Context-Aware Convolutional Neural Network for Grading of Colorectal Cancer Histology Images

Muhammad Shaban, Ruqayya Awan, Muhammad Moazam Fraz, *Senior Member, IEEE*, Ayesha Azam, Yee-Wah Tsang, David Snead, and Nasir M. Rajpoot, *Senior Member, IEEE*

***Abstract*— Digital histology images are amenable to the application of convolutional neural networks (CNNs) for analysis due to the sheer size of pixel data present in them. CNNs are generally used for representation learning from small image patches (e.g. 224 × 224) extracted from digital histology images due to computational and memory constraints. However, this approach does not incorporate high-resolution contextual information in histology images. We propose a novel way to incorporate a larger context by a context-aware neural network based on images with a dimension of 1792 × 1792 pixels. The proposed framework first encodes the local representation of a histology image into high dimensional features then aggregates the features by considering their spatial organization to make a final prediction. We evaluated the proposed method on two colorectal cancer datasets for the task of cancer grading. Our method outperformed the traditional patch-based approaches, problem-specific methods, and existing context-based methods. We also presented a comprehensive analysis of different variants of the proposed method.**

***Index Terms*— Computational pathology, deep learning, context-aware convolutional networks, cancer grading.**

Manuscript received December 3, 2019; accepted January 23, 2020. Date of publication February 3, 2020; date of current version June 30, 2020. The work of Muhammad Shaban was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant 1829583. The work of Nasir M. Rajpoot was supported in part by The Alan Turing Institute through the EPSRC under Grant EP/510129/1. This work was partly supported by the Medical Research Council (MRC) award for the MICAHIL project [MR/P015476/1]. *(Corresponding author: Nasir M. Rajpoot.)*

Muhammad Shaban and Ruqayya Awan are with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: m.shaban@warwick.ac.uk; r.awan.1@warwick.ac.uk).

Muhammad Moazam Fraz is with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K., and also with the National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan (e-mail: moazam.fraz@warwick.ac.uk).

Ayesha Azam is with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K., and also with the Department of Pathology, University Hospitals Coventry and Warwickshire, Coventry CV2 2DX, U.K. (e-mail: ayesha.azam@warwick.ac.uk).

Yee-Wah Tsang and David Snead are with the Department of Pathology, University Hospitals Coventry and Warwickshire, Coventry CV2 2DX, U.K.

Nasir M. Rajpoot is with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K., also with the Department of Pathology, University Hospitals Coventry and Warwickshire, Coventry CV2 2DX, U.K., and also with the The Alan Turing Institute, London NW1 2DB, U.K. (e-mail: n.m.rajpoot@warwick.ac.uk).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.2971006

I. Introduction

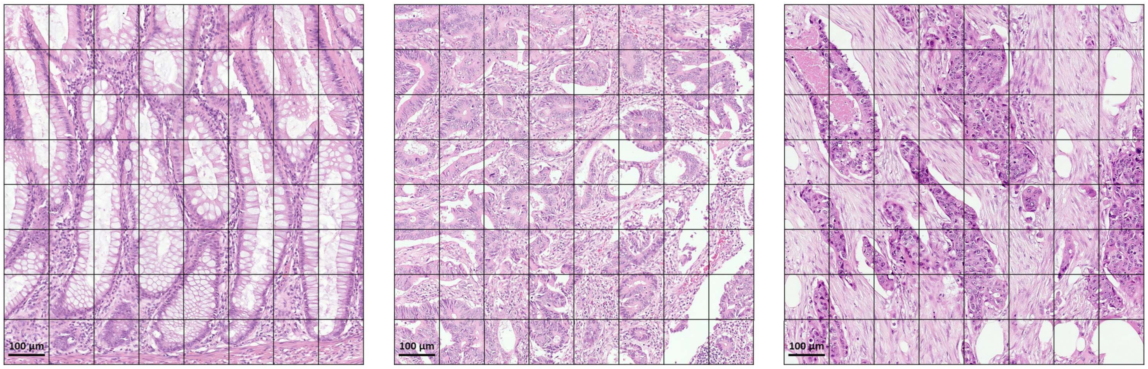
Histology slides are used by pathologists to analyze the micro-anatomy of cells and tissues through a microscope. However, recent technological developments in digital imaging solutions [1] have digitized the histology slides (histology images) to enable the pathologists to do the same analysis over the computer screen. These histology images are significantly larger than natural images. Each image contains tens of thousands of cells and each cell nucleus usually takes around 50×50 square pixels at the highest magnification level (e.g. $40\times$). The digitization process results in an explosion of data which leads to new avenues of research for machine learning and deep learning communities.

Convolutional neural networks (CNNs) have been widely used to achieve the state-of-the-art results for different histology image analysis tasks such as nuclei detection and classification [2]–[4], metastasis detection [5]–[7], tumor segmentation [8] and cancer grading [9]–[11]. Each task requires a different amount of contextual information, for instance, cell classification needs only high-resolution cell appearance along with little neighboring tissue whereas tumour detection and segmentation rely on a larger context covering multiple cells simultaneously. Due to tumour heterogeneity, cancer grading requires high-resolution cell information as well as the contextual spatial organization of cells in the tumour microenvironment (TME). Most existing CNN based methods applied to histology images follow a patch based approach to train different models which tend to ignore contextual information due to memory constraints. Although these models are often trained on a large number of image patches extracted from histology images, often spatial relationships between neighbouring patches are ignored. Due to the lack of large contextual information, the inference is independent of underlying tissue architecture and it is performed based on the limited context captured by individual patches. This approach works well for problems where contextual information is relatively less important for prediction. However, contextual information becomes vital in problems where diagnostic decisions are made on the basis of underlying tissue architecture such as cancer grading.

In this paper, we consider colorectal cancer (CRC) grading to demonstrate the significance of context-aware CNNs in cancer histology image analysis. CRC is the fourth most common cause of cancer-related deaths worldwide [12]. The

0278-0062 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.



Normal Low Grade High Grade

Fig. 1. Three visual field regions of colorectal tissue which highlight the importance of larger context for correct grading. Each box shows the 224 × 224 pixel context captured by a standard patch classifier at 20× magnification.

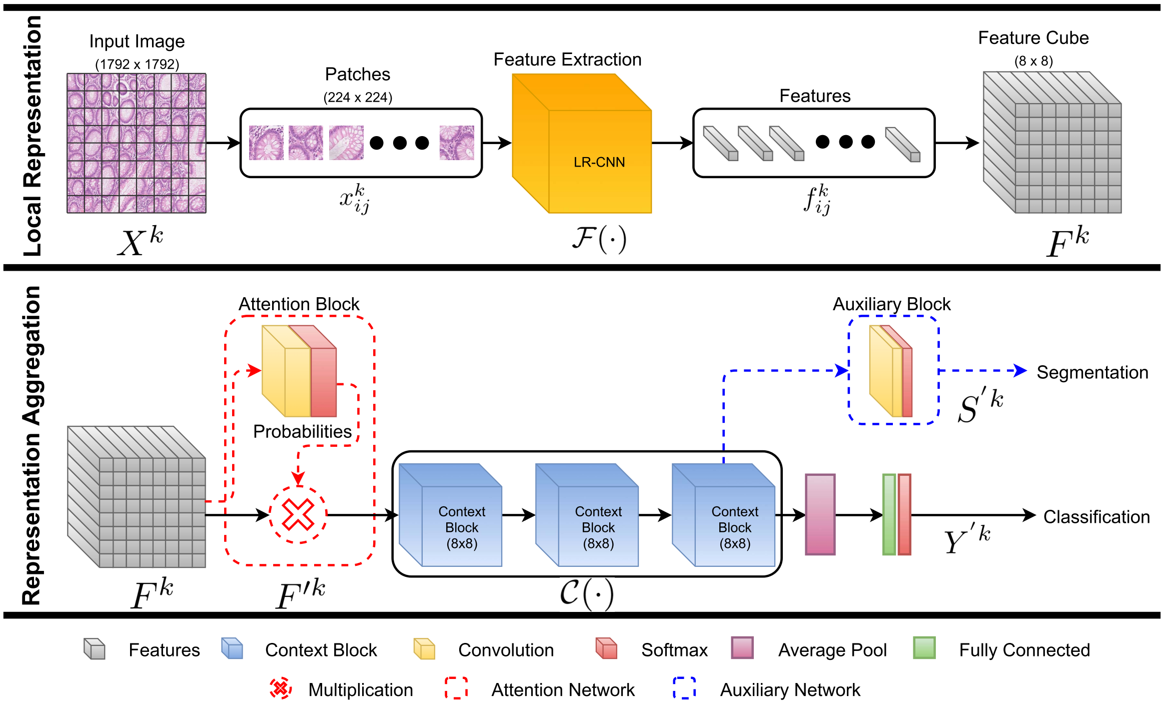


Fig. 2. Flow diagram of the proposed context-aware framework for CRC grading. The top row shows local representation learning. The bottom row illustrates the network architecture for representation aggregation learning which consists of multiple context blocks and other standard layers. Dashed lines represent the blocks of a specific network design whereas solid lines represent the common blocks (see Table I for notations).

grade of CRC is determined by pathologists by collective analysis of individual cancer cells' abnormality and their spatial organization as a distorted glandular structure in the histology image. Several studies on the prognostic significance of CRC adopted a two-tiered grading system to reduce the inter-observer variability [13], [14], merging the well and moderately differentiated glands into a low-grade tumor and classifying tissue with poorly and undifferentiated glands as a high-grade tumor. In this work, we consider diagnostic regions captured from CRC histology images containing enough context to reliably predict the cancer grade (see Figure 1). We refer to them as visual fields in this paper as selected by an expert pathologist. A CNN based method for CRC grading requires an image with large contextual information to capture cell organization for accurate grading.

We propose a novel framework for context-aware learning of histology images. The proposed framework first learns the Local Representation by a CNN (LR-CNN) and then aggregate the contextual information through a representation aggregation CNN (RA-CNN), as shown in Figure 2. The proposed framework takes a large size image (1792 × 1792) as an input unlike the usual input image size (224 × 224) of standard patch classifiers. The input image is then divided into small patches (224 × 224) in sliding window fashion with no-overlap. The LR-CNN takes the patches as input and converts them into high-dimensional feature vectors where the length of feature vectors depends on the choice of LR-CNN network. These feature vectors are arranged in the form of a feature-cube using the same spatial arrangement in which the corresponding patches were extracted. This feature-cube

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

is then fed into the RA-CNN to make predictions based on both high-resolution feature representation and spatial context. The proposed context-aware framework is flexible enough to incorporate any state-of-the-art image classifier as LR-CNN for local representation learning with the RA-CNN. We present detailed results and show that our proposed framework achieves superior performance over traditional patch-based approaches and existing context-aware methods. Moreover, the proposed framework also outperforms the methods designed specifically for CRC grading using handcrafted features based on gland architecture. Our main contributions in this paper are as follows:

- We propose a novel framework for context-aware learning from large high-resolution input images.
- The proposed framework is highly flexible since it can leverage any state-of-the-art network design for local representation learning.
- We explore different context-aware learning and training strategies to examine the framework's ability to learn the contextual information.
- We report the results of comprehensive experiments (with 100+ network models) and comparisons to demonstrate the superiority of the proposed context-aware learning framework over traditional patch-based methods and existing context-aware learning methods.

II. Related Work

Related work is divided into two subsections: methods related to context-aware learning and some problem specific methods on cancer grading.

A. Context-Aware Learning

In the literature, various approaches have been presented to incorporate the contextual information for the classification of histology images. Some researchers [15]–[17] used image down-sampling, a common practice followed in natural image classification, to capture the context from a larger histology image. However, this approach is not suitable for problems where cell information is as important as the context. Adaptive patch sampling [18] and discriminative patch selection [19] from histology images is another way to integrate the sparse context. These methods are not capable of capturing small regions of interest at high resolution e.g. tumor cells and their local contextual arrangement. Some methods [6], [20]–[22] leverage the multi-resolution nature of histology images and use multi-resolution based classifiers to capture context. These multi-resolution approaches only consider a small part of an image at high resolution and the remaining part at low resolutions to make a prediction. Therefore, these approaches lack the contextual information of cellular architecture at high resolution in a histology image.

Recently, some works [23]–[26] have used larger high-resolution patches to improve the segmentation of histology images. Zanjani *et al.* [25] and Li and Ping [26] used a CNN based feature extractor followed by a Conditional Random Field (CRF) for context learning. The latter is end-to-end trainable with a patch size of 672 × 672, considerably smaller than the patch size used in the proposed method. Agarawalla *et al.* [23] and Kong *et al.* [24] used a 2D Long Short-Term Memory (LSTM) instead of CRF to improve tumor segmentation. Some works [27], [28] used larger patches at high resolution for the task of context-based classification. Awan *et al.* [27] proposed a context-aware network for breast cancer classification. They used standard SVM to learn the context from the CNN based features of the patches extracted from a high-resolution image. Due to the nature of the final classifier, this work is only capable of capturing a limited context. Bejnordi *et al.* [28] proposed a similar approach for breast tissue classification. They trained their network in two steps. In the first step, they used a small patch size and in the second step, they fixed the weights of half of the network to feed a larger patch for training the remaining half of the network. Their network also suffers from a limited context problem as they managed to train a network with the largest patch size of $1,024 \times 1,024$ pixels with small batch size (10 patches). Sirinukunwattana *et al.* [29] presented a systematic comparison of different context-aware methods to highlight the importance of context-aware learning.

As opposed to the aforementioned methods, our proposed method is different in a network design such that it is flexible enough to accommodate any state-of-the-art CNN architecture for representation learning and a custom CNN based architecture for representation aggregation. The representation learning and aggregation are stacked together for context-aware learning with larger image size, 1792×1792 and a typical batch size of 64 images.

B. Problem Specific Method

A number of automated methods for objective grading of breast, prostate, and colorectal cancer [11], [30]–[32] have been proposed in the literature. For instance, in [30]–[32], a linear classifier is trained with handcrafted features based on the glandular morphology for prostate cancer grading. Awan *et al.* [11] presented a method for two-tier CRC grading based on the extent of deviation of the gland from its normal shape (circular/elliptical). They proposed a novel Best Alignment Metric (BAM) for this purpose. As a pre-processing step, CNN based gland segmentation was performed, followed by the calculation of BAM for each gland. For every image, average BAM was considered as a feature along with two more features inspired by BAM values. Finally, an SVM classifier was trained using this feature set for CRC grading.

Our proposed method differs from these existing methods in two ways. First, it does not depend on the intermediate step of gland segmentation making it independent of segmentation inaccuracies. Second, the proposed method is entirely based on a deep neural network which makes this framework independent of cancer type. Therefore, the proposed framework could be used for other context-based histology image analysis problems.

III. The Proposed Method

The proposed framework for context-aware grading consists of two stacked CNNs as shown in Fig 2. The first network, LR-CNN, converts the high-resolution information of an image into high dimensional feature-cube through patch based feature

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

TABLE I
ENUMERATION OF SYMBOLS USED IN THE PAPER

Symbol	Description	Symbol	Description
D	Image dataset	X^k	k^{th} image
K	Number of images	Y^k	Label of k^{th} image
C	Number of Classes	S^k	Mask of k^{th} image
$\mathbf{X}$	Set of all images	$\mathbf{Y}$	Labels of $\mathbf{X}$
$\mathbf{S}$	Masks of $\mathbf{X}$	d	Patch dataset
M	Patches in an image column	i	$1, \ldots, M$
N	Patches in an image row	j	$1, \ldots, N$
x^k_{ij}	ij^{th} patch of X^k	y^k_{ij}	Label of x^k_{ij} patch
$\mathcal{F}(\cdot)$	Feature extractor	f^k_{ij}	Features of x^k_{ij}
L_f	Fully connected layer	L^g_p	Global pooling layer
$L^{a \times a}_c$	$a \times a$ convolution layer	L_s	Softmax layer
$\rightarrow$	Transition between layers	$\bullet$	Preceding layer's output
$\otimes$	Hadamard product	$\oplus$	Feature Concatenation
$\mathcal{B}(\cdot)$	Context-block	$\mathcal{C}(\cdot)$	Context-Net
$\mathbf{F}$	Feature of $\mathbf{X}$	$\mathbf{F}'$	Weighted Feature of $\mathbf{X}$
$\mathbf{Y}'$	Predicted labels of $\mathbf{X}$	Y'^k	Predicted label of X^k
$\mathbf{S}'$	Predicted Masks of $\mathbf{X}$	S'^k	Predicted Mask of X^k
W^k	k^{th} image weight	θ	Learnable Parameters
$\mathcal{L}_{cls}$	Classification cost function	$\mathcal{L}_{wgt}$	Weighted cost function
$\mathcal{L}_{seg}$	Segmentation cost function	$\mathcal{L}_{joint}$	Joint cost function

extraction. The second network, RA-CNN aggregates the learned representation in order to learn the spatial context from the feature-cube to make a prediction. We leverage the power of traditional patch classifiers to learn local representation from individual patches. However, we explore different network architectures for context block in RA-CNN for context-aware learning. Moreover, different training strategies are explored to find the optimal configurations of the context-aware grading framework. The following section explains each building block of the proposed framework in details. The notations used to describe each building blocks are summarized in Table I.

A. Network Input

The input to our framework is an image (X^k) from a dataset, $D = \{X^k, Y^k, S^k; k = 1, \ldots K\}$, of large high resolution images which consists of K images with corresponding labels $Y^k \in \{1, \ldots, C\}$ for classification into C classes and coarse patch level segmentation masks $S^k \in \{1, \ldots, C\}$ for multi-task learning. Each image is divided into $M \times N$ patches of same size where x^k_{ij} and y^k_{ij} represent the ij^{th} patch of k^{th} image and its corresponding label, respectively. We used a patch dataset, $d = \{(x^k_{ij}, y^k_{ij}), | \; x^k_{ij} \in X^k, y^k_{ij} \in Y^k\}$, which consists of patches and their corresponding labels for pre-training of LR-CNN.

B. Local Representation Learning

The first part of the proposed framework encodes an input image X^k into a feature-cube F^k. All the input images are processed through the LR-CNN in a patch based manner. The proposed framework is flexible enough to use any state-of-the-art image classifier for local representation learning (LR-CNN) such as ResNet50 [33], MobileNet [34], Inception [35], or Xception [36]. This flexibility also enables it to use pre-trained weights in case of a limited dataset. Moreover, it is possible to train the LR-CNN independently before plugging it into the proposed framework, enabling it to learn meaningful representation [37] which leads to early convergence of the context-aware learning part of the framework.

C. Feature Pooling

The spatial dimensions of the output feature f^k_{ij} of a patch x^k_{ij} may vary depending on the input patch dimensions and the network architecture for feature extraction. A global feature pooling layer is employed to get a similar dimensional feature vector for all variations of the proposed framework. Both average and max global pooling strategies are explored. After global pooling, features of all patches are rearranged in the same spatial order ($M \times N$) as extracted patches to construct the feature-cube F^k for context-aware learning. The depth of the feature-cube depends on the choice of LR-CNN. For the sake of generality, we will represent the output of our LR-CNN as follows,

$$\mathbf{F} = \mathcal{F}(\mathbf{X}, \theta_{\mathcal{F}}) \rightarrow L^g_p(\cdot) \tag{1}$$

where $\mathcal{F}$ represents the fully convolutional part of the LR-CNN and acts as a feature extractor whereas $\mathbf{X}$ is the batch of images and $\mathbf{F}$ is the local feature representation of $\mathbf{X}$ after pooling L^g_p, which could be a global average or max pooling layer. The operator ($\rightarrow$) provides the output of the preceding layer to the following layer and the operator ($\cdot$) represents the output of the preceding layer.

D. Feature Attention

As the input to the proposed framework has a relatively large spatial dimension, there may be some part of the image that may not have any significance for the prediction of the image label. We introduce an attention block which gives less weight to insignificant features and vice-versa. This attention block takes feature-cube as input and learns the weights for the feature-cube. Hadamard product (element-wise product) is taken between the weights and input feature-cube to increase the impact of more important areas of an image in label prediction and vice-versa. The weighted feature-cube $\mathbf{F}'$ is defined as:

$$\mathbf{F}' = L^{1\times1}_c(\mathbf{F}, \theta_c) \rightarrow L_s(\cdot) \otimes \mathbf{F}, \tag{2}$$

where $L^{1\times1}_c$ and θ_c represent the 1×1 convolution layer and its parameters, respectively. L_s denotes the softmax layer and the operator $\otimes$ is used to represent Hadamard product.

E. Context Blocks

Since the LR-CNN is used to encode the important patch-based image representation into a feature-cube, the main aim of the context block (CB) is to learn the spatial context within the feature cube. The CB learns the relation between the features of the image patches considering their spatial location. We propose three different CB architectures, each with different complexity and capability to capture the context information. First CB, $\mathcal{B}_1(\cdot)$, is comprised of a 3×3 convolution layer followed by ReLU activation and batch normalization. Second CB, $\mathcal{B}_2(\cdot)$, uses residual block [33] architecture with two different filter sizes. It consists of three convolution layers each followed by a batch normalization and ReLU activation. The first and last layers are with 1×1

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

convolution filter to squeeze and expand the feature depth. The output feature-maps of the last layer are concatenated with the input features-maps which makes its final output. The $\mathcal{B}_2(\cdot)$ is defined as:

$$\begin{aligned}\mathcal{B}_2(\mathbf{F}',\theta_{\mathcal{B}_2}) &= [L_c^{1\times1}(\mathbf{F}',\theta_{\mathcal{B}_2^1}) \rightarrow L_c^{3\times3}(\cdot,\theta_{\mathcal{B}_2^2})\\ &\quad \rightarrow L_c^{1\times1}(\cdot,\theta_{\mathcal{B}_2^3})] \oplus \mathbf{F}', \end{aligned} \tag{3}$$

where $L_c^{1\times1}$ and $L_c^{3\times3}$ denote the convolution layers with 1×1 and 3×3 filter sizes; $\theta_{\mathcal{B}_2^1}$, $\theta_{\mathcal{B}_2^2}$, and $\theta_{\mathcal{B}_2^3}$ are the parameters of different convolution layers and $\theta_{\mathcal{B}_2}$ represents parameter of the whole context block for brevity. The operator $\oplus$ represents the concatenation of feature-maps.

Unlike the previous two context blocks, our third CB processes the input feature-maps in parallel with different filter sizes to capture context from varying receptive fields. Similar to the blocks in [35], it consists of multiple 1×1 and 3×3 convolution layers each followed by a batch normalization and ReLU activation. A 3×3 average pooling layer $L_p^{3\times3}$ is also used to average the local context information. The CB, $\mathcal{B}_3$, is defined as:

$$\begin{aligned}\mathcal{B}_3(\mathbf{F}',\theta_{\mathcal{B}_3}) &= [L_c^{1\times1}(\mathbf{F}',\theta_{\mathcal{B}_3^1}) \rightarrow L_c^{3\times3}(\cdot,\theta_{\mathcal{B}_3^2})\\ &\quad \rightarrow L_c^{3\times3}(\cdot,\theta_{\mathcal{B}_3^3})] \oplus [L_c^{1\times1}(\mathbf{F}',\theta_{\mathcal{B}_3^4})]\\ &\quad \oplus[L_c^{1\times1}(\mathbf{F}',\theta_{\mathcal{B}_3^5}) \rightarrow L_c^{3\times3}(\cdot,\theta_{\mathcal{B}_3^6})]\\ &\quad \oplus[L_p^{3\times3}(\mathbf{F}') \rightarrow L_c^{1\times1}(\cdot,\theta_{\mathcal{B}_3^7})], \end{aligned} \tag{4}$$

where $\theta_{\mathcal{B}_3^1}$ to $\theta_{\mathcal{B}_3^7}$ are the parameters of different convolution layers and $\theta_{\mathcal{B}_3}$ represents parameter of the whole context block for the sake of notational simplicity.

F. Representation Aggregation for Context Learning

A cascaded set of three context blocks ($\mathcal{C}(\cdot)$) of the same type ($\mathcal{B}_1$,$\mathcal{B}_2$, or $\mathcal{B}_3$) is used in RA-CNN. These context blocks are explained in section III-E. The output of $\mathcal{C}(\cdot)$ is followed by a global average pooling layer, a fully connected layer, and a softmax layer to make the final prediction in the required number of classes. The final prediction $\mathbf{Y}'$ from the features of input images $\mathbf{X}$ is computed as:

$$\mathbf{Y}' = \mathcal{C}(\mathbf{F}',\theta_{\mathcal{C}}) \rightarrow L_p^g(\cdot) \rightarrow L_f(\cdot,\theta_{f'}) \rightarrow L_s(\cdot), \tag{5}$$

where $\theta_{\mathcal{C}}$ and $\theta_{f'}$ represent the parameters of all context blocks and the fully connected layer in RA-CNN, respectively. The proposed framework is trained with categorical cross-entropy loss based cost function $\mathcal{L}_{cls}(\cdot)$ which is defined as:

$$\mathcal{L}_{cls}(\mathbf{Y},\mathbf{Y}') = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} Y_c^k \log_2(Y_c^{'k}), \tag{6}$$

where Y_c^k and $Y_c^{'k}$ are the ground truth and predicted probabilities of k^{th} image for c^{th} class.

G. Auxiliary Block

The proposed framework is designed for the classification of large input images. Therefore, the label of an input image may depend on a set of different primitive structures (such as glands) and their spatial organization. We proposed an auxiliary block to exploit these primitive structures. This auxiliary block acts as a patch based segmentation of the primitive structures in an input image (k) and outputs a coarse patch based segmentation mask ($S^{'k}$). The segmentation masks ($\mathbf{S}'$) of input images $\mathbf{X}$ from their features $\mathbf{F}'$ is defined as:

$$\mathbf{S}' = \mathcal{C}(\mathbf{F}',\theta_{\mathcal{C}}) \rightarrow L_c^{1\times1}(\cdot,\theta_{c'}) \rightarrow L_s(\cdot), \tag{7}$$

where $L_c^{1\times1}$ is a convolution layer with $\theta_{c'}$ parameters. The addition of auxiliary block enables the proposed framework to learn in a multi-task setting [38]–[41] where both tasks share the same base network which helps to overcome the issue of representation bias and overfitting. The loss function for one task acts as a regularizer for the other tasks. The weights are optimized based on joint loss($\mathcal{L}_{joint}$) which consist of $\mathcal{L}_{cls}$ and segmentation-map based loss function ($\mathcal{L}_{seg}$). Both $\mathcal{L}_{seg}$ and $\mathcal{L}_{joint}$ are defined as:

$$\mathcal{L}_{seg}(\mathbf{S},\mathbf{S}') = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} S_c^k \log_2(S_c^{'k}), \tag{8}$$

$$\begin{aligned}\mathcal{L}_{joint}(\mathbf{Y},\mathbf{Y}',\mathbf{S},\mathbf{S}') &= \alpha \times \mathcal{L}_{cls}(\mathbf{Y},\mathbf{Y}')\\ &\quad + (1-\alpha) \times \mathcal{L}_{seg}(\mathbf{S},\mathbf{S}'), \end{aligned} \tag{9}$$

where α is a hyper-parameter which defines the contribution of both loss functions in the final loss. Similar to patch classifier, the loss function ($\mathcal{L}_{joint}$) is minimized with RMSprop optimizer [42].

H. Training Strategies

We trained the proposed framework in four different ways for the sake of completeness in experimentation. First, the proposed framework is trained without attention block and by minimizing the $\mathcal{L}_{cls}(\cdot)$ loss only. This configuration is represented by solid line blocks in Fig 2. Second, the same configuration as first but trained with a sample-based weighted loss function, $\mathcal{L}_{wgt}(\cdot)$, which give more weight to the image patches with relatively less region of interest (glandular region) as compared to the background. The weight of an image and $\mathcal{L}_{wgt}(\cdot)$ are defined as follow,

$$W^k = \begin{cases} \frac{1}{R_{roi}^k}, & \text{if } R_{roi}^k > \alpha \\ \frac{1}{\alpha}, & \text{otherwise} \end{cases} \tag{10}$$

$$\mathcal{L}_{wgt}(\mathbf{Y},\mathbf{Y}') = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} W^k Y_c^k \log_2(Y_c^{'k}), \tag{11}$$

where R_{roi}^k and W^k represent the ratio of the region of interest and the weight of the k^{th} image. The α is the ratio threshold, selected empirically as 0.10, sets the upper limit of an image weight. Third, multi-task learning based training with the help of an auxiliary block by using joint classification and segmentation loss, $\mathcal{L}_{joint}$. Last, training using the same joint loss but with attention-based feature-cube to amplify the contribution of more important features in the feature-cube. The network configuration of this strategy is represented by both solid and dotted lines blocks in Fig 2. We termed these strategies as *standard*, *weighted*, *auxiliary*, and *attention*, respectively.

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DISTRIBUTION OF VISUAL FIELDS OF DIFFERENT CLASSES FOR BOTH DATASET

Dataset	Normal	Low Grade	High Grade	Total
CRC [11]	71	33	35	139
Extended CRC	120	120	60	300

IV. DATASETS & PERFORMANCE MEASURES

In this section, we explain the dataset details used for training and evaluation of the proposed framework and metrics for performance evaluation.

A. Datasets

The proposed framework is evaluated on two colorectal datasets[1] in order to demonstrate its capabilities. The first colorectal cancer (CRC) dataset was used by Awan *et al.* [11] for exactly the same task of colorectal grading. It is comprised of visual fields extracted from 38 Hematoxylin and Eosin (H&E) stained Whole Slide Images (WSIs) of colorectal cancer cases based on a two-tier grading system [13], [14]. The dataset consists of 139 visual fields with an average size of 4548 × 7520 pixels obtained at 20× magnification. These visual fields are classified into three different classes (normal, low grade, and high grade) based on the organization of glands in the visual fields by the expert pathologist. We extend this dataset with more visual fields extracted from another 68 H&E stained WSIs using the same criteria. Our extended colorectal cancer (Extended CRC) dataset consists of 300 visual fields with an average size of 5000 × 7300 pixels. A detailed distribution of the visual fields of different grades is presented in Table II for both datasets. We follow 3-fold cross validation for a fair comparison of the proposed method with the method presented in [11]. All visual fields extracted from one case only lies in one fold and we use one fold for training, one for validation (hyper-parameter tuning) and one for the testing to do strong cross-validation on extended CRC dataset. Patches of two different sizes 224 × 224 and 1792 × 1792 pixels are extracted for the training of traditional patch classifiers and our proposed framework, respectively. A background class is introduced to handle the patches with no or little glandular regions. For each class in each fold, we extracted 30 000 patches for patch classification and 6000 overlapping patches for context-aware classification using random rotation and flipping based augmentation for both datasets.

B. Performance Measures

We have used two metrics, the accuracy and Rank-sum measure, for performance evaluation. The average accuracy refers to the percentage of visual fields classified correctly whereas weighted accuracy is the sum of accuracy of each class weighted by the number of samples in that class. Rank-sum based evaluation metric is used to summarize the accuracy of different models trained using a specific setting in order to compare models trained with different context-blocks and LR-CNNs. Different colors are used to represent different rank for better illustrative visualization as shown in Table IV and V. The orange color indicates the best performing method whereas green and blue colours indicate that the results are within 97.5% and 95% of the best performing method, respectively. The rank for these colors are: orange = 1, green = 2, blue = 3, and no colour = 4. The lowest rank-sum represents the best performance.

TABLE III
ACCURACY COMPARISON OF FOUR PATCH CLASSIFIERS

Network	Fold-1	Fold-2	Fold-3	Mean	SD
ResNet50	93.48	93.62	89.13	92.08	2.08
MobileNet	93.48	**95.74**	89.13	**92.78**	2.74
Inception-v3	**95.65**	91.49	86.96	91.37	3.55
Xception	93.48	91.49	**91.30**	92.09	**0.98**

V. EXPERIMENTAL RESULTS

The results of the different variants of the proposed framework are presented to show the superior performance of all the variants over simple patch based methods. These variations include the use of four different state-of-the-art classifiers for local representation learning in LR-CNN; spatial dimensionality reduction through average and max-pooling; the usage of three different context-blocks in RA-CNN; and four different training strategies. By employing different combinations of above-mentioned variations, we trained around 100 models in total for each fold on the CRC dataset. The details of experimental evaluation are given in following subsections.

A. Experimental Setup

The CRC visual fields are divided into patches of size 1792 × 1792, and the label of each patch is predicted using the proposed framework with a stride of 224 × 224. To avoid redundant processing of the same region, the visual fields are processed with LR-CNN to get representation features of each local region. Afterwards, RA-CNN is applied in a sliding window manner to aggregate local representation for context-aware predictions. Through this approach, we process a visual field with a 64 times bigger context as compared to standard patch classifier with only 10% additional processing time. The overall label of a visual field is derived from counting the most predicted class (majority voting), excluding background class in the visual field. Note that, all the reported results are calculated on visual fields.

B. LR-CNN Based Classifiers

Four different LR-CNNs are trained using ResNet50 [33], Inception [35], MobileNet [34], and Xception [36] with patch size of 224 × 224 to get the baseline patch based classification results. The ResNet-50 [33] and Inception network are the winner of Image-Net [43] challenge in 2015 and 2016, respectively. MobileNet is a lightweight network with just 3 million parameters whereas Xception network uses separable convolutions which results in a significant reduction in computational complexity. The performance of these classifiers for CRC grading is reported in Table III. Although the performance of all classifiers is comparable, MobileNet shows superior performance with the highest mean accuracy. On the

[1] Both CRC and Extended CRC datasets are publicly available at this page: https://warwick.ac.uk/fac/sci/dcs/research/tia/data/

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
RANK-SUM BASED COMPARISON OF THREE DIFFERENT CONTEXT-AWARE NETWORKS WITH STANDARD PATCH CLASSIFIERS. THE ORANGE, GREEN, AND BLUE REPRESENTS THE RANK 1, 2 AND 3, RESPECTIVELY

LR-CNN (Avg)	Baseline	RA-CNN 1	RA-CNN 2	RA-CNN 3
ResNet50	92.08±2.08	94.25±2.70	92.08±2.08	93.51±3.10
MobileNet	92.78±2.74	93.52±3.55	93.52±1.78	94.25±2.70
InceptionV3	91.37±3.55	94.23±3.71	94.96±2.72	95.68±1.78
Xception	92.09±0.98	94.96±2.72	94.96±2.72	95.68±3.55
Rank-sum	10	7	8	**5**

TABLE V
ROBUSTNESS ANALYSIS OF FEATURE EXTRACTORS ACROSS DIFFERENT METHODS. THE ORANGE, GREEN, AND BLUE REPRESENTS THE RANK 1, 2 AND 3, RESPECTIVELY

Methods	ResNet50 (%)	MobileNet(%)	InceptionV3(%)	Xception(%)
RA-CNN 1 (Avg)	94.25±2.70	93.52±3.55	94.23±3.71	94.96±2.72
RA-CNN 1 (Max)	93.52±1.87	93.51±3.10	94.23±2.07	93.54±3.03
RA-CNN 2 (Avg)	92.08±2.08	93.52±1.78	94.96±2.72	94.96±2.72
RA-CNN 2 (Max)	95.68±3.55	93.52±3.55	92.80±2.72	93.54±3.03
RA-CNN 3 (Avg)	93.51±3.10	94.25±2.70	95.68±1.78	95.68±3.55
RA-CNN 3 (Max)	94.23±2.07	92.82±2.01	94.25±2.70	94.96±2.72
Rank-sum	12	12	10	**8**

other hand, Xception classifier shows consistent performance across three folds with the lowest standard deviation (SD).

C. RA-CNN Based Context-Aware Learning

We experimented with three context-blocks, $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$, to train three different variations of RA-CNN, which we termed as RA-CNN 1, RA-CNN 2, and RA-CNN 3. These three RA-CNN classifiers are trained separately with all four LR-CNNs as explained in section III-F, hence giving 12 different combinations of the context-aware network. The rank-sum method is used to compare the performance of these networks with each other and also with the LR-CNNs. The results in table IV shows that context-aware networks achieve superior performance as compare to standard patch based classifiers (LR-CNNs). The RA-CNN 3 achieves the best Rank-sum (lowest) which shows its robustness across different representation learning networks. The other two context-aware networks also show comparable performance by remaining in the 97.5% of the best performer.

D. Local Representation Robustness

We also conducted different experiments to analyze the robustness of local representation learned by different LR-CNNs. These LR-CNNs are used in combination with different RA-CNNs for context learning along with different feature pooling strategies. Each LR-CNN is used to training three RA-CNNs with both global average and global max pooled feature-cubes. The table V compares the results using Rank-sum based measure. It can be observed that the Xception model turns-out as the most robust feature extractor in LR-CNNs with the best rank-sum score of 8. The Inception model shows comparable results to the best performer as its network design has significant overlap with Xception architecture.

E. Training Strategies

We experimented with four different context related training strategies (*Standard*, *Weighted*, *Auxiliary* and *Attention*) to explored their impact on overall performance. The details of each training strategy are given in Section III-H. Table VI shows the comparison of these training strategies for Xception based LR-CNN. Each entry in the table contains the average accuracy across three RA-CNNs for particular feature pooling (shown in rows) and the training strategies (in columns). Attention-based training shows the superior results for max-pooled features whereas standard training strategy achieves comparable performance for average-pooled features. However, auxiliary loss based training remains robust for both pooling types and achieves the best overall accuracy. More importantly, each model shows superior performance than the baseline LR-CNN classifier as shown in Fig. 3. The graphical illustration of 24 experiments using the best performing LR-CNN is shown in Fig. 3. The results obtained with different combinations of feature pooling type, the context blocks in RA-CNN and the training strategies used for the experiments are illustrated in the bar-chart format for better visual comparison. The accuracy obtained by Xception based LR-CNN is considered as the baseline for comparative analysis. Bar-charts for results with other LR-CNNs are given in supplementary material.

TABLE VI
COMPARISON FOR DIFFERENT TRAINING STRATEGIES BASED ON AVERAGE ACCURACY ACROSS THREE RA-CNNS WITH XCEPTION BASED FEATURES

Feature (Pooling)	Standard	Weighted	Auxiliary	Attention
Xception (Max)	94.01	94.49	94.73	**95.21**
Xception (Avg)	**95.20**	94.72	94.72	94.00
Mean	94.61	94.60	**94.72**	94.61

F. Result Summary

The gist of detailed experimentation and comparisons is that larger contextual information helps in better automated grading of colorectal cancer and the proposed approach demonstrated the ability to capture larger context. In practice, Xception based LR-CNN is the most robust feature extractor for context-aware learning and RA-CNN 3 showed robustness to most of the feature extraction methods. Attention based training strategy is suitable for both RA-CNN 1 and RA-CNN 3 with max pooling features. Last but not least, all proposed variations of context-aware framework perform better than the baseline Xception based classifier.

VI. COMPARATIVE RESULTS

The results of the best performing context-aware method are compared with state-of-the-art approaches on both datasets. These approaches are categorized into domain oriented methods, traditional patch based classifiers, and context-aware methods. The brief description of these approaches and comparative analysis is presented in the following subsections.

A. Domain Oriented Methods

Awan *et al.* [11] presented a two-step problem specific method for CRC grading as explained in Section II-B. They experimented with two different feature sets which we refer to as BAM-1 and BAM-2 in this paper. BAM-1 comprises

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

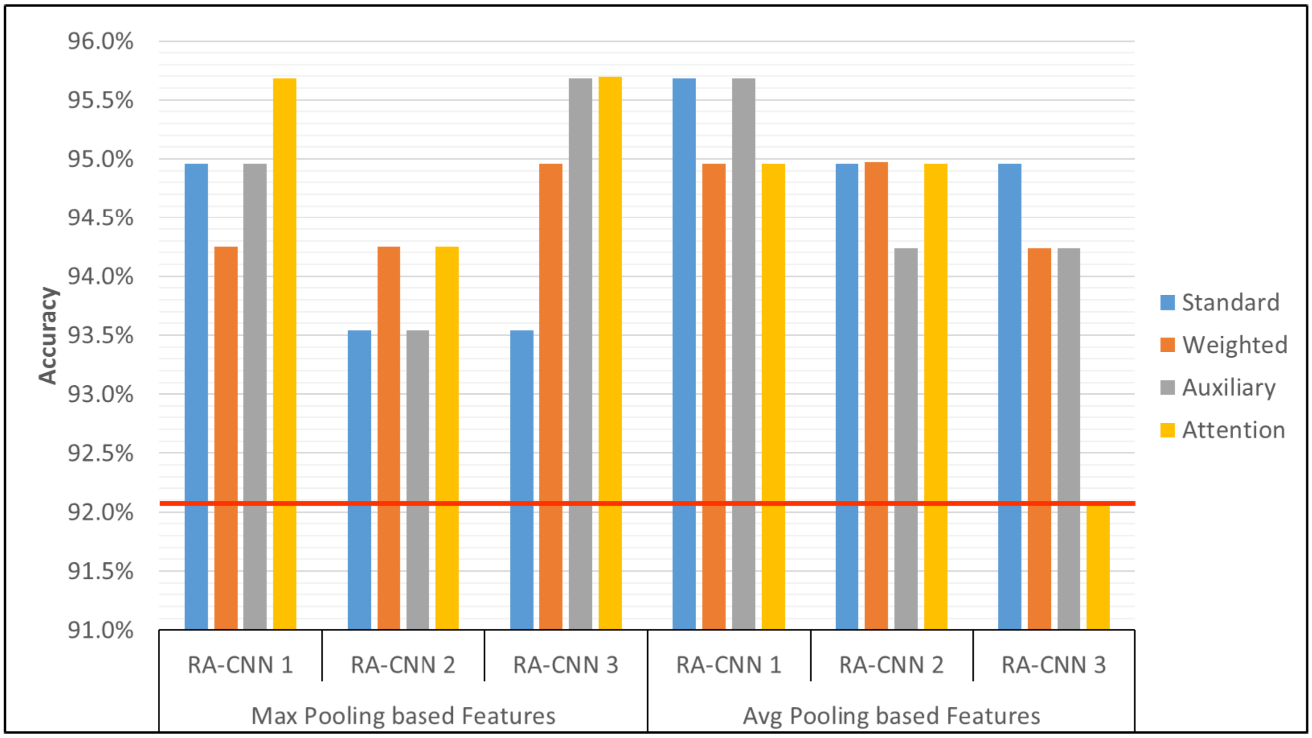


Fig. 3. Results of 24 experiments using best performing local representation features (Xception). Legend represents the different training strategies whereas different bars represents the results for three context-aware networks with max and average pooling based features. Red line indicates the baseline accuracy of patch based Xception classifier.

TABLE VII

AVERAGE ACCURACY BASED GRADING COMPARISON OF PROPOSED CONTEXT-AWARE METHOD WITH STATE-OF-THE-ART METHODS ON CRC DATASET

ID	Methods	Binary (%)	Three-class (%)
1	BAM - 1 [11]	95.70±2.10	87.79±2.32
2	BAM - 2 [11]	97.12±1.27	90.66±2.45
3	ResNet50 [33]	98.57±1.01	92.08±2.08
4	MobileNet [34]	97.83±1.77	92.78±2.74
5	InceptionV3 [35]	98.57±1.01	91.37±3.55
6	Xception [36]	98.58±2.01	92.09±0.98
7	CNN-SVM [44]	96.44±3.61	92.12±3.57
8	CNN-LR [44]	98.58±2.01	93.52±0.07
9	CNN-LSTM [29]	96.44±3.61	89.96±3.54
10	Proposed	**99.28±1.25**	**95.70±3.04**

of average BAM and BAM entropy while BAM-2 comprises of an additional feature known as regularity index. They evaluated their method using only average accuracy based measures for both binary and 3-class grading. We reported the results presented by the author in their paper [11] on CRC dataset to avoid any retraining bias and compared using average accuracy based measure for a fair comparison. Their method achieved good accuracy for binary grading, normal vs cancer, however, it lakes the robustness required for multi-class grading of CRC visual fields whereas the proposed method achieved superior performance on both tasks (see Table VII).

B. Patch-Based Classifiers

The results for four standard patch classifiers on both datasets are presented in Table VII and VIII. There is a slight difference in the ranking of these classifiers on both datasets. However, Xception classifier remains consistent in terms of low variance in performance on both datasets. We further experimented with different patch sizes using Xception classifier on the Extended CRC dataset. The results show that the significant change in the patch size without any modification in the network architecture leads to a decrease in the performance as can be seen in Table VIII for Xception network. The performance of all the patch based classifiers is below the performance of proposed method.

C. Context-Aware Methods

The decision fusion based methods [44], [45] can be loosely considered as context-aware methods if used to predict the visual field labels through the aggregation of patch predictions. We compared our method with the two approaches used by Hou *et al.* [44] on the Extended CRC dataset. They used Support Vector Machine (SVM) with RBF kernel (CNN-SVM) and Logistic regression (CNN-LR) for decision fusion from the class histogram of patch probabilities. We used the best performing patch classifiers for each dataset (MobileNet for CRC, Xception for Extended CRC) to get the patch probabilities. The CNN-LR shows some performance improvements over the best performing patch classifiers but this performance is still below the proposed method on both datasets (see Table VII and VIII). The CNN-SVM method does not perform as good as the simple majority voting based patch classifier. A similar performance pattern can be observed in the Hoe *et al.* paper [44] for the task of Glioma classification. We believe that the major difference between these simple decision fusion and context-aware methods is the ability to adjust the prediction of a patch using

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
ACCURACY BASED GRADING COMPARISON OF PROPOSED CONTEXT-AWARE METHOD WITH STATE-OF-THE-ART METHODS ON THE EXTENDED CRC DATASET

ID	Methods	Patch Size	Binary Classification		3-Class Classification	
			Average (%)	Weighted (%)	Average (%)	Weighted (%)
1	ResNet50 [33]	224x224	95.67±2.05	95.69±1.53	86.33±0.94	80.56±1.04
2	MobileNet [34]	224x224	95.33±2.49	95.42±2.23	84.33±3.30	77.78±4.83
3	InceptionV3 [35]	224x224	93.67±1.89	94.31±1.57	84.67±1.70	81.11±1.97
4	Xception [36]	224x224	96.67±2.05	96.80±1.71	86.33±0.94	81.39±1.71
5	Xception [36]	112x112	92.00±3.27	92.22±2.64	81.33±3.40	74.72±4.53
6	Xception [36]	448x448	97.00±2.83	97.08±2.36	**86.67±0.94**	80.42±1.25
7	CNN-SVM [44]	224x224	96.00±0.82	96.39±0.86	82.00±1.63	76.67±2.97
8	CNN-LR [44]	224x224	96.33±1.70	96.39±1.37	**86.67±1.25**	82.50±0.68
9	CNN-LSTM [29]	1792x1792	95.33±2.87	94.17±3.58	82.33±2.62	83.89±2.08
10	Proposed	1792x1792	**97.67±0.94**	**97.64±0.79**	**86.67±1.70**	**84.17±2.36**

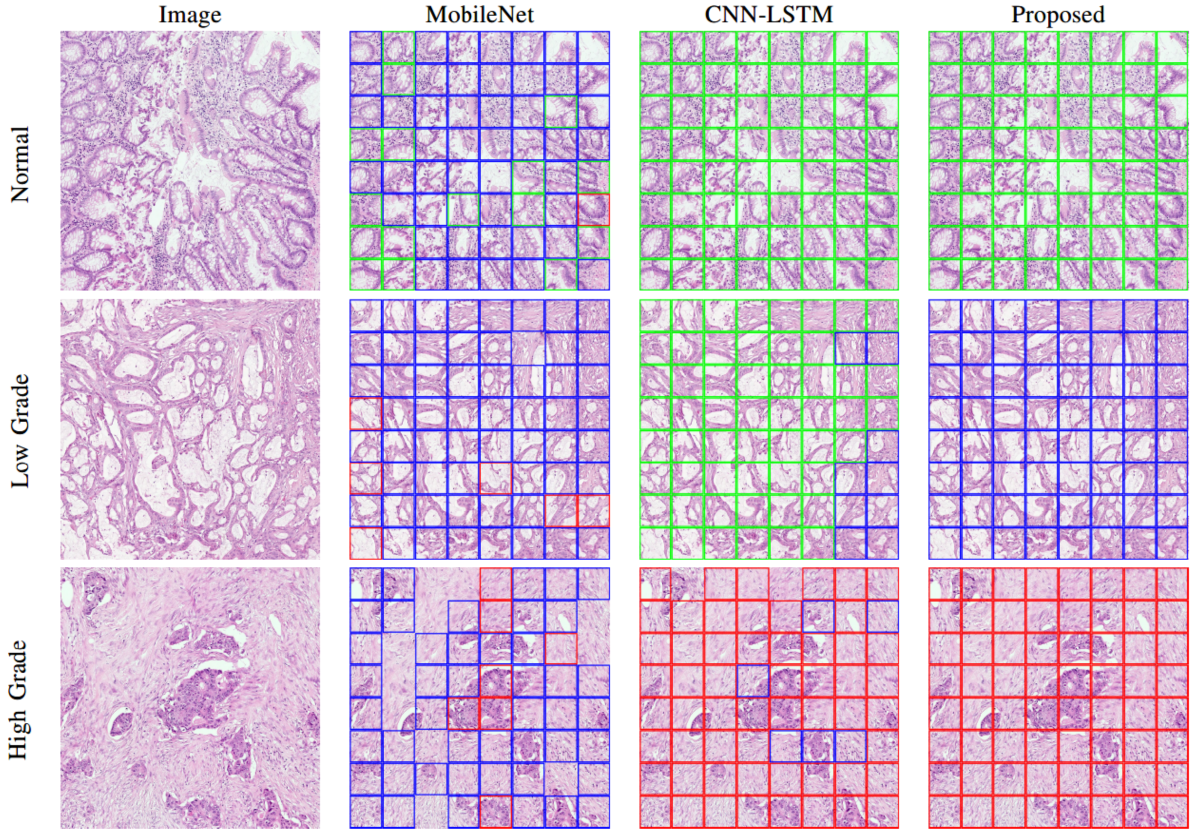


Fig. 4. Visual results on CRC grading dataset are shown for patch classifier, existing context, and the proposed method on an image of size 1792×1792. The stride size for context networks is equal to the size of patch (224×224) used for patch classifier. Green, blue and red colors of overlaid rectangular boxes show the normal, low and high-grade predictions respectively, whereas empty box areas represent non-glandular/background regions.

its neighbourhood information. The decision fusion based methods only use predicted patch probabilities whereas as context-aware methods have access to the high dimensional features of neighbouring patches.

We also compared our method with an LSTM based context-aware method (CNN-LSTM) proposed in a systemic study on context-aware learning by Sirinukunwattana *et al.* [29] using prostate and breast cancer datasets. They used LSTM to capture the context from CNN features of four downsampled versions (1×, 2×, 4×, and 8×) of the input patch. The code is publicly available by the authors of the paper [29] and we use that code to retrain the method on both datasets for a fair comparison. Our best performing context-aware method outperformed the CNN-LSTM method on both datasets (see Table VII and VIII). This performance improvement could be attributed to the proposed method's ability to use high resolution input patch without any downsampling for context learning, unlike CNN-LSTM. Moreover, we used a relatively more powerful CNN network (e.g Xception) for LR-CNN for feature extraction whereas Sirinukunwattana *et al.* opted for a light-weight network for feature extraction to make their network end-to-end.

D. The Proposed Method

The different variants of the proposed method have shown comparable performance but we consider our best performing context-aware configuration for comparative analysis. The best

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

performance is achieved by RA-CNN 1 trained with attention based training strategy on max pooled features. It shows 3.61% and 2.78% better performance gain over simple patch classifiers on both CRC (Table VII) and Extended CRC (Table VIII) datasets, respectively. We also investigated the performance based on the coarse patch-based segmentation using RA-CNN 1 trained with auxiliary training strategy on the Extended CRC dataset. Although, it achieves the weighted accuracy of 87.50%, it has a high variance of 5.14% across three folds of the Extended CRC dataset. Therefore, we did not consider it as our benchmark for comparative analysis in Table VIII.

E. Visual Comparison

The visual comparison of best performing patch classifier, Sirinukunwattana *et al.* (CNN-LSTM) and the proposed method on three different images with normal, low and high grades are shown in Figure 4. Patch classifier's prediction is quite irregular for any given image due to the lake of contextual information. The predictions of CNN-LSTM are relatively smooth but it predicts the wrong label for the low-grade image which might be due to the use of a low-resolution images for context learning. However, the proposed method predictions are smooth and consistent with the ground truth labels.

VII. Conclusion

In this paper, we present a novel context-aware deep neural network for cancer grading, which is able to incorporate 64 times larger context than standard CNN based patch classifiers. The proposed network is well-suited for the CRC grading task which relies on recognizing abnormalities in glandular structures. These clinically significant structures vary in size and shape that cannot be captured efficiently with standard patch classifiers due to computational and memory constraints. The proposed context-aware network is comprised of two stacked CNNs. The first LR-CNN is used for learning the local representation of the histology image. The learned local representation is then aggregated considering its spatial pattern by RA-CNN. The proposed context-aware model is evaluated on two datasets for CRC grading. A comprehensive analysis of different variations of the proposed model is presented and compared with existing approaches in the same evaluation setting. The qualitative and quantitative results demonstrate that our method outperformed the patch based classification methodologies, the domain-oriented techniques, and existing context-based methods. This approached is suitable for cancer analysis which requires large contextual information in the histology images. This includes Gleason grading in prostate cancer and tumor growth pattern classification in lung cancer. Moreover, this approach can further be extended to perform downstream analysis at the digital whole slide image level for patient survival analysis.

References

[1] L. Pantanowitz, N. Farahani, and A. Parwani, "Whole slide imaging in pathology: Advantages, limitations, and emerging perspectives," *Pathol. Lab Med. Int.*, vol. 7, p. 23, Jun. 2015.

[2] N. A. Koohababni, M. Jahanifar, A. Gooya, and N. Rajpoot, "Nuclei detection using mixture density networks," in *Machine Learning in Medical Imaging* (Lecture Notes in Computer Science), vol. 11046, Y. Shi, H. I. Suk, and M. Liu, Eds. Cham, Switzerland: Springer, 2018, pp. 241–248, doi: 10.1007/978-3-030-00919-9_28.

[3] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[4] T.-H. Song, V. Sanchez, H. Ei Daly, and N. M. Rajpoot, "Simultaneous cell detection and classification in bone marrow histology images," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1469–1476, Jul. 2019.

[5] B. E. Bejnordi *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[6] N. A. Koohbanani, T. Qaisar, M. Shaban, J. Gamper, and N. Rajpoot, "Significance of hyperparameter optimization for metastasis detection in breast histology images," in *Computational Pathology and Ophthalmic Medical Image Analysis* (Lecture Notes in Computer Science), vol. 11039, D. Stoyanov *et al.*, Eds. Cham, Switzerland: Springer, 2018, pp. 139–147, doi: 10.1007/978-3-030-00949-6_17.

[7] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P.-A. Heng, "Fast ScanNet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1948–1958, Aug. 2019.

[8] T. Qaiser, K. Sirinukunwattana, K. Nakane, Y.-W. Tsang, D. Epstein, and N. Rajpoot, "Persistent homology for fast tumor segmentation in whole slide histology images," *Procedia Comput. Sci.*, vol. 90, pp. 119–124, 2016.

[9] E. Arvaniti *et al.*, "Automated Gleason grading of prostate cancer tissue microarrays via deep learning," *Sci. Rep.*, vol. 8, Aug. 2018, Art. no. 12054.

[10] N. Ing *et al.*, "Semantic segmentation for prostate cancer grading by convolutional neural networks," *Med. Imag., Digit. Pathol.*, vol. 10581, Mar. 2018, Art. no. 105811B.

[11] R. Awan *et al.*, "Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 16852.

[12] B. W. Stewart and C. P. Wild, "World cancer report 2014," Int. Agency Res. Cancer, Lyon, France, Jan. 2014, vol. 3, no. 1, pp. 392–402.

[13] W. Blenkinsopp, S. Stewart-Brown, L. Blesovsky, G. Kearney, and L. Fielding, "Histopathology reporting in large bowel cancer," *J. Clin. Pathol.*, vol. 34, no. 5, pp. 509–513, 1981.

[14] J. Jass *et al.*, "The grading of rectal cancer: Historical perspectives and a multivariate analysis of 447 cases," *Histopathology*, vol. 10, no. 5, pp. 437–459, May 1986.

[15] S. S. Chennamsetty, M. Safwan, and V. Alex, "Classification of breast cancer histology image using ensemble of pre-trained neural networks," in *Image Analysis and Recognition* (Lecture Notes in Computer Science), vol. 10882, A. Campilho, F. Karray, and B. ter Haar Romeny, Eds. Cham, Switzerland: Springer, 2018, pp. 804–811, doi: 10.1007/978-3-319-93000-8_91.

[16] M. Kohl, C. Walz, F. Ludwig, S. Braunewell, and M. Baust, "Assessment of breast cancer histology using densely connected convolutional networks," in *Image Analysis and Recognition* (Lecture Notes in Computer Science), vol. 10882, A. Campilho, F. Karray, and B. ter Haar Romeny, Eds. Cham, Switzerland: Springer, 2018, pp. 903–913, doi: 10.1007/978-3-319-93000-8_103.

[17] I. Koné and L. Boulmane, "Hierarchical resnext models for breast cancer histology image classification," in *Image Analysis and Recognition* (Lecture Notes in Computer Science), vol. 10882, A. Campilho, F. Karray, and B. ter Haar Romeny, Eds. Cham, Switzerland: Springer, 2018, pp. 796–803, doi: 10.1007/978-3-319-93000-8_90.

[18] X. Wang *et al.*, "Weakly supervised learning for whole slide lung cancer image classification," *Med. Imag. Deep Learn.*, vol. 1, pp. 1–10, Apr. 2018.

[19] A. Cruz-Roa *et al.*, "High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196828.

[20] N. Alsubaie, M. Shaban, D. Snead, A. Khurram, and N. Rajpoot, "A multi-resolution deep learning framework for lung adenocarcinoma growth pattern classification," in *Medical Image Understanding and Analysis.* (Communications in Computer and Information Science), vol. 894, M. Nixon, S. Mahmoodi, and R. Zwiggelaar, Eds. Cham, Switzerland: Springer, 2018, pp. 3–11, doi: 10.1007/978-3-319-95921-4_1.

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.

[21] A. Bentaieb, H. Li-Chang, D. Huntsman, and G. Hamarneh, "A structured latent model for ovarian carcinoma subtyping from histopathology slides," *Med. Image Anal.*, vol. 39, pp. 194–205, Jul. 2017.

[22] Y. Liu *et al.*, "Detecting cancer metastases on gigapixel pathology images," Mar. 2017, *arXiv:1703.02442*. [Online]. Available: https://arxiv.org/abs/1703.02442

[23] A. Agarwalla, M. Shaban, and N. M. Rajpoot, "Representation-aggregation networks for segmentation of multi-gigapixel histology images," Jul. 2017, *arXiv:1707.08814*. [Online]. Available: https://arxiv.org/abs/1707.08814

[24] B. Kong, X. Wang, Z. Li, Q. Song, and S. Zhang, "Cancer metastasis detection via spatially structured deep network," in *Information Processing in Medical Imaging* (Lecture Notes in Computer Science), vol. 10265, M. Niethammer *et al.*, Eds. Cham, Switzerland: Springer, 2017, pp. 236–248, doi: 10.1007/978-3-319-59050-9_19.

[25] F. G. Zanjani *et al.*, "Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces," *Med. Imag., Digit. Pathol.*, vol. 10581, Mar. 2018, Art. no. 105810I.

[26] Y. Li and W. Ping, "Cancer metastasis detection with neural conditional random field," Jun. 2018, *arXiv:1806.07064*. [Online]. Available: https://arxiv.org/abs/1806.07064

[27] R. Awan, N. A. Koohbanani, M. Shaban, A. Lisowska, and N. Rajpoot, "Context-aware learning using transferable features for classification of breast cancer histology images," Feb. 2018, *arXiv:1803.00386*. [Online]. Available: https://arxiv.org/abs/1803.00386

[28] B. E. Bejnordi *et al.*, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *J. Med. Imag*, vol. 4, no. 4, p. 1, Dec. 2017.

[29] K. Sirinukunwattana, N. K. Alham, C. Verrill, and J. Rittscher, "Improving whole slide segmentation through visual context—A systematic study," Jun. 2018, *arXiv:1806.04259*. [Online]. Available: https://arxiv.org/abs/1806.04259

[30] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Proc. 5th IEEE Int. Symp. Biomed. Imag., Nano Macro*, May 2008, pp. 284–287.

[31] R. Farjam, H. Soltanian-Zadeh, K. Jafari-Khouzani, and R. A. Zoroofi, "An image analysis approach for automatic malignancy determination of prostate pathological images," *Cytometry*, vol. 72B, no. 4, pp. 227–240, Jul. 2007.

[32] K. Nguyen, B. Sabata, and A. K. Jain, "Prostate cancer grading: Gland segmentation and structural features," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 951–961, May 2012.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[35] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Apr. 2017, *arXiv:1610.02357*. [Online]. Available: https://arxiv.org/abs/1610.02357

[37] Z. Zhang, Y. Xie, F. Xing, M. Mcgough, and L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6428–6436.

[38] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.* New York, NY, USA: ACM, 2008, pp. 160–167.

[39] X. Chu, W. Ouyang, W. Yang, and X. Wang, "Multi-task recurrent neural network for immediacy prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3352–3360.

[40] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," Jun. 2018, *arXiv:1806.08730*. [Online]. Available: https://arxiv.org/abs/1806.08730

[41] A. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3712–3722.

[42] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6A overview of mini-batch gradient descent," Coursera, Mountain View, CA, USA, Tech. Rep. 8, 2012, vol. 14.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[44] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2424–2433.

[45] M. M. Kokar, J. A. Tomasik, and J. Weyman, "Data vs. decision fusion in the category theory framework," in *Proc. FUSION*, 2001, pp. 1–6.

Authorized licensed use limited to: RMIT University Library. Downloaded on May 28,2021 at 12:49:48 UTC from IEEE Xplore. Restrictions apply.