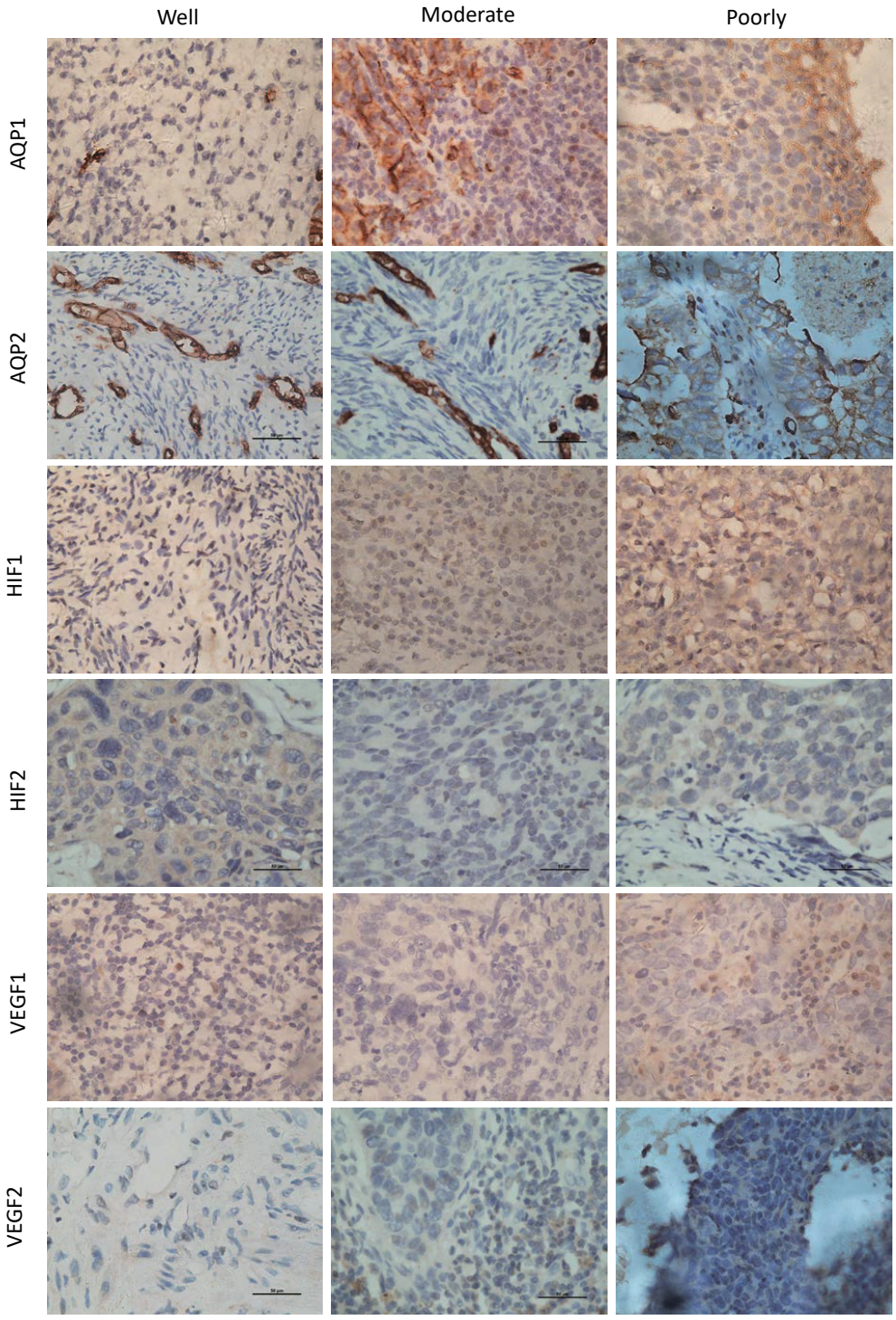

FIGURE 4. An example of the six cervical IHC datasets. The first, second and last column show histopathological images of well, moderate and poorly differentiation stages of cervical cancer, respectively. The images of lines one to six belong to the six databases of AQP1, AQP2, HIF1, HIF2, VEGF1, and VEGF2, respectively.

Then, the classification results of the selected feature-classifier combinations and the generated unary potential in the image-level are shown in Fig. 6, where the tags on the horizontal axis denote the selected combinations and the optimized unary potential. For example, "RGBGray" means the color features extracted from R, G, B channels and gray-level images; "LINEAR" means the kernel function of SVM classifier is a linear function; the number after "ANN" means the number of hidden layers; the number after the "RF" refers to the index of the tree, because we represent the number of

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.2019.2924467, IEEE Access

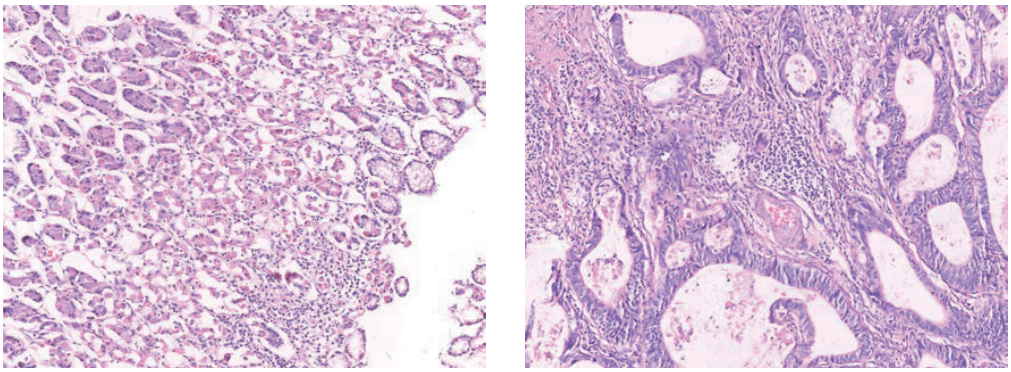

FIGURE 5. An example of the gastric HE image dataset. The left one is normal and the right one is abnormal.

TABLE 2. Experimental data settings (Gastric HE Dataset). The first column shows the differentiation stages. The second to the forth columns show usage of data, respectively. The last column is the number of normal images, the number of abnormal images and the total number of images in the dataset for this dataset.

Stage \ Usage	Train	Validation	Test	Sum
Normal	25	25	50	100
Abnormal	25	25	50	100
Sum	50	50	100	200

trees as 2^n $(n = 1, 2, ..., 11)$.

From Fig. 6, we find that the classification result is significantly improved by the unary potential than all the single feature-classifier combinations. Especially, the unary potential can increase the classification accuracy in a range of 10.22% to 30.27%, with an increasing rate from 15.13% to 49.60%. These data fully reflect the validity of the unary potential model.

Comparing the results of different staining datasets (AQP1, HIF1 and VEGF1) given by Dr. A and Dr. B, the combination of RGBGray features and RF classifiers is better, and a combination of RGBGray-RF is combined in the unary potential model. However, this rule does not hold in the dataset given by Dr. B and Dr. C. In another three datasets (AQP2, HIF2, and VEGF2), there is not only a combination of RGBGray-RF but also a combination of GLCM-ANN and DAISY-ANN. In addition, comparing the results of different sets of data for the same kind of staining, the accuracy of the second set of data is lower than that of the first set. This is related to the method, time, experience, and location of the slice selected by different doctors.

In addition, we also used the gastric HE dataset to do the same experiments. Fig. 7 shows the classification results of the selected feature-classifier combinations and the generated unary potential in the image-level. We can find that the classification result is significantly improved by the in the unary potential than the all single feature-classifier combinations, where an accuracy of 94.00% is achieved, exceeding the best single result of 85.15%. In this dataset, the unary potential increases the accuracy by 8.85%, with an increasing rate of 10.39%. From these two data of increasing quantity and increasing rate, the result of gastric HE dataset is not as high as that of cervical IHC datasets, because the single classification accuracy rate of gastric HE dataset reaches 85.19%, even higher than most cervical IHC unary potential results, and the upside is not large, and the accuracy of the unary potential reaches 94.00%, which proves that the unary potential of the proposed MHCRF model is valid.

2) Evaluation of Binary Potential

First, in order to select effective feature-classifier combinations to general our binary potential, we compare the 133 accuracies of single classifiers on all the six cervical IHC validation sets in the patch-level, and the classification result of AQP1 validation set is shown in TABLE 4 as an example.

Then, the classification results of the selected features and the generated potentials are in the image-level shown in Fig. 8, where the horizontal axis denotes the optimized combination of features and the pre-classifiers. Further, "RGBGray" means the color features extracted from R, G, B channels and gray-level images. And the number after "ANN" means the number of hidden layers. Moreover, the number after the "RF" refers to the index of the tree, because we represent the number of trees as 2^n $(n = 1, 2, ..., 11)$.

From Fig. 8, we find that the classification result is significantly improved by the in the binary potential than the all single feature-classifier combinations. Especially, the binary potential can increase the classification accuracy in a range of 7.07% to 12.95%, with an increasing rate from 9.63% to 20.23%. These data fully reflect the validity of the binary potential model. From the two data of increment and growth rate, the effect of unary potential is much better than the binary potential. This is because in the same dataset, the classification accuracy of a single combination under the unary framework is basically lower than that under the binary framework, and the increasing space of the unary potential is bigger than that of the binary potential.

Comparing the results of different staining datasets

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.2019.2924467, IEEE Access

TABLE 3. An example of patch-level pre-classification accuracies of unary potential on AQP1 validation set. The first two columns show the types of classifiers. The third to the last columns show the extracted features. The numbers in red bold are related to 11 selected feature-classifier combinations. (Unit:%)

Classifier	Type	RGBGray	SIFT	DAISY	GLCM	HOG	Inception-V3	VGG-16
SVM	RBF	34.62	41.67	34.62	34.62	36.86	34.62	34.62
	Linear	56.23	43.06	37.69	43.59	36.79	34.54	47.26
ANN	1 hidden layer	61.72	44.59	48.47	**53.56**	41.95	41.06	52.10
	2 hidden layers	62.75	44.84	49.32	53.03	41.84	40.03	**52.35**
	3 hidden layers	61.47	45.23	48.29	**53.63**	41.63	40.99	51.60
	4 hidden layers	61.22	44.55	47.72	**53.63**	41.24	40.95	51.60
	5 hidden layers	60.90	44.20	48.01	51.99	40.53	40.42	50.21
	6 hidden layers	59.33	43.23	48.29	**53.92**	41.03	39.78	50.71
RF	$n = 1$	52.53	39.14	42.24	46.37	36.04	33.97	44.80
	$n = 2$	54.91	39.25	42.34	46.87	36.04	30.52	46.72
	$n = 3$	56.98	40.85	42.41	48.50	36.79	38.82	48.22
	$n = 4$	60.86	41.06	42.34	47.76	37.86	39.71	49.07
	$n = 5$	63.28	42.34	44.59	50.11	39.85	36.89	51.18
	$n = 6$	62.93	42.81	45.09	49.79	39.46	40.74	51.82
	$n = 7$	63.14	43.87	43.70	50.32	40.74	40.35	51.32
	$n = 8$	**63.43**	43.30	44.84	50.25	41.45	39.49	**52.46**
	$n = 9$	**63.89**	43.87	44.44	50.43	42.24	39.81	52.14
	$n = 10$	**63.92**	43.80	44.20	49.86	42.81	39.64	52.24
	$n = 11$	**63.64**	43.38	44.62	50.57	43.55	39.85	**52.42**

TABLE 4. An example of patch-level pre-classification accuracies of binary potential on AQP1 validation set. The first two columns show the types of classifiers. The third to the last columns show the extracted features. The numbers in red bold are related to 11 selected feature-classifier combinations. (Unit:%)

Classifier	Type	RGBGray	SIFT	DAISY	GLCM	HOG	Inception-V3	VGG-16
SVM	RBF	34.62	34.62	34.37	34.62	45.33	41.38	44.12
	Linear	62.82	45.16	37.86	42.55	40.14	39.53	59.05
ANN	1 hidden layer	**71.83**	50.71	49.39	59.22	45.05	42.70	**62.32**
	2 hidden layers	**71.40**	51.07	49.54	**61.86**	45.19	42.84	61.25
	3 hidden layers	70.30	50.75	50.18	**61.04**	46.72	43.87	61.29
	4 hidden layers	**71.37**	52.42	50.36	**61.61**	45.76	44.69	**62.39**
	5 hidden layers	70.33	51.42	48.97	**61.97**	46.01	43.87	**62.14**
	6 hidden layers	**71.30**	50.28	49.68	59.54	46.65	43.98	61.65
RF	$n = 1$	50.04	43.41	45.58	52.88	37.43	35.83	51.57
	$n = 2$	56.45	43.30	44.69	56.13	37.78	38.14	54.27
	$n = 3$	58.55	46.83	45.23	54.34	42.66	38.96	57.23
	$n = 4$	61.82	45.62	45.83	54.56	42.02	40.81	59.79
	$n = 5$	61.93	46.58	46.05	57.05	42.27	41.17	59.15
	$n = 6$	62.46	46.87	46.12	58.01	45.37	42.31	59.79
	$n = 7$	63.18	48.86	47.01	57.26	44.62	42.70	59.69
	$n = 8$	63.00	47.58	46.40	56.91	45.73	41.45	60.79
	$n = 9$	63.35	47.61	47.15	57.94	45.87	41.35	60.65
	$n = 10$	63.53	47.72	46.76	57.55	44.80	41.70	60.51
	$n = 11$	63.75	47.61	47.04	57.76	45.73	41.52	60.54

(AQP1, HIF1 and VEGF1) given by Dr. A and Dr. B, the combination of RGBGray features and ANN classifiers is better, and a combination of RGBGray-ANN is combined in the binary potential model. However, this rule does not hold in the dataset given by Dr. B and Dr. C. In another three datasets (AQP2, HIF2, and VEGF2), there is not only a combination of RGBGray-ANN but also a combination of GLCM-ANN and VGG16-RF. In addition, comparing the results of different sets of data for the same kind of staining, the accuracy of the second set of data is lower than that of the first set. This case occurs the same as the unary potential, where different doctors have different methods, skills, experience, and judgements.

Furthermore, we also used the gastric HE dataset to do the same experiments. Fig. 9 shows the classification results of the selected feature-classifier combinations and the generated binary potential in the image-level. We can find that the classification result is significantly improved by the binary potential than all the single feature-classifier combinations, where an accuracy of 96.00% is achieved, exceeding the best single result of 89.85%. In this dataset, the binary potential increases the accuracy by 6.15%, an increasing rate of 6.84%. From the two data of increasing quantity and increasing rate, the result of gastric HE dataset is not as good as that of cervical IHC dataset, because the single classification accuracy rate of gastric HE dataset reaches 89.85%, even higher than most cervical IHC binary potential results, and the upside is not large, and the accuracy of the binary potential reaches 96.00%, which proves that the binary potential part of the proposed MHCRF model is valid.

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.2019.2924467, IEEE Access

FIGURE 6. A comparison of image-level classification accuracies between selected feature-classifier combinations and the generated unary potential on the six cervical IHC validation sets.

3) Experimental Results and Analysis of Unary and Binary Potentials

Fig. 10 shows the confusion matrices for image-level classification results for six cervical IHC datasets, including unary potentials on validation sets, binary potentials on validation sets, the MHCRF framework on validation sets, and the MHCRF framework on test sets.

Comparing the four confusion matrices of the same dataset in Fig. 10, we find that the accuracy of the validation set classification results is generally greater than or equal to the larger values in unary and binary. Only in HIF1 the validation set results are close to the smaller values in unary and binary.

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.2019.2924467, IEEE Access

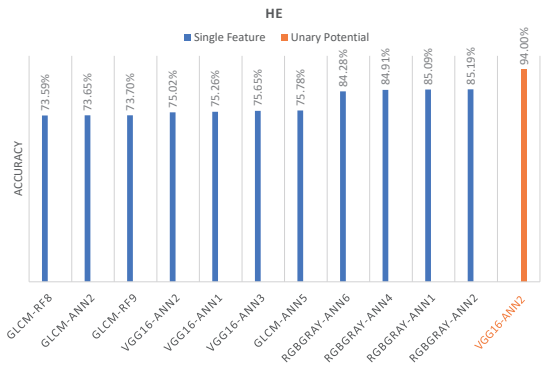

FIGURE 7. A comparison of image-level classification accuracies between selected feature-classifier combinations and the generated unary potential on the gastric HE validation set.

The test set results are generally about 5% to 10% lower than the validation set results, and occasionally 1% to 4% higher (AQP1 and VEGF2).

Furthermore, we use the result of AQP1 as an example to analyse the experimental performance of unary and binary potentials. From TABLE 3 and TABLE 4, we can find that several feature-classifier combinations have good performance, such as the 11 selected combinations. However, there are some classification performance of combinations is not good, only higher than 33.33% (the probability of randomly guessing). On one hand, color feature combinations almost more efficient than other combinations, which shows that the cervical IHC staining can enhance the image color information effectively. In texture feature combinations, GLCM feature combinations are relatively good, because it can describe the cervical cell distribution trend very robustly. In DL feature combinations, the VGG-16 combinations are significantly better than the Inception-V3 combinations. This is because we set the patch size to 100-by-100 pixels, there are only one or two nuclei in a patch, and this case is more suitable for the ability of VGG-16 network. So, texture features and DL features can behave as well as color features. On the other hand, for the same feature, SVMs perform the worst in different classifiers, and ANN and RF each have their own strengths. In the unary potential, ANNs and RFs have similar performance. In the binary potential, ANNs' performance is significantly better. This is because SVMs need to find a hyperplane to achieve classification in the case of multi-classification, and its effect is not as good as that of ANNs' backpropagation network and RF forest structure. Moreover, the dataset size of from 2800 to 5200 patches is more advantageous for ANNs and RFs. In contrast, although the models of the other five datasets have different details, their frameworks and processes are consistent.

In the gastric HE dataset, we also show the confusion matrices in Fig. 11. Since the gastric HE dataset is a binary classification, the four confusion matrices are also 3×3 sized. The results of both unary and binary are very good, reaching 94% and 96%. The validation set results are close to a larger value of 96%. The test set results were not much reduced, and the accuracy rate was 93%. The reasons why these four results of the gastric HE classification are much higher than those of the cervical IHC are as follows. First, the gastric HE dataset only needs to perform the two classification task, while the cervical IHC datasets need to perform the three classification task. Second, for computers, it is possible that gastric HE staining methods are easier to distinguish and identify than cervical IHC staining. Third, the quality of the image will also affect the results of the classification. Last, in terms of the amount of data for a single dataset, the amount of data in the gastric HE dataset is approximately twice that of the other six cervical IHC datasets.

C. EVALUATION OF THE PROPOSED MHCRF MODEL

From these results, it can be seen that although the combined result of the unary and binary potentials has the same performance on the validation set, it has a better classification performance on the test set, showing a huge potential of the proposed method.

In order to further prove the effective classification ability of our method in the CHIC work, besides the evaluation with accuracy in Fig. 10, we also calculate another four evaluation criteria in TABLE 5, including recall (sensitivity), precision, specificity and F1-score, of each differentiation stage. These four criteria are defined in Eq. (5), Eq. (6), Eq. (7) and Eq. (8) [36].

$$\text{Recall} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

$$\begin{aligned} \text{F1} - \text{score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \end{aligned} \tag{8}$$

where TP is the True Positive (positive sample is predicted to be positive), TN is the True Negative (negative sample is predicted to be negative), FP is the False Positive (negative sample is predicted to be positive), FN is the False Negative (positive sample is predicted to be negative), Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances (high recall means that an algorithm returned most of the relevant results), Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances (high precision means that an algorithm returned substantially more relevant results than irrelevant ones), Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such, and the F1-score (also F-score or F-measure) is a measure of a accuracy of a test (it considers both the precision and the recall of the test to compute the score).

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.2019.2924467, IEEE Access

FIGURE 8. A comparison of image-level classification accuracies between selected feature-classifier combinations and the generated binary potential on the six cervical IHC validation sets.

From TABLE 5, considering the overall performance of four criteria, we can find that the result of AQP1 dataset obtains the highest evaluations with the values of 87.87% recall, 87.96% precision, 94.03% specificity and 87.86% F1-score. In contrast, the result of AQP2 achieves the lowest evaluations with the values of 69.88% recall, 69.31% precision, 84.66% specificity and 69.45% F1-score. Meanwhile, the results of the other four datasets have overall evaluations around 76% recall, 79% precision, 88% specificity, 76% F1-score. Hence, the results mentioned above show the effectiveness and stability of our MHCRF model.

Meanwhile, we show the receiver-operating characteristic (ROC) curves of these six cervical IHC datasets in Fig. 12. The ROC curve analyses the binary classification model, so the three-classification model in the figure draws three ROC curves, with blue representing high differentiation, green

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.2019.2924467, IEEE Access

TABLE 5. Evaluation of each differentiation stage in six cervical IHC test sets. (Unit:%)

Dataset	Differentiation stage	Recall (Sensitivity)	Precision	Specificity	F1-Score
AQP1	Well	94.12	88.89	93.94	91.43
	Moderate	88.24	93.75	96.97	90.91
	Poorly	81.25	81.25	91.18	81.25
	Mean	87.87	87.96	94.03	87.86
AQP2	Well	56.25	64.29	83.33	60.00
	Moderate	76.47	72.22	82.76	74.29
	Poorly	76.92	71.43	87.88	74.07
	Mean	69.88	69.31	84.66	69.45
HIF1	Well	86.67	86.67	93.94	86.67
	Moderate	64.71	91.67	96.77	75.86
	Poorly	100	76.19	84.38	86.49
	Mean	83.79	84.84	91.70	83.01
HIF2	Well	57.89	91.67	97.22	70.97
	Moderate	94.74	56.25	61.11	70.59
	Poorly	58.82	90.91	97.37	71.43
	Mean	70.49	79.61	85.23	70.99
VEGF1	Well	64.71	78.57	91.18	70.97
	Moderate	56.25	64.29	85.71	60.00
	Poorly	100	78.26	84.85	87.80
	Mean	73.65	73.71	87.25	72.92
VEGF2	Well	64.29	75.00	91.43	69.23
	Moderate	81.25	61.90	75.76	70.27
	Poorly	84.21	100	100	91.43
	Mean	76.58	78.97	89.06	76.98

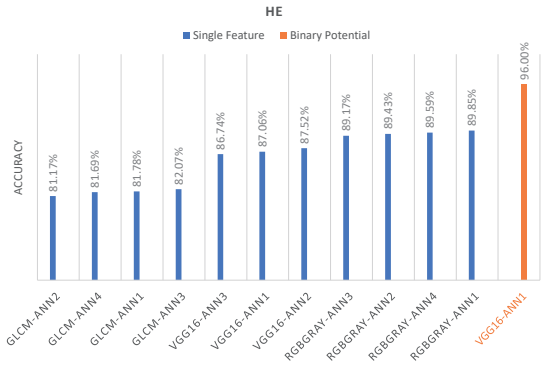

FIGURE 9. A comparison of image-level classification accuracies between selected feature-classifier combinations and the generated binary potential on the gastric HE validation set.

representing moderate differentiation, and red representing poor differentiation. The area under the Curve of ROC (AUC ROC) is often used to evaluate classifier performance. The larger the AUC value, the higher the correct rate. From Fig. 12, we can find that the three curves of the validation set are similar to the three curves of Binary potential, in general. However, the well differentiated and poorly differentiated curves of the validation set in the AQP2 dataset are close to the Binary potential, while the mid-differentiation curve is in a state of being between the two under the influence of the Unary potential. This is a good explanation for the validity of our proposed MHCRF model.

Similarly, we extract the corresponding four evaluation criteria and ROC curves in the gastric HE dataset. TABLE 6 shows the four evaluation criteria of gastric HE test set. And Fig. 13 shows the four ROC curves with two-classification.

TABLE 6. Evaluation of each differentiation stage in gastric HE test set. (Unit:%)

Recall (Sensitivity)	Precision	Specificity	F1-Score
92.00	93.88	94.00	92.93

D. MISCLASSIFICATION ANALYSIS

Fig. 14 shows some misclassification examples of the classification results by the MHCRF model on the six cervical IHC test sets. Meanwhile, misclassification examples of gastric HE dataset is shown in Fig. 15. According to our analysis and speculation, the reasons for image classification errors are as follows:

- First, because the contents of the cervical histopathological images are complex, where the characteristics and properties between various differentiation stages are not always obviously different, resulting in a difficulty of image feature extraction.
- Second, the applied binary potential layout has a small coverage and cannot effectively contain spatial information.
- Thirdly, because our method is a weakly supervised learning approach, we only use image-level labels to train our MHCRF model, some contents of differentiation stages are mixed in one image, resulting in the training information is not clear enough for the ML algorithm.

Additionally, considering Fig. 14, because moderate differentiation is the stage between well and poorly differentiations, most of the misclassification cases occur between well and moderate, or poorly and moderate stages. In contrast, well and poorly differentiations have fewer misclassification cases.

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.2019.2924467, IEEE Access

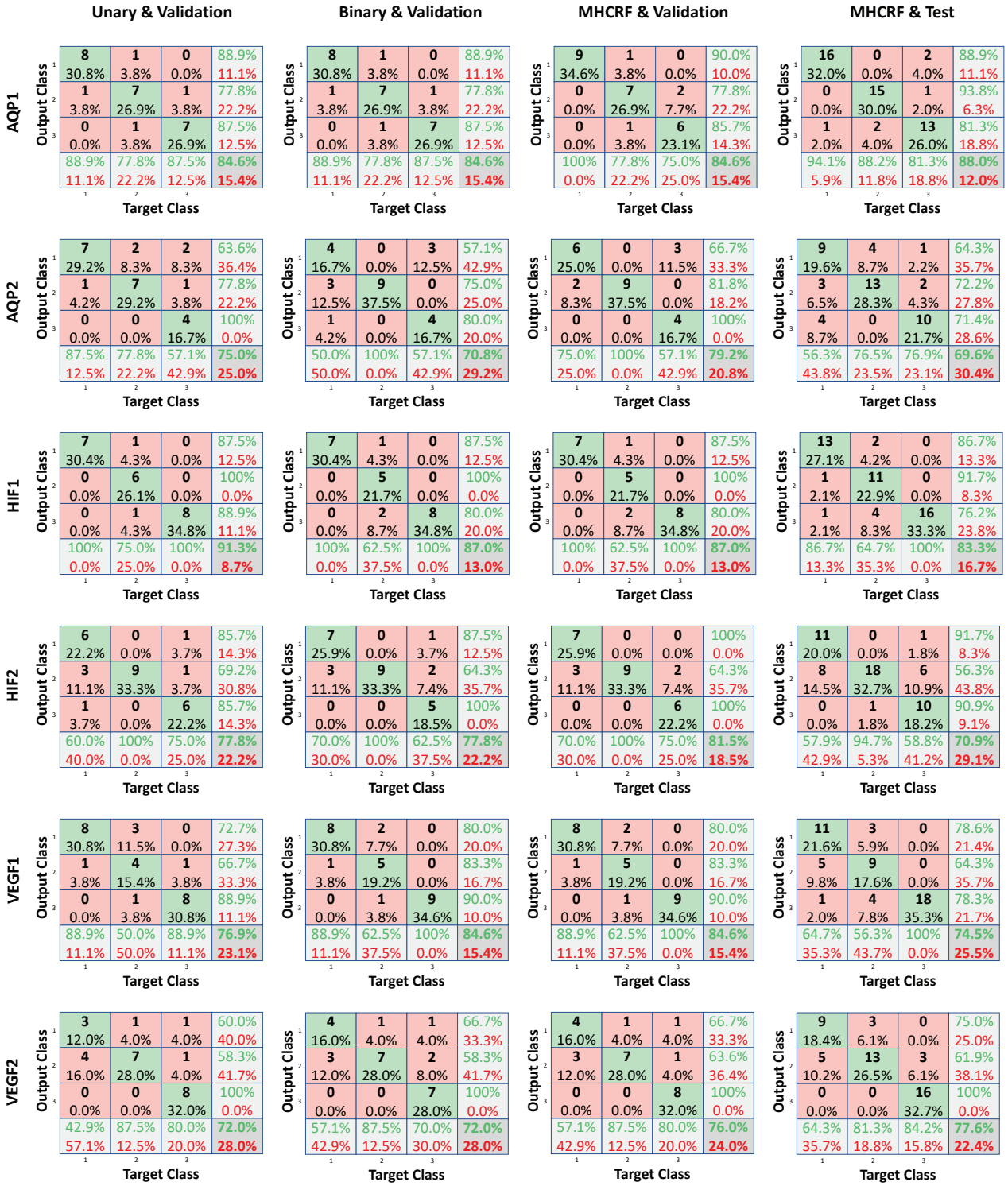

FIGURE 10. Classification results using the six cervical IHC datasets. The first column to the fourth column represent the confusion matrices for the classification results of unary potentials on validation sets, binary potentials on validation sets, the MHCRF framework on validation sets, and the MHCRF framework on test sets, respectively.

E. COMPUTATIONAL TIME

Finally, we briefly describe computational time of our MHCRF classification method. In our experiment, we use a workstation with Intel® Core^{TM} i7-7700 CPU with 3.60 GHz, 32 GB RAM and GeForce GTX 1080 8GB. Regarding the training time on six cervical IHC datasets and one gastric

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.2019.2924467, IEEE Access

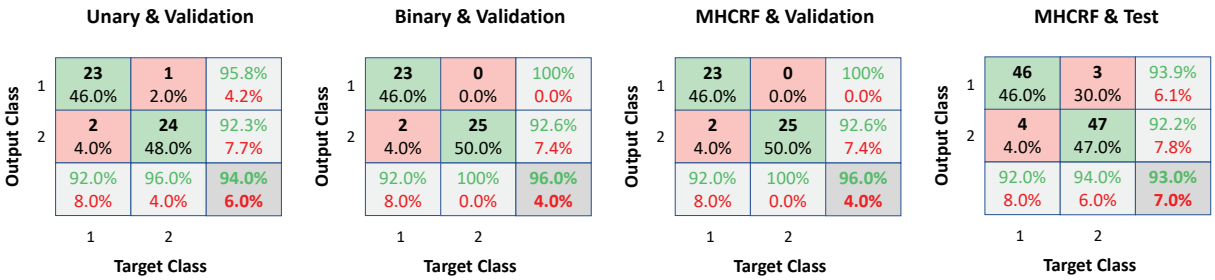

FIGURE 11. Classification results using the gastric HE dataset. The first column to the fourth column represent the confusion matrices for the classification results of unary potential on validation set, binary potential on validation set, the MHCRF framework on validation set, and the MHCRF framework on test set, respectively.

HE dataset, the mean time is about 12.5 hours. For the test time, the detailed information is shown in TABLE 7.

TABLE 7. The test time of six cervical IHC datasets and one gastric HE dataset.

Dataset	Total Time	Image Numbers	Average time
AQP1	112.33 s	50	2.24 s
AQP2	17.43 s	46	0.38 s
HIF1	155.20 s	48	3.24 s
HIF2	7.72 s	55	0.14 s
VEGF1	111.82 s	51	2.19 s
VEGF2	149.40 s	49	3.05 s
HE	23.23 s	100	0.23 s
Mean	–	–	1.64 s

From TABLE 7, although the training time is 12.5 hours, the mean test time for one image is 1.64 s, showing the feasibility of our MHCRF method in the practical clinical fields.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a weakly supervised MHCRF model to classify the cervical histopathological images into well, moderate and poorly differentiation stages. The proposed MHCRF method not only considers the classical color and texture features, but also combines the state-of-the-art deep learning techniques into the framework. Furthermore, this MHCRF model build both unary and binary potentials to describe the spatial relationship between the image locations. In the experiment, the proposed method is tested on the six cervical IHC datasets and obtains an around overall classification accuracy of 77.32% and the highest one of the six is 88%, showing the effectiveness and potential of the method. In addition, we carry out extended experiments on a gastric HE dataset, achieving overall accuracy of 93%, which can fully demonstrate the generalization ability of our MHCRF model.

In our future work, we plan to increase the amount of data in a single dataset, allowing the same doctors to expand the data. Then although we have tested our MHCRF model on gastric cancer, we will test it on more cancer types, such as breast cancer and liver cancer. Meanwhile, we have not optimized individual features or classifiers yet, but we will adjust the relevant parameters to make the classifiers at their best status. In addition, we will try to use more types of features and classification algorithms to improve the weakly supervised MHCRF model.

ACKNOWLEDGEMENTS

We thank B.E. Changhao Sun and B.Sc. Muhammad Mamunur Rahaman from the MIaMIA Group for their support in additional experiments and proofreading, respectively.

REFERENCES

[1] S. Aly and A. Mohamed. Unknown-Length Handwritten Numeral String Recognition Using Cascade of PCA-SVMNet Classifiers. IEEE Access, 7:52024–52034, 2019. II-B0b

[2] Y. Artan, M. Haider, D. Langer, and et al. Prostate Cancer Localization with Multispectral MRI Using Cost-Sensitive Support Vector Machines and Conditional Random Fields. IEEE Transactions on Image Processing, 19(9):2444–2455, 2010. II-C

[3] J. Berek and N. Hacker. Berek and Hacker's Gynecologic Oncology. Wolters Kluwer/Lippincott Williams & Wilkins Health, America, 2010. II-A

[4] F. Bray, J. Ferlay, I. Soerjomataram, and et al. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians, 68(6):394–424, 2018. I, II-A

[5] L. Breiman. Random Forests. Machine Learning, 45(1):5–32, 2001. II-B0b

[6] Y. Cai, Y. Li, C. Qiu, and et al. Medical Image Retrieval Based on Convolutional Neural Network and Supervised Hashing. IEEE Access, 7:51877–51885, 2019. II-B0b

[7] K. Chang, T. Lin, L. Shih, and et al. Analysis and Prediction of the Critical Regions of Antimicrobial Peptides Based on Conditional Random Fields. PloS One, 10(3):e0119490, 2015. II-C

[8] C. Cortes and V. Vapnik. Support-Vector Networks. Machine Learning, 20(3):273–297, 1995. II-B0b

[9] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In Proc. of CVPR 2005, volume 1, pages 886–893, 2005. II-B0a, III-B2

[10] D. Decoste and B. Schölkopf. Training Invariant Support Vector Machines. Machine Learning, 46(1-3):161–190, 2002. II-B0b

[11] M. Fahey, L. Irwig, and P. Macaskill. Meta-analysis of PapTest Accuracy. American Journal of Epidemiology, 141(7):680–689, 1995. II-A

[12] N. Feng, S. Xu, Y. Liang, and et al. A Probabilistic Process Neural Network and Its Application in ECG Classification. IEEE Access, 7:50431–50439, 2019. II-B0b

[13] P. Guo, K. Banerjee, R. Stanley, and et al. Nuclei-Based Features for Uterine Cervical Cancer Histology Image Analysis With Fusion-Based Classification. IEEE Journal of Biomedical and Health Informatics, 20(6):1595–1607, 2016. II-B0a

[14] R. Gupta. Conditional Random Fields. Unpublished Report, IIT Bombay, 2006. III-A

[15] J. Hammersley and P. Clifford. Markov Fields on Finite Graphs and Lattices. Unpublished, 1971. III-A

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.2019.2924467, IEEE Access

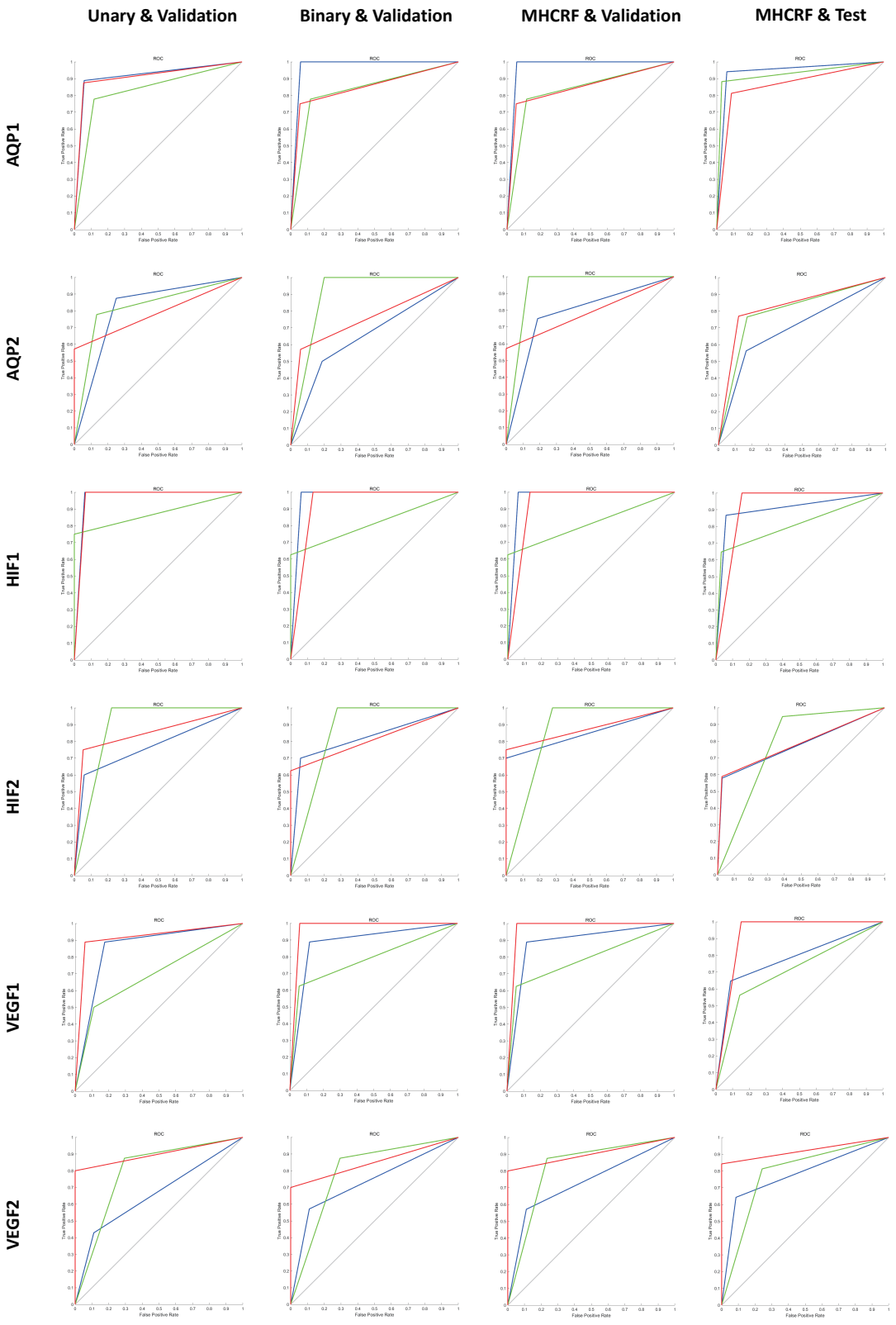

FIGURE 12. ROC curves using the six cervical IHC datasets. The blue line, the green line and the red line represent well, moderate and poorly differentiation, respectively. The first column to the fourth column represent the confusion matrices for the classification results of unary potential on validation set, binary potential on validation set, the MHCRF model on validation set, and the MHCRF model on test set, respectively.

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.2019.2924467, IEEE Access

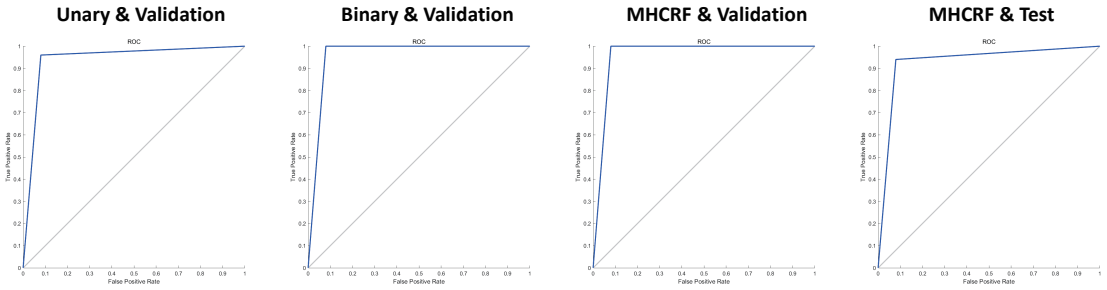

FIGURE 13. ROC curves using the gastric HE datasets. The first to the fourth represent the confusion matrices for the classification results of unary potential on validation set, binary potential on validation set, the MHCRF framework on validation set, and the MHCRF framework on test set, respectively.

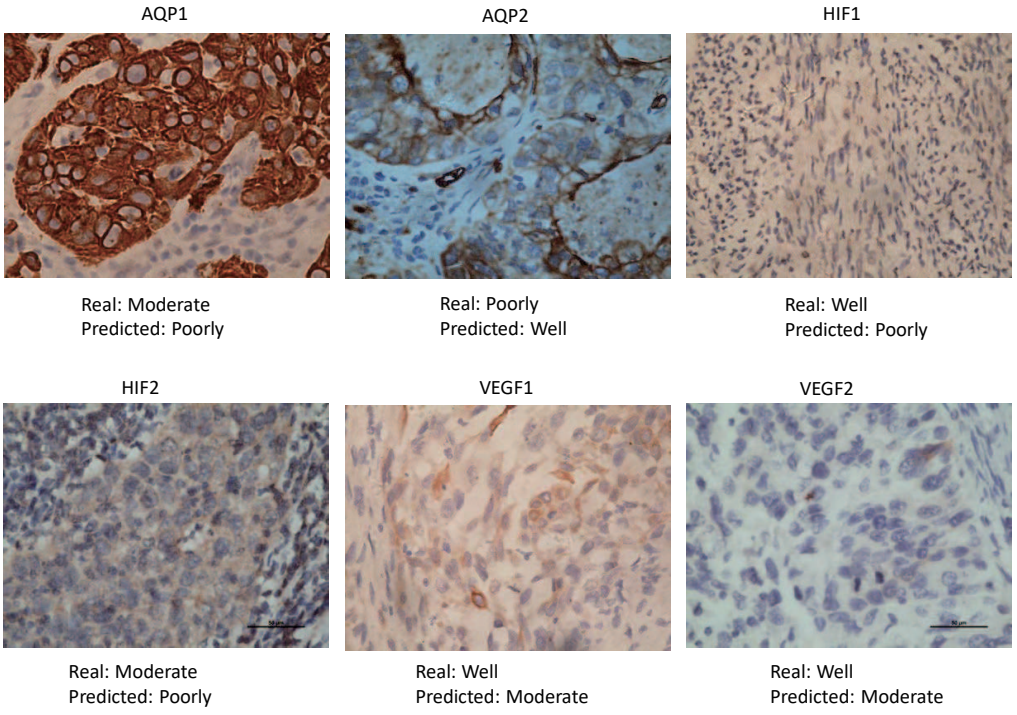

FIGURE 14. An example of the classification results using the six cervical IHC datasets.

[16] R. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, 6:610–621, 1973. II-B0a

[17] L. He, L. Long, S. Antani, and et al. Histology Image Analysis for Carcinoma Detection and Grading. Computer Methods and Programs in Biomedicine, 107(3):538–556, 2012. II-B0a

[18] X. He, R. Zemel, and M. Carreira-Perpiñán. Multiscale Conditional Random Fields for Image Labeling. In Proc. of CVPR 2004, volume 2, pages II–II, 2004. II-C

[19] T. Ho. Random Decision Forests. In Proc. of ICDAR 1995, volume 1, pages 278–282, 1995. II-B0b

[20] L. Hou, D. Samaras, T. Kurc, and et al. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. In Proc. of CVPR 2016, pages 2424–2433, 2016. I, III-B1

[21] M. Huh, P. Agrawal, and A. Efros. What Makes ImageNet Good for Transfer Learning? ArXiv Preprint, page online, 2016. III-B2

[22] H. Kekre, S. Thepade, T. Sarode, and et al. Image Retrieval Using Texture Features extracted from GLCM, LBG and KPE. International Journal of Computer Theory and Engineering, 2(5):695, 2010. II-B0a, III-B2

[23] T. Kohonen. An Introduction to Neural Computing. Neural Networks, 1(1):3–16, 1988. II-B0b

[24] S. Kosov, K. Shirahama, C. Li, and et al. Environmental Microorganism Classification Using Conditional Random Fields and Deep Convolutional Neural Networks. Pattern Recognition, 77:248–261, 2018. II-C

[25] S. Kumar and M. Hebert. Discriminative Random Fields. International Journal of Computer Vision, 68(2):179–201, 2006. III-B2, III-B3

[26] V. Kumar and S. Robbins. Robbins Basic Pathology. Saunders/Elsevier, America, 2007. II-A

[27] J. Lafferty, A.McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. of ICML 2001, pages 282–289, 2001. II-C, III-A

[28] J. Lee, S. Lim, and C. Ahn. Automotive ECU Data-Based Driver's Propensity Learning Using Evolutionary Random Forest. IEEE Access, 7:51899–51906, 2019. II-B0b

[29] C. Li, K. Shirahama, and M. Grzegorzek. Environmental Microorganism Classification Using Sparse Coding and Weakly Supervised Learning. In Proc. of EMC@ICMR 2015, pages 9–14, 2015. II-B0b

[30] D. Lowe. Object Recognition from Local Scale-invariant Features. In Proc. of ICCV 1999, volume 2, pages 1150–1157, 1999. II-B0a

[31] D. Lowe. Distinctive Image Features from Scale-invariant Keypoints. International Journal of Computer Vision, 60(2):91–110, 2004. II-B0a, III-B2

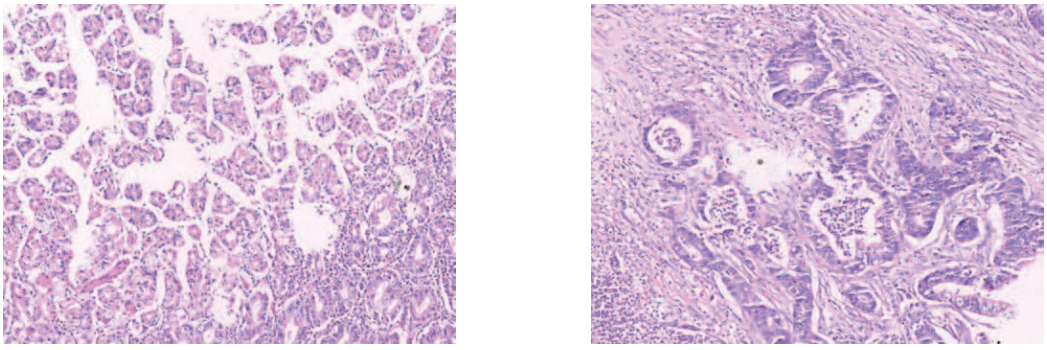

FIGURE 15. An example of the classification results using the gastric HE dataset.

[32] D. Mary, V. Anandan, and K. Srinivasagan. An Effective Diagnosis of Cervical Cancer Neoplasia by Extracting the Diagnostic Features Using CRF. In Proc. of ICCEET 2012, pages 563–570, 2012. II-C

[33] C. Novak and S. Shafer. Anatomy of A Color Histogram. In Proc. of CVPR 1992, pages 599–605, 1992. II-B0a

[34] I. Nyirjesy. Conization of Cervix. https://emedicine.medscape.com/article/270156-overview, 2015. II-A

[35] S. Park, D. Sargent, R. Lieberman, and et al. Domain-Specific Image Analysis for Cervical Neoplasia Detection Based on Conditional Random Fields. IEEE Transactions on Medical Imaging, 30(3):867–878, 2011. II-C

[36] D. Powers. Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1):37–63, 2011. IV-C

[37] J. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez-Jimenez. UPGMpp: a Software Library for Contextual Object Recognition. In Proc. of REACTS 2015, 2015. II-C

[38] B. Settles. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In Proc. of JNLPBA 2004, pages 104–107, 2004. II-C

[39] F. Sha and F. Pereira. Shallow Parsing with Conditional Random Fields. In Proc. of HLT-NAACL 2003, pages 134–141, 2003. II-C

[40] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. ArXiv Preprint, page online, 2014. I, III-B2

[41] F. Suard, A. Rakotomamonjy, A. Bensrhair, and et al. Pedestrian Detection Using Infrared Images and Histograms of Oriented Gradients. In Proc. of IV 2006, pages 206–212, 2006. II-B0a

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, and et al. Rethinking the Inception Architecture for Computer Vision. In Proc. of CVPR 2016, pages 2818–2826, 2016. I, III-B2

[43] S. Theodoridis and K. Koutroumbas. Chapter 4 - Nonlinear Classifiers. In S. Theodoridis and K. Koutroumbas, editors, Pattern Recognition (Fourth Edition), pages 151–260. Academic Press, Boston, 2009. II-B0b

[44] E. Tola, V. Lepetit, and P. Fua. A Fast Local Descriptor for Dense Matching. In Proc. of CVPR 2008, volume 00, pages 1–8, 2008. II-B0a, III-B2

[45] E. Tola, V. Lepetit, and P. Fua. Daisy: An Efficient Dense Descriptor Applied to Wide-baseline Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(5):815–830, 2010. II-B0a

[46] X. Wang, J. Wu, and H. Yang. Robust Image Retrieval Based on Color Histogram of Local Feature Regions. Multimedia Tools and Applications, 49(2):323–345, 2010. II-B0a, III-B2

[47] Y. Wang and J. Rajapakse. Contextual Modeling of Functional MR Images with Conditional Random Fields. IEEE Transactions on Medical Imaging, 25(6):804–812, 2006. II-C

[48] J. Wright, S. Gagnon, R. Richart, and et al. Treatment of Cervical Intraepithelial Neoplasia Using the Loop Electrosurgical Excision Procedure. Obstetrics and Gynecology, 79(2):173–178, 1992. II-A

[49] Z. Zhang and C. Lin. Pathological Image Classification of Gastric Cancer Based on Depth Learning. ACM Transactions on Intelligent Systems and Technology, 45(11A):263–268, 2018. IV-A1

[50] C. Zhao, X. Zhou, L. Sui, and et al. Cervical Cancer Screening and Clinical Management: Cytology, histology, Colposcopy. Beijing Science and Technology Press, China, 2017. I, IV-A1

[51] S. Zheng, S. Jayasumana, B. Romera-Paredes, and et al. Conditional Random Fields as Recurrent Neural Networks. In Proc. of ICCV 2015, 2015. III-A