Journal Pre-proof

Cellular Community Detection For Tissue Phenotyping In Colorectal Cancer Histology Images

Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, Nasir Rajpoot

PII: S1361-8415(20)30061-X
DOI: https://doi.org/10.1016/j.media.2020.101696
Reference: MEDIMA 101696

To appear in: *Medical Image Analysis*

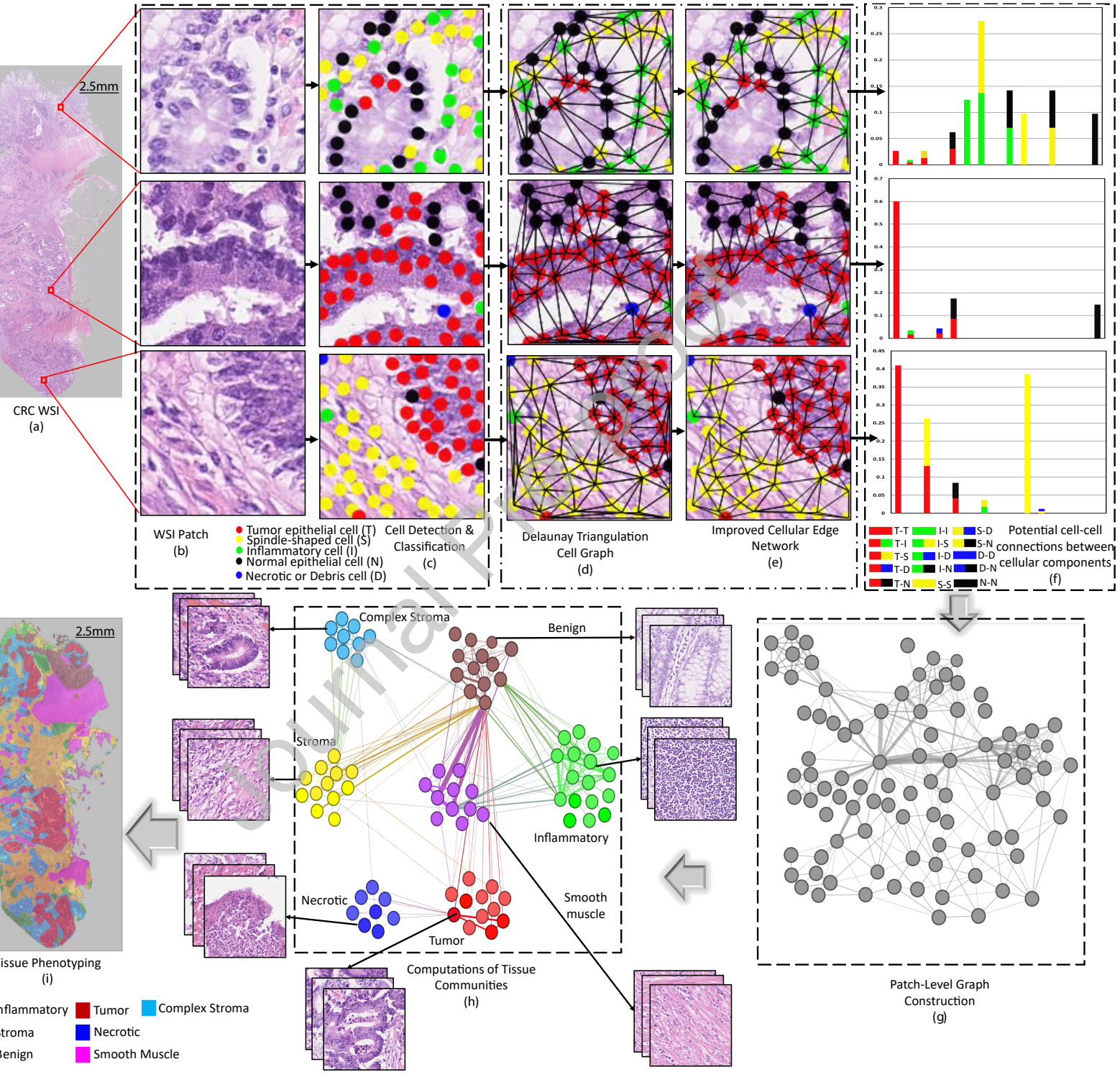
Received date: 12 September 2019
Revised date: 18 February 2020
Accepted date: 2 April 2020

Please cite this article as: Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, Nasir Rajpoot, Cellular Community Detection For Tissue Phenotyping In Colorectal Cancer Histology Images, *Medical Image Analysis* (2020), doi: https://doi.org/10.1016/j.media.2020.101696

Highlights

- We pose the problem of identifying tissue phenotypes as a community detection problem in histological landscape where each community represents a distinct tissue phenotype, for example tumor, benign, stroma, inflammatory, complex stroma, and smooth muscle. To the best of our knowledge, the formulation of tissue phenotyping as community detection has not been done before.
- We propose geodesic density gradients for tissue phenotyping, a novel way of phenotyping tissue segments in large multi-gigapixel WSIs of histology slides. We show that it results in significant performance improvement.
- Instead of using texture features to represent a patch of WSI, we employ potential interactions between various types of cells as representative features. These features are biologically more meaningful and better capture the distribution of different types of cells in the histology patch.
- We propose a new large-scale dataset for tissue phenotyping. It consists of 280K patches extracted from 20 WSIs of CRC slides stained with H&E. Each WSI contains exhaustive region level annotations of seven distinct tissue phenotypes labelled by experienced pathologists. This dataset (CRC-TP) will be released with the publication of this manuscript.
- A dataset for Cell Classification (CC) has been extended to include five distinct cell types: tumor epithelial, normal epithelial, necrotic, spindle-shaped, and inflammatory cells. The extended dataset (CRC-CC) will also be made publicly available with the publication of this manuscript.

Graphical Abstract



Medical Image Analysis (2020)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Cellular Community Detection For Tissue Phenotyping In Colorectal Cancer Histology Images

Sajid Javed[a], Arif Mahmood[b], Muhammad Moazam Fraz[a], Navid Alemi Koohbanani[a], Ksenija Benes[c], Yee-Wah Tsang[c], Katherine Hewitt[c], David Epstein[d], David Snead[c], Nasir Rajpoot[a,*]

[a]*Department of Computer Science, University of Warwick, Coventry, CV4 7AL, U.K.*
[b]*Department of Computer Science, Information Technology University, Lahore, Pakistan.*
[c]*Department of Pathology, University Hospitals Coventry and Warwickshire, Walsgrave, Coventry, CV2 2DX, U.K.*
[d]*Mathematics Institute, University of Warwick, Coventry, CV4 7AL, U.K.*

ARTICLE INFO

Article history:
Received 1 May 2013
Received in final form 10 May 2013
Accepted 13 May 2013
Available online 15 May 2013

Communicated by S. Sarkar

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Computational Pathology, Tissue Phenotyping, Tumor Microenvironment, Cellular Communities.,

ABSTRACT

Classification of various types of tissue in cancer histology images based on the cellular compositions is an important step towards the development of computational pathology tools for systematic digital profiling of the spatial tumor microenvironment. Most existing methods for tissue phenotyping are limited to the classification of tumor and stroma and require large amount of annotated histology images which are often not available. In the current work, we pose the problem of identifying distinct tissue phenotypes as finding communities in cellular graphs or networks. First, we train a deep neural network for cell detection and classification into five distinct cellular components. Considering the detected nuclei as nodes, potential cell-cell connections are assigned using Delaunay triangulation resulting in a cell-level graph. Based on this cell graph, a feature vector capturing potential cell-cell connection of different types of cells is computed. These feature vectors are used to construct a patch-level graph based on chi-square distance. We map patch-level nodes to the geometric space by representing each node as a vector of geodesic distances from other nodes in the network and iteratively drifting the patch nodes in the direction of positive density gradients towards maximum density regions. The proposed algorithm is evaluated on a publicly available dataset and another new large-scale dataset consisting of 280K patches of seven tissue phenotypes. The estimated communities have significant biological meanings as verified by the expert pathologists. A comparison with current state-of-the-art methods reveals significant performance improvement in tissue phenotyping.

1. Introduction

Tumor microenvironment (TME) plays a crucial role in the development of intra-tumor heterogeneity (ITH) (Marusyk et al. (2012)). It is, therefore, vital that we develop ways to systematically profile spatial characteristics of the TME in order to better understand tumor heterogeneity and consequently exploit it for therapeutic gain (Alizadeh et al. (2015)). Computational pathology is a rapidly emerging discipline (van der Laak et al. (2018)), spurred by the recent revolution in digital pathology (DP) imaging which has been shown to be non-inferior to glass slide based visual assessment by pathologists for routine diagnostic purposes (Snead et al. (2016)), concerned with the development of computational algorithms for the processing and analysis of DP images. Automatic tissue phenotyping, identifi-

*Corresponding author: Tel.: +44-24-7657-3795; fax: +0-000-000-0000;
e-mail: n.m.rajpoot@warwick.ac.uk (Nasir Rajpoot)

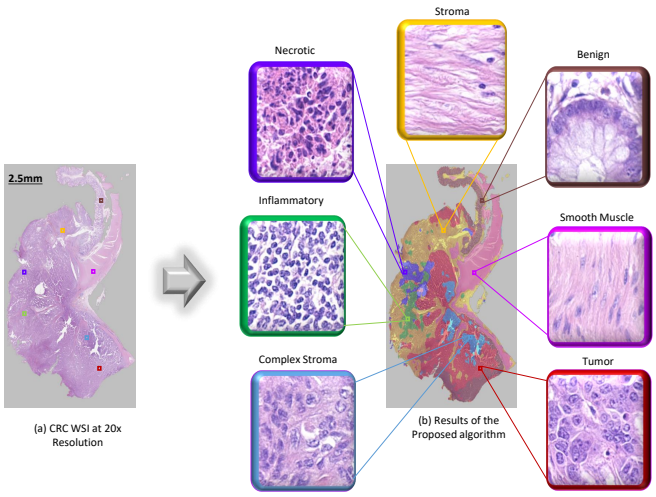


Fig. 1: A sample whole slide image of colorectal cancer (CRC) from CRC-TP dataset. Different tissue components including tumor, stroma, smooth muscle, necrotic, inflammatory, complex stroma, and benign tissue images are shown in different colors.

cation, and localization of a diverse range of tissue types, in digitized whole-slide images (WSIs) of tissue slides stained with routine Hematoxylin & Eosin (H&E) dyes can serve as a building block towards the development of computational pathology tools for systematic digital profiling of the spatial TME (Kather et al. (2016); Madabhushi and Lee (2016); Nalisnik et al. (2017); Sari and Gunduz-Demir (2018); Sirinukunwattana et al. (2018) and can be employed for cancer grading and prognostication Nalisnik et al. (2017); Sirinukunwattana et al. (2018); Huijbers et al. (2012); Louis et al. (2015)).

Because of the importance of this problem in computational pathology, a number of approaches have been proposed for the automatic identification of tissue phenotypes (Kather et al. (2016, 2019); Nalisnik et al. (2017); Sirinukunwattana et al. (2018); Bianconi et al. (2015); Huang et al. (2017b); Lazebnik et al. (2006); Linder et al. (2012); Sarkar and Acton (2018); Srinivas et al. (2014); Tamura et al. (1978); Vu et al. (2016); Wright et al. (2009); Xu et al. (2016, 2017)). Texture analysis is a commonly used approach for tissue phenotyping (Kather et al. (2016); Bianconi et al. (2015); Linder et al. (2012); Tamura et al. (1978)), whereby texture features such as local binary patterns and Gabor features of different histology images are computed to train classifiers which are then used to predict distinct tissue types. For instance, Sarkar and Acton (2018) recently proposed a saliency guided dictionary approach where Gabor features were extracted for histology image classification. Bianconi et al. (2015) proposed five different kinds of perception-based texture features, while Linder et al. (2012) reported a simple SVM classifier trained on a set of local binary patterns and contrast measure features. Although texture-based methods may be attractive due to their simplicity, texture features do not fully capture the biological significance of tissue types resulting in performance degradation (Kather et al. (2016)).

In recent years, a growing number of deep learning methods have also been proposed to classify WSIs into distinct tissue types (Nalisnik et al. (2017); Huang et al. (2017b); Xu et al. (2016, 2017); Janowczyk and Madabhushi (2016)). Xu et al. (2016) proposed a fully supervised deep CNN model for segmentation and classification of epithelial and stromal regions in histology images. Huang et al. (2017b) proposed an unsupervised domain adaptation deep network for segmenting histology images into meaningful regions. Most deep learning methods for tissue phenotyping share a common denominator which is their need for large amount of annotated histology data for training which may be tedious to obtain (Huang et al. (2017b); Janowczyk and Madabhushi (2016)). Another shortcoming of most existing literature is that although encouraging results were reported in these studies, most of the current methods are limited to the discrimination of tumor epithelium and stroma only (Bianconi et al. (2015), Huang et al. (2017b), Linder et al. (2012), Xu et al. (2016)). Like most solid tumors, colorectal cancer (CRC) tissue does not consist of only tumor and stroma components (Kather et al. (2016)). It also contains a complex rich mix of several other tissue phenotypes including smooth muscle, inflammatory, necrotic, complex stroma, and benign tissue, as shown in Fig. 1.

In this paper, we propose the concept of cellular communities comprising of different types of cells and pose the problem of tissue phenotyping as a cellular community detection problem. The premise is that spatially adjacent cells are more likely to receive intercellular signals from each other than from cells that are further away. It is also well established that the intercellular signalling between various types of cells in the microenvironment can lead to the progression of cancer (Alberts et al. (2015)). In clinical practice, pathologists consider the spatial distributions of different cellular components while identifying complex tissue phenotypes, such as the complex stroma.

Community detection methods have attracted a good deal of attention in the literature for understanding real-world complex networks in recent years, see for instance (Fortunato (2010); Harenberg et al. (2014); Mahmood et al. (2017)). The edges and nodes in a network are often inhomogeneous, resulting in groups of nodes with higher concentration of edges known as communities that share many common attributes and similar behaviour. Different tissue types such as stroma, tumor, and necrotic etc., also form local cellular communities which can play an important role in the interpretation of WSIs as shown in Fig. 1. We propose a novel semi-supervised community detection algorithm for automatic recognition of distinct tissue phenotypes in a colon cancer WSI. We first construct local cell-cell networks using potential cell-cell connection between cells as features and whereby adjacent cells are connected with each other while distant cells are not, taking into account the various ranges of cell signalling. Dividing a WSI into several thousand patches, we then construct a patch-level graph for the WSI using the cell-cell connection frequencies as features. Finally, we identify tissue phenotypes by mining in the patch-level graph for cellular communities that are biologically meaningful and clinically relevant.

A major limitation of most community detection methods is the presence of a relatively high number of inter-community edges which makes the detection of communities difficult (For-

tunato (2010)). To address this problem, we propose to map the patch-level network nodes to the geometric space by representing each node as a vector of geodesic distances from other nodes in the network. The geodesic density gradient is then computed in the geodesic space and nodes are drifted towards maximum density regions (Mahmood et al. (2017)). After the convergence of the network in the geometric space, simple K-means clustering algorithm is used to assign community labels to each patch (see Fig. 1). The nodes in each community represent biologically meaningful tissue components which are distinct from the other communities. An earlier version of this work was presented at the MICCAI Computational Pathology workshop (Javed et al. (2018)). The main contributions of this work are as follows:

1. Instead of using texture features to represent a patch of WSI, we consider the potential cell-cell connections between various types of cells as representative features of a patch. These features are biologically more meaningful and better capture the distribution of different types of cells in the histology patch.

2. We pose the problem of identifying tissue phenotypes as a community detection problem in histological landscape where each community represents a distinct tissue phenotype, for example tumor, benign, stroma, inflammatory, complex stroma, and smooth muscle. To the best of our knowledge, the formulation of tissue phenotyping as community detection has not been done before. The use of geodesic density gradients for tissue phenotyping is also novel and has resulted in significant performance improvement.

3. We propose a new large-scale dataset for tissue phenotyping which consists of 280K patches extracted from 20 WSIs of CRC slides stained with H&E. Each slide is taken from a different patient. Each WSI contains exhaustive region-level annotation of seven distinct tissue phenotypes labelled by experienced pathologists (**KB** and **KH**). The dataset has two different testing and training settings including patch-level separation and patient-level separation. This CRC Tissue Phenotyping (CRC-TP) dataset will soon be publicly released.

4. An existing dataset known as CRCHistoPhenotypes [1](Sirinukunwattana et al. (2016)) for Cell Detection and Classification (CDC) has been extended to include five distinct cell types: tumor epithelial, normal epithelial, necrotic, spindle-shaped, and inflammatory cells. This dataset also contains patch-level and patient-level separations between training and testing splits. The extended dataset named as CRC-CDC will soon be made publicly available.

The proposed algorithm is evaluated on two independent datasets including colon cancer tissue dataset (Kather et al. (2016)) and our proposed CRC-TP dataset and compared with 27 recent state-of-the-art methods. The results demonstrate the superiority of the proposed algorithm over the existing methods by a significant margin.

The rest of this paper is organized as follows. Recent literature on tissue phenotyping is given in Section 2. Section 3 describes the proposed algorithm in detail. Experiments and results are discussed in Section 4, and finally conclusions and future directions are given in Section 5.

2. Related Work

In the past few years, many studies have investigated histology image classification problem (Bianconi et al. (2015); Huang et al. (2017b); Kather et al. (2016); Lazebnik et al. (2006); Linder et al. (2012); Nalisnik et al. (2017); Sarkar and Acton (2018); Sirinukunwattana et al. (2018); Srinivas et al. (2014); Tamura et al. (1978); Vu et al. (2016); Wright et al. (2009); Xu et al. (2016, 2017)). Many excellent surveys have also been contributed in this direction (Irshad et al. (2014); Janowczyk and Madabhushi (2016); Komura and Ishikawa (2018); Madabhushi and Lee (2016); Qaiser et al. (2018); Veta et al. (2014)). Existing tissue phenotyping approaches can be broadly categorized into texture-based methods (Bianconi et al. (2015); Kather et al. (2016); Kothari et al. (2013); Linder et al. (2012); Tamura et al. (1978)), sparse representation methods (Sarkar and Acton (2018); Srinivas et al. (2014); Vu et al. (2016)), and deep learning methods (Bejnordi et al. (2018); Du et al. (2018); Huang et al. (2017b); Nalisnik et al. (2017); Xu et al. (2016, 2017)).

Texture-based methods estimate the local texture around a pixel of the histology image to alleviate the effect of heterogeneity (Bianconi et al. (2015); Kather et al. (2016); Kothari et al. (2013); Linder et al. (2012); Tamura et al. (1978)). These features consist of Local Binary Patterns (LPB), Gabor features, lower and higher order histogram features, gray level co-occurrence matrix at different directions, and perception-based features. Texture features of different histology images are first estimated, and then they are used to train SVM classifiers for predicting tissue phenotypes. Tamura et al. (1978) proposed five different perception-based features including coarseness, contrast, directionality, line-likeness, and roughness. Bianconi et al. (2015) exploited these perception features for tissue phenotyping. Kothari et al. (2013) proposed Fourier shape-based descriptor for the identification of retinal tumor in images. Linder et al. (2012) proposed to use LBP with contrast measure features. Encouraging results were reported in these studies. However, the studies presented in (Bianconi et al. (2015)) and (Linder et al. (2012)) were limited for the identification of tumor epithelium and stromal tissue phenotypes. To address this deficiency, Kather et al. (2016) recently proposed to use six different types of texture-based descriptors for the classification of eight different tissue phenotypes in colon cancer histology images. Although, the discrimination performance improved, the texture descriptors do not fully capture the biological significance of the tissue components, hence this method is not very accurate in identifying tumors with complex stroma and mucosa (Kather et al. (2016)).

[1]https://warwick.ac.uk/TIAlab/data/crchistolabelednucleihe/

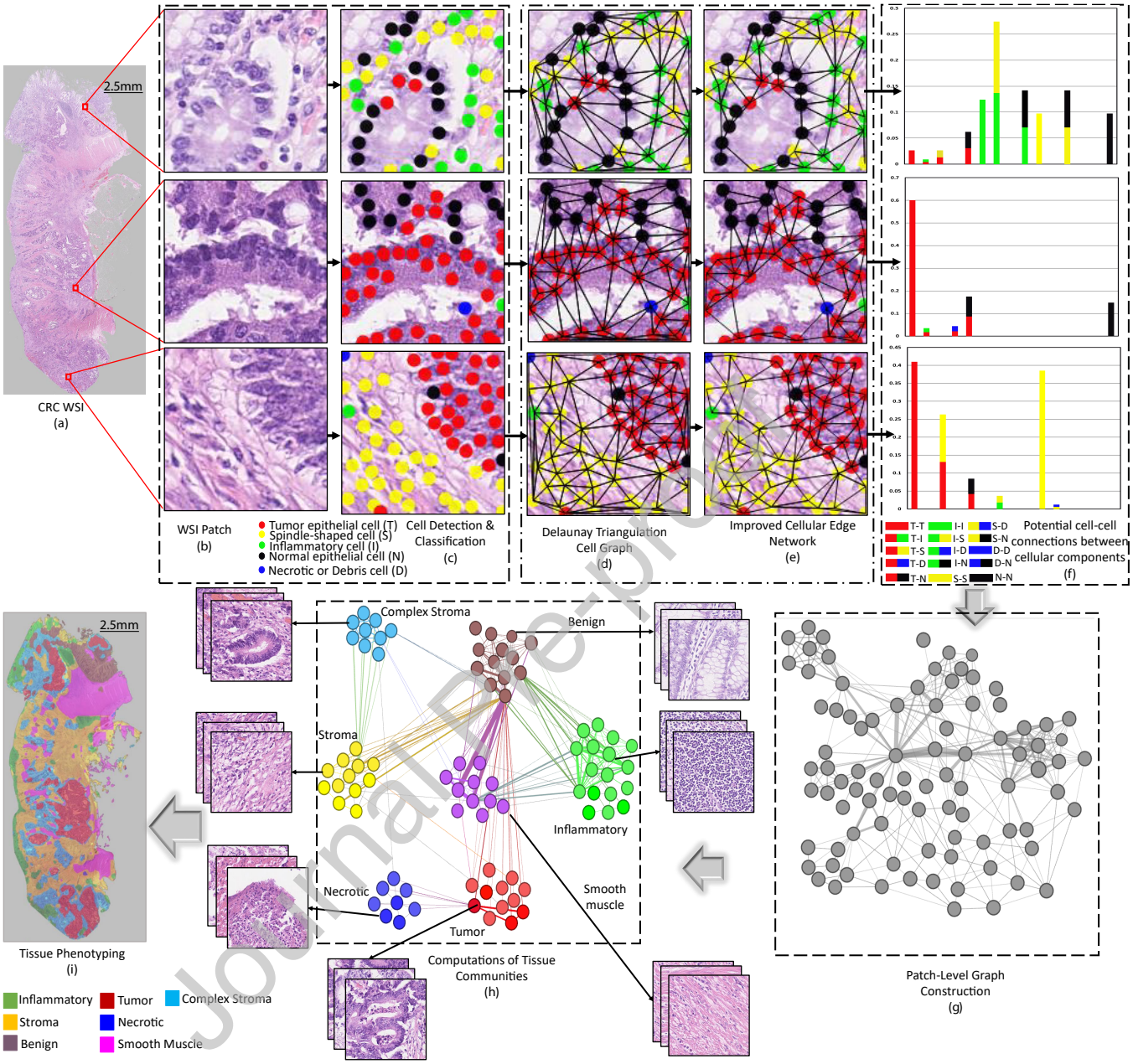


Fig. 2: The framework of our proposed tissue phenotyping algorithm using community detection. (a) An example CRC WSI taken from CRC-TP dataset; (b) A patch of size 150 × 150 pixels (extracted at 20x magnification level in our CRC-TP dataset); (c) Results of the cell detection and classification method where red, green, blue, yellow, and black colors represent Tumor epithelial (T), Inflammatory (I), Debris or necrotic (D), Spindle-shaped (S), and Normal epithelial (N) cells; (d) Construction of cell graph using Delaunay triangulation where nodes represent the spatial locations of cellular components and edges represent potential cell-cell connections of cellular components; (e) Reduction in heterogeneity by removing cell-cell connections between distant cells; (f) Computation of 15 dimensional feature vector from each cell graph capturing potential cell-cell connections and distribution of cell nuclei; The bars of single color show cell-cell connection between similar types of cell, while the bars of two colors represent cell-cell connection between two different cellular components; For example, red bars show T to T cell-cell connection and red-black bars show T to N cell-cell connection; (g) Patch-based graph construction, where each node is a feature vector of the corresponding cell graph as shown in (d); (h) Proposed community detection algorithm where different colors represent local tissue communities; The nodes drift iteratively towards local maximum geodesic density regions resulting in a reduced intra-class distance and increased inter-class distance; (i) Results of the proposed algorithm where the local tissue communities are overlaid on WSI.

Sparse representation approaches encode a histology image as a sparse linear combination of basis functions or dictionary atoms (Lazebnik et al. (2006); Sarkar and Acton (2018); Srinivas et al. (2014); Vu et al. (2016); Wright et al. (2009)). For each tissue phenotype, a different dictionary is learned and based on the representation error, tissue phenotypes of test images are identified. Srinivas et al. (2014) proposed a multi-channel dictionary using the RGB tissue features. Vu et al.

(2016) proposed a dictionary learning technique trained on increasing the inter-class and decreasing the intra-class variability. Sarkar and Acton (2018) recently proposed a saliency-guided sparse representation approach for multi-class tissue phenotypes. Results reported in these studies are promising however, the dictionaries are trained by using the color and texture features resulting in the performance degradation similar to the texture-based approaches.

Recently, Deep CNN (DCNN) based methods have also been proposed for tissue phenotyping (Bejnordi et al. (2018); Du et al. (2018); Huang et al. (2017b); Nalisnik et al. (2017); Xu et al. (2016, 2017)). DCNN models learn the rich hierarchy of convolutional features for each class and then predict the tissue type. Xu et al. (2016) proposed a DCNN model for classifying breast cancer histology images. Their network comprised of two convolutional layers, two max-pooling layers, and two fully connected layers followed by a soft-max layer. Du et al. (2018) and Huang et al. (2017b) proposed DCNN models incorporating the notion of domain adaptation in the AlexNet and GoogleNet. Xu et al. (2017) improved the AlexNet model for the segmentation and classification of histology images. Bejnordi et al. (2018) proposed three DCNN models for classifying breast cancer WSIs. The first network was trained to classify WSI into fat, stroma, and epithelium tissues. The second DCNN processed the stromal regions and predicted the complex stroma regions. The third DCNN was trained to classify invasive cancer in the WSIs. These studies produced better results in many complex situations however, these methods are limited to binary classification including tumor epithelium and stroma. Moreover, these methods require large amounts of labelled training histology data, which may not always be available. In contrast, we propose a semi-supervised algorithm which does not require any labelled training data for the classification of tissue phenotypes.

Most of the existing approaches consider binary classification only and rely on texture features. In contrast, we observe that if the potential cell-cell connections between cellular components can be exploited as a discriminator, the performance of tissue phenotyping can be significantly improved in the presence of complex tissue structure. Moreover, we propose the tissue classification problem as identifying network communities. To the best of our knowledge, no similar method has previously been reported for tissue classification.

3. Tissue Phenotyping via Community Detection

In the proposed tissue phenotyping algorithm, a given WSI is divided into non-overlapping patches, and in each patch, we classify cells using a deep neural network. In this study, we have used a patch size of 150×150 pixels at $20\times$ resolution from each WSI. Based on the cell-cell connections and distribution of different cellular components in each patch, we compute patch-level feature vectors which are then used to compute a patch-level graph. In this graph, each node represents a locality contained by a patch. Based on the connections between different nodes, the patch-level graph is divided into seven histology communities. A schematic diagram of the overall proposed algorithm is shown in Fig. 2. The proposed approach consists of four main steps including cell detection and classification, cell graph construction and computation of cell-cell connections features, construction of patch-level graph, and computation of tissue phenotype communities using a community detection algorithm. In the following subsections, each of these steps are explained in more detail.

3.1. Cell Nuclei Identification

In this work, potential cell-cell connections between different cellular components has been used as features which are then used for identifying tissue communities. In order to compute potential cell-cell connections, we first identify different types of cells in each histology patch referred to as locality. For this purpose, we use Spatially Constrained Convolutional Neural Network (SC-CNN) proposed by Sirinukunwattana et al. (2016) and pre-trained Tunable Shape Priors CNN (TSP-CNN) proposed by Tofighi et al. (2019) for cell detection. For the training of SC-CNN for cell detection, nuclei centres were manually marked. A probability map was generated such that maximum probability was assigned to the centroid pixels. For the other pixels, the probability decreases as the distance from the centroid increases. Using this probability map, the detection network is trained to assign an appropriate probability to each pixel in the test patch for being a nuclei centroid.

The classification SC-CNN network proposed by Sirinukunwattana et al. (2016) was able to classify only four classes including Epithelial, Miscellaneous, Inflammatory, and Fibroblast. In the current work, we extended the classification network to predict five distinct classes including Tumor epithelial (T), Spindle-shaped (S), Debris or necrotic (D), Normal epithelial (N), and Inflammatory (I) cells. Multiple shifted patches are extracted around each detected nuclei location which are used for the training of the classification network. The classification network comprises of two convolution layers and two max-pooling layers with a stride of 2×2, two fully connected layers followed by the classification layer and the probability for each label is predicted using soft-max layer. For a test nuclei, multiple shifted patches are extracted and classified using the network and the class label of the test nuclei is computed from a weighted sum of all the probability maps of the shifted patches. A patch having a larger distance from the detected nuclei is assigned smaller weight compare to a patch closer to the nuclei. The output of the network is a set of five different types of cell nuclei shown in Fig. 2 (c).

3.2. Cell Graph Construction

For each patch $\mathbf{X}_i \in \mathbb{R}^{p\times p}$ (patch size is 150×150 at 20× magnification level),), we construct a cellular graph such that the vertices correspond to the spatial locations of cells and the edges are assigned using Delaunay triangulation (Fig. 2 (d)). The Delaunay triangulation estimates a triangle for each cell by finding two nearest cells and inserts edges among the three cells. We observe that cells on the opposite sides of tissue constituent white space also known as lumen and endothelium known as micro-vessels do not communicate to each other. To avoid these edges, we use a distance threshold between the cells. The edges between cells which are at a distance larger

than a threshold are discarded as shown in Fig. 2 (e). By removing these edges, the problem of heterogeneity within the edges is also reduced.

For each patch, we compute a feature vector by computing 15 potential cell-cell connections between cellular components including T to T (red), T to I (red and green), T to S (red and yellow), T to D (red and blue), T to N (red and black), I to I (green), I to S (green and yellow), I to D (green and blue), I to N (green and black), S to S (yellow), S to D (yellow and blue), S to N (yellow and black), D to D (blue), D to N (blue and black), and N to N (black) as shown in Fig. 2 (f). The cell-cell connection features are computed as the frequency of each cell-cell connection in a given cell graph:

$$\mathbf{h}_i = \texttt{Cell-CellConnectionFeat}(\mathbf{A}_i^{cg}, l_i^{cg}), \quad (1)$$

where $\mathbf{A}_i^{cg}$ is the adjacency matrix of cell graph of i_{th} patch and l_i^{cg} is the cell labels for each node in the same cell graph, and $\mathbf{h}_i \in \mathbb{R}^m$ represents distribution of cellular components in the cell-graph, where $m = 15$. We create an input data matrix for each WSI as

$$\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n] \in \mathbb{R}^{n \times m}, \quad (2)$$

where n denotes the number of patches in the WSI.

3.3. Patch-Level Graph Construction

Using the cell graph feature vectors, we construct an undirected graph $\mathbf{G}_p = (\mathbf{V}, \mathbf{A})$ such that each vertex $\mathbf{v}_i$ corresponds to $\mathbf{h}_i$ in the feature matrix $\mathbf{H}$. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is computed by employing chi-squared distance as:

$$A(i,j) = \exp\Big(-\frac{1}{\sigma} \sum_{k=1}^{m} \frac{(\mathbf{h}_{i,k} - \mathbf{h}_{j,k})^2}{\mathbf{h}_{i,k} + \mathbf{h}_{j,k}} \Big), \quad (3)$$

where $\mathbf{h}_i \in \mathbb{R}^m$ and $\mathbf{h}_j \in \mathbb{R}^m$ are two feature vectors, and σ is a weight decay control parameter. The adjacency matrix $\mathbf{A}$ represents a weighted graph whereby the weight between two vertices quantifies the closeness or the similarity between the corresponding cell graphs. Using the adjacency matrix $\mathbf{A}$, we compute a geodesic distance matrix $\mathbf{G}$.

$$\mathbf{G} = \text{All-Pairs-Shortest-Paths}(\mathbf{A}). \quad (4)$$

The geodesic distance is more meaningful in case the data is distributed on a nonlinear manifold. In such cases, the chi-squared distance between two features may be small but corresponding geodesic distance may be large. We assume that the network represented by the adjacency matrix $\mathbf{A}$ in Eq. (3) is fully connected. We represent each network node $\mathbf{h}_i \in \mathbb{R}^m$ using its shortest distances from all other nodes in the network $\mathbf{g}_i \in \mathbb{R}^n$ also known as geodesic distances using Eq. (4). Note that the geodesic distance computation acts as a kernel projecting the feature vector to a higher dimensional space. This projection results in better separation between different clusters in the tissue phenotype network. The distance between two geodesic vectors $\mathbf{g}_i$ and $\mathbf{g}_j$ in $\mathbf{G} = [\mathbf{g}_1, \cdots, \mathbf{g}_n]$ is defined as

$$d_{i,j} = \sqrt{(\mathbf{g}_i - \mathbf{g}_j)^\top \mathbf{W} (\mathbf{g}_i - \mathbf{g}_j)}, \quad (5)$$

where $\mathbf{W} \in R^{n \times n}$ is a diagonal matrix containing weights for each dimension of the geodesic distance vector. These weights are adjusted such that the local and global structure of the network becomes equally important (Mahmood et al. (2017)).

$$W(p,q) = \begin{cases} \frac{1}{2}, \text{ if } p = q = \{i,j\}, \\ \frac{1}{n-2}, \text{ if } p = q \neq \{i,j\}, \\ 0, \text{ if } p \neq q. \end{cases} \quad (6)$$

Thus, the weight of the shortest distances corresponding to $p = q = \{i, j\}$ becomes 1.00 and the weight of the remaining shortest distances, which are $n - 2$, also collectively becomes 1.00. So, we ensure a balance between direct distances and indirect distances. Using this definition of distances in the geodesic space, we compute cellular communities in the patch-level graph as described below.

3.4. Computing Cellular Communities

In the patch level graph, instead of considering each network node as a discrete point in the geodesic space, we consider it yielding a continuous density function. As an example, a density at a point $\mathbf{s}$ induced by a node $\mathbf{g}_i$ is given by

$$K(\mathbf{s}|\mathbf{g}_i, \sigma_g) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma_g^n} \exp\Big(\frac{-(\mathbf{g}_i - \mathbf{s})^\top \mathbf{W} (\mathbf{g}_i - \mathbf{s})}{2\sigma_g^2} \Big), \quad (7)$$

The parameter σ_g is the bandwidth of the kernel function in geodesic space. By varying σ_g, we can vary the probability density induced by a node at a particular distance from that node. Each network node is assumed to induce its density in the whole geodesic space. The probability density at point $\mathbf{s}$ induced by all network nodes gets superimposed. The resulting density is given by

$$f(\mathbf{s}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma_g^n} \sum_{i=1}^{n} K\Big(\frac{\sqrt{\mathbf{W}}(\mathbf{s} - \mathbf{g}_i)}{\sigma_g} \Big). \quad (8)$$

The cumulative density function as defined by Eq. (8) varies across the space. We intend to drift the network nodes towards the higher density regions. Each density region corresponds to a particular tissue phenotype in the WSI. For this purpose, we compute the gradient of the cumulative density function as follows

$$\nabla f(\mathbf{s}) = \frac{\sqrt{\mathbf{W}}}{(2\pi)^{\frac{n}{2}} \sigma_g^n} \sum_{i=1}^{n} \nabla K\Big(\frac{\sqrt{\mathbf{W}}(\mathbf{s} - \mathbf{g}_i)}{\sigma_g} \Big), \quad (9)$$

where ∇ is a gradient operator with respect to each of the dimensions of the space. Using the values of K from Eq. (7) and differentiating it with respect to $\mathbf{g}$ as:

$$\nabla f(\mathbf{s}) = \frac{\sqrt{\mathbf{W}}}{(2\pi)^{\frac{n}{2}} \sigma_g^n} \sum_{i=1}^{n} \Big(\frac{\sqrt{\mathbf{W}}(\mathbf{s} - \mathbf{g}_i)}{\sigma_g^2} K(\mathbf{s}|\mathbf{g}_i, \sigma_g) \Big), \quad (10)$$

where $\nabla f(\mathbf{s})$ is the estimate of the average density gradient pointing in the direction of the maximum increase in density. If

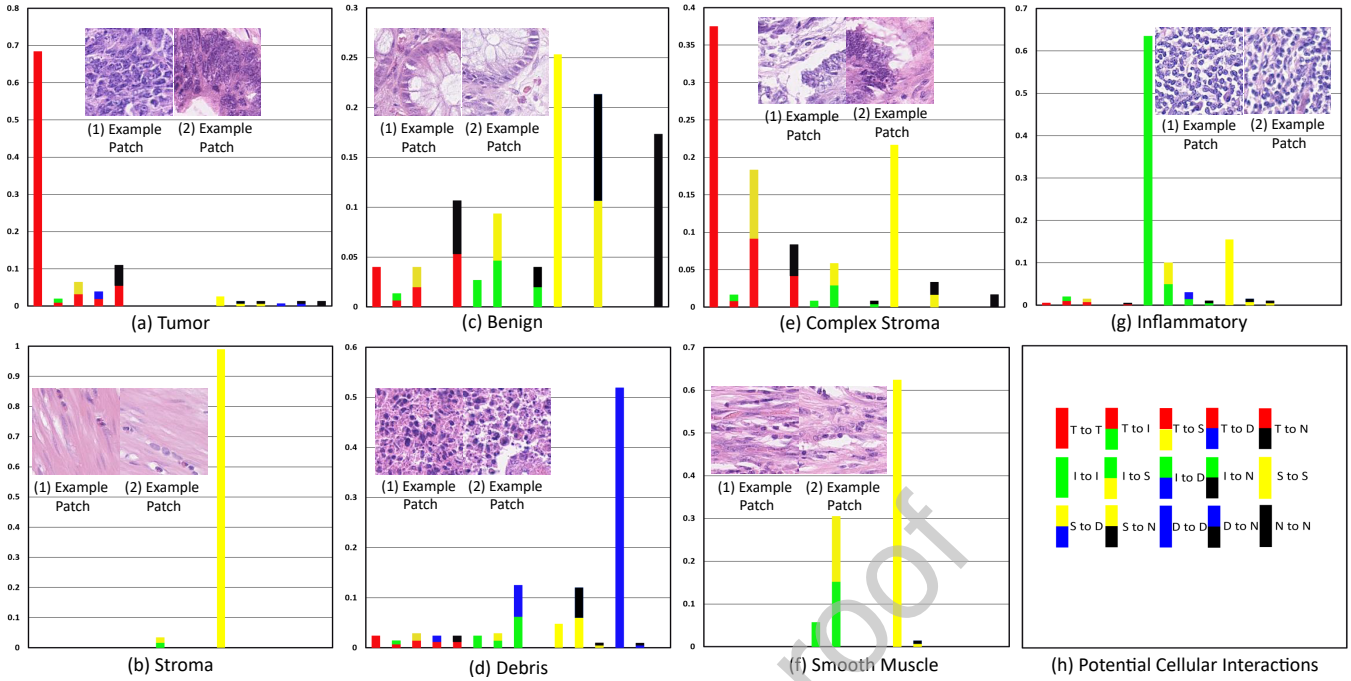


Fig. 3: Representative cluster centres of seven distinct tissye phenotypes using our proposed algorithm. The representative tissue images are also shown in each tissue phonotype.

each network node is drifted towards positive density gradient, then nodes will converge towards maximum density regions. In these regions, the density gradient will approach to zero. It is because density will be the same in all directions. Assuming $\mathbf{s}$ to be the current estimate of a node, setting $\nabla f(\mathbf{s}) = 0$, we get the new estimate as follow:

$$\Delta\mathbf{s} = \frac{\sum_{i=1}^{n} \mathbf{s} \exp((\mathbf{g}_i - \mathbf{s})^\top \mathbf{W}(\mathbf{g}_i - \mathbf{s})/2\sigma_g^2)}{\sum_{i=1}^{n} \exp((\mathbf{g}_i - \mathbf{s})^\top \mathbf{W}(\mathbf{g}_i - \mathbf{s})/2\sigma_g^2)}. \quad (11)$$

Eq. (11) is repeatedly applied to each node of the network to get updated node position $\mathbf{g}_i^{k+1} = \mathbf{g}_i^k + \Delta\mathbf{s}$, where $\mathbf{g}_i^k$ is the current position of the node in k^{th} iteration and $\mathbf{g}_i^{k+1}$ is the updated position $k+1$ iteration. It results in each node iteratively drifting towards local density maximum. The nodes are assumed to converge to the final positions when the cumulative drift r_{k+1} becomes less than a threshold.

$$r_{k+1} = \sum_{i=1}^{n} \|\mathbf{g}_i^{k+1} - \mathbf{g}_i^k\|_2. \quad (12)$$

In maximum density regions, gradients become very small therefore most of the nodes may not converge to a single point in space, instead, most of the nodes stop at different close-by positions. Therefore, in order to obtain a discrete community labels, we apply K-means algorithm on the final positions $\mathbf{g}_i^{k+1}$ of nodes.

$$\ell = \text{K-means}(\{\mathbf{g}_i^{k+1}\}_{i=1}^{n}, c_l), \quad (13)$$

where c_l is the number of tissue communities, and $\ell \in R^n$ is the community label vector. Each cluster indicates a discrete community corresponding to a particular tissue phenotype. Using the community labels found by Eq. (13), we compute the geometric centres for each tissue community for the cell connectivity features given by Eq. (2). These geometric centres $\mathbf{c}_j$ are considered as representative samples of each tissue phenotype. Figs. 3 (a)-(h) show these representative samples obtained from each cluster centre. The tissue patches belonging to each cluster are presented to the experienced pathologists. The computed clusters are biologically meaningful and the pathologists assigned a distinct tissue phenotype to each cluster including tumor, stroma, complex stroma, smooth muscle, debris, benign, and inflammatory. Algorithm 1 describes each step of the proposed method. The predicted community labels are compared with the ground truth labels of each patch using three different clustering quality measures including normalized mutual information, adjusted rank index, and purity as discussed in Section 4.4 below.

4. Experiments and Evaluations

The proposed Tissue Phenotyping using Community Detection (TPCD) algorithm is evaluated both quantitatively and qualitatively on two different CRC datasets including Colon Cancer Tissue (CCT) dataset (Kather et al. (2016)) and our newly proposed CRC-TP dataset which has two versions. The first version has patch-level separation between testing and training data while the second version has patient-level separation as specified below. The results of tissue phenotyping algorithm are compared with 27 state-of-the-art methods including 12 published methods, 4 deep neural networks-based methods, 5 Graph CNN-based (GCN) methods, and 7 variants of the proposed algorithm. Since, the cell detection and classification are

Algorithm 1: Proposed Tissue Phenotyping Algorithm.

Input: n image patches $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n]$, where each $\mathbf{X}_i \in \mathbb{R}^{p \times p}$.
Ouput: Representative tissue communities such as Tumor ($\mathbf{c}_t \in \mathbb{R}^m$), Stroma ($\mathbf{c}_s \in \mathbb{R}^m$), Debris ($\mathbf{c}_d \in \mathbb{R}^m$), Inflammatory ($\mathbf{c}_{in} \in \mathbb{R}^m$), Smooth muscle ($\mathbf{c}_{sm} \in \mathbb{R}^m$), Benign ($\mathbf{c}_b \in \mathbb{R}^m$), and Complex stroma ($\mathbf{c}_{cs} \in \mathbb{R}^m$)
Step 1: Cell detection and classification on each $\mathbf{X}_i \in \mathbb{R}^{p \times p}$.
Step 2: Construct cell-level graph.
Step 2: Compute $\mathbf{h}_i$ using Eq. (1).
Step 3: Compute $\mathbf{H}$ using Eq. (2).
Step 4: Compute $\mathbf{A} \in \mathbb{R}^{n \times n}$ using Eq. (3).
Step 5: Compute $\mathbf{G} \in \mathbb{R}^{n \times n}$ using Eq. (4).
while *not converged* **do**
  1. Compute drift for each node using Eq. (11).
  2. Check convergence according to Eq. (12)
end
Step 6: Compute community labels using Eq. (13).

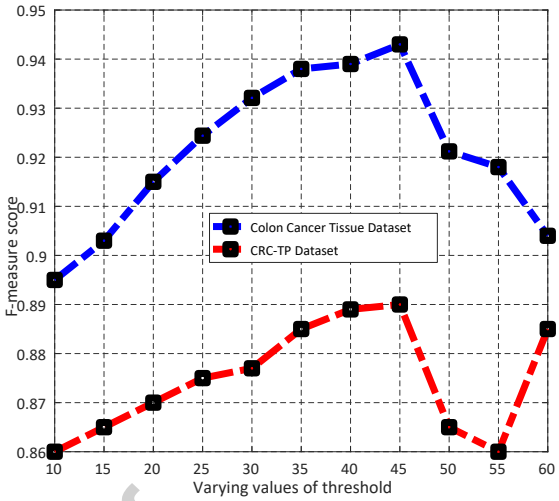


Fig. 4: Varying values of threshold to remove distant edges in cellular graph.

pre-processing steps for our proposed tissue phenotyping algorithm, therefore, we also discuss the performance of different existing methods on our newly proposed CRC-CDC dataset.

4.1. *State-of-the-art Compared Methods*

4.1.1. *Comparison with existing Tissue Phenotyping Methods*

We compared our proposed algorithm with the following methods: K-Medoids clustering with Chi-square Distance (KM-CD) (Sirinukunwattana et al. (2018)), Subspace Clustering (Elhamifar and Vidal (2013)), Dictionary Learning with KL Divergence (DL+kldiv) (Mairal et al. (2012)), Sparse representation-based Compression Distance (SCD) (Guha and Ward (2014)), Sparse Representation-based Classification (SRC) (Wright et al. (2009)), Best Five Features with SVM classification (B5F-SVM) (Kather et al. (2016)), Best Six Features with SVM classification (B6F-SVM) (Kather et al. (2016)), Discriminative Features Oriented Dictionary learning (DFOD) (Vu et al. (2016)), Simultaneous Sparsity model for Histopathological Image Representation and Classification (SHIRC) (Srinivas et al. (2014)), Spatial Pyramid Matching (SPM) (Lazebnik et al. (2006)), Saliency-based Dictionary Learning with Smoothness constraints (SDLs) (Sarkar and Acton (2018)), and SVM-CNN (Xu et al. (2017)). All implementations are obtained from the original authors and we used the default parameters as proposed by the original authors. We implemented SVM-CNN method for multi-class tissue classification (Xu et al. (2017)). We extracted deep features from the fully connected layer 2 (fc-2) of AlexNet (Krizhevsky et al. (2012)) and then we trained linear SVM classifier for tissue phenotyping.

4.1.2. *Tissue Phenotyping Using Deep Neural Networks*

We compared our methods with the four deep neural networks including Mobile DCNN (MobileNet) (Howard et al. (2017)), deep Residual CNN-50 (ResNet50) (He et al. (2016)), ResNet101 (He et al. (2016)), and DenseNet (Huang et al. (2017a)). These networks were pertained on the ImageNet database (Deng et al. (2009)). We replaced the classification layer and fine-tuned these networks with stochastic gradient descent with a momentum of 0.8. To gauge the performance of these network architectures, we randomly divided the tissue phenotyping datasets into 70% training set and 30% testing set. We trained all networks on a desktop workstation with two NVidia Titan Xp GPUs with a mini-batch size of 256 and a learning rate of 3×10^{-4} for 130 epochs. In all cases, rotational invariance was achieved through data augmentation with random horizontal and vertical flips of the training images. Images were re-sized to the neural network input size if necessary.

4.1.3. *Tissue Phenotyping using GCN Methods*

The Graph CNN (GCN) methods compute the node embedding of the graph which are then used for node classification in transductive as well as inductive learning manners. The labels of the training nodes, feature vectors, and input graph are fed to GCN methods. We compared our proposed algorithm with five GCN methods including GCN with fast localized Spectral Filtering (GCN-SF) (Defferrard et al. (2016)), Semi-Supervised classification with GCN (SSC-GCN) (Kipf and Welling (2017)), GCN for web-scale Recommender Systems (GCN-RS) (Ying et al. (2018)), Deep Graph Infomax (DGI) (Veličković et al. (2019)), and GCN with Arma filters (GCN-AF) (Bianchi et al. (2019)). These methods are compared with Euclidean distance-based graph as well as our proposed cell features-based patch-level graph given by Eq. (3). For Euclidean distance-based graph construction, the deep features are extracted using the VGG-16 model. The features are compressed using PCA and the resulting feature vector of dimension 128 is obtained for each patch, which is then used for the distance computation as suggested by (Li et al. (2018)). The implementations of compared GCN methods are taken from the

original authors [2] and training is performed by using the recommended parameters in the relevant publications.

4.1.4. *Proposed Algorithm Variants and Settings*

Different steps of the proposed algorithm are evaluated by designing six variants including TPCD-1, our previous study referred as TPCD-2 (Javed et al. (2018)), TPCD-3, TPCD-Hist, SVM-Cellfeatures, TPCD-4, and TPCD-CG. In TPCD-1 algorithm, geodesic distance computation is skipped and the network as given by Eq. (3) is directly used for further processing. TPCD-2 and TPCD-3 algorithms are similar with the only difference that cells at a relatively larger distance also communicate in TPCD-2 which results in increased heterogeneity. In TPCD-3 algorithm, these cell-cell connections are removed using a threshold on the physical distance between different cellular components as previously described in Sec. 3.2. In TPCD-Hist, the histogram of cell types is used as a feature vector while the remaining processing is similar to TPCD-3. TPCD-4 is similar to TPCD-3 except for the cell detection component instead of using SC-CNN method proposed by Sirinukuwattana *et al.* (Sirinukunwattana et al. (2016)), we have employed a recently proposed cell detection method known as TSP-CNN proposed by (Tofighi et al. (2019)). SVM-Cellfeatures consists of SVM classifier using cell-cell connections features. In SVM-CellFeatures, SVM-CNN, B5F-SVM, and B6F-SVM methods, we train the SVM classifier and we used 10-fold cross validation. Both datasets were randomly divided in 10 parts, and 10 rounds of training and testing were performed. For each subdivision a different 10% subset of the dataset was used for testing while the other 90% was used for training.

In addition, we also implemented TPCD algorithm on each Cell-level Graph (TPCD-CG) without exploiting cell classification information. From each patch, we constructed cell-level graph using Delaunay triangulation and then the feature vector corresponding to three structural properties including average degree, average clustering coefficient, and diameter (Dorogovtsev and Mendes (2002)) of the cell-level graph is computed. TPCD algorithm is then used to compute the distinct tissue phenotypes.

In our experiments, we used a threshold of 45-pixel distance to remove the distant edges in TPCD-3 algorithm. Performance variation is investigated by varying this threshold from 10 to 60 in steps of 5 as shown in Fig. 4. The best performance is observed for a threshold of 45. In Eq. 12, the cumulative drift r_{k+1} is bounded to be more than 0.003. For $r_{k+1} < 0.003$, further iterations are stopped.

4.2. *Datasets*

4.2.1. *Proposed Cell Detection and Classification Dataset (CRC-CDC)*

An earlier study of cell detection and classification was performed on CRCHistoPhenotypes dataset[3] consisting of 100 H & E stained histology images obtained from 9 patients (Sirinukunwattana et al. (2016)). This dataset had four cell classes including epithelial, inflammatory, miscellaneous, and fibroblast. The epithelial class contained both normal epithelial and tumor-epithelial cells. Therefore, tissue phenotyping based on these classes resulted in the same community label for the tumor and the benign phenotypes which is an undesired result. In order to differentiate tumor from benign tissue phenotype, we have to re-label normal epithelial cells and tumor epithelial cells separately. In the current study, we extend the CRCHistoPhenotypes dataset to 256 H & E stained images of CRC obtained from 20 different patients and containing five distinct cell classes including tumor epithelial, normal epithelial, spindle-shaped, inflammatory cells, and necrotic. The extended version is named as CRC-CDC dataset in which each visual field contains 500×500 pixels extracted at 20x magnification level. For the annotations purpose, the same protocol was used as reported by the previous study (Sirinukunwattana et al. (2016)). Manual annotations of cell nuclei are made by experienced pathologists (**YT** and **KH**) and partly by a research fellow under the supervision of the same pathologists. After full annotations, each annotated nuclei was reviewed by both of the pathologists; therefore refining their own and each others annotations. Annotating the data in this way ensured that minimal nuclei were missed in the annotation process. However, we cannot avoid inevitable few pixel difference between the annotation and the true nuclei centre. A total of 38,984 nuclei are marked at the centre for detection purposes. Out of these, 30,531 nuclei have associated class labels. In total, there are 7,231 tumor epithelial cells, 6,545 normal epithelial cells, 5,712 spindle-shaped cells, 6,971 inflammatory cells, and 4,072 necrotic cells.

To test the generalization of the cell detection and classification network SC-CNN (Sirinukunwattana et al. (2016)), two experimental settings are used. In the first experiment, 70% nuclei are randomly selected for training and the remaining 30% nuclei are used for testing. In the second experiment, patient-level separation is maintained by keeping the images from 14 patients as training data while the images of the remaining 6 patients are used for testing data.

4.2.2. *Colon Cancer Tissue (CCT) Dataset*

The CCT dataset contains eight different types in human CRC histology obtained from H&E stained slides of CRC samples (Kather et al. (2016)). The tissue categories are manually annotated and overlapping patches of size 150×150 extracted from these samples. The 8 categories are: tumor, stroma, complex structured stroma, lymphocytes, debris, mucosa, adipose, and background. Due to a lack of cellular structure, the background and adipose classes are not considered in our experiments. Sample images from the remaining 6 tissue classes are shown in Fig. 5. There are a total of 3,750 images in these 6 classes, with 625 images per class.

4.2.3. *Proposed CRC Tissue Phenotyping (CRC-TP) Dataset*

This dataset consists of 280K patches extracted from 20 WSIs of CRC stained with H & E taken from our local University Hospitals Coventry and Warwickshire (UHCW) for tissue phenotyping. The 20 WSIs are obtained from 20 different patients. Each WSI is manually region-level annotated by ex-

[2] https://github.com/rusty1s/pytorch_geometric

[3] https://warwick.ac.uk/fac/sci/dcs/research/tia/data/crchistolabelednucleihe/

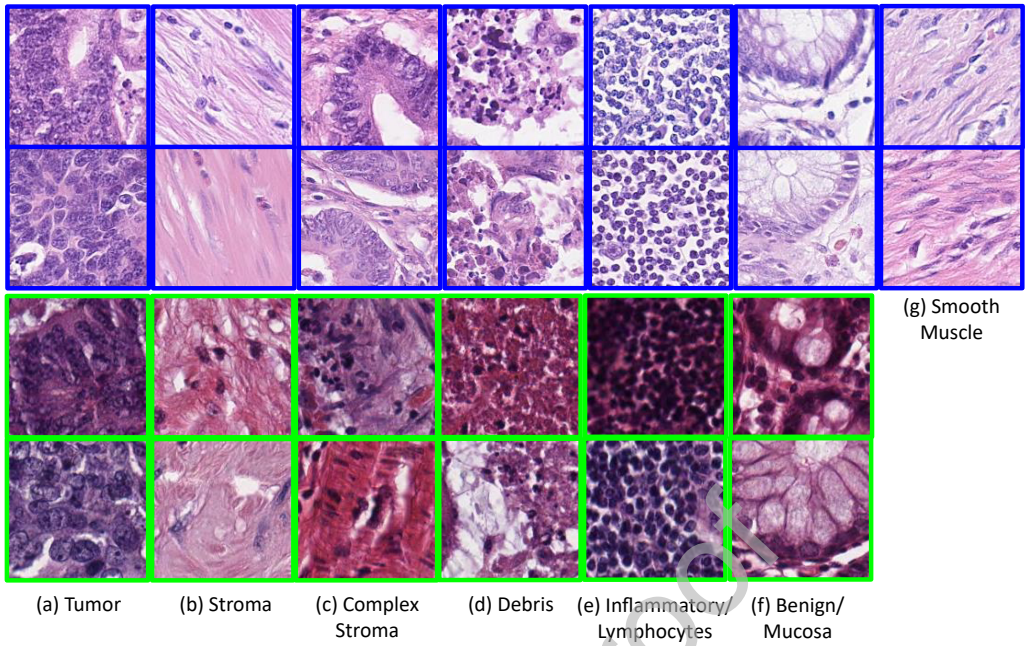


Fig. 5: Sample images of six and seven tissue phenotypes from CCT dataset (Kather et al. (2016)) and our newly proposed CRC-TP dataset. From left to right; (a) Tumor, (b) Stroma, (c) Complex Stroma, (d) Debris, (e) Inflammatory or Lymphoytes, (f) Benign or Mucosa, and (g) Smooth muscle. The blue boundary line shows the sample images of CRC-TP dataset while green boundary line shows the samples images of CCT dataset.

Table 1: Comparative performance of nuclei detection and classification in terms of average F_1 score for detection and weighted average $\widehat{F}$ score for classification on CRCHistoPhenotypes and CRC-CDC datasets. The combined performance represents the performance of both nuclei detection and classification. The two best results are shown in red and blue colors, respectively.

Datasets	Methods	Detection Performance (F_1 score)	Classification Performance (4 Nuclei Classes) (Weighted Average $\widehat{F}$ score)	Combined Performance (Nuclei Detection + 4 Nuclei Classes) (Weighted Average $\widehat{F}$ score)
CRCHisto Dataset (Sirinukunwattana et al. (2016)) (4 Nuclei Classes) (29,756 Annotated Nuclei)	SC-CNN (Sirinukunwattana et al. (2016))	0.80	0.78	**0.69**
	TSP-CNN(Tofighi et al. (2019))	0.85	-	-
	TSP-CNN+SC-CNN	-	-	**0.73**
	ResNet50 (He et al. (2016))	-	0.74	-
	DenseNet (Huang et al. (2017a))	-	0.70	-
	SC-CNN+ResNet50	-	-	0.66
	SC-CNN+DenseNet	-	-	0.61
	TSP-CNN+ResNet50	-	-	0.69
	TSP-CNN+DenseNet	-	-	0.67
		Detection Performance (F_1 score)	Classification Performance (5 Nuclei Classes) (Weighted Average $\widehat{F}$ score)	Combined Performance (Nuclei Detection + 5 Nuclei Classes) (Weighted Average $\widehat{F}$ score)
CRC-CDC Dataset (Proposed) (5 Nuclei Classes) (38,984 Annotated Nuclei)	SC-CNN (Sirinukunwattana et al. (2016))	0.83	0.86	**0.78**
	TSP-CNN(Tofighi et al. (2019))	0.87	-	-
	TSP-CNN+SC-CNN	-	-	**0.80**
	ResNet50 (He et al. (2016))	-	0.77	-
	DenseNet (Huang et al. (2017a))	-	0.76	-
	SC-CNN+ResNet50	-	-	0.71
	SC-CNN+DenseNet	-	-	0.68
	TSP-CNN+ResNet50	-	-	0.74
	TSP-CNN+DenseNet	-	-	0.72
		Detection Performance (F_1 score)	Classification Performance (5 Nuclei Classes) (Weighted Average $\widehat{F}$ score)	Combined Performance (Nuclei Detection + 5 Nuclei Classes) (Weighted Average $\widehat{F}$ score)
CRC-CDC Dataset (Proposed) (5 Nuclei Classes) (38,984 Annotated Nuclei) (Patient-Level Separation)	SC-CNN (Sirinukunwattana et al. (2016))	0.82	0.83	**0.75**
	TSP-CNN(Tofighi et al. (2019))	0.86	-	-
	TSP-CNN+SC-CNN	-	-	**0.79**
	ResNet50 (He et al. (2016))	-	0.71	-
	DenseNet (Huang et al. (2017a))	-	0.69	-
	SC-CNN+ResNet50	-	-	0.67
	SC-CNN+DenseNet	-	-	0.65
	TSP-CNN+ResNet50	-	-	0.70
	TSP-CNN+DenseNet	-	-	0.69

pert pathologists (**KB and KH**) for seven distinct tissue phenotypes. Out of 20 WSIs, the tumor regions were marked from five WSIs, stroma from three WSIs, complex stroma from four WSIs, smooth muscle from two WSIs, Inflammatory from three WSIs, Benign from four WSIs, while the Debris regions were marked from four WSIs. Using these boundaries, patches were

extracted and each patch was assigned a unique label based on majority of its content. Each patch and its label were then inspected by the same pathologists and verified correctness of the patch and its label. Patches containing significant pixels from more than one phenotype were discarded. Therefore, in the resulting dataset, patch of a particular phenotype mostly contains one tissue phenotype however, we cannot avoid the presence of small percentage of other phenotypes in addition to the identified label. Overal, the dataset consists of 50K patches each for Tumor (Tu), Stroma (St), Complex Stroma (CS), and Smooth Muscle (SM) phenotypes. Each of the Benign (Be) and Inflammatory (In) phenotypes consist of 30K patches while the Debris (De) class consists of 20K patches. Following the Kather et al. (2016), the patch size is fixed to 150×150 pixels extracted at $20\times$ magnification level and the patches are non-overlapping. Fig. 5 shows some sample tissue images from the proposed dataset.

To test the generalization of the proposed tissue phenotyping algorithm and compared methods, two experimental settings are used. In the first experiment, 70% patches of each tissue phenotype are randomly selected for training and remaining 30% are used for testing. In the second experiment, patient-level separation is maintained by keeping 14 patients data for training and remaining 6 patients data for testing. The number of patches are kept same in both experiments.

4.3. *Cell Detection and Classification Performance on CRC-CDC Dataset*

For cell detection and classification, we compare the performance of the SC-CNN and TSP-CNN methods on CRCHistoPhenotypes dataset and on our proposed CRC-CDC dataset with nuclei-level separation and with patient-level separation. The SC-CNN has two different networks one for nuclei detection and one for nuclei classification as discussed in Sec. 3.1 while, TSP-CNN has only detection network therefore, we also combined SC-CNN classification network with TSP-CNN to get the combined detection and classification performance. In addition to SC-CNN cell classification network, we also evaluated the performance of ResNet50 and DenseNet for cell classification. The SC-CNN detection network is re-trained on CRC-CDC dataset while TSP-CNN [4] was pre-trained on CRCHistoPhenotypes dataset. The classification networks including SC-CNN, ResNet50, and DenseNet are trained on CRC-CDC dataset for five distinct nuclei classes.

For SC-CNN networks, we use input patch size of 27×27 pixels containing a single cell, cropped by keeping the nuclei at the centre position. We also use data augmentations in which we rotate patches (0^0, 90^0, 180^0, and 270^0) and flip along vertical and horizontal axis to make the networks orientation independent. We also extracted multiple patches for the same nuclei at shifted locations to make the networks shift invariant and to improve the cell localization. For network training, we used cross entropy loss function with stochastic gradient descent with momentum of 0.9, 120 epochs, and learning rate was set as 10^{-3}. For ResNet50 and DenseNet, the input patch size is enlarged as required by the respective network.

We followed the same two-fold cross validation procedure for performance evaluation as suggested by (Sirinukunwattana et al. (2016)). The nuclei detected within 6-pixel distance from the ground truth locations are considered as True Positives (TP). The nuclei detection performance is evaluated using F_1 measure score as:

$$F_1 = 2 \times \frac{Precision \times Recal}{Precision + Recall}, \text{ where}$$
$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{14}$$

and False Positives (FP) are incorrectly detected nuclei, while False Negatives (FN) are miss-detected nuclei. The aim is to maximize F_1 measure so that its value is close to one.

Table 1 shows the performance of nuclei detection in terms of F_1 score averaged over all test images. For CRCHistoPhenotypes dataset, TSP-CNN has obtained the highest F_1 score of 0.85 while SC-SNN has obtained 0.80 F_1 score. On CRC-CDC dataset with nuclei-level separation experiment, TSP-CNN has obtained 0.87 while SC-CNN has obtained 0.83 average F_1 score. For the CRC-CDC dataset with patient-level separation, TSP-CNN has obtained 0.86 while SC-CNN has obtained 0.82 average F_1 score. As compared to nuclei-level separation CRC-CDC, the performance is 1% less for both methods which demonstrates that the patient-level separation has posed an equal challenge for both methods.

To evaluate the cell classification performance, the weighted average $\widehat{F}$ score is used as follows:

$$\widehat{F} = \frac{\sum_i^{c_l} n_i F_{1(i)}}{n}, \tag{15}$$

where c_l is the number of cell classes, n_i is the number of test samples in i-th class, and n is the total number of test samples. Table 1 shows the comparison of cell classification performance using ground truth cell detection as well as combined detection and classification performed by compared methods.

On CRCHistoPhenotypes dataset for 4 nuclei classes, SC-CNN has obtained 0.78 while ResNet50 has obtained 0.74 weighted average $\widehat{F}$ score using ground truth nuclei annotations. The combined performance of TSP-CNN+SC-CNN is 0.73 while SC-CNN has obtained 0.69 weighted average $\widehat{F}$ score. The combined performance is reduced because the nuclei detection by respective networks is performed instead of using ground truth annotations.

On CRC-CDC dataset with nuclei-level separation for five nuclei classes including Tumor epithelial (T), Normal epithelial (N), Inflammatory (I), Spindle-shaped (S), and Debris (D), the SC-CNN has obtained 0.86 and ResNet50 has obtained 0.77 weighted average $\widehat{F}$ score using ground truth nuclei annotations. However, the combined performance of TSP-CNN+SC-CNN has remained 0.80 while the second best performing method is SC-CNN by obtaining weighted average $\widehat{F}$ score of 0.78.

For the CRC-CDC dataset with patient-level separation, SC-CNN has obtained 0.83 and ResNet50 has obtained 0.71

[4] http://php.scripts.psu.edu/mqt5352/SP-CNN/SP-CNN.php

weighted average $\widehat{F}$ score using ground truth nuclei annotations. In case of combined performance, TSP-CNN+SC-CNN has obtained 0.79 while SC-CNN has obtained 0.75 weighted average $\widehat{F}$ score. For patient-level separation, the performance of SC-CNN has reduced by 3% while TSP-CNN+SC-CNN is reduced by 1% compared to nuclei-level separation combined performance results.

The combined performance of cell detection and classification has remained best for TSP-CNN+SC-CNN while SC-CNN has remained the second best performing method. Both of these methods are used as a pre-processing steps for the proposed tissue phenotyping algorithm.

4.4. *Clustering Performance*

The clustering performance of the proposed algorithm is evaluated using three different clustering measures including Normalized Mutual Information (NMI) Schütze et al. (2008), Adjust Rand Index (ARI) Zhao and Karypis (2004), and Purity Zhao and Karypis (2004) on CCT and CRC-TP with patch-level separation datasets. The NMI is computed as follows:

$$\text{NMI} = \frac{\sum_{i=1}^{c_l}\sum_{j=1}^{c_f} m_{i,j}\log(nm_{i,j}/m_i m_j)}{\sum_{i=1}^{c_l} m_i \log(m_i/n)\sum_{j=1}^{c_f} m_j \log(m_j/n)}, \tag{16}$$

where c_l are the number of classes in the ground truth, c_f are the number of found classes, $m_{i,j}$ is the two dimensional joint probability of ground truth and the found classes, m_i is the marginal probability of ground truth, and m_j is the marginal probability of found classes. A higher value of NMI shows better clustering performance of an algorithm. The ARI represents the percentage of TP and True Negative (TN) decisions over testing samples as defined below:

$$ARI(c_l, c_f) = \frac{1}{c_l}\sum_{i=1}^{c_l}\frac{TP_i + TN_i}{n}, \tag{17}$$

The ARI value is in the range of [0,1] and higher values represent better clustering performance. Similarly, the Purity measure represents the percentage of the total number of nodes clustered correctly. Let $\Omega = \{w_1, \cdots, w_k\}$ be the computed clustered labels and $C = \{c_1, \cdots, c_k\}$ be ground truth class labels, the purity is defined as below:

$$Purity(\Omega, C) = \frac{1}{n}\sum_{k}\max_{j}|w_k \cap c_j|, \tag{18}$$

where $|w_k \cap c_j|$ represents the number of nodes in the intersection of w_k and c_j.

Table 2 shows the performance comparison of our proposed tissue phenotyping algorithms with other state-of-the-art methods on CCT dataset having six tissue phenotypes. On the average, TPCD-4 has remained the best performer for all the three clustering measures NMI, ARI, and Purity, while TPCD-3 has remained the second best performer. It is because of the better performance of TSP-CNN for cell detection in TPCD-4 algorithm compared to SC-CNN in TPCD-3. By considering the tissue phenotype-wise performance, TPCD-4 has remained the best performer for the Tumor, Stroma, and Debris tissue components on all three measures. For the complex stroma tissue phenotype, TPCD-4 obtained the best performance for NMI and ARI while for Purity measure SPM has remained the best performer. For Mucosa, TPCD-4 performed best for ARI and Purity measures while for NMI, B6F-SVM has remained the best performer. For the Lymphocytes tissue phenotype, TPCD-3 and TPCD-4 both remained the best performers for ARI and Purity while for NMI, B6F-SVM performed best.

Table 3 shows the performance comparison of different clustering methods on CRC-TP dataset having seven distinct tissue phenotypes. On the average, TPCD-4 obtained the best performance while TPCD-3 remained the second best performer on all three measures. In terms of tissue phenotype-wise performance, TPCD-4 has remained best performer for Tumor and Complex Stroma phenotypes for all three measures. For Stroma tissue phenotype, TPCD-4 has remained best for NMI and ARI measures while for Purity, ResNet achieved the best performance. For the Benign class, TPCD-4 performed best for NMI and ARI, while TPCD-2 performed best in terms of Purity. For Debris class, TPCD-4 remained best for NMI and ARI while for Purity the TPCD-2 remained the best performer. For inflammatory class, the TPCD-4 performed best in terms of NMI while TPCD-2 remained best in terms of ARI and Purity. For the Smooth Muscle tissue phenotype, TPCD-4 performed best in terms of NMI and Purity while TPCD-2 performed best in terms of ARI and Purity.

4.5. *Performance Comparison on CCT Dataset*

We compare the performance of the proposed algorithms with the current state-of-the-art methods in terms of F_1 score for tissue phenotyping. In CCT dataset, all tissue classes have an equal number of instances therefore, the average F_1 and weighted average $\widehat{F}$ scores remain the same. Since, this dataset has only tissue phenotype labels at patch-level therefore, the cell detection and classification is performed by using SC-CNN network trained on CRC-CDC dataset. In order to remove the stain differences between CRC-CDC and CCT datasets, we have used the Macenko method for stain normalization as a pre-processing step Macenko et al. (2009).

Table 4 shows the comparative performance in terms of average F_1 score of six tissue phenotypes on CCT dataset. The proposed algorithms TPCD-4 and TPCD-3 have remained the best performers by achieving 94.5% and 94.0% average F_1 score. The TPCD-2 has obtained average F_1 score of 92.5% while the nearest competitors are DenseNet and B6F-SVM which obtained 89.5% and 89.7%. For tumor phenotype, the GCN method, GCN-AF, has obtained 0.86 F_1 score using deep features-based Euclidean distance graph and using our proposed cell features-based graph (Eq. 3), the GCN-AF has obtained 0.88 F_1 score. All the compared methods have obtained less than 0.90 F_1 score except for DenseNet and ResNet101 both obtaining 0.91 F_1 score. The proposed variants TPCD-2, TPCD-3, and TPCD-4 have obtained 0.92, 0.95, and 0.95 F_1 score, respectively.

For Stroma phenotype, majority of the compared methods have obtained less than 0.90 F_1 score except KM-CD (0.92).

Table 2: Clustering performance of the proposed algorithm in terms of NMI, ARI, and Purity measures on CCT dataset (Kather et al. (2016)) and its comparison with state-of-the-art methods. The best performer is shown in red and the best second best performer is shown in blue color, respectively.

Methods	Clustering Measures	Tumor	Stroma	Complex Stroma	Mucosa	Debris	Lymphocytes	Average
ResNet	NMI	0.67	0.79	0.69	0.68	0.74	0.70	0.71
	ARI	0.91	0.93	**0.96**	0.95	0.90	0.94	0.93
	Purity	0.88	0.95	0.79	0.80	0.88	0.92	0.87
DL-KLdiv	NMI	0.70	0.68	0.74	0.66	0.68	0.62	0.68
	ARI	0.85	0.82	0.84	0.82	0.87	0.86	0.84
	Purity	0.77	0.80	0.61	0.58	0.77	0.78	0.71
MobileNet	NMI	0.71	0.65	0.59	0.61	0.78	0.67	0.66
	ARI	0.88	0.90	0.93	0.90	0.89	0.92	0.90
	Purity	0.77	0.65	0.63	0.67	0.69	0.63	0.67
SCD	NMI	0.66	0.56	0.57	0.65	0.69	0.68	0.63
	ARI	0.85	0.86	0.87	0.90	0.88	0.86	0.87
	Purity	0.48	0.77	0.50	0.62	0.60	0.64	0.60
B5F-SVM	NMI	0.70	0.74	0.67	0.69	0.75	0.77	0.72
	ARI	0.90	0.93	0.92	0.95	0.94	0.93	0.92
	Purity	0.88	0.82	0.80	0.88	0.82	0.91	0.85
B6F-SVM	NMI	0.71	0.70	0.79	**0.82**	0.74	**0.87**	0.77
	ARI	**0.93**	0.96	0.95	**0.97**	**0.96**	0.95	0.95
	Purity	0.88	0.82	0.80	0.89	0.81	0.91	0.85
SDLs	NMI	0.68	0.75	0.72	0.64	0.71	0.76	0.71
	ARI	0.85	0.87	0.89	0.90	0.92	0.91	0.89
	Purity	0.72	0.80	0.70	0.74	0.91	0.87	0.79
DFOD	NMI	0.73	0.67	0.64	0.71	0.78	0.80	0.72
	ARI	0.89	0.91	0.94	0.90	0.92	0.94	0.91
	Purity	0.81	0.90	0.81	0.87	**0.97**	0.73	0.84
SPM	NMI	0.69	0.70	0.62	0.66	0.68	0.65	0.66
	ARI	0.79	0.82	0.77	0.83	0.81	0.80	0.80
	Purity	0.89	0.94	**0.90**	0.76	0.82	0.83	0.85
SRC	NMI	0.54	0.59	0.69	0.54	0.52	0.60	0.58
	ARI	0.82	0.80	0.78	0.83	0.80	0.86	0.81
	Purity	0.76	0.88	0.57	0.65	0.70	0.83	0.73
KM-CD	NMI	**0.77**	0.72	0.75	0.72	0.78	0.82	0.76
	ARI	0.87	0.90	0.92	0.95	0.94	**0.97**	0.92
	Purity	0.90	0.93	0.67	0.82	0.87	0.97	0.86
Subspace Clustering	NMI	0.62	0.51	0.69	0.51	0.49	0.52	0.55
	ARI	0.80	0.83	0.81	0.85	0.80	0.87	0.82
	Purity	0.70	0.45	0.61	0.71	0.60	0.67	0.62
SHIRC	NMI	0.69	0.71	0.77	0.80	0.79	0.78	0.75
	ARI	0.84	0.87	0.85	0.82	0.88	0.90	0.86
	Purity	0.83	0.76	0.75	0.88	0.78	0.84	0.80
TPCD-2	NMI	0.71	0.73	0.81	0.76	0.84	0.80	0.77
	ARI	0.92	0.91	0.95	0.94	**0.96**	**0.97**	0.94
	Purity	**0.95**	0.96	0.78	**0.91**	0.96	**0.98**	0.92
Proposed TPCD-3	NMI	0.76	**0.78**	**0.85**	0.80	**0.87**	0.84	**0.81**
	ARI	**0.98**	**0.98**	**0.96**	**0.97**	**0.99**	**0.99**	**0.97**
	Purity	**0.98**	**0.97**	0.83	**0.94**	**0.98**	**0.99**	**0.94**
Proposed TPCD-4	NMI	**0.78**	**0.80**	**0.86**	**0.81**	**0.89**	**0.85**	**0.83**
	ARI	**0.98**	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**	**0.98**
	Purity	**0.98**	**0.98**	**0.85**	**0.94**	**0.98**	**0.99**	**0.95**

The proposed algorithms TPCD-1, TPCD-2, TPCD-3, and TPCD-4 have obtained 0.92, 0.94, 0.95, and 0.96 F_1 score, respectively. The Complex Stroma is one of the difficult tissue phenotypes for discriminating it from the tumor class. The DenseNet has produced the best F_1 score of 0.89 while our proposed algorithm variants SVM-CellFeatures, TPCD-3, and TPCD-4 have obtained 0.87, 0.87, and 0.88 F_1 score, respectively. The proposed algorithms TPCD-3 and TPCD-4 are successful in obtaining comparative performance by leveraging the potential cell-cell connections between cellular components while most existing methods are suffered from performance degradation due to texture features which are not able to handle rich tissue heterogeneity.

In the case of Mucosa tissue, B6F-SVM has obtained the best performance of 0.94 while our proposed algorithms TPCD-2, TPCD-3, and TPCD-4 have obtained 0.90, 0.92, and 0.93 F_1 sore, respectively. Most of the existing methods have achieved less than 0.90 F_1 score for Mucosa tissue which shows that Mucosa tissue pose a significant challenge to all the compared methods. The Debris and Lymphocytes are well differentiated phenotypes therefore; these classes do not pose a significant challenge for the majority of the compared methods. The TPCD-2, TPCD-3, and TPCD-4 algorithms have obtained the best F_1 scores of 0.96, 0.97, and 0.97, respectively, for Debris and 0.97, 0.98, and 0.98, respectively, for Lymphocytes. In Debris, the nearest competitor is ResNet101 obtaining 0.92 while in case of Lymphocytes KM-CD obtained 0.95 compared to 0.98 obtained by TPCD-3 and TPCD-4.

The proposed variant TPCD-Hist which uses number of cell-types as a feature vector is also not able to obtain the comparative performance. The better performance achieved by our proposed algorithm variants is mainly leveraged by proposed cell-cell connections features.

Table 3: Clustering performance of the proposed algorithm in terms of NMI, ARI, and Purity measures on CRC-TP dataset and its comparison with state-of-the-art methods. The best performer is shown in red and the best second best performer is shown in blue color, respectively.

Methods	Clustering Measures	Tumor	Stroma	Complex Stroma	Benign	Debris	Inflammatory	Smooth Muscle	Average
ResNet	NMI	0.71	**0.82**	0.73	0.82	0.78	0.72	0.73	0.75
	ARI	0.93	0.94	**0.97**	0.96	0.92	**0.96**	0.94	0.94
	Purity	0.90	**0.96**	0.82	0.81	0.90	**0.93**	**0.95**	0.89
DL-KLdiv	NMI	0.72	0.69	0.76	0.69	0.71	0.64	0.67	0.69
	ARI	0.88	0.84	0.85	0.84	0.89	0.89	0.93	0.87
	Purity	0.79	0.82	0.63	0.61	0.79	0.81	0.83	0.75
MobileNet	NMI	0.75	0.68	0.62	0.65	0.81	0.71	0.75	0.71
	ARI	0.90	0.92	0.95	0.91	0.92	0.93	0.95	0.81
	Purity	0.79	0.68	0.66	0.72	0.75	0.67	0.67	0.70
SCD	NMI	0.68	0.59	0.59	0.68	0.72	0.72	0.74	0.67
	ARI	0.88	0.89	0.91	0.93	0.92	0.90	0.90	0.90
	Purity	0.51	0.79	0.53	0.64	0.62	0.67	0.69	0.63
B5F-SVM	NMI	0.72	0.75	0.69	0.72	0.77	0.79	0.82	0.74
	ARI	0.92	0.94	0.95	0.95	0.96	0.94	0.94	0.94
	Purity	0.90	0.84	0.82	0.90	0.86	**0.93**	**0.95**	0.88
B6F-SVM	NMI	0.73	0.69	0.76	0.80	0.71	0.84	0.85	0.76
	ARI	0.90	0.91	0.94	0.94	0.91	0.89	0.91	0.91
	Purity	0.89	0.84	0.82	0.85	0.78	0.87	0.89	0.84
SDLs	NMI	0.67	0.70	0.68	0.61	0.68	0.71	0.73	0.68
	ARI	0.81	0.85	0.87	0.86	0.84	0.89	0.90	0.86
	Purity	0.74	0.82	0.69	0.72	0.86	0.83	0.85	0.78
DFOD	NMI	0.74	0.65	0.66	0.69	0.74	0.78	0.80	0.72
	ARI	0.85	0.87	0.91	0.86	0.87	0.91	0.93	0.88
	Purity	0.80	0.89	0.79	0.84	**0.92**	0.71	0.73	0.81
SPM	NMI	0.64	0.67	0.59	0.62	0.66	0.62	0.63	0.63
	ARI	0.77	0.80	0.74	0.79	0.78	0.78	0.80	0.78
	Purity	0.86	0.90	0.87	0.74	0.79	0.80	0.82	0.82
SRC	NMI	0.56	0.57	0.71	0.52	0.50	0.57	0.61	0.57
	ARI	0.80	0.79	0.80	0.85	0.82	0.83	0.84	0.81
	Purity	0.78	0.85	0.59	0.68	0.73	0.85	0.87	0.76
KM-CD	NMI	**0.81**	0.74	0.77	0.74	0.80	0.84	0.85	0.79
	ARI	0.90	0.92	0.94	0.94	0.96	0.95	0.96	0.93
	Purity	0.89	0.88	0.69	0.84	0.86	**0.93**	**0.95**	0.86
Subspace Clustering	NMI	0.59	0.48	0.64	0.49	0.50	0.50	0.53	0.53
	ARI	0.77	0.79	0.75	0.74	0.77	0.78	0.80	0.77
	Purity	0.68	0.48	0.59	0.68	0.56	0.64	0.63	0.60
SHIRC	NMI	0.71	0.72	0.76	0.78	0.75	0.74	0.74	0.74
	ARI	0.81	0.84	0.80	0.80	0.86	0.87	0.90	0.84
	Purity	0.85	0.79	0.77	0.90	0.82	0.86	0.90	0.84
TPCD-2	NMI	0.75	0.74	0.85	0.79	0.85	0.82	0.84	0.80
	ARI	0.94	0.92	0.96	0.95	**0.97**	**0.97**	**0.98**	0.95
	Purity	**0.96**	**0.95**	0.81	**0.93**	**0.97**	**0.95**	**0.96**	**0.93**
Proposed TPCD-3	NMI	0.78	0.80	**0.86**	**0.84**	**0.89**	**0.85**	**0.87**	**0.84**
	ARI	**0.98**	**0.97**	0.95	**0.97**	**0.98**	**0.96**	0.94	**0.96**
	Purity	0.95	0.93	**0.91**	**0.91**	0.89	0.91	**0.95**	**0.92**
Proposed TPCD-4	NMI	**0.83**	**0.84**	**0.89**	**0.85**	**0.91**	**0.88**	**0.89**	**0.87**
	ARI	**0.99**	**0.99**	**0.98**	**0.99**	**0.98**	**0.96**	**0.95**	**0.97**
	Purity	**0.98**	0.94	**0.92**	**0.91**	0.90	0.92	**0.96**	**0.93**

4.6. *Performance Comparison on CRC-TP Dataset*

The evaluations on this dataset are performed in two different settings including patch-level separation and patient-level separation as discussed in section 4.2.3.

4.6.1. *Evaluation with Patch-Level Separation*

Table 5 shows the performance comparison in terms of F_1-score (F_1) for seven distinct tissue phenotypes and weighted average $\widehat{F}$-score ($\widehat{F}$) over all classes with other state-of-the-art methods. The proposed algorithms TPCD-3 and TPCD-4 have performed best on CRC-TP dataset with $\widehat{F}$ of 0.91 and 0.89. The removal of distant cell-cell connections in TPCD-3 has reduced heterogeneity and therefore improved the accuracy by 1.0% as compared to TPCD-2. TPCD-1 has obtained 0.84 $\widehat{F}$ which is still competitive with compared methods. The inclusion of Geodesic distance in TPCD-4 algorithm has caused upto 7.0% increase in accuracy as compared to TPCD-1, therefore Geodesic distance is an important step in our proposed algorithm. Among the compared methods, ResNet101 and DenseNet have achieved an $\widehat{F}$ of 0.87 and 0.86 $\widehat{F}$ close to TPCD-2.

In tumor phenotype, our proposed algorithm TPCD-4 has obtained 0.96 and TPCD-3 has obtained 0.95 which are significantly larger than the compared methods. The nearest competitors are TPCD-2 and ResNet101 both obtaining 0.93 $\widehat{F}$. In stroma phenotype, TPCD-4 and TPCD-3 are the best performing algorithms obtaining a $\widehat{F}$ of 0.94 and 0.93 while our proposed TPCD-2 has obtained 0.90 $\widehat{F}$. The nearest competitor is TPCD-1 which has obtained 0.89 $\widehat{F}$ while among the compared methods ResNet101 obtained 0.87 $\widehat{F}$. In complex stroma phenotype, both TPCD-4 and TPCD-3 algorithms have obtained 0.87 $\widehat{F}$. It is because of the cell detection performance of SC-CNN in TPCD-3 algorithm approached to the performance of

Table 4: Comparative performance of multi-class tissue classification on Colon Cancer Tissue (CCT) dataset (Kather et al. (2016)). Performance is presented in terms of F_1 score for each tissue phenotype and average F_1 score in (%) for all tissue components. The two best results are shown in red and blue fonts respectively.

Classical Methods	Tumor	Stroma	Complex	Mucosa	Debris	Lympho	F_1
KM-CD (Sirinukunwattana et al. (2018))	0.85	0.92	0.71	0.81	0.91	0.95	85.1
Subspace Clustering (Elhamifar and Vidal (2013))	0.62	0.68	0.50	0.56	0.68	0.70	62.5
DL-KLdiv (Mairal et al. (2012))	0.80	0.83	0.64	0.76	0.86	0.70	76.4
SCD (Guha and Ward (2014))	0.59	0.63	0.71	0.69	0.65	0.78	67.4
SRC (Wright et al. (2009))	0.63	0.70	0.73	0.76	0.80	0.79	73.5
B5F-SVM (Kather et al. (2016))	0.86	0.85	0.84	0.86	0.84	0.86	85.2
B6F-SVM (Kather et al. (2016))	0.87	0.88	0.85	**0.94**	0.90	0.90	89.7
DFOD (Vu et al. (2016))	0.81	0.88	0.79	0.80	0.84	0.78	81.3
SHIRC (Srinivas et al. (2014))	0.79	0.81	0.80	0.82	0.80	0.81	80.3
SPM (Lazebnik et al. (2006))	0.85	0.83	0.87	0.80	0.85	0.79	83.6
SDLs (Sarkar and Acton (2018))	0.76	0.80	0.67	0.71	0.91	0.91	79.3
Deep Learning Methods	**Tumor**	**Stroma**	**Complex**	**Mucosa**	**Debris**	**Lympho**	**F_1**
DenseNet (Huang et al. (2017a))	0.91	0.88	**0.89**	0.87	0.91	0.92	89.5
SVM-CNN (Xu et al. (2017))	0.81	0.78	0.80	0.79	0.82	0.80	80.0
ResNet50 (He et al. (2016))	0.83	0.82	0.84	0.82	0.85	0.84	83.7
ResNet101 (He et al. (2016))	0.91	0.88	**0.88**	0.86	0.92	0.91	89.2
MobileNet (Howard et al. (2017))	0.73	0.72	0.71	0.71	0.80	0.75	73.2
Euclidean distance-based Deep GCN Methods	**Tumor**	**Stroma**	**Complex**	**Mucosa**	**Debris**	**Lympho**	**F_1**
GCN-RS (Ying et al. (2018))	0.82	0.80	0.78	0.73	0.82	0.84	79.8
DGI (Veličković et al. (2019))	0.84	0.82	0.80	0.78	0.84	0.83	81.8
GCN-SF (Defferrard et al. (2016))	0.72	0.70	0.68	0.75	0.78	0.80	73.8
SSC-GCN (Kipf and Welling (2017))	0.62	0.59	0.64	0.65	0.72	0.70	65.3
GCN-AF (Bianchi et al. (2019))	0.86	0.85	0.83	0.80	0.85	0.84	83.8
Cell features-based Deep GCN Methods	**Tumor**	**Stroma**	**Complex**	**Mucosa**	**Debris**	**Lympho**	**F_1**
GCN-RS(Ying et al. (2018))	0.85	0.83	0.82	0.78	0.85	0.86	83.1
DGI (Veličković et al. (2019))	0.86	0.84	0.84	0.80	0.86	0.85	84.1
GCN-SF (Defferrard et al. (2016))	0.75	0.72	0.71	0.77	0.81	0.83	76.5
SSC-GCN (Kipf and Welling (2017))	0.66	0.61	0.66	0.68	0.74	0.73	68.0
GCN-AF (Bianchi et al. (2019))	0.88	0.87	0.86	0.82	0.87	0.85	85.8
Proposed Algorithms	**Tumor**	**Stroma**	**Complex**	**Mucosa**	**Debris**	**Lympho**	**F_1**
TPCD-CG	0.69	0.66	0.64	0.68	0.72	0.74	68.8
SVM-CellFeatures	0.85	0.83	0.87	0.80	0.85	0.79	83.1
TPCD-1	0.85	0.92	0.71	0.81	0.91	0.95	85.2
TPCD-2 (Javed et al. (2018))	**0.92**	0.94	0.83	0.90	**0.96**	**0.97**	92.5
TPCD-Hist	0.86	0.82	0.85	0.82	0.90	0.92	86.1
TPCD-3	**0.95**	**0.95**	0.87	0.92	**0.97**	**0.98**	**94.0**
TPCD-4	**0.95**	**0.96**	**0.88**	**0.93**	**0.97**	**0.98**	**94.5**

TSP-CNN in TPCD-4 algorithm for complex stroma phenotype. The nearest competitors are TPCD-2, ResNet101, and DenseNet each obtaining 0.84 $\widehat{F}$.

In Benign tissue phenotype, TPCD-4 and TPCD-3 algorithms have obtained 0.90 and 0.89 $\widehat{F}$ while TPCD-2 has obtained 0.86 $\widehat{F}$. In this case, normal epithelial to normal epithelial cell-cell connections are observed quite higher on the micro-vessels, therefore the removal of distant edges was not helpful in this case. The nearest competitor is ResNet101 which obtained 0.86 $\widehat{F}$. In Debris tissue, the DenseNet and B5F-SVM have achieved the best performance of 0.91 $\widehat{F}$, while TPCD-4 has obtained 0.90 $\widehat{F}$. The nearest competitors are TPCD-2 and ResNet50 both obtaining 0.88 $\widehat{F}$. In Inflammatory tissue type, ResNet101 obtained the best performance of 0.95 $\widehat{F}$, while DenseNet and B5F-SVM methods obtained 0.92 $\widehat{F}$. Our proposed algorithm TPCD-4 obtained $\widehat{F}$ of 0.88 for inflammatory tissue. In the case of Smooth Muscle phenotype, TPCD-4 and TPCD-2 obtained 0.88 and 0.87 $\widehat{F}$ while TPCD-3 obtained 0.86 $\widehat{F}$. The removal of distant cellular edges has shown performance degradation in this tissue component. An improved cell detection performance with TPCD-2 would have resulted in further performance improvement like TPCD-4.

Table 5: Comparative performance of multi-class tissue phenotyping on CRC-TP dataset using patch-level separation between training and testing splits. Performance is presented in terms of F_1 score for each tissue phenotype and weighted average $\widehat{F}$ score for all tissue components. The two best results are shown in red and blue fonts respectively.

Methods	Tu	St	CS	Be	De	In	SM	$\widehat{F}$ score
KM-CD (Sirinukunwattana et al. (2018))	0.72	0.79	0.62	0.73	0.80	0.78	0.79	0.73
Subspace Clustering (Elhamifar and Vidal (2013))	0.48	0.62	0.45	0.46	0.64	0.65	0.63	0.55
DL-KLdiv (Mairal et al. (2012))	0.62	0.65	0.60	0.79	0.73	0.76	0.70	0.68
SCD (Guha and Ward (2014))	0.60	0.61	0.55	0.69	0.81	0.79	0.69	0.65
B5F-SVM (Kather et al. (2016))	0.86	0.77	0.73	0.75	**0.91**	**0.92**	0.78	0.80
SRC (Wright et al. (2009))	0.73	0.75	0.65	0.60	0.85	0.66	0.64	0.69
DFOD (Vu et al. (2016))	0.84	0.81	0.73	0.71	0.78	0.74	0.74	0.77
SHIRC (Srinivas et al. (2014))	0.78	0.75	0.61	0.65	0.68	0.78	0.69	0.71
SPM (Lazebnik et al. (2006))	0.82	0.80	0.70	0.85	0.83	0.84	0.74	0.79
SDLs (Sarkar and Acton (2018))	0.86	0.83	0.70	0.72	0.81	0.80	0.70	0.77
Deep Learning Methods	**Tu**	**St**	**CS**	**Be**	**De**	**In**	**SM**	**$\widehat{F}$ score**
ResNet50 (He et al. (2016))	0.81	0.81	0.78	0.81	0.88	0.87	0.85	0.82
ResNet101 (He et al. (2016))	0.93	0.87	**0.84**	0.86	0.86	**0.95**	0.80	0.87
DenseNet (Huang et al. (2017a))	0.84	0.86	**0.84**	0.82	**0.91**	**0.92**	0.85	0.86
SVM-CNN (Xu et al. (2017))	0.80	0.78	0.80	0.73	0.84	0.86	0.79	0.80
MobileNet (Howard et al. (2017)	0.79	0.79	0.68	0.81	0.76	0.82	0.76	0.77
Euclidean distance-based Deep GCN Methods	**Tu**	**St**	**CS**	**Be**	**De**	**In**	**SM**	**$\widehat{F}$ score**
GCN-RS (Ying et al. (2018))	0.71	0.69	0.65	0.68	0.73	0.76	0.75	0.71
DGI (Veličković et al. (2019))	0.74	0.70	0.64	0.70	0.72	0.78	0.77	0.72
GCN-SF(Defferrard et al. (2016))	0.59	0.60	0.54	0.64	0.68	0.66	0.61	0.60
SSC-GCN (Kipf and Welling (2017))	0.60	0.63	0.58	0.60	0.65	0.63	0.60	0.61
GCN-AF (Bianchi et al. (2019))	0.78	0.80	0.76	0.75	0.81	0.80	0.77	0.78
Cell features-based Deep GCN Methods	**Tu**	**St**	**CS**	**Be**	**De**	**In**	**SM**	**$\widehat{F}$ score**
GCN-RS (Ying et al. (2018))	0.74	0.72	0.69	0.73	0.77	0.75	0.79	0.74
DGI (Veličković et al. (2019))	0.77	0.72	0.69	0.74	0.75	0.81	0.82	0.75
GCN-SF (Defferrard et al. (2016))	0.66	0.63	0.60	0.66	0.71	0.69	0.64	0.65
SSC-GCN (Kipf and Welling (2017))	0.63	0.64	0.62	0.65	0.69	0.68	0.63	0.64
GCN-AF (Bianchi et al. (2019))	0.79	0.83	0.78	0.78	0.83	0.84	0.80	0.80
Deep Learning Methods	**Tu**	**St**	**CS**	**Be**	**De**	**In**	**SM**	**$\widehat{F}$ score**
SVM-CellFeatures	0.80	0.74	0.82	0.72	0.81	0.80	0.78	0.78
TPCD-1	0.90	0.89	0.80	0.80	0.85	0.82	0.80	0.84
TPCD-2 (Javed et al. (2018))	0.93	0.90	**0.84**	0.86	0.88	0.84	**0.87**	0.88
TPCD-Hist	0.85	0.83	0.82	0.78	0.81	0.82	0.79	0.82
TPCD-3	**0.95**	**0.93**	**0.87**	**0.89**	0.87	0.85	0.86	**0.89**
TPCD-4	**0.96**	**0.94**	**0.87**	**0.90**	**0.90**	0.88	**0.88**	**0.91**

4.6.2. *Evaluation with Patient-Level Separation*

Table 6 shows the tissue phenotyping performance comparison on CRC-TP dataset with patient-level separation. Overall, TPCD-4 and TPCD-3 obtained 0.84 and 0.83 $\widehat{F}$. Compared to patch-level separation results on Table 5, the performance of all the compared methods is significantly reduced. It is because the testing dataset is completely unseen by the cell detection and classification networks in the proposed TPCD algorithms which caused accuracy degradation of 6.0% for TPCD-2, TPCD-3, and TPCD-4. Compared to ResNet101, DenseNet, and GCN-AF with cell features-based graph construction, the accuracies are reduced by 6.0%, 7.0%, and 4.0%, respectively which are also in the same range as compared to proposed algorithms. Moreover, the TPCD-3 and TPCD-4 algorithms have performed better than the other compared methods on Tu, St, CS, Be, and De tissue phenotypes, respectively. Overall, the patient-level separation is more challenging compared to the patch-level separation across training and testing data.

4.7. *Visual Evaluation*

The qualitative classification results are thoroughly examined by experienced pathologist (**KB**) and found to match with manual assessment. The results of the proposed algorithm are overlaid on the WSI taken from CRC-TP dataset as shown in Figs.

Table 6: Comparative performance of multi-class tissue phenotyping on CRC-TP dataset using patient-level separation between training and testing splits. Performance is presented in terms of F_1 score for each tissue phenotype and weighted average $\widehat{F}$ score for all tissue components. The two best results are shown in red and blue fonts respectively.

Methods	Tu	St	CS	Be	De	In	SM	$\widehat{F}$ score
KM-CD (Sirinukunwattana et al. (2018))	0.66	0.75	0.58	0.68	0.76	**0.82**	**0.82**	0.72
B5F-SVM (Kather et al. (2016))	0.78	0.71	0.69	0.75	0.81	0.80	0.73	0.74
DFOD (Vu et al. (2016))	0.78	0.77	0.68	0.65	0.71	0.69	0.64	0.71
SHIRC (Srinivas et al. (2014))	0.70	0.71	0.57	0.59	0.62	0.72	0.62	0.66
SDLs (Sarkar and Acton (2018))	0.80	0.75	0.64	0.67	0.77	0.73	0.62	0.71
Deep Learning Methods	Tu	St	CS	Be	De	In	SM	$\widehat{F}$ score
ResNet101 (He et al. (2016))	0.86	0.82	0.79	0.78	0.81	**0.84**	0.77	0.81
DenseNet (Huang et al. (2017a))	0.80	0.79	0.77	0.76	0.82	**0.82**	0.79	0.79
Euclidean distance-based Deep GCN Methods	Tu	St	CS	Be	De	In	SM	$\widehat{F}$ score
GCN-RS (Ying et al. (2018))	0.63	0.59	0.58	0.62	0.66	0.70	0.70	0.63
DGI (Veličković et al. (2019))	0.65	0.63	0.59	0.66	0.67	0.72	0.70	0.64
GCN-SF (Defferrard et al. (2016))	0.52	0.54	0.49	0.58	0.60	0.59	0.55	0.54
SSC-GCN (Kipf and Welling (2017))	0.52	0.57	0.51	0.54	0.55	0.56	0.53	0.54
GCN-AF (Bianchi et al. (2019))	0.72	0.73	0.68	0.64	0.76	0.75	0.72	0.71
Cell features-based Deep GCN Methods	Tu	St	CS	Be	De	In	SM	$\widehat{F}$ score
GCN-RS (Ying et al. (2018))	0.66	0.64	0.60	0.65	0.70	0.67	0.71	0.66
DGI (Veličković et al. (2019))	0.72	0.66	0.64	0.66	0.68	0.75	0.76	0.70
GCN-SF Defferrard et al. (2016))	0.60	0.54	0.55	0.59	0.64	0.60	0.57	0.58
SSC-GCN (Kipf and Welling (2017))	0.55	0.56	0.54	0.58	0.60	0.60	0.57	0.56
GCN-AF (Bianchi et al. (2019))	0.72	0.75	0.74	0.71	0.79	0.78	0.72	0.74
Deep Learning Methods	Tu	St	CS	Be	De	In	SM	$\widehat{F}$ score
SVM-CellFeatures	0.76	0.69	0.77	0.66	0.75	0.77	0.72	0.73
TPCD-1	0.85	0.83	0.76	0.77	0.80	0.78	0.76	0.80
TPCD-2 (Javed et al. (2018))	0.85	0.84	0.79	0.81	0.80	0.79	**0.82**	**0.82**
TPCD-Hist	0.81	0.77	0.77	0.73	0.76	0.75	0.72	0.76
TPCD-3	**0.88**	**0.85**	**0.82**	**0.85**	**0.82**	0.80	0.79	**0.83**
TPCD-4	**0.87**	**0.87**	**0.85**	**0.83**	**0.83**	**0.82**	**0.83**	**0.84**

6 and 7. Non-overlapping patches of 150×150 pixels from the test WSI with patient-level separation are extracted and phenotyped by the TPCD-4 algorithm. The predicted labels are shown by a distinct colors which are overlaid on the original WSI (Fig. 6 (b) and Fig. 7 (b)). The resulting color-coded classification maps are smoothed by a median filter to remove the blocky effects from patch-level classification.

4.8. Computational Complexity

Execution times are compared on a machine with Intel core i9 processor and 128GB RAM. The average cell detection time is 0.21 sec for 150×150 patch and classification time is 0.08 sec. On a same patch size, the Delaunay triangulation takes 0.2 sec while the feature vector extraction takes 0.04 seconds. Complexity of 2-D Delaunay triangulation is $O(clog(c))$, where c is the number of cells detected in a patch. Complexity of graph construction is $O(mn^2)$, where n is the number of patches and m is the size of feature vector. Since the chi-square distance is symmetric, each patch pair distance computation is required only once reducing the overall computation to half. An all pair shortest distance algorithm proposed by Pettie and Ramachandran (2005) has time complexity of $O(mnlog(\alpha(n,s)))$, where s is the number of edges in the patch graph and $\alpha(n,s)$ is a slowly growing function. Jiang et al. (2011) have proposed a relatively faster algorithm which takes $O(\mu(n)loglog(n))$ time on a random scale free network with n vertices. Our implementation of the algorithm used in this step takes $O(mtn^2)$ where t is the number of iterations. We observe that the algorithm converges in less than 5 iterations.

5. Conclusions

In this work, a novel semi-supervised cellular community detection algorithm is proposed for tissue phenotyping based on

cell detection and classification, and clustering of image patches into biologically meaningful communities. First deep neural networks are used for cell detection and classification and then based on potential cell-cell connections between these cells, feature vectors are computed at the patch level. These feature vectors are then used to construct a patch level network using chi-square distance such that each node is a patch in WSI and edges have weights inversely proportional to the distance between the feature vectors. In this network, geodesic distances are computed which are then used to compute node clusters such that each cluster corresponds to a particular tissue phenotype. The proposed algorithm has exhibited better performance than end-to-end deep learning methods as well as several existing algorithms based on handcrafted features.

We showed that the proposed approach was able to achieve better performance mainly because it uses both deep learning and handcrafted features which complement each other. Also the proposed potential cell-cell connections features are biologically more meaningful than the texture-based features used in most existing methods. The concept of constructing a graph and then using geodesic distance for community detection has also significantly contributed to the performance. It is because the graph based approaches work well even if the underlying classes are not linearly separable. The geodesic distance has also performed similar to kernels projecting data to higher dimensional spaces such that the classes become linearly separable. Owing to all these novel steps, the proposed algorithm was able to achieve superior classification accuracy on an existing as well as newly proposed large scale tissue phenotyping dataset.

This new dataset will soon be made publicly available with two experimental settings including patch-level separation and patient-level separation between training and testing splits. Currently, we have used five distinct cell classes including tumor epithelial, normal epithelial, spindle-shaped, necrotic, and inflammatory. Addition of further cellular components such as blood cells may result in performance improvements and also reveal more micro-level tissue communities. The proposed algorithm can potentially be used on large number of WSIs of different cohorts for separating tissue communities. Tissue phenotyping in a WSI can aid with understanding the contents of the WSI and form the basis of comprehensive digital profiling of spatial patterns in the tumor microenvironment associated with cancer subtypes in terms of survival and clinical outcomes.

Acknowledgment

This work was supported by the Medical Research Council [MR/P015476/1].

References

Alberts, B., Roberts, K., Lewis, J., Bray, D., Hopkin, K., Johnson, A.D., Walter, P., Raff, M., 2015. Essential cell biology. Garland Science.

Alizadeh, A.A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., et al., 2015. Toward understanding and exploiting tumor heterogeneity. Nature medicine 21, 846.

Bejnordi, B.E., Mullooly, M., Pfeiffer, R.M., Fan, S., Vacek, P.M., Weaver, D.L., Herschorn, S., Brinton, L.A., van Ginneken, B., Karssemeijer, N., et al., 2018. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. Modern Pathology , 1.

Bianchi, F.M., Grattarola, D., Livi, L., Alippi, C., 2019. Graph Neural Networks with distributed ARMA filters. CoRR .

Bianconi, F., Álvarez-Larrán, A., Fernández, A., 2015. Discrimination between tumour epithelium and stroma via perception-based features. Neurocomputing 154, 119–126.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering, in: Adv. NIPS.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: IEEE CVPR.

Dorogovtsev, S.N., Mendes, J.F., 2002. Evolution of networks. Advances in physics 51, 1079–1187.

Du, Y., Zhang, R., Zargari, A., Thai, T.C., Gunderson, C.C., Moxley, K.M., Liu, H., Zheng, B., Qiu, Y., 2018. Classification of tumor epithelium and stroma by exploiting image features learned by deep convolutional neural networks. An. of BE , 1–12.

Elhamifar, E., Vidal, R., 2013. Sparse subspace clustering: Algorithm, theory, and applications. IEEE T-PAMI 35, 2765–2781.

Fortunato, S., 2010. Community detection in graphs. Phy. Rep. 486, 75–174.

Guha, T., Ward, R.K., 2014. Image similarity using sparse representation and compression distance. IEEE T-M 16, 980–987.

Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., Samatova, N., 2014. Community detection in large-scale networks: a survey and empirical evaluation. WIR: Comp. Stat. 6, 426–439.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: IEEE CVPR.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 .

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017a. Densely connected convolutional networks, in: IEEE CVPR.

Huang, Y., Zheng, H., Liu, C., Ding, X., Rohde, G.K., 2017b. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. IEEE J-BHI 21, 1625–1632. doi:10.1109/JBHI.2017.2691738.

Huijbers, A., Tollenaar, R., v Pelt, G., Zeestraten, E., Dutton, S., McConkey, C., Domingo, E., Smit, V., Midgley, R., Warren, B., et al., 2012. The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: validation in the VICTOR trial. Annals of Oncology 24, 179–185.

Irshad, H., Veillard, A., Roux, L., Racoceanu, D., 2014. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review-current status and future potential. IEEE RBE 7, 97–114.

Janowczyk, A., Madabhushi, A., 2016. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. JPI 7.

Javed, S., Fraz, M.M., Epstein, D., Snead, D., Rajpoot, N.M., 2018. Cellular community detection for tissue phenotyping in histology images, in: Computational Pathology and Ophthalmic Medical Image Analysis. Springer, pp. 120–129.

Jiang, X., Wang, H., Tang, S., Ma, L., Zhang, Z., Zheng, Z., 2011. A new approach to shortest paths on networks based on the quantum bosonic mechanism. N. J. of Phy. 13, 013022.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine 16, e1002730.

Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. Scientific reports 6, 27988.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. ICLR .

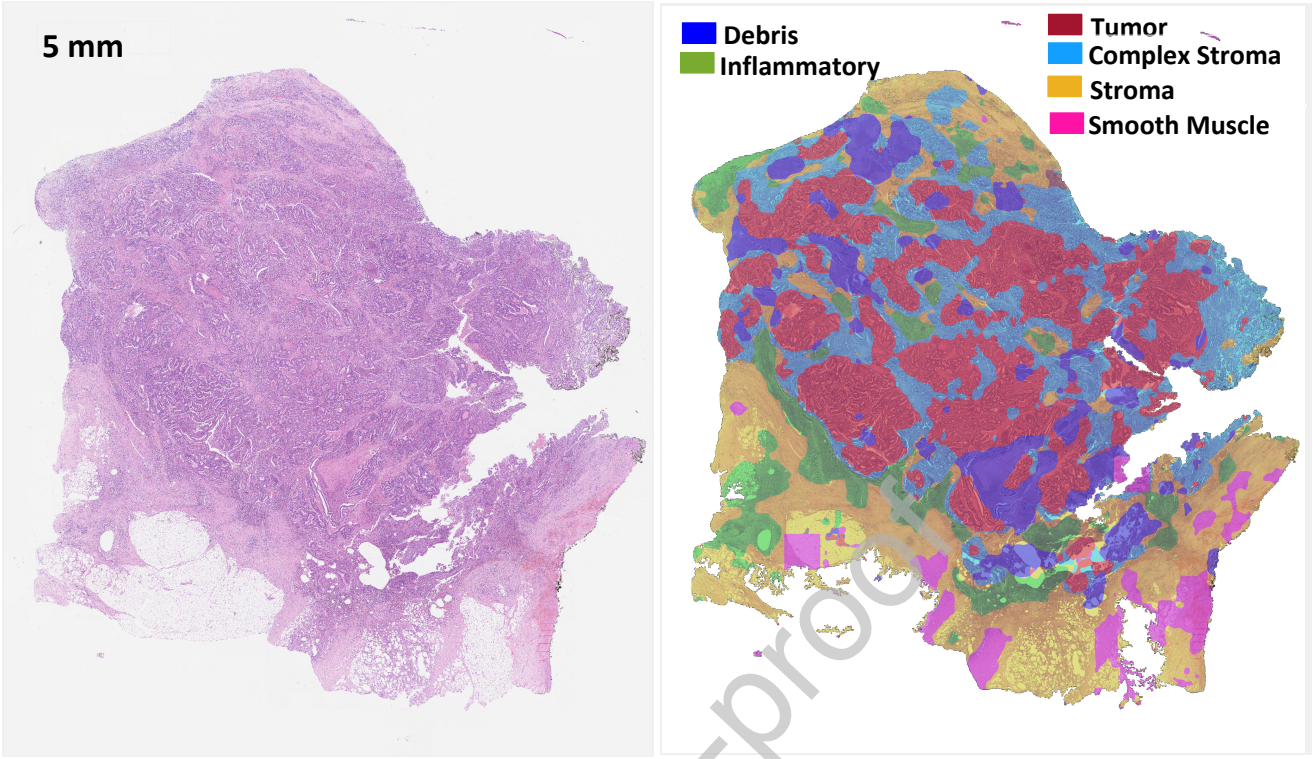
Komura, D., Ishikawa, S., 2018. Machine learning methods for histopathological image analysis. Comp. and Struc. Bio. J. 16, 34–42.

Kothari, S., Phan, J.H., Young, A.N., Wang, M.D., 2013. Histological image classification using biologically interpretable shape-based features. BMC MI 13, 9.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: NIPS.

van der Laak, J., Rajpoot, N., Vossen, D., 2018. The promise of computational pathology. The Pathologist , 16–26.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial



(a) Original Image (a) Original Image

Fig. 6: **Qualitative Evaluation**: The results of the proposed tissue phenotyping algorithm TPCD-4 on an unseen test WSI taken from our proposed CRC-TP dataset (patient-level separation) overlaid on the original WSI. The color map is manually examined by experienced pathologist and found to be matching with manual assessment.

pyramid matching for recognizing natural scene categories. in: IEEE CVPR, pp. 2169–2178.

Li, R., Yao, J., Zhu, X., Li, Y., Huang, J., 2018. Graph CNN for Survival Analysis on Whole Slide Pathological Images, in: Springer MICCAI.

Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordling, S., Haglund, C., Ahonen, T., Pietikäinen, M., Lundin, J., 2012. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagnostic pathology 7, 22.

Louis, D.N., Feldman, M., Carter, A.B., Dighe, A.S., Pfeifer, J.D., Bry, L., Almeida, J.S., Saltz, J., Braun, J., Tomaszewski, J.E., et al., 2015. Computational pathology: a path ahead. Archives of pathology & laboratory medicine 140, 41–50.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis, in: IEEE ISBI, pp. 1107–1110.

Madabhushi, A., Lee, G., 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. MIA 33, 170–175.

Mahmood, A., Small, M., Al-Maadeed, S.A., Rajpoot, N., 2017. Using geodesic space density gradients for network community detection. IEEE T-KDE 29, 921–935. doi:10.1109/TKDE.2016.2632716.

Mairal, J., Bach, F., Ponce, J., 2012. Task-driven dictionary learning. IEEE T-PAMI 34, 791–804.

Marusyk, A., Almendro, V., Polyak, K., 2012. Intra-tumour heterogeneity: a looking glass for cancer? Nature Reviews Cancer 12, 323.

Nalisnik, M., Amgad, M., Lee, S., Halani, S.H., Vega, J.E.V., Brat, D.J., Gutman, D.A., Cooper, L.A., 2017. Interactive phenotyping of large-scale histology imaging data with histomicsml. Scien. Rep. 7, 14588.

Pettie, S., Ramachandran, V., 2005. A shortest path algorithm for real-weighted undirected graphs. SIAM J. on Comp. 34, 1398–1431.

Qaiser, T., Mukherjee, A., Reddy Pb, C., Munugoti, S.D., Tallam, V., Pitkäaho, T., Lehtimäki, T., Naughton, T., Berseth, M., Pedraza, A., et al., 2018. Her 2 challenge contest: a detailed assessment of automated her 2 scoring algorithms in whole slide images of breast cancer tissues. Histopathology 72, 227–238.

Sari, Gunduz-Demir, C., 2018. Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. IEEE T-MI , 1.

Sarkar, R., Acton, S.T., 2018. Sdl: Saliency-based dictionary learning framework for image similarity. IEEE T-IP 27, 749–763.

Schütze, H., Manning, C.D., Raghavan, P., 2008. Introduction to information retrieval, in: Proceedings of the international communication of association for computing machinery conference.

Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.W., Snead, D.R., Cree, I.A., Rajpoot, N.M., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE T-MI 35, 1196–1206.

Sirinukunwattana, K., Snead, D., Epstein, D., Aftab, Z., Mujeeb, I., Tsang, Y.W., Cree, I., Rajpoot, N., 2018. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. Scien. Rep. 8, 13692.

Snead, D.R., Tsang, Y.W., Meskiri, A., Kimani, P.K., Crossman, R., Rajpoot, N.M., Blessing, E., Chen, K., Gopalakrishnan, K., Matthews, P., et al., 2016. Validation of digital pathology imaging for primary histopathological diagnosis. Histopathology 68, 1063–1072.

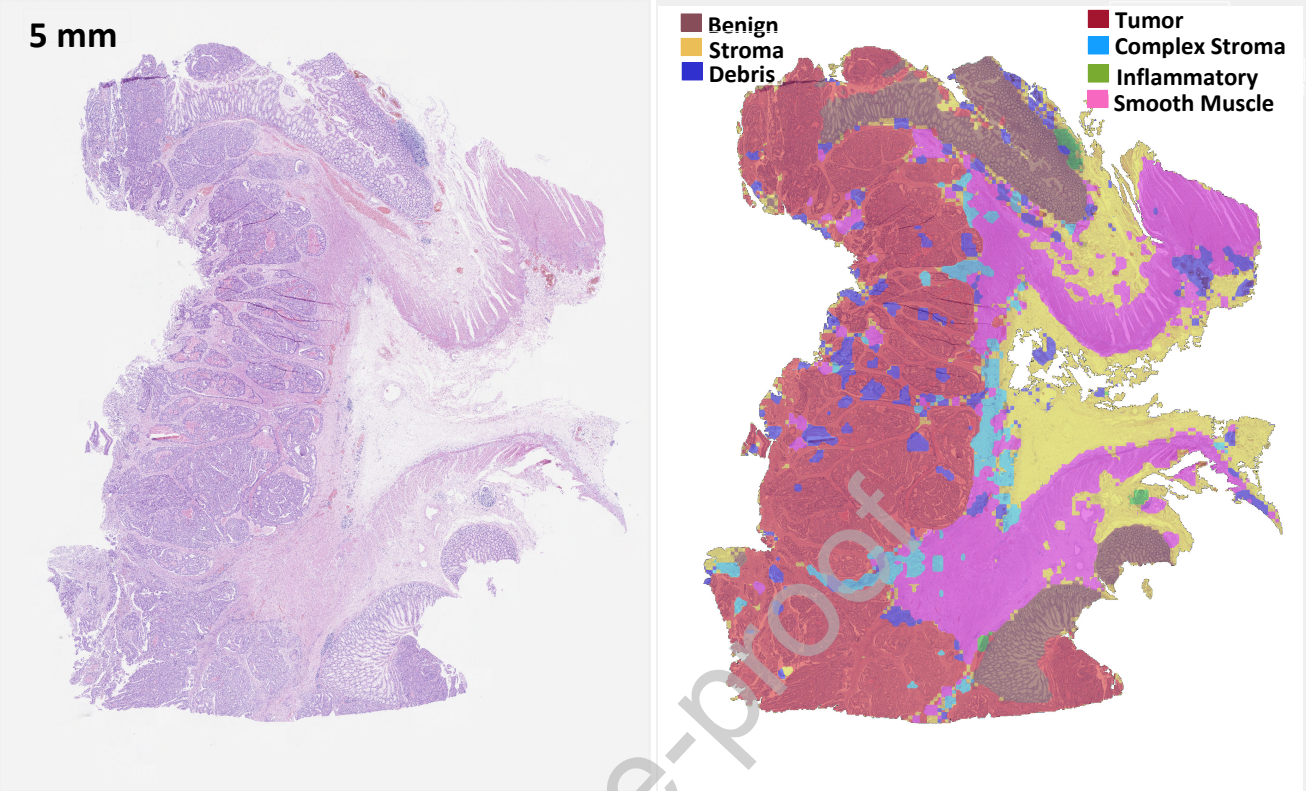
Srinivas, U., Mousavi, H.S., Monga, V., Hattel, A., Jayarao, B., 2014. Simultaneous sparsity model for histopathological image representation and classification. IEEE T-MI 33, 1163–1179.

Tamura, H., Mori, S., Yamawaki, T., 1978. Textural features corresponding to visual perception. IEEE T-SMC 8, 460–473.

Tofighi, M., Guo, T., Vanamala, J.K.P., Monga, V., 2019. Prior Information Guided Regularized Deep Learning for Cell Nucleus Detection. IEEE T-MI 38, 2047–2058.

Veličković, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D., 2019. Deep Graph Infomax, in: ICLR.

Veta, M., Pluim, J.P., Van Diest, P.J., Viergever, M.A., 2014. Breast cancer histopathology image analysis: A review. IEEE T-BE 61, 1400–1411.



(a) Input WSI (b) Classification Results

Fig. 7: **Qualitative Evaluation**: The results of the proposed tissue phenotyping algorithm TPCD-4 on an unseen test WSI taken from our proposed CRC-TP dataset (patient-level separation) overlaid on the original WSI. The color map is manually examined by experienced pathologist and found to be matching with manual assessment.

Vu, T.H., Mousavi, H.S., Monga, V., Rao, G., Rao, U.A., 2016. Histopathological image classification using discriminative feature-oriented dictionary learning. IEEE T-MI 35, 738–751.

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE T-PAMI 31, 210–227.

Xu, J., Luo, X., Wang, G., Gilmore, H., Madabhushi, A., 2016. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 191, 214–223.

Xu, Y., Jia, Z., Wang, L.B., Ai, Y., Zhang, F., Lai, M., Eric, I., Chang, C., 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC BI 18, 281.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L., Leskovec, J., 2018. Graph convolutional neural networks for web-scale recommender systems, in: ACM-ICKDDM.

Zhao, Y., Karypis, G., 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning 55, 311–331.

Conflict of Interest and Authorship Conformation Form

Article Title: **Cellular Community Detection for Tissue Phenotyping In Colorectal Cancer Histology Images**

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

- The following authors have a following academic affiliations with organizations

Author's name	Affiliation
Sajid Javed	University of Warwick
Arif Mahmood	Information Technology University
Muhammad Moazam Fraz	University of Warwick
Navid Alemi Koohbanani	University of Warwick
Ksenija Benes	UHCW
Yee-Wah Tsang	UHCW
Katherine Hewitt	UHCW
David Epstein	University of Warwick
David Snead	UHCW
Nasir Rajpoot	University of Warwick

Conflict of Interest and Authorship Conformation Form

Article Title: **Cellular Community Detection for Tissue Phenotyping In Colorectal Cancer Histology Images**

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

- The following authors have a following academic affiliations with organizations

Author's name	Affiliation
Sajid Javed	University of Warwick
Arif Mahmood	Information Technology University
Muhammad Moazam Fraz	University of Warwick
Navid Alemi Koohbanani	University of Warwick
Ksenija Benes	UHCW
Yee-Wah Tsang	UHCW
Katherine Hewitt	UHCW
David Epstein	University of Warwick
David Snead	UHCW
Nasir Rajpoot	University of Warwick