Cervical Intraepithelial Neoplasia Grading from Prepared Digital Histology Images

Brid BROSNAN*, Inna SKARGA-BANDUROVA*,1, Tetiana BILOBORODOVA and Illia SKARHA-BANDUROV

*Oxford Brookes University, UK
bSaarland University of Applied Science HTW Saar, Saarbrücken, Germany
*cUniversity of Oxford, UK

Abstract. The paper proposes an integrated approach to the automated diagnosis of cervical intraepithelial neoplasia (CIN) in epithelial patches extracted from digital histology images. Experiments were conducted to determine the most suitable deep learning model for the dataset and fuse patch predictions to decide the final CIN grade of the histology samples. Seven candidate CNN architectures were assessed in this study. Three fusion methods were applied to the best CNN classifier. The model ensemble, combined CNN classifier and highest performing fusion method achieved an accuracy of 94.57%. This result shows significant improvement over the state-of-the-art classifiers for cervical cancer histopathology images. It is hoped that this work will contribute towards further research to automate diagnosis of CIN from digital histopathology images.

Keywords. Cervical intraepithelial neoplasia (CIN), histology image, deep learning, convolutional neural network (CNN)

1. Introduction

Cervical cancer is a significant global challenge recorded as the fourth most common cancer among women globally, with an estimated 604,000 new cases and 342,000 deaths in 2020 [1]. Cervical intraepithelial neoplasia (CIN) precedes invasive cervical cancer. CIN is graded 1, 2 or 3. CIN 1 has a high probability of regression. CIN 2 and CIN 3 are considered precancerous lesions. CIN 3 is synonymous with severe dysplasia and carcinoma in situ [2]. Early stage diagnosis and treatment can cure cervical cancer.

The advent of AI and deep learning in digital pathology has yielded notable improvements in diagnostic proficiency [3]. However, the application of AI-based solutions to biomedical image data still faces several challenges, among which are scarcity of publicly available data; ethical constraints; availability of experts to label data and interobserver variability. Moreover, histopathological images themselves pose some unique challenges. Image appearance is highly variable due to slide preparation, the nonstandard shape of the biopsied tissue and size of the epithelium regions, as well as the presence of artefacts, e.g., stains, ink markers, tapes, and blurred regions [4, 5]. Hence, wide adoption of deep learning models in digital pathology is limited. This project aims to detect CIN grade from digital histology images, addressing some of these data issues.

1 Inna Skarga-Bandurova, School of Engineering, Computing and Mathematics, Oxford Brookes University, Wheatley Campus, Oxford, OX33 1HX, UK; E-mail: iskarga-bandurova@brookes.ac.uk.
1. Methods

Figure 1 provides the technical context for stratified squamous epithelium (SSE) detection, CIN classification and diagnosis from histology images used in this study. The proposed strategy comprises preparatory phase where regions of interest (RoI) are automatically extracted from histology images, followed by CIN classification and decision-making phase.

In Phase I, a methodology that integrated both geometric and image-based techniques was utilized to automate the detection of stratified squamous epithelium (SSE) fragments in histology images. This approach was further used to generate suitable patches, thereby creating a unique SSE dataset. Resulting patches were evaluated and checked for any inconsistencies. Phase I has been comprehensively detailed in [6]. In Phase II, the focus of this paper, the research protocol necessitated the inclusion of supplementary patches to increase sample variety. Samples from MTCHI dataset [7], used with the authority of [7 8], were manually prepared to enhance the CHI histology patch image dataset. The histology image analysis included a thorough evaluation of the classification, integration, and CIN grade decision-making. This process required fusion of constituent patch predictions to determine the final CIN grade of the histology sample. Seven state-of-the-art Convolutional Neural Network (CNN) architectures were tested to define the benchmark model and three fusion strategies for assigning final CIN grade to the histology images were implemented and compared.

The CHI dataset [6] includes SSE patches obtained from glass slides images, 357 patch images from 38 CIN 1 and 443 patch images from 58 CIN 2 grade. The MTCHI samples were derived from 80 digital whole slide images (WHI). Of the 80 MTCHI images, those without unanimously agreed pathologists’ CIN grade label, and those without sufficient SSE, were excluded from this study. Of the 49 remaining images, 21 were classified as normal, 1 as CIN1, 10 as CIN2, and 17 as CIN3. These images were prepared by extracting 1114 SSE patch images: 425, Normal; 16, CIN 1; 128, CIN 2; and 545, CIN 3. Exclusion of ‘edge patch’ images during pre-processing further reduced the dataset. Ultimately, 1715 images were used in the study. All images were resized 200 × 200 pixels.
2. Results

Repeated refinement of the data, and training parameters, enabled achievement of 92.14% accuracy for CNN model [5], adapted for multiclass classification, further selected as a benchmark (Fig. 2). The benchmark classifier mean F-score (92.19%) indicates excellent agreement between the model’s predictions and the true values. Mean kappa (89.4%) and weighted kappa (90.83%) indicate very good agreement. Kappa is slightly lower than accuracy, reflecting the small class imbalance. Individual CIN class assessment indicates F-score is highest for CIN 3 (96.2%) and lowest for CIN 2 (88.8%). These results suggest the small class imbalance is not critical in this study, where CIN 2 is the majority class followed by CIN 3, Normal and CIN 1.

Finally, three fusion methods were implemented to fuse patch class predictions and determine the final CIN grade. A1 assigns the most severe predicted CIN grade to the sample and indicates a Type I error. CNN + A1 ensemble accuracy is 90.7%. A2 assigns the most prevalent predicted CIN grade to the sample. CNN + A2 ensemble achieves the highest diagnostic accuracy (94.57%), F-score (94.36%), kappa (92.23%) and weighted kappa (93.57%) on our dataset. This indicates A2 is best at predicting the individual human pathologist’s expert opinion of the grade of the histology samples used in this study. A3 builds on A2, by using the CNN model class probabilities to determine predicted class confidence. CNN + A3 ensemble accuracy is 93.8%.

CNN + A2 per-class performance metrics are provided in Table 1. Table 2 shows a comparison of final model accuracy with state-of-the-art classifiers for cervical cancer histopathology images.

Table 1. CNN + A2 Per-class Performance Metrics

<table>
<thead>
<tr>
<th>Class</th>
<th>Sensitivity / Recall / TPR</th>
<th>Specificity</th>
<th>Precision</th>
<th>FPR</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIN 1</td>
<td>0.9714</td>
<td>0.9787</td>
<td>0.9444</td>
<td>0.0213</td>
<td>0.9577</td>
<td>0.9751</td>
</tr>
<tr>
<td>CIN 2</td>
<td>0.9123</td>
<td>0.9861</td>
<td>0.9811</td>
<td>0.0139</td>
<td>0.9455</td>
<td>0.9492</td>
</tr>
<tr>
<td>CIN 3</td>
<td>1.0000</td>
<td>0.9821</td>
<td>0.8947</td>
<td>0.0179</td>
<td>0.9444</td>
<td>0.9911</td>
</tr>
<tr>
<td>Non-neoplasm</td>
<td>0.9500</td>
<td>0.9817</td>
<td>0.9048</td>
<td>0.0183</td>
<td>0.9268</td>
<td>0.9658</td>
</tr>
</tbody>
</table>

Table 2. Comparison of model classification accuracy with state-of-the-art CIN classifiers

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>F-score (%)</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConvNet feature extraction and classification. Fusion.</td>
<td>77.25</td>
<td></td>
<td>Almubarak et al. [10]</td>
</tr>
<tr>
<td>Fusion-based, Hybrid Deep Learning Manual &amp; CNN</td>
<td>80.72</td>
<td></td>
<td>Almubarak et al. [9]</td>
</tr>
<tr>
<td>Deep CIN</td>
<td>88.50</td>
<td>88.00</td>
<td>Sornapudi et al. [11]</td>
</tr>
<tr>
<td>End-to-end framework - Automated ROI Detection</td>
<td>85.00</td>
<td>84.20</td>
<td>Sornapudi et al. [5]</td>
</tr>
<tr>
<td>CNN &amp; A2 fusion</td>
<td>94.57</td>
<td>94.36</td>
<td>Ours</td>
</tr>
</tbody>
</table>
3. Discussion and conclusions

The experiment findings show that, although the training dataset class imbalance was relatively low, stratifying the validation data improved interpretation and provided more stable results. Data augmentation during model fitting increased accuracy by over 7.5%. All fusion methods identified the actual CIN 3 cases. CNN + A2 fusion is the best model to detect CIN 1 and CIN 2 in this dataset. However, fusion method comparison was limited due to the small number of samples with conflicting patch predictions. Individual misclassification findings include a non-neoplasm sample, misclassified as CIN 3 by all fusion methods. One case was particularly extreme because three different patch classes are predicted, none of which are the actual sample class. Summing up the above, although overall performance is very good, the potential application of the classifier in medical diagnostics needs further investigation. Assessment of each CIN class before fusion shows that the CNN model best detects CIN 3. CIN 3 precision and recall are 94.2% and 98.4% respectively, which are higher than CIN 1 and CIN 2. Expert validation of the classifier output would provide more confidence in the model prior to fusing the patch results and assigning the fused predicted CIN grade to the parent sample. Ultimately, it is hoped that both stages of the complex histology image analysis project will be combined, enabling CIN grade identification directly from the histology sample image. It is envisaged that borderline predictions will be presented in a decision support system, automating what can be done readily by machines and freeing up scarce pathologist resources to make clinical decisions best left to human experts.

References