Automatic CIN grades prediction of sequential cervigram image using LSTM with multistate CNN features

Zijie Yue, Shuai Ding*, Member, IEEE, Weidong Zhao, Hao Wang, Jie Ma, Youtao Zhang, Member, IEEE, Yanchun Zhang

***Abstract*—Cervical cancer ranks as the second most common cancer in women worldwide. In clinical practice, colposcopy is an indispensable part of screening for cervical intraepithelial neoplasia (CIN) grades and cervical cancer but exhibits high misdiagnosis rate. Existing computer-assisted algorithms for analyzing cervigram images have neglected that colposcopy is a sequential and multistate process, which is unsuitable for clinical applications. In this work, we construct a cervigram-based recurrent convolutional neural network (C-RCNN) to classify different CIN grades and cervical cancer. Convolutional neural networks (CNN) are leveraged to extract spatial features. We develop a sequence-encoding module to encode discriminative temporal features and a multistate-aware convolutional layer to integrate features from different states of cervigram images. To train and evaluate the performance of C-RCNN, we leveraged a dataset of 4,753 real cervigrams and obtained 96.13% test accuracy with a specificity and sensitivity of 98.22% and 95.09%, respectively. Areas under each receiver operating characteristic curves (AUC) are above 0.94, proving that visual representations and sequential dynamics can be jointly and effectively optimized in the training phase. Comparative analysis demonstrated the effectiveness of the proposed C-RCNN against competing methods, showing significant improvement over only focusing on a single frame. This architecture can be extended to other applications in medical image analysis.**

***Index Terms*—Endoscopy, Cervix, Computer-aided detection and diagnosis, Machine learning, Neural network**

I. Introduction

CERVICAL cancer is a serious disease that threatens women’s health worldwide. As one of the four most common cancers among women [1], cervical cancer ranks second in terms of cancer fatality rate among women who are approximately 15–44 years of age [2]–[4]. Early detection can effectively help prevent cervical cancer through screening cervical intraepithelial neoplasia (CIN). The possibility of developing cervical cancer can be reduced by receiving appropriate treatment. According to the World Health Organization, detection results can be divided into CIN1 (mild), CIN2 (moderate), CIN3 (severe), and cervical cancer [1]. One important goal in clinical examination is to classify among CIN1, CIN2/3, and cancer.

Colposcopy is an indispensable part of screening for cervical cancer [5], [6]. Colposcopy enhances sensitivity in detecting high-grade CIN lesions and early invasive cancer [7]. Currently, colposcopy is marked as the gold standard for detecting precancerous lesions of the cervix [6]. A colposcopy mainly consists of three parts. First, the physiological saline test involves the use of a cotton swabbed with physiological saline to wipe the surface of the cervix. Second, acetic acid solution is applied to the cervix, which causes abnormal cells of the cervix uteri to turn white gradually [8], which is called acetic-white epithelium. Given that this process lasts for approximately 3 min, a series of images will be collected in each 30-second interval to reflect the dynamic changes of lesions. A green lens will be used to observe and capture the vascular location with its varicosity [9]. Finally, compound iodine solution will be used as the third reagent, and a negative result indicates suspicious lesions [10]. Although colposcopy is the most widely used screening method at present, its accuracy depends largely on the subjective experience of doctors. Even senior experts demonstrates only 48% specificity on clinical examination [11]. This circumstance not only causes a large amount of unnecessary expense to patients but also wastes medical resources.

To overcome these challenges, many studies have been dedicated for analyzing cervigrams automatically. For example, seven classic classifiers, such as support vector machine (SVM), K-nearest neighbor algorithm (KNN), and linear regression (LR), have been used as baselines to evaluate performance on a cervigram dataset [12]. Moreover, the features of colposcopy images have been integrated with PAP/HPV test results, and a classification accuracy of 88.91% through neural networks has

This work is fully supported by the National Natural Science Foundation of China Nos. 91846107, 71571058, and 71690235, Anhui Provincial Science and Technology Major Project Nos. 17030801001, and 18030801137, and the Fundamental Research Funds for the Central Universities No. PA2019GDQT0021.

Z.Yue, S.Ding and H.Wang are with the School of Management, Hefei University of Technology. (e-mail:q164910798@gmail.com; dingshuai@hfut.edu.cn; waynehfut@mail.hfut.edu.cn).

W.Zhao and J.Ma are with the department of gynecology, First Affiliated Hospital of Science and Technology of China. (e-mail: victorzhao@163.com; mj77927@163.com).

Y.Zhang is with the department of Computer Science, University of Pittsburgh, Pittsburgh, USA. (e-mail: zhangyt@cs.pitt.edu).

Y.Zhang is with Centre for Applied Informatics, Victoria University, Melbourne, Australia. (e-mail: yanchun.zhang@vu.edu.au).

Both S.Ding and J.Ma are correponding authors.

been achieved [13]. However, existing analysis toward cervigrams only focused on a single frame, which shows the most obvious changes in the acetic-white test, ignoring the fact that the test results are determined by the dynamic characteristics of the lesions. Furthermore, existing studies have not considered the other two states of images, which contribute greatly to the doctor's judgment. The clinical effectiveness of such methods is limited by the input data.

The current paper presents the construction of C-RCNN to address sequential and multistate cervigrams. C-RCNN is a computer-assisted algorithm that can be applied in routine colposcopy and help physicians for auxiliary diagnosis. We leverage not only CNN models to consider the spatial features of each frame but also Long Short-Term Memory (LSTM) to incorporate the extraction of temporal features into the overall architecture. Moreover, we integrate features of three different cervigram states together and train the network to classify normal, CIN1, CIN2/3, and cervical cancer cervigrams. The experimental results are compared with those of some competing classifiers on the same dataset, whereas test accuracy, sensitivity, specificity, missed diagnosis rate, misdiagnosis rate, and AUC are tested to evaluate performance.

The main contributions of this work are summarized as follows:

1) A novel CIN grade and cervical cancer classification method is presented to analyze sequential and multistate cervigram images automatically. In contrast with alternative methods, C-RCNN can effectively encode spatio-temporal information and extract high-level representations in a data-driven way.

2) To obtain dynamic information of lesions precisely, we designed a sequence encoding module based on LSTM and a constructed voting layer that can produce more discriminative temporal features than previously.

3) To integrate multistate features in the overall architecture, a concatenate layer and a multistate-aware convolutional layer are presented for integrating and realizing dimensionality reduction for features from different states.

4) Our proposed method is evaluated with 4,753 real cervigram images. Our achieved results outperform other approaches by a significant margin. The overall test accuracy reaches 96.13% with a specificity and sensitivity of 98.22% and 95.09%, respectively. The AUC is above 0.94 for each ROC curve.

The remainder of this paper is organized as follows. Section 2 provides an overview on the related work for computer-assisted methods, especially deep learning algorithms for cervigram image analysis. Section 3 provides detailed explanations of our proposed C-RCNN. Experimental results, evaluation metrics, and comparisons with other methods are discussed in Section 4. Section 5 concludes the paper with future research directions.

II. RELATED WORK

A. *Computer-Assisted Algorithms for Cervigrams Analysis*

Computer-assisted algorithms can effectively improve the classification performance of cervigrams, and can be a potentially widely accessible automatic screening method for cervical cancer. Over the past decade, previous studies [14], [15] employed hand-crafted features to represent images and combined them with other examination information, including PAP/HPV test results and patient's past medical history, to calculate classification probabilities. Instead of only considering color features, adding texture and other varied types of hand-crafted features can increase the accuracy in recognizing important vascular patterns [16]. A filter bank of texture models has been used for recognizing punctuation and mosaicism, resulted in a positioning accuracy exceeding 95%. However, the selection of hand-crafted features requires intensive professional knowledge, and manually designing proper features that are fusible across different modalities is difficult [13]. By contrast, neural networks, SVM, KNN, linear discriminant analysis, and decision trees can be applied to analyze cervigram images automatically [17]. For example, linear SVM and KNN have been employed to classify cervigrams, and the accuracy obtained using a neural network is higher than that of traditional methods [13]. Deep learning methods to automatically determine the characteristics of lesions exhibit improved performance [12], [13]. The use of deep learning methods in medical image recognition is a current research trend [18]–[24].

B. *Sequential Medical Image Analysis*

Similar to cervigram images, medical images, such as endoscopic videos and MRI pictures, are composed of sequential images. The temporal features contained in sequential medical images have been analyzed. For the segmentation of brain images [25], LSTM has been used instead of 3D CNN for MR image analysis, achieving better results than existing methods. The complementary information of visual and temporal features learned from a CNN with LSTM has been utilized, producing additional discriminative spatio-temporal features to largely boost the surgical workflow recognition problem [26]. A deep learning framework combines convolution layers, deconvolution layers, and LSTM layer for identifying and classifying breast cancer images [27]. These methods inspired us to use CNN with LSTM for analyzing sequential images. This paper presents the construction of C-RCNN, which extracts the spatio-temporal features of cervigram images to improve the classification accuracy of each CIN grade and cervical cancer.

III. METHODOLOGY

A. *Overall*

In this section, the proposed computer-assisted algorithm is introduced briefly. Our proposed C-RCNN is an essential part of a clinical decision support system for cervical cancer screening, which should also contain the cervigram acquisition module to shoot multistate image sequences, as well as the output module to display the classification results to physicians for auxiliary diagnosis. An overview of C-RCNN is shown in Fig. 1. Based on the multistate and sequential images obtained

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/JBHI.2019.2922682, IEEE Journal of Biomedical and Health Informatics

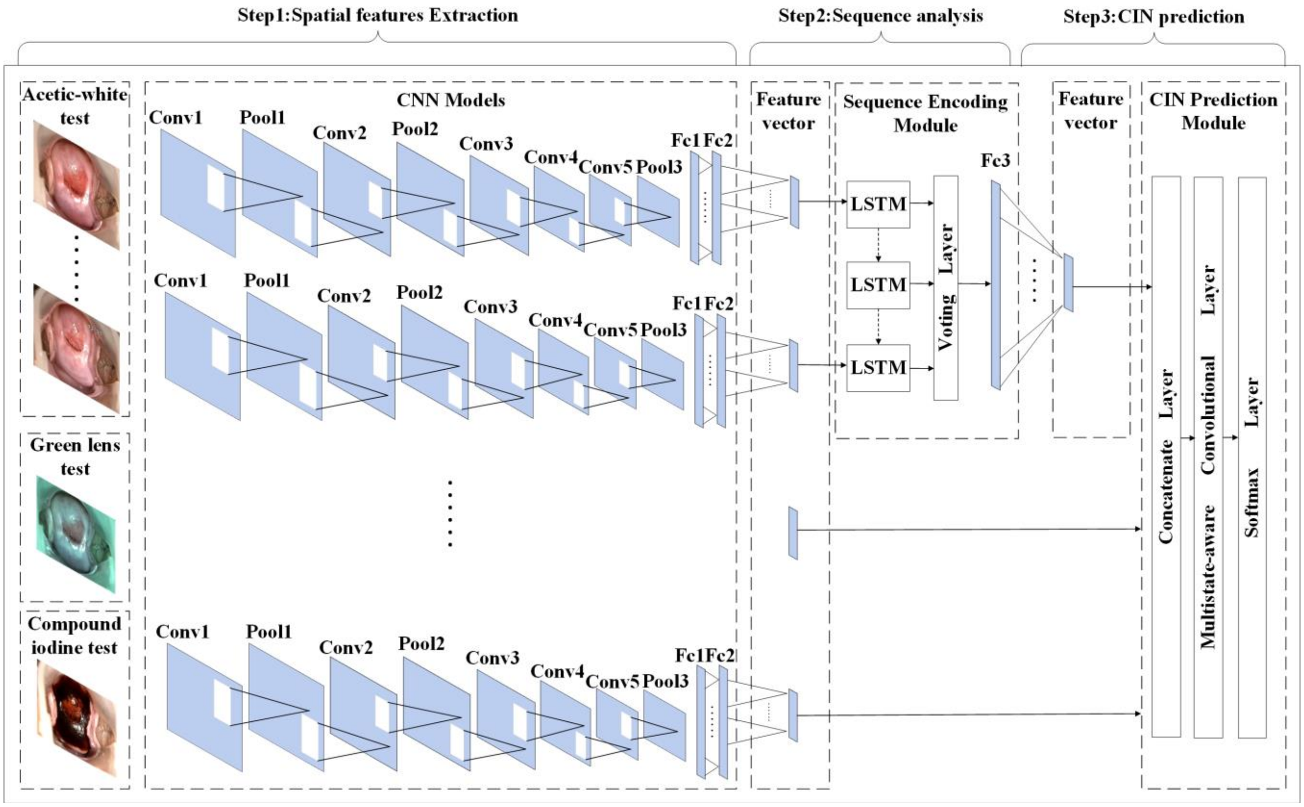


Fig.1. Overall architecture for proposed methodology.

in the course of the colposcopy, C-RCNN focuses on extracting the spatial and temporal features among the image sequences. Moreover, C-RCNN integrates the features of three different states during colposcopy together. This method can obtain classification results of different CIN grades and cervical cancer by using the three following steps:

In step 1, for the three different states of colposcopy images, including sequential images photographed after applying acetic acid to the cervix uteri epithelium and images photographed after using the green lens and compound iodine solution, several constructed CNN models are leveraged to extract spatial features of each frame.

In step 2, a sequence encoding module is applied to detect and locate lesions inside patch sequence candidates and extract the temporal information of each frame by using the sequential images of the acetic-white test, while a voting layer is constructed to produce additional discriminative features.

In step 3, a concatenate layer is utilized to integrate feature vectors to a feature matrix. A multistate-aware convolutional layer is then constructed, which is used for realizing the dimensionality reduction of feature vectors generated in different states. Finally, the classification results of different CIN grades, as well as cervical cancer can be obtained by the Softmax layer.

In general, our proposed C-RCNN can be applied to routine colposcopy and provide reliable support to reduce the misdiagnosis and missed diagnosis rates.

B. *Multistate CNN for Spatial Features Extraction*

Extracting discriminative spatial features of each colposcopy image is the first step of C-RCNN. For a given image sequence, a high-level feature encoding stage is required to properly understand and extract the visual characteristics of the lesions that are present in a specific frame. Compared with previous methods, such as extracting hand-crafted features or using simple neural network, CNNs exhibit better performance [28]. On this regard, a CNN architecture is construct for tackling this crucial task. Note that the CNN models constructed for different states are trained independently, indicating that they do not share weights and parameters.

The selection of CNN architectures depends on the classification requirements and the amount of resources that are occupied [29]. AlexNet [30] has shown good results in the classification of various datasets, we selected its network hierarchy and attempted to improve performance by modifying the hyperparameters to extract additional discriminative spatial features. The specific information on the proposed CNN is shown in Table I, which is composed of several convolutional and pooling layers. The output feature maps of each convolutional layer are calculated using the following equation:

$$x_j^l = Relu(pooling_{average}(\sum x_i^{l-1} * k_{ij}) + b_j^l), \qquad (1)$$

where x_j^l is the j_{th} feature map generated by the convolutional layer l, x_i^{l-1} is the i_{th} feature map of the previous convolutional layer $l-1$, k_{ij} represents the i_{th} trained convolution kernel, b_j^l is the additive bias, $pooling_{average}$ is the average-pooling operation while $*$ represents the convolution operation, and $Relu$ is the activation function.

TABLE I
SPECIFIC INFORMATION OF THE PROPOSED CNN ARCHITECTURE FOR SPATIAL FEATURE EXTRACTION

Layers	Conv1	Pool1	Conv2	Pool2	Conv3	Conv4	Con5	Pool3	Fc1	Fc2
Kernel	7*7	3*3	5*5	3*3	3*3	3*3	3*3	3*3	-	-
Stride	4	2	1	2	1	1	1	2	-	-
Channel	96	96	128	128	256	256	128	128	512	128

Finally, the fully connected layer reduces all of the feature maps into a one-dimensional spatial feature vector y.

For the sequential images of the acetic-white test, the images are input into the CNN model frame by frame to ensure that the spatial features of each frame can be extracted directly. The generated vector sequence $\{y_1 \dots y_n\}$ is then passed to the sequence encoding module. For the images photographed after the green lens and compound iodine solution are used, only their discriminative spatial features are considered, and the feature vector $\{z_2, z_3\}$ is passed to the concatenate layer, which is constructed before the multistate-aware convolutional layer.

C. Sequence Encoding Module Based on LSTM

Given that the acetic-white test is a dynamic process in that the color of abnormal cells gradually turns white, the discrimination of analyzing a single image can only obtain a static appearance of the spatial information, and the CIN grades cannot be inferred from the characteristics of the lesions contained in only one frame. Therefore, the analysis of the acetic-white test must account for the dependency relationships between the successive frames and extract sequential dynamic features frame by frame, which will improve the discrimination accuracy of the overall result.

Compared with the traditional RNN, LSTM can learn long-term dependencies, and the gradient does not tend to vanish when trained with back propagation through time due to its special architecture [31]. The combination of CNN and LSTM achieves good results in video caption generation and target recognition [32], [33], which can also be applied to medical image analysis because it can discover more discriminative spatio-temporal features [26], [34] than previously. Thus, LSTM is used according to our sequential-level images.

In this section, we design a sequence-encoding module based on LSTM. The architecture of this module consists of an LSTM layer, a voting layer, and a fully connected layer, which has the input sequence $\{y_1 \dots y_n\}$ generated by the CNN. After extracting temporal features, the spatio-temporal information can be well encoded in the output vector z_1. The LSTM layer employs 256 LSTM units and dropout is adopted in our encoding module to prevent overfitting.

The LSTM layer is composed of a series of LSTM units. Each unit employs an input gate i_t, a forget gate f_t, an output gate o_t, and a memory cell c_t. Here, i_t controls how much new information enters the unit and alters the state of the memory cell. The variable f_t controls what to be remembered and what to be forgotten, c_t is a summation of the incoming information, and o_t allows the state of the memory cell to affect the current hidden state or other units. These unique structures enable the LSTM to capture long-term temporal dynamics and overcome the vanishing gradient problem. The calculation formulas of LSTM are as follows:

$$i_t = \sigma(W_{ri} r_t + W_{hi} h_{t-1} + b_i),$$
$$f_t = \sigma\left(W_{rf} r_t + W_{hf} h_{t-1} + b_f\right),$$
$$o_t = \sigma(W_{ro} r_t + W_{ho} h_{t-1} + b_o),$$
$$g_t = \varphi\left(W_{rg} r_t + W_{gh} h_{t-1} + b_g\right),$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t,$$
$$h_t = o_t \otimes \mathrm{\varphi}(c_t),$$

For the input sequence $\{y_1 \dots y_n\}$, y_t represents the input at time step *t*, and the output h_{t-1} of the previous moment is used to calculate the current unit states. The LSTM units account for the previous state during the calculation of the current gates, where σ represents the hard-sigmoid nonlinear activation function $\sigma(a) = \frac{1}{1+e^{-a}}$ to normalize the values, φ is the Tanh nonlinear activation function $\varphi(a) = \frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$ that maps real values to (-1, 1), $\otimes$ is an elementwise multiplication that involves computations with gates, and the sets $\{W\}$ and $\{b\}$ represent the weight matrix and bias, respectively.

For the output sequence $\{h_1, h_2, \dots, h_n\}$ of the LSTM layer, existing research works always directly treat the last hidden state h_n as the final descriptor. However, for the problem of cervigram classification, previous states also contain valuable information for doctors to make diagnoses. Thus, we averaged all state's output to produce additional discriminative results and improve classification accuracy by our constructed voting layer.

$$z_1 = \left(\textstyle\sum_{t=1}^{n} h_t\right)/\, n \tag{2}$$

Therefore, for the output vector sequence $\{h_1, h_2, \dots, h_n\}$ of the LSTM layer, we integrate and reduce its dimensionality to the final spatio-temporal feature z_1 by the voting layer and a constructed fully connected layer.

D. Multistate-Aware Convolutional Layer for CIN Grade Prediction

After the above two steps, the CNN-LSTM part of the C-RCNN architecture has fully extracted the spatio-temporal features within each state; however, to generate the colposcopy results, the differences between the successive states [9], [10] also affect the final diagnosis, which must be considered. The characteristics of each state must be considered to make a final prediction.

For this reason, a concatenate layer is leveraged to integrate feature vectors $\{z_1, z_2, z_3\}$ to a feature matrix Z. Moreover, a multistate-aware convolutional layer is constructed to reduce the dimensionality and extract multi-state features of Z. This constructed convolutional layer is used to learn the differences between different state features and to refine state-level

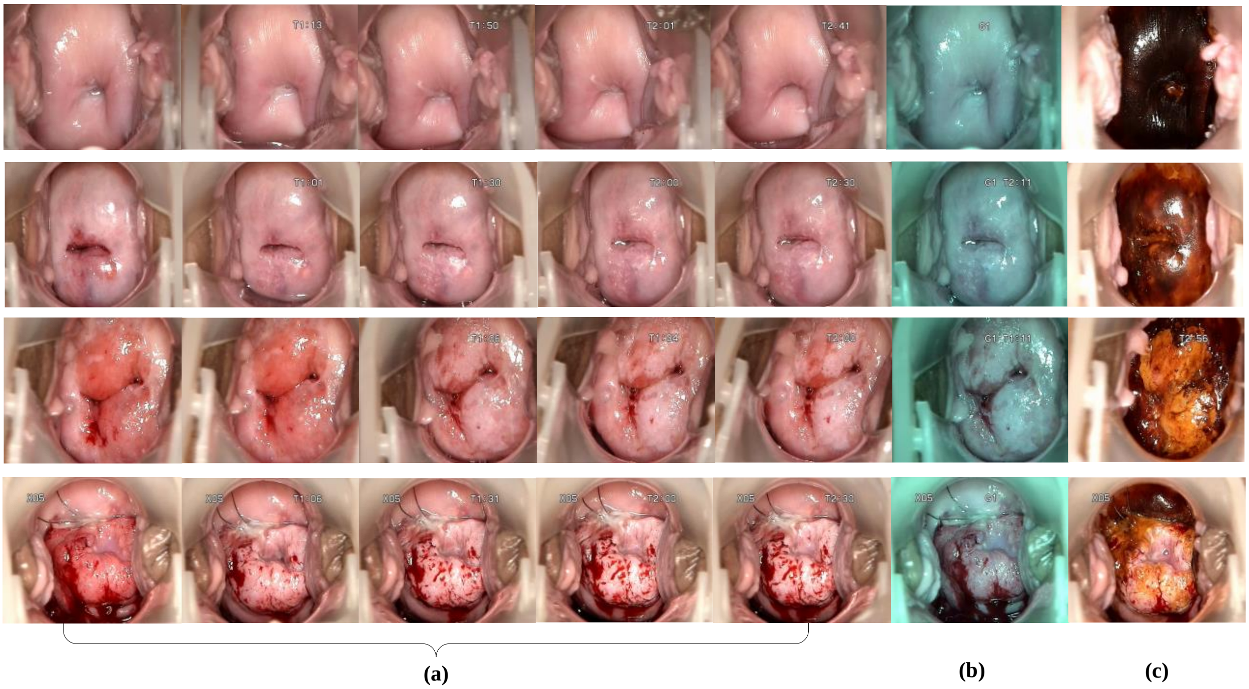


Fig.2. Four cases of colposcopy data, which are Normal, CIN1, CIN2/3, Cervical Cancer respectively. (a) Sequential images photographed after applying acetic acid to the cervix uteri epithelium. (b) Image photographed after using green lens. (c) Image photographed applying compound iodine solution.

semantic information. The final state-level descriptor vector X is generated after the convolutional layer. Finally, the prediction probability of CIN grades and cervical cancer is yielded by forwarding X to the Softmax layer. The calculation formula is:

$$p = Softmax(UX + B), \tag{3}$$

where p represents the prediction probability of each class, the sets $\{U\}$ and $\{B\}$ represent the weight matrix and bias matrix, respectively.

In the architecture of the C-RCNN, we integrate the above three parts seamlessly. We closely integrate different modules and encode additional discriminative spatio-temporal features with both the static performance and dynamic process considered in the design of our network architecture. We take full account of the characteristics of the lesions in different states and produce classification results in different CIN grades and cervical cancer.

IV. EXPERIMENTS

To evaluate the performance of the proposed C-RCNN, a series of experiments is performed based on clinical cervigrams to extensively validate our method in this section. First, our dataset and evaluation metrics are introduced. The experimental setup is then presented. Finally, we show the experimental results and comparisons with competing methods.

A. Dataset and Evaluation Metrics

Our experiments are conducted on a dataset of 679 colposcopy cases from July 2013 to February 2017 at the First Affiliated Hospital of Science and Technology of China. All cervigram images are performed as part of patients' routine clinical practice. No exclusion criteria based on age or race is employed. Each case contains 5 sequential images of an acetic-white test with sequence markers, an image photographed using the green lens, and an image photographed after applying the compound iodine solution. Four cases are shown in Fig. 2. A total of 4,753 real clinical cervigrams with a resolution of 640*480 are selected for training and testing. Before training, during March 2017 to September 2017, all cervigram images were labelled by four gynecologists with over 20 years of clinical experience. Our dataset includes 282 normal, 129 CIN1, 196 CIN2/3, and 72 cervical cancer cases. The distribution of patients' age is statistically analyzed and is shown in Table II.

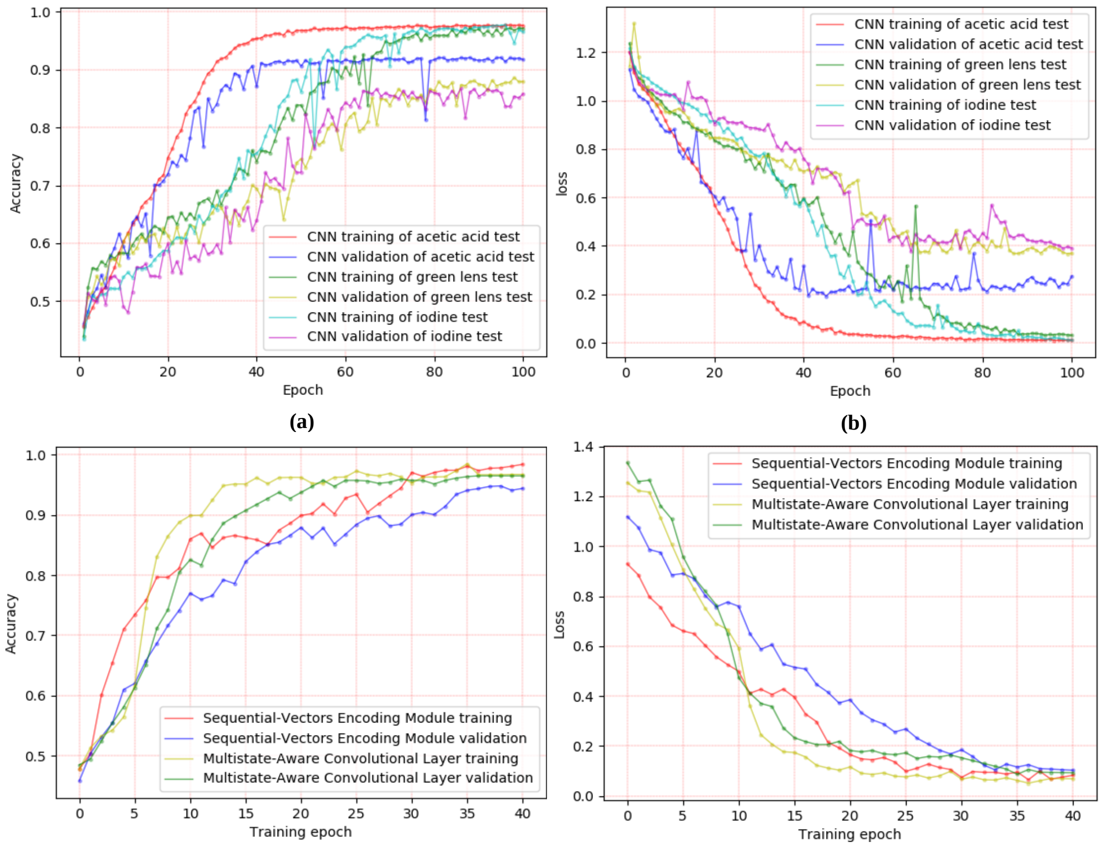
TABLE II
PATIENT AGE DISTRIBUTION IN OUR DATASET

Category	<21	21-29	30-40	41-65	>65	Total
Normal	22	50	83	101	26	282
CIN 1	12	21	31	43	22	129
CIN2/3	10	15	63	71	37	196
Cancer	0	0	8	53	11	72
Total	44	86	185	268	96	679

To analyze the performance of our C-RCNN, we measure six evaluation metrics that are widely adopted in the medical diagnosis field, namely sensitivity (Se), specificity (Sp), missed diagnosis rate (β), misdiagnosis rate (α), test accuracy, and AUC. The valuation metrics of top five metrics are defined as:

$$Sp = \frac{|true\ negative|}{|true\ negative| + |false\ positive|},$$

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/JBHI.2019.2922682, IEEE Journal of Biomedical and Health Informatics



(a) (b)

(c) (d)

Fig.3. The loss-value curves and accuracy curves of training set and validation set. (a) The CNN accuracy curves of three different states. (b) The CNN loss curves of three different states. (c) The accuracy curves of sequence encoding module and multistate-aware convolutional layer. (d) The loss curves of sequence encoding module and multistate-aware convolutional layer.

$$Se = \frac{|true\ positive|}{|true\ positive| + |false\ negative|},$$

$$\alpha = 1 - Sp,$$

$$\beta = 1 - Se,$$

$$Accuracy = \frac{|correctly\ classified\ patient\ cases|}{|test\ cases|},$$

The first four evaluation metrics are the gold standards established for clinical diagnosis. In our experiments, the average values of each evaluation metric in all categories are finally calculated as the final descriptors. True positive refers to the set of patients who fall into the positive class and are correctly classified, false negative refers to the set of patients who fall into the positive class but are misclassified as negative, and true negative and false positive are similarly defined.

ROC curves are typically used in binary classification to study the output of a classifier, which typically feature true positive rate on the Y axis and false positive rate on the X axis. In our experiments, we consider the output as a binary, that is, one of these multi-class is treated as positive class and others as negative class, and extend the ROC curve to multi-class or multi-label classification so one ROC curve can be drawn per label. Another evaluation measure for multi-class classification is macro-averaging, which assumes equal weight to the classification of each label.

B. Experimental Setup

Before the training phase of the C-RCNN, our dataset is split following the train–validation–test pattern [35]. For our experiments, 5-fold cross validation is adopted to evaluate the experimental performances. The entire dataset is randomly divided into five subsets, and the training sets consist of possible combinations of three of these five subsets for cross validation for a total of $C_5^3 = 10$ training sets. Another randomly selected subset is used as a validation set to help adjust the hyperparameters in the training phase. The remaining subset constitutes a test set to finally evaluate the performance of C-RCNN. Therefore, for 5-fold cross validation, ten experiments are conducted, and the average values of each metric are considered as our experimental results. After splitting, our training set contains 2,849 clinical images, whereas both the validation and test sets contain 952 cervigrams. To train a robust deep model on a small dataset, we leverage the data augmentation technique that is widely used in the area of computer vision and pattern recognition to augment our training set from the aspects of rotation, horizontal flip, and vertical flip. The augmented training set contains 14,245 cervigrams. The cross-entropy and stochastic gradient descent strategy are selected to calculate the loss and fine tune the parameters. All of the weights and biases are determined through training after

TABLE III
THE EXPERIMENTAL RESULTS OF DIFFERENT PARTS IN THE PROPOSED C-RCNN ARCHITECTURE

Classifier	Accuracy	Sp	Se	α	β
CNN for acetic acid test	90.91%	96.54%	90.57%	3.46%	9.43%
CNN-LSTM for acetic acid test	94.21%	97.66%	93.63%	2.34%	6.37%
CNN for green lens test	88.81%	95.51%	86.47%	4.49%	13.53%
CNN for iodine test	86.65%	91.11%	83.54%	8.89%	16.46%
C-RCNN	**96.13%**	**98.22%**	**95.09%**	**1.78%**	**4.91%**

TABLE IV
PERFORMANCE COMPARISON BETWEEN OUR PROPOSED METHOD AND THE COMPETING APPROACHES

Approach	Accuracy	Sp	Se	α	β
2D-CNN	86.45%	90.24%	84.00%	9.76%	16.00%
3D-CNN+2D-CNN[38], [39]	91.90%	95.08%	88.91%	4.92%	11.09%
C-RCNN	**96.13%**	**98.22%**	**95.09%**	**1.78%**	**4.91%**

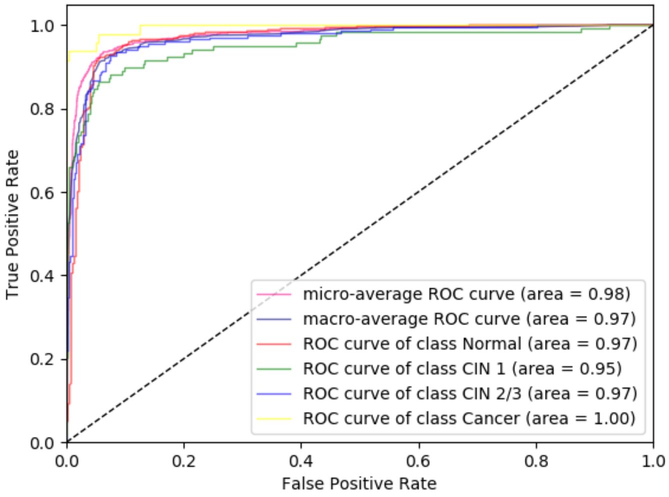


Fig.4. ROC curves of proposed C-RCNN for multiclass classification.

random initialization. We set up the training batch size as 64 and the epochs as 100.

All experiments are performed on a machine running Linux OS with an Intel Xeon @ 2.16 GHz CPU, an NVIDIA GeForce Titan X 4-way graphics card, and 128 GB of RAM. The Keras framework is implemented.

C. Results and Discussion

In this section, the experimental results are presented. The performance of the proposed C-RCNN architecture and each module are analyzed in detail. To verify the effectiveness of our method while considering discriminatively spatio-temporal features, we compare existing methods, such as SVM, KNN, and 3D-CNN, as peer competitors. Note that 5-fold cross validation is adopted to evaluate the experimental performances, and the average value of each metric is considered as our experimental results.

1) Experimental Performance Analysis: Given that the proposed C-RCNN aims to classify cases with different CIN grades and cervical cancer, we leverage our clinical cervigram dataset for experiments. The experimental results are shown in Table III. The test accuracy of the acetic-white test by using our constructed CNN architecture achieves 90.91%. On this basis, the addition of the sequence encoding module increases the test accuracy to 94.21%, proving that accounting for discriminative spatio-temporal features is beneficial. The overall test accuracy reaches 96.13% with a specificity and sensitivity of 98.22% and 95.09%, respectively, because the multistate-aware convolutional layer integrates the features of three states, which considers the characteristics of lesions after the green lens and compound iodine solution are used. The experimental results show that considering the multistate information in this problem is beneficial to the determination of the final results, and the performance of the overall architecture is more remarkable than that of each independent state.

We calculate the loss values and fine tune the network parameters through the back-propagation algorithm in the training phase. The loss-value curves and the accuracy curves of our experiments are shown in Fig. 3. We conclude that the CNN accuracy curves of three different states reach the highest points around the 80th epoch, whereas the loss-value curve of our constructed sequence encoding module tends to be stable at the 35th epoch. The structure of the multistate-aware convolutional layer is simple in such a way that we can obtain the optimal solution as soon as possible.

Considering that our dataset is an imbalanced dataset and to prove that we obtain accurate classification results in each class, we show the ROC curves in Fig. 4. Note that we focus on a multi-class classification problem, and the analysis must be performed per class, that is, one-versus-all scheme. Four basic ROC curves, as well as micro-averaging ROC curve and macro-averaging ROC curve, can be drawn. The experimental results shown in Fig. 4 reveal that for each ROC curve, the AUC is above 0.94, which indicates that our approach exhibits accurate classification performance for each class, and both spatial and temporal features can be extracted by our methodology. Second, the AUC of the cervical cancer class reaches the highest of 1.00, which can be concluded that our C-RCNN demonstrates improved performance for identifying cancer lesions. Compared with other classes, the AUC of CIN1 class is the lowest because the precancerous lesions come in many types

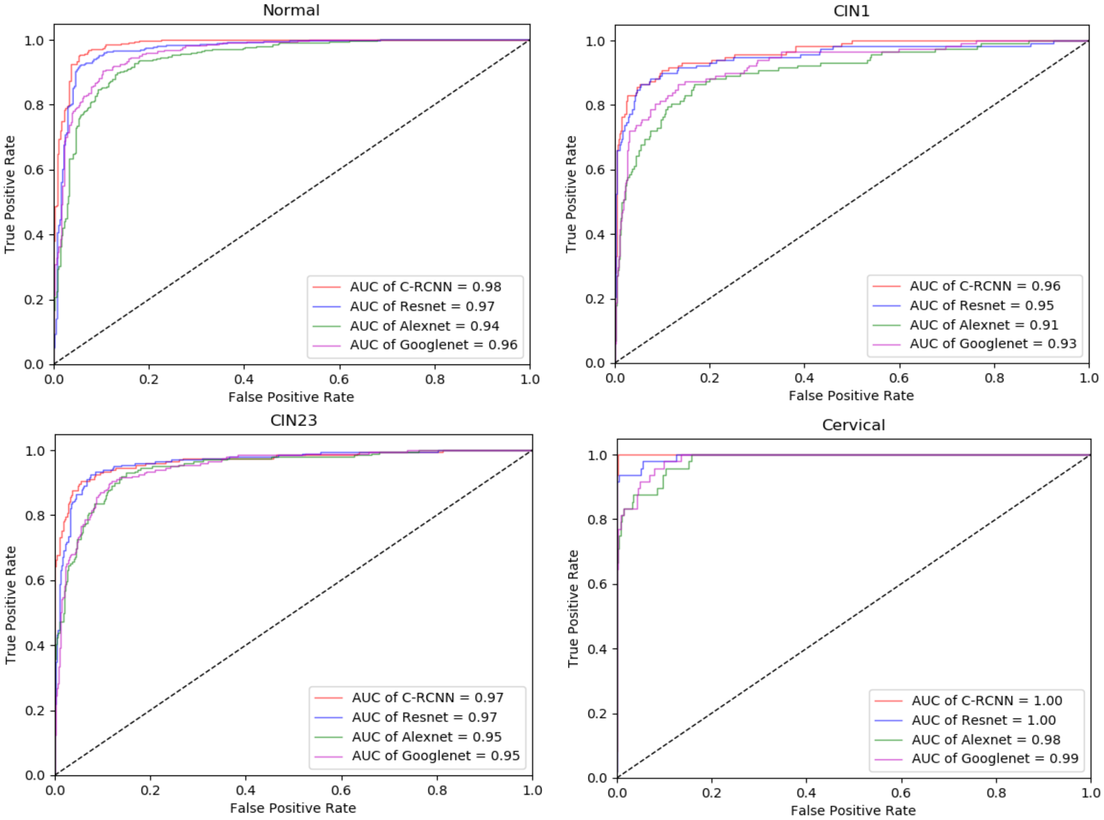
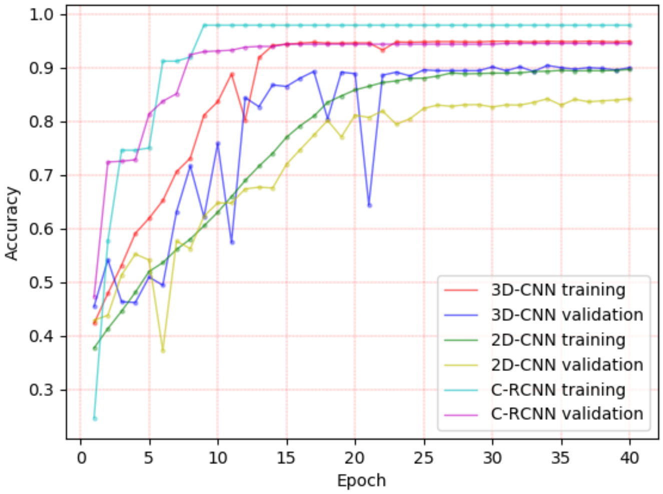


Fig.6. ROC analysis for the proposed C-RCNN and other competing CNN models: Alexnet, Googlenet and Resnet. The analysis was performed per class (one-versus-all).



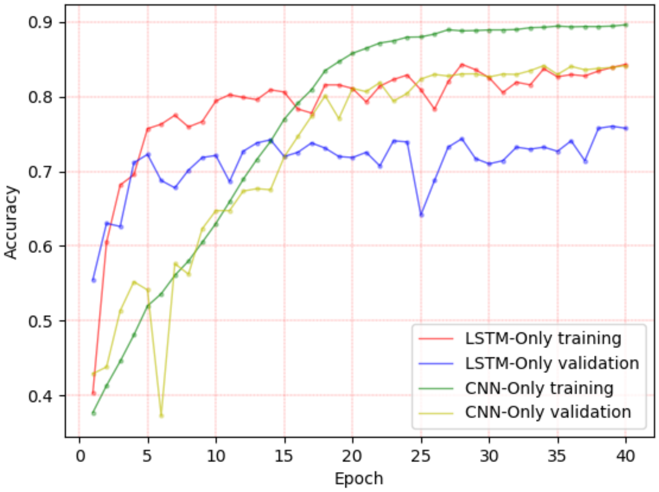


Fig.5. The accuracy curves between 2D-CNN, 3D-CNN and our proposed C-RCNN.

Fig.7. The accuracy curves after removing the CNN model and the LSTM model separately.

and shapes, whereas our approach may not have learned each characteristic well. Most precancerous lesions are hardly to be observed and learned from computer-aided diagnostic algorithms. Finally, the areas under micro-averaging ROC curve and macro-averaging ROC curve are 0.98 and 0.97, respectively, which further explains that the superiority of the proposed method after averaging over all considered classes is also significant.

2) Comparison to the Competing Methods: To further evaluate the performance of the proposed C-RCNN, we compare it with traditional machine learning methods, including SVM and KNN, which are high-efficiency methods for performing classification in cervigrams. SVM is leveraged to implement classification [36], whereas a domain-specific automated image analysis framework is proposed with a KNN for the detection of cervical cancer and CIN grades [37]. The test accuracy is 58.30% and 56.19%, respectively, which is much lower than we expected. We also select some neural network structures that are commonly used in medical image analysis, such as 3D-CNN, to reconstruct a deep learning model. Inspired by previous studies [38], [39], we leverage 3D-CNN instead of LSTM to extract the temporal features among the

TABLE V
COMPARISON OF THE PROPOSED CNN ARCHITECTURE WITH OTHER CNNS

	Accuracy	Sp	Se	α	β
AlexNet	86.75%	89.96%	83.40%	10.04%	16.60%
GoogLe Net	90.44%	92.37%	87.63%	7.63%	12.37%
ResNet	93.21%	96.27%	91.88%	3.73%	8.12%
C-RCNN	**96.13%**	**98.22%**	**95.09%**	**1.78%**	**4.91%**

TABLE VI
PERFORMANCE COMPARISON BETWEEN THE CNN-LSTM ARCHITECTURE AND THE CNN OR LSTM MODELS

	Accuracy	Sp	Se	α	β
CNN-Only	85.76%	91.18%	84.06%	8.82%	15.94%
LSTM-Only	74.55%	86.41%	70.19%	13.59 %	29.81%
CNN-LSTM	**94.21%**	**97.76%**	**93.63%**	**2.34%**	**6.37%**

image sequences, and 2D-CNN is selected to reduce the dimensionality of the multistate features in the end. In addition, we attempt to use CNN to classify different CIN grades and cervical cancer directly without considering the characteristics of the lesions, which differ from state to state. The same dataset is used to train each classifier, and 5-fold cross validation is adopted to evaluate their performances. As shown in Table IV, the test accuracy is 86.45% and 91.90%. The accuracy curves between our method and the competing approaches are shown in Fig. 5.

The experimental results show that the performance of existing methods appear to be ineffective when considering the images photographed after applying the green lens and compound iodine solution. Such methods are not applicable to analyzing multistate cervigram images, which can improve the performance actually. Existing methods cannot address sequential images and do not account for the dynamic characteristics of the lesions, except for 3D-CNN. We can conclude that the competing methods do not apply to our dataset and that they are unsuitable for clinical examination.

3) Effect of CNN parameters for spatial feature extraction: One of the most important tasks in C-RCNN is to extract discriminative spatial features. Although increasing the number of network layers can improve the CNNs' performance, it will also considerably require prolonged training time and occupy additional computing resources. In our experiments, several parameters of the classical CNNs' architectures, including AlexNet, GoogLeNet, and ResNet, are adjusted to extract additional discriminative features. Finally, the obtained CNN architecture based on the AlexNet meets our requirements.

We compare the performances of different architectures that are constructed with some classical CNN models. Note that only the CNN models utilized are different but exhibits the same LSTM and following layers. Cross validation is adopted, whereas each metric is averaged. The results are shown in Table V. Our constructed C-RCNN exhibits a test accuracy of 96.13% with a specificity and sensitivity of 98.22% and 95.09%, respectively, which is clearly superior to others. For a detailed comparison at different operating points, we also perform ROC analysis for some classical CNN architectures and the proposed C-RCNN. The comparisons are conducted for each of the considered classes, and the analysis is performed using a one-versus-all scheme. As shown in Fig. 6, the proposed C-RCNN achieves the highest AUC on each of the four classes whereas the architecture with Resnet shows a competitive performance. The performances of AlexNet and GoogLeNet are not as ideal as we expect, especially on the class of CIN1. The results of the analysis confirm that our constructed architecture show remarkably superior performance against other methods and is suitable for the problem of cervical cancer screening.

4) Effect of each part of CNN-LSTM: To evaluate each part of the CNN-LSTM network's contribution to the overall result, we remove the CNN model and the LSTM model separately. As shown in Table VI and Fig. 7, for the part of CNN-only, we remove the LSTM layer and substitute it with a fully connected layer. The feature vector is directly classified by the Softmax layer with a test accuracy of 85.76%. For the part of LSTM-only, we construct a deep LSTM model that can capture a high level of sequence information by stacking LSTM layers. Each layer in the LSTM is a hierarchy that receives the hidden state of the previous layer as input. We utilize the entire image into the LSTM model directly and construct a fully connected layer to complete the classification, which obtains a test accuracy of 74.55%. By contrast, the application of CNN-LSTM is superior in all evaluation metrics, from which it can be inferred that using the CNN-LSTM architecture can improve the performance better than using the CNN or LSTM model separately. Considering not only the static performance but also the dynamic characteristics of the lesions is helpful in the CIN grade and cervical cancer classification.

V. CONCLUSIONS AND FUTURE WORK

In this paper, focusing on the problem of the low specificity of colposcopy examination, we propose a novel computer-assisted algorithm that is different from designing hand-crafted features to describe the visual appearance and dynamic changes or traditional machine learning methods. The proposed discriminative learning model, C-RCNN, exhibits a key novelty in using CNN to extract the spatial features while leveraging LSTM to extract the temporal features. We also consider multistate cervigrams generated in clinical examination and then integrate them all to reduce the dimension. In this manner, not only the visual representations and sequential dynamics can be jointly and effectively optimized in the training phase, but also differences between different state features can be learned to refine high-level semantic information. Compared with the competing methods, the C-RCNN shows improved performance in terms of the specificity, sensitivity, missed diagnosis rate, misdiagnosis rate, test accuracy, and AUC. Importantly, the proposed C-RCNN is a computer-assisted algorithm that can be applied into the colposcopy routine because its input data is the clinical image sequences generated during colposcopy. In clinical practice, C-RCNN is a core part of the clinical decision support system that also requires cervigram acquisition module to obtain pictures of the cervix, as well as an output module to display the classification results to physicians for auxiliary diagnosis.

In our future work, we plan to leverage different variations of RNN, such as GRU and bidirectional LSTM. We also plan to achieve object recognition tasks toward our cervigram dataset. Given that our proposed method is mainly used for the classification of different CIN grades and cervical cancer, no specific outputs of the lesion positions or lesion types can provide doctors with intuitive judgments. Therefore, we plan to try YOLO, Faster R-CNN, and other object recognition methods. We will extend this architecture to other applications in medical image analysis, such as gastroscopy and capsule gastroscopy.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their detailed and thoughtful feedback which improved the quality of this paper significantly. This work is fully supported by the National Natural Science Foundation of China Nos. 91846107, 71571058, and 71690235, Anhui Provincial Science and Technology Major Project Nos. 17030801001, and 18030801137, and the Fundamental Research Funds for the Central Universities No. PA2019GDQT0021.

REFERENCES

[1] WHO and ICO, “Human Papillomavirus and Related Diseases Report - WORLD,” *HPV Inf. Cent.*, no. Albania, pp. 1–138, 2014.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018,” *CA. Cancer J. Clin.*, vol. 68, no. 1, pp. 7–30, 2018.

[3] M. H. Forouzanfar *et al.*, “Breast and cervical cancer in 187 countries between 1980 and 2010: A systematic analysis,” *Lancet*, vol. 378, no. 9801, pp. 1461–1484, 2011.

[4] A. I. Ojesina *et al.*, “Landscape of genomic alterations in cervical carcinomas,” *Nature*, vol. 506, no. 7488, pp. 371–375, 2014.

[5] C. R. Eheman *et al.*, “National Breast and Cervical Cancer Early Detection Program data validation project,” *Cancer*, vol. 120, no. SUPPL. 16, pp. 2597–2603, 2014.

[6] T. Denkçeken *et al.*, “Elastic light single-scattering spectroscopy for the detection of cervical precancerous ex vivo,” *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 123–127, 2013.

[7] D. G. Ferris, M. Schiffman, and M. S. Litaker, “Cervicography for triage of women with mildly abnormal cervical cytology results,” *Am. J. Obstet. Gynecol.*, vol. 185, no. 4, pp. 939–943, 2001.

[8] H. Greenspan *et al.*, “Automatic detection of anatomical landmarks in uterine cervix images,” *IEEE Trans. Med. Imaging*, vol. 28, no. 3, pp. 454–468, 2009.

[9] A. Milbourne *et al.*, “Results of a pilot study of multispectral digital colposcopy for the in vivo detection of cervical intraepithelial neoplasia,” *Gynecol. Oncol.*, vol. 99, no. 3 SUPPL., pp. 67–75, 2005.

[10] M. Segondy *et al.*, “Performance of careHPV for detecting high-grade cervical intraepithelial neoplasia among women living with HIV-1 in Burkina Faso and South Africa: HARP study,” *Br. J. Cancer*, vol. 115, no. 4, pp. 425–430, 2016.

[11] M. F. Mitchell, D. Schottenfeld, G. Tortolero-Luna, S. B. Cantor, and R. Richards-Kortum, “Coloposcopy for the diagnosis of squamous intraepithelial lesions-A meta-analysis,” *Obstet. Gynecol.*, vol. 91, no. 4, pp. 626–631, 1998.

[12] T. Xu *et al.*, “Multi-feature based benchmark for cervical dysplasia classification evaluation,” *Pattern Recognit.*, vol. 63, no. January 2016, pp. 468–475, 2017.

[13] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, “Multimodal deep learning for cervical dysplasia diagnosis,” in *MICCAI*, 2016, pp. 115–123.

[14] D. Song *et al.*, “Multimodal entity coreference for cervical dysplasia diagnosis,” *IEEE Trans. Med. Imaging*, vol. 34, no. 1, pp. 229–245, 2015.

[15] T. Xu, X. Huang, E. Kim, L. R. Long, and S. Antani, “Multi-test cervical cancer diagnosis with missing data estimation,” in *SPIE Medical Imaging*, 2015, vol. 56, p. 94140X.

[16] S. Gordon, G. Zimmerman, and H. Greenspan, “Image segmentation of uterine cervix images for indexing in PACS,” in *Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems*, 2004, no. May, p. 298.

[17] Y. Jusman, S. C. Ng, and N. A. Abu Osman, “Intelligent screening systems for cervical cancer,” *Sci. World J.*, vol. 2014, 2014.

[18] G. Litjens *et al.*, “A Survey on Deep Learning in Medical Image Analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[19] R. Zhang *et al.*, “Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features from Nonmedical Domain,” *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 1, pp. 41–47, 2017.

[20] J. X. Qiu, H. J. Yoon, P. A. Fearn, and G. D. Tourassi, “Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports,” *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 1, pp. 244–251, 2018.

[21] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[22] Z. Yan *et al.*, “Multi-Instance Deep Learning: Discover Discriminative Local Anatomies for Bodypart Recognition,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1332–1343, 2016.

[23] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA - J. Am. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[24] H. Wang, S. Ding, D. Wu, Y. Zhang, and S. Yang, “Smart connected electronic gastroscope system for gastric cancer screening using multi-column convolutional neural networks,” *Int. J. Prod. Res.*, vol. 7543, pp. 1–12, 2018.

[25] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber, “Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation,” *Comput. Sci.*, pp. 1–9, 2015.

[26] Y. Jin *et al.*, “SV-RCNet: Workflow Recognition from Surgical Videos using Recurrent Convolutional Network,” *IEEE Trans. Med. Imaging*, vol. 37, no. 5, pp. 1114–1126, 2017.

[27] 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. Image Process.*, vol. 27, no. 5, pp. 2189–2200, 2018.

[28] B. Microbiana *et al.*, “Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[29] Z. Luo, L. Liu, J. Yin, Y. Li, and Z. Wu, “Deep learning of graphs with ngram convolutional neural networks,” *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2125–2139, 2017.

[30] A. Krizhevsky, I. Sutskever, and H. Geoffrey E., “ImageNet Classification with Deep Convolutional Neural Networks,” in *NIPS*, 2012, pp. 1–9.

[31] F. Wu *et al.*, “Temporal interaction and causal influence in community-based question answering,” *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2304–2317, 2017.

[32] J. Donahue *et al.*, “Long-Term Recurrent Convolutional Networks for Visual Recognition and Description,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, 2017.

[33] N. Lu, Y. Wu, L. Feng, and J. Song, “Deep Learning for Fall Detection: 3D-CNN Combined with LSTM on Video Kinematic Data,” *IEEE J. Biomed. Heal. Informatics*, vol. 2194, no. c, 2018.

[34] C. Xu, L. Xu, Z. Gao, S. Zhao, H. Zhang, and Y. Zhang, “Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture,” *Med. Image Anal.*, vol. 50, pp. 82–94, 2018.

[35] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.

[36] E. Njoroge, S. R. Alty, M. R. Gani, and M. Alkatib, “Classification of cervical cancer cells using FTIR data.,” in *EMBS*, 2006, pp. 5338–5341.

[37] S. Y. Park, D. Sargent, R. Lieberman, and U. Gustafsson, “Domain-specific image analysis for cervical neoplasia detection based on conditional random fields,” *IEEE Trans. Med. Imaging*, vol. 30, no. 3, pp. 867–878, 2011.

[38] Q. Dou *et al.*, “3D deeply supervised network for automated segmentation of volumetric medical images,” *Med. Image Anal.*, vol. 41, pp. 40–54, 2017.

[39] Q. Dou *et al.*, “Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.