Knowledge Distillation in Histology Landscape by Multi-Layer Features Supervision

Sajid Javed, Arif Mahmood, Talha Qaiser, and Naoufel Werghi, *Senior Member, IEEE*

***Abstract*—Automatic tissue classification is a fundamental task in computational pathology for profiling tumor micro-environments. Deep learning has advanced tissue classification performance at the cost of significant computational power. Shallow networks have also been end-to-end trained using direct supervision however their performance degrades because of the lack of capturing robust tissue heterogeneity. Knowledge distillation has recently been employed to improve the performance of the shallow networks used as student networks by using additional supervision from deep neural networks used as teacher networks. In the current work, we propose a novel knowledge distillation algorithm to improve the performance of shallow networks for tissue phenotyping in histology images. For this purpose, we propose multi-layer feature distillation such that a single layer in the student network gets supervision from multiple teacher layers. In the proposed algorithm, the size of the feature map of two layers is matched by using a learnable multi-layer perceptron. The distance between the feature maps of the two layers is then minimized during the training of the student network. The overall objective function is computed by summation of the loss over multiple layers combination weighted with a learnable attention-based parameter. The proposed algorithm is named as Knowledge Distillation for Tissue Phenotyping (KDTP). Experiments are performed on five different publicly available histology image classification datasets using several teacher-student network combinations within the KDTP algorithm. Our results demonstrate a significant performance increase in the student networks by using the proposed KDTP algorithm compared to direct supervision-based training methods.**

***Index Terms*—Knowledge distillation, features distillation, histology image classification, tissue phenotyping.**

I. Introduction

The development of modern slide scanners for capturing multi-gigapixel Whole Slide Images (WSIs) has enabled significant growth of computational pathology [9], [13], [30], [35], [45], [49]. In clinical practice, these WSIs are considered as a gold standard for better cancer grading, improved diagnoses, and prognosis [47]. These WSIs have been leveraged by many machine learning techniques to facilitate clinicians and pathologists to assess the degree of malignancy of cancer by automatically analyzing the tumor micro-environment [47], [54]. A typical WSI may contain tens of thousands of pixels at the highest magnification level. Such enormous sizes of WSIs pose significant challenges to machine learning techniques due to the increased demand for computational power and storage capacity. To handle this challenge, WSIs have often been divided into patches of relatively smaller size which are then processed by the machine learning techniques as shown in Fig. 1. The main aim of machine learning techniques is to assist pathologists in in improving their diagnosis performance by increasing reproducibility and reducing inter-observer variations [11], [28], [37], [53].

Automatic tissue phenotyping in histology images is one of the important tasks in computational pathology [4], [19], [20], [44]. One of its aims is to learn cancer biomarkers within the tumor-infiltrating lymphocytes landscape for better cancer diagnosis, grading, prognosis, and evaluating response-to-treatment [13], [22], [30], [32]. It also has an important role in profiling intra-tumor heterogeneity, epigenetics, and cancer progression [34]. Four different examples of tissue classifications are shown in Fig. 1. This fundamental problem has been addressed by many machine learning researchers. However, wide variations of textures, tissue structure, and heterogeneity in histology images pose significant challenges to machine learning techniques [22]. In order to capture such heterogeneity, deep neural networks have been employed which require a large number of annotated training samples to learn rich feature representations. Such deep neural networks have obtained excellent results however, expensive computational resources are also required in addition to the huge volume of annotated training WSIs [45], [47]. Therefore, such tissue classification tools are not feasible on devices with limited resources, e.g., embedded devices.

In order to reduce the amount of training data as well as computational resources, knowledge distillation techniques have recently been proposed that effectively train a lightweight student network from a heavyweight teacher network [14]. The generalization ability of the student model can be improved by

Manuscript received 3 September 2022; revised 28 November 2022; accepted 30 December 2022. Date of publication 17 January 2023; date of current version 5 April 2023. This work was supported by the Khalifa University of Science and Technology under Faculty Start-Up Grant FSU-2022-003 Award 8474000401. The Work of Naoufel Werghi was supported by ASPIRE, UAE, under Grant AARE20-279. *(Corresponding author: Sajid Javed.)*

Sajid Javed and Naoufel Werghi are with the Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi 1227768, UAE (e-mail: sajid.javed@ku.ac.ae; naoufel.werghi@ku.ac.ae).

Arif Mahmood is with the Department of Computer Science, Information Technology University, 54000 Lahore, Pakistan (e-mail: arif.mahmood@itu.edu.pk).

Talha Qaiser is with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: talha.qaiser@warwick.ac.uk).

Digital Object Identifier 10.1109/JBHI.2023.3237749

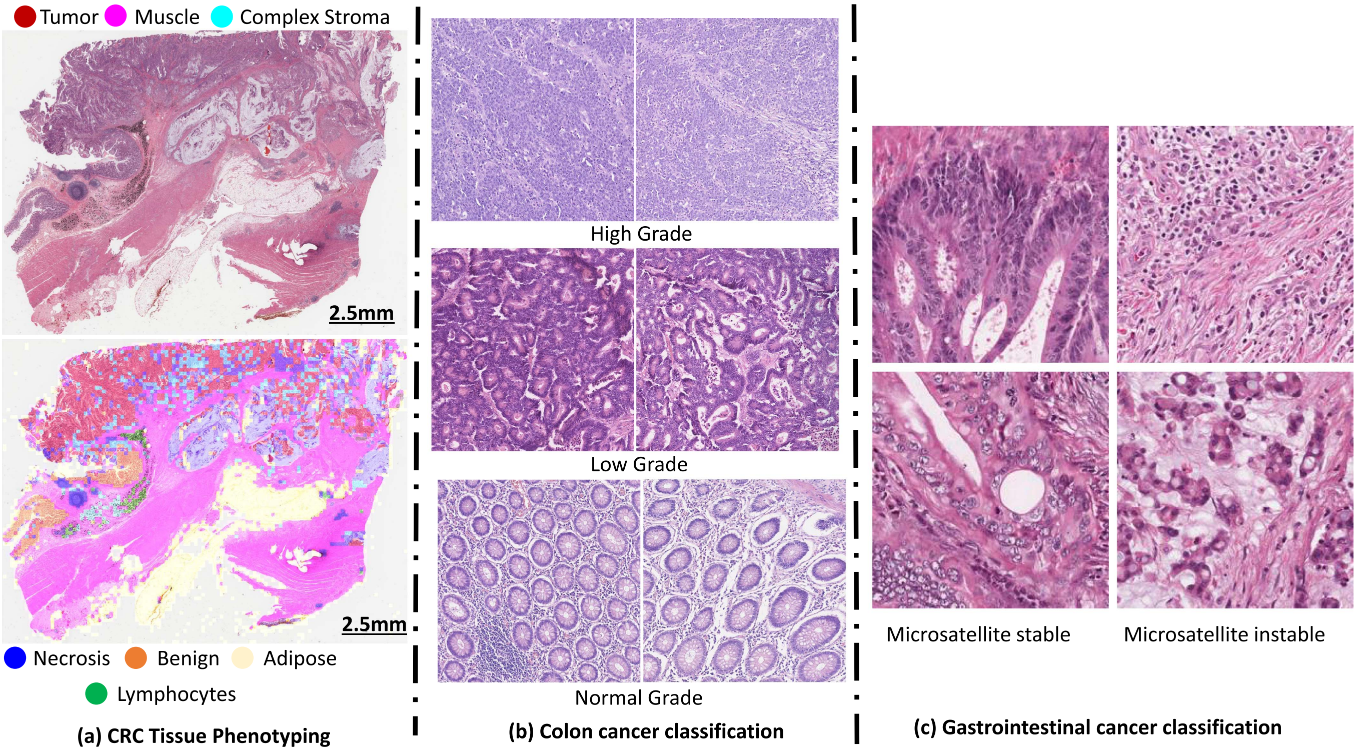

Fig. 1. Different applications of the tissue classification in computational pathology: (a) Tissue phenotyping of ColoRectal Cancer (CRC) in Whole Slide Image (WSI) selected from TCGA [46] on the top and corresponding maps of tissue classes on the bottom. (b) Colon cancer classification showing high grade, low grade, and normal grade tissue images from top to bottom [41]. (c) Gastrointestinal cancer classification showing tissue images of micro-satellite stable on the left and micro-satellite instable on the right [23].

training it to mimic the feature representations and matching the predictions of the teacher model. Recently, such techniques have also obtained significant attention in the machine learning community for object classification [6], action recognition [31], and object tracking applications [42]. In computational pathology, knowledge distillation can reduce the resource requirements at the inference time thus improving the response time and reducing the cost of equipment. Also, the generalization capability of the student model trained with knowledge distillation is much better than the one trained on just the tissue phenotyping data. It is because the teacher network leverages the benefits of large-scale datasets such as ImageNet for pre-training. This knowledge is then transferred to the student model in the context of tissue phenotyping resulting in improved performance. However, the strength of knowledge distillation techniques is not fully explored in computational pathology research for the purpose of training lightweight models for tissue classification.

In the current work, we bridge this research gap by proposing a novel knowledge distillation algorithm for histology image classification task. Most initial knowledge distillation techniques proposed using prediction of the teacher model to be used as a target for the student model [17], [57]. Although it produces good results, however, the information is quite abstract at the final layer. The student model gets only the opportunity to learn the information kept by the final layers ignoring rich information contained in the intermediate layers of the teacher model [14]. In order to exploit this information, feature-map-based knowledge distillation has been proposed in which the feature maps of a student layer are matched with the feature maps of a particular teacher layer [1], [14]. These methods improved from the initial knowledge distillation work however, the supervision provided by the teacher network is still limited to the number of layers in the student network. In order to handle this drawback, we propose multi-layer supervision for a single student layer. More precisely, we propose each student layer be supervised by multiple teacher layers providing better knowledge distillation compared to the existing feature-based techniques.

In the proposed algorithm, a lightweight student network is trained to mimic the rich feature representations of a heavyweight teacher network which is pre-trained using conventional schemes. Each layer in the student model gets supervision from multiple layers of the teacher model. The existing feature map-based knowledge distillation techniques proposed consecutive teacher layers to supervise the corresponding student layers in the same order. We however propose a distributed supervision covering the whole spectrum of feature maps in the teacher model. It is obtained by providing backward links from the latter teacher layers to the earlier student layers such that fewer student layers cover most of the teacher layers. The multi-layer supervision enables the student layers to encode rich information which was not possible from direct training of the student model.

In order to make the distributed supervision more effective, an attention mechanism is exploited which facilitates better knowledge distillation from multiple teacher layers to a single student layer. For this purpose, we compute self-similarity between different student and teacher layers separately. The similarity matrices are then non-linearly transformed into queries and keys such that the overall algorithm performance improves [48]. This is obtained by using two different fully connected neural networks. The queries and keys are then projected to obtain self-attention weights which appropriate the supervision of different teacher layers to a particular student layer. These self-attention weights transfer the rich semantic information contained in the later layers of the teacher model to the earlier layers of the student model through knowledge distillation resulting in significant performance improvements.

The proposed algorithm is dubbed as Knowledge Distillation for Tissue Phenotyping (KDTP). A large number of experiments are performed on five different tissue classification datasets [8], [19], [22], [23], [41] using many combinations of the teacher-student models with KDTP algorithm. In some of these combinations, we observe the improved performance of the student model even beyond that of the teacher model. For instance, ResNet-18 when used as the student network and trained using our proposed KDTP algorithm has consistently outperformed the ResNet-50 used as the teacher model. This demonstrates the effectiveness of using the proposed algorithm for the tissue classification task. The main contributions of the current work are as follows:

1) In this work, we improve the tissue classification performance using a knowledge distillation algorithm that includes both multi-layer supervisions as well as response-level distillation.
2) A novel multi-layer self-attention-based feature maps distillation is proposed which facilitates multiple teacher layers to supervise a single student layer.
3) A novel forward links-based distributed knowledge distillation is proposed which distributes the teacher supervision to each student layer.
4) Extensive teacher-student model combinations are tested on five different datasets to validate the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows: Section II presents a literature review on tissue phenotyping and knowledge distillation methods. Section III explains the proposed algorithm in detail. Section IV presents the exhaustive experimental evaluations while Section V draws the conclusion and future directions of the current work.

II. Literature Review

We divide the related work into two different sections to briefly summarize state-of-the-art tissue phenotyping and knowledge distillation methods.

A. Tissue Phenotyping Methods

The classical tissue phenotyping approaches compute local texture features such as local binary patterns and Gabor features which are then used for classifier training [2], [24], [40]. For instance, Kather's et al., used a multi-texture feature analysis method for tissue phenotyping of eight distinct classes in CRC histology images [24]. Similarly, a dictionary learning-based approach utilizing Gabor features has also been proposed by Sarkar et al., [40]. Other classical texture features-based approaches are also proposed in [2].

Moving towards deep learning era, state-of-the-art tissue phenotyping performance has been advanced [22], [45], [52]. In end-to-end deep learning-based methods, a Convolutional Neural Network (CNN) is trained on a set of training images for the task of patch-based tissue phenotyping. For instance, Bejnordi et al., employed three different networks to classify stromal and epithelium tissues from breast cancer WSIs [12]. In some other methods, the CNN's have also only been used as a features extractor component to training a classifier [3]. For instance, AlexNet architecture [27] was used for deep features extraction by Yu et al., [52]. The extracted features are then used to train a linear support vector machine classifier for histology image segmentation. Some studies also consider fine-tuning the existing trained networks [22], [40]. For instance, Kathers et al., fine-tuned a VGG-19 network [43] on nine distinct tissue classes for the estimation of tumor-stroma scores which is then used for a large-scale study for survival prediction analysis [22]. Some other researchers have recently proposed biologically more meaningful features based on cellular interactions for tissue classification [19], [20]. Han et al. proposed weakly supervised semantic segmentation method [15]. They used patch-level labels for the estimation of pixel-level labels using weak supervision for tissue semantic segmentation. Li et al. proposed a pyramidal deep broad learning method for tissue classification [29].

In the current work, we propose to fine-tune a student network using a pre-trained teacher network for the purpose of tissue phenotyping. Our approach is based on multiple types of knowledge distillation supervision including multi-layer feature maps-based and network prediction-based supervision. To the best of our knowledge, no such knowledge distillation technique containing multi-layer feature supervision has previously been proposed for tissue phenotyping in histology images.

B. Knowledge Distillation Techniques

Earlier knowledge distillation methods relied on the predictions of a larger teacher neural network to distill knowledge to the student network [14], [17]. For instance, Hinton et al., used the predictions of the teacher network as soft targets for the student network for image classification problem [17]. Zagoruyko et al., used the attention maps of the teacher network to train the student network [55]. Thus the attention is transferred from the teacher to the student improving the student's classification performance. Wang et al., also transferred attention using selected features for knowledge distillation [50]. The importance of the features is dynamically established during the knowledge distillation step. Chen et al., used the logits for the knowledge transfer in object detection task [7]. Zhang et al., employed the heatmaps generated by the teacher model for knowledge distillation to the student network in human pose estimation task [56]. Zhang et al., extended the idea of using a single teacher network towards using multiple teachers or students [57]. His work proposed mutual learning of multiple deep networks using logits. These early studies reported improved performance in different tasks however, these methods rely on the final output of the teacher network which is difficult for the student network to learn especially at the initial and intermediate layers [14].

In addition to the teacher network response, feature maps at the intermediate layers have also been used for distilling knowledge to the student network [1]. A variety of feature-based knowledge distillation methods have been proposed in the literature [1], [10], [14]. For instance, Romero et al., directly matched the feature activations of the teacher model and the student model for knowledge distillation [1]. Passalis and Tefas distilled knowledge by using the probability distribution in the features space [38]. Kim et al., proposed an improved form of the intermediate representation for better knowledge transfer using feature maps [26]. Jin et al., used the concept of hint layers to better supervise the student model [21]. Chen et al., proposed to adaptively assign attention weights to different teacher layers which are then used to distill knowledge to student model in a cross-layer manner [6].

Despite significant progress in knowledge distillation research, its applications in computational pathology are quite sparse [5], [25], [36]. Chaudhury et al., proposed mutual learning of teacher and student networks for breast cancer classification [5]. Marini et al., also proposed a knowledge distillation method for Gleason score classification in prostate cancer images [36]. Recently, Dipalma et al., proposed resolution-based distillation for improving histology image classification [10]. Ke et al., used the self-distillation model for identifying patch-level MSI and MSS in histology images [25]. In contrast to these knowledge distillation approaches, we propose multi-layer supervision for each student layer which is distributed on multiple teacher layers. The student supervision is distributed over all intermediate layers of the larger teacher model by exploiting an attention mechanism. To the best of our knowledge, our proposed algorithm is novel not only in computational pathology applications but also in general knowledge distillation research.

III. Proposed Methodology

In this section, we explain the proposed Knowledge Distillation for Tissue Phenotyping (KDTP) algorithm in detail. Our KDTP algorithm consists of one deeper teacher network and one shallower student network. Both networks are pre-trained on ImageNet dataset [39] for the natural image classification task. Both networks are then fine-tuned on different histology image datasets for classification tasks shown in Fig. 1. The fine-tuned student network is then further trained using the proposed knowledge distillation algorithm consisting of teacher response supervision as well as intermediate representations supervision. Our proposed knowledge distillation algorithm is shown in Fig. 2. The details of the proposed algorithm are discussed in the following subsections.

1) Teacher Response-Based Knowledge Distillation: Let $\mathbf{X} = \{\mathbf{d}_i, \mathbf{y}_i\}_{i=1}^{n}$ be the training dataset consisting of n tissue instances from c distinct tissue classes, where d_i is the feature vector and y_i is the corresponding ground-truth label in the form of one hot-encoded vector. The teacher and student logits are normalized using a soft-max layer with a softening parameter $0 < \sigma \leq 1.0$ to get their respective responses as [14]:

$$r_{s/t}(i,j) = \frac{\exp(\sigma g(i,j))}{\sum_{i=1}^{c} \exp(\sigma g(i,j))}, \tag{1}$$

where $g(i,j)$ is the logit corresponding to j-th class and i-th instance in a batch and $r_{s/t}$ is the normalized response of student or teacher model. The response $r_t \in \mathbb{R}^{b\times c}$ of the teacher model for each input tissue instance d_i in batch b is used to supervise the corresponding student response $r_s \in \mathbb{R}^{b\times c}$ using Kullback Leibler (KL) divergence as [14]:

$$KL(r_s||r_t) = \sum_{i=1}^{b}\sum_{j=1}^{c} r_s(i,j)\log\frac{r_s(i,j)}{r_t(i,j)}, \tag{2}$$

where b is mini-batch size. In addition to KL, the multi-class cross-entropy classification loss function $\mathcal{L}_{CE}(y_i, r_s)$ is also used to train the student network [14], [17]. The combined response-based loss function to be minimized while training the student network is given as follows:

$$\mathcal{L}_r = \mathcal{L}_{CE}(y_i, r_s) + \frac{1}{\alpha^2}KL(r_s||r_t), \tag{3}$$

where α is the hyper-parameter that is used to ensure the relative importance of both loss terms.

2) Teacher Intermediate Representations-Based Knowledge Distillation: The intermediate representation supervision is obtained by minimizing some distance measures between the feature maps of the teacher and the student at intermediate layers. Let $s_p(i) \in \mathbb{R}^{c_p\times h_p\times w_p}$ and $t_q(i) \in \mathbb{R}^{c_q\times h_q\times w_q}$ be the feature maps of the p-th student layer and q-th teacher layer for i-th tissue image, where c, h, and w represent the number of channels, height, and the width of the respective feature maps. The feature-based knowledge distillation is obtained by minimizing the following loss function [14], [17]:

$$\mathcal{L}_{pq}(i) = ||f_s(\mathbf{s}_p(i)) - f_t(\mathbf{t}_q(i))||_2, \tag{4}$$

where $f_s(\cdot)$ and $f_t(\cdot)$ are the transformation functions to match the spatial dimensions of the teacher and student features map using the pooling operations and Multi-Layer Perceptron (MLP). Since a single student layer gets supervision from multiple teacher layers, therefore, the overall features-based distillation loss is given by:

$$\mathcal{L}_{IR} = \sum_{p=1}^{p_s}\sum_{q=1}^{q_t}\sum_{i=1}^{b} n_{p,q}(i)\mathcal{L}_{pq}(i), \tag{5}$$

where p_s and q_t are the total number of layers in the student and teacher networks. The parameter $n_{p,q}(i)$ is an attention map for each position in a batch and is learned in an end-to-end manner as discussed in the following section. The overall objective function of the proposed algorithm is given by:

$$\mathcal{L}_{total} = \mathcal{L}_r + \frac{1}{\beta}\mathcal{L}_{IR}, \tag{6}$$

where β is a hyper-parameter to be learned on the training dataset.

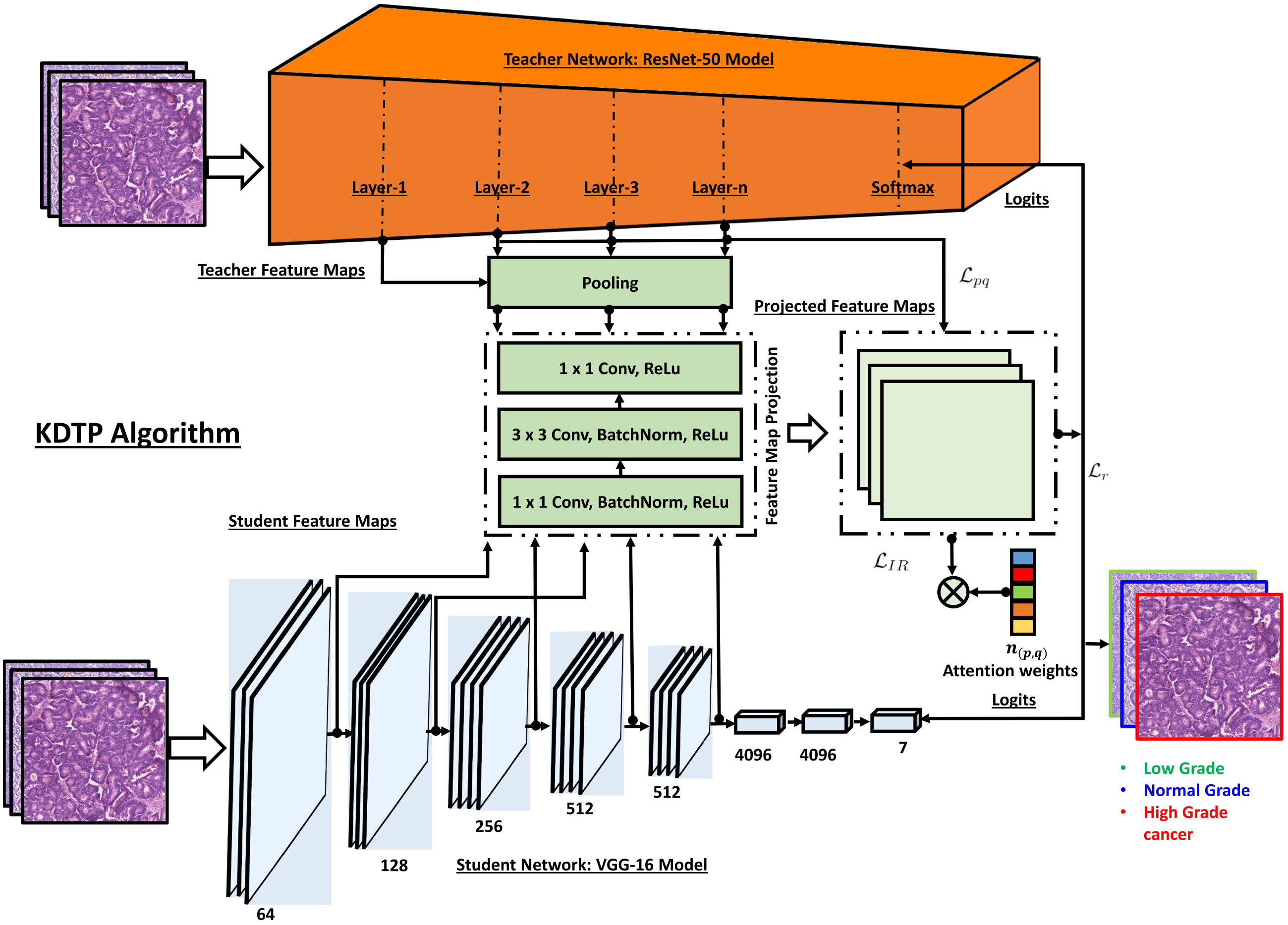

Fig. 2. System diagram of the proposed Knowledge Distillation for Tissue Phenotyping (KDTP) algorithm. As an example, a teacher network ResNet-50, and a student network VGG-16 are shown. Teacher-student feature maps are first projected to a common subspace representation and then the attention weights $n_{p,q}$ are learned in an end-to-end manner. Only the student network is re-trained by mimicking features from the teacher model.

A. Learning Transformation Functions

In order to compute the similarity between teacher and student intermediate representations as shown by (4), the spatial dimensions of the two layers should be transformed to a common subspace using $f_s(\cdot)$ and $f_t(\cdot)$. These transformations are obtained by first applying a pooling operation on the larger height and width of any of the two layers to match it with the smaller dimension. Then, we use an MLP to reduce the larger number of channels to the smaller ones in the two layers. This MLP consists of three sequential layers each comprising 3×3 convolutional filters. The number of filters in each layer is selected such that the number of target channels is obtained at the output. The weights of these MLPs are learned in an end-to-end fashion while training the overall network. In order to reduce the number of these MLPs for the deeper teacher-student combinations, we employ them at the block level instead of the layer level.

B. Attention Mechanism

Layer semantics in a deep neural network varies with the depth. Earlier layers provide local semantics while the later layers provide global context. For effective feature map-based knowledge distillation, an attention mechanism is required to guide the supervision process. It is required to identify the effectiveness of a particular teacher layer to be the supervisor of a specific student layer. For this purpose, we compute the similarity between the feature maps for each teacher layer and the student layer within a particular batch. More specifically, the feature map at the p-th student layer $s_p \in \mathbb{R}^{c_p \times h_p \times w_p}$ is vectorized as $\mathbb{R}^{c_p h_p w_p \times 1}$. For the i-th instance of the tissue image within the batch b, the similarity map is given by:

$$\rho_s(i,p) = S_p^\top s_p(i), \tag{7}$$

where S_p is the matrix of feature maps at layer p for a full batch and $\rho_s(i,p) \in \mathbb{R}^b$ is the similarity of i-th instance $s_p(i)$ with all other vectors in the batch b. Similarly, for the teacher network, the similarity of the same image i within the same batch b for the q-th layer is given by:

$$\rho_t(i,q) = T_q^\top t_q(i), \tag{8}$$

where T_q is the matrix of feature maps at layer q for a full batch and $\rho_t(i,q) \in \mathbb{R}^b$ is the similarity of i-th instance $t_q(i)$ with all other vectors in a batch. The batch similarity vectors are then transformed using two different fully connected networks

$\theta_s(i,p) = FC_s(\rho_s(i,p))$ and $\theta_t(i,q) = FC_t(\rho_s(i,q))$, where $\theta_s(i,p) \in \mathbb{R}^z$ and $\theta_t(i,q) \in \mathbb{R}^z$ are transformed similarity vectors having dimension $z < b$, the batch size. Each of these fully connected networks shares their parameters across all batches and all images. These fully connected networks are learned in an end-to-end fashion to minimize student network loss.

Motivated by self-attention mechanisms in [48], [51], the attention weights are computed by using the exponential of the similarity function as:

$$a_{p,q}(i) = \exp(\theta_s(i,p)^\top \theta_t(i,q)), \tag{9}$$

the sum of attention weights for a particular student layer across all teacher layers is given by:

$$S_p(i) = \sum_{q=1}^{q_t} a_{p,q}(i), \tag{10}$$

where q_t is the number of teacher layers. The normalized attention weights are then given by:

$$n_{p,q}(i) = \frac{a_{p,q}(i)}{S_p(i)}, \tag{11}$$

In this formulation, the sum of normalized attention weights for a particular instance and fixed student layer turns out to be one across all teacher layers. This will ensure that the feature magnitude is not amplified due to the usage of attention weights. The attention weight $n_{p,q}(i)$ is then used in (5) for the computation of intermediate representation loss.

IV. Experimental Evaluations

In this section, we evaluate the proposed KDTP algorithm on five publicly available benchmark datasets including Invasive Ductal Carcinoma (IDC) classification in breast cancer histology images [8], Colon cancer classification into high, low, and normal grades [41], tissue phenotyping using CRC-TP dataset [19], Kather's Colon Cancer dataset [22], and classification of microsatellite stability/instability in gastrointestinal cancer [23]. The results are compared with several baseline individual teacher and student networks as discussed in the following sub-sections.

A. Teacher-Student Architectures

We employ a number of teacher-student combinations based on well-known deep networks including VGG [43], ResNet [16], MobileNet [18], and ShuffleNet [33] for evaluation. For rigorous evaluation of the proposed algorithm, the shallow and deeper versions of these networks are employed in our experiments. For the case of the student network, we employed VGG-8, VGG-13, VGG-19, ShuffleNetV1, ShuffleNetV2, and MobileNetV2. For the case of the teacher network, we used ResNet-8, ResNet-34, ResNet-50, VGG-19, and ShuffleNetV2.

All networks are pre-trained on the ImageNet dataset. We first fine-tuned teacher models on the aforementioned histology image datasets for the tissue classification task. Also, the last layer of each student model is fine-tuned to a particular histology image dataset while the rest of the network weights are kept frozen. These networks are then used for the evaluation of the proposed KDTP algorithm We set a momentum of 0.9 in all our experiments for network training using stochastic gradient descent. We also employed data augmentation techniques including horizontally and vertically flipped images, rotation using five different angles, and image blurring. We set the initial learning rate as 0.01 and batch size of 64 in all architectures. We fine-tuned all the student and teacher models using 240 epochs.

B. Training Details of KDTP

To minimize our proposed KDTP loss function ((6)), we set the hyper-parameter β to 2.5×10^{-3} and the softening parameter σ in (3) to 0.25. The transformation functions used in (5) consist of a stack of three layers with 1×1, 3×3, and 1×1 convolutions to match the dimensions of teacher and student feature maps. The transformation functions are learned in an end-to-end manner. The fully connected layers of the attention mechanism FC_s and FC_t are also learned in an end-to-end manner.

C. Datasets

1) CRC-TP Dataset [19]: CRC-TP dataset is proposed by Javed et al., consisting of 280 k patches belonging to seven distinct tissue classes including tumor, stroma, complex stroma, smooth muscle, necrotic, normal benign, and lymphocytes. The dataset is generated using 20 H & E stained WISs of 20 distinct CRC patients. Each patch in this dataset consists of 150×150 pixels extracted at 20× magnification level. We employed the same training and testing splits of the seven tissue phenotypes provided by the respective authors.

2) Breast Cancer Dataset [8]: This dataset is proposed by Cruz-Roa et al., and used to classify positive and negative patches of IDC. It consists of 277,524 patches extracted at 40× resolution level from 162 WSIs. The size of each patch is 50×50 pixels. Out of those, 198,738 patches belong to negative IDC and 78,786 patches belong to positive IDC. Our algorithm is evaluated on this dataset for binary classification problems using 70% of training and 30% of testing patches.

3) Kather's Colon Cancer Dataset (KCCD) [22]: It contains nine different tissue classes: Muscle, Normal colon mucosa (Norm), Tumor colorectal adenocarcinoma epithelium (Tumor), background, adipose, Mucus, Lymphocytes (Lympho), Debris, and Complex stroma distributed over 100 K training samples and 7.18 K testing samples. Each sample has a resolution of 224×224 pixels and is extracted at 20× magnification level.

4) Gastrointestinal Cancer Classification Dataset [23]: This dataset contains 218,578 unique tissue patches derived from histological images of gastric cancer patients in the TCGA cohort [46]. All images are derived from formalin-fixed paraffin-embedded (FFPE) diagnostic slides. This dataset is used for the binary classification of Micro-satellite Instablity (MSI) and stability (MSS). The training and testing images of each class are provided by the original author. The training and testing splits of MSI consist of 50285 and 27904 unique tissue images. The MSS training and testing splits contain 50285 and 90104 samples.

5) Colorectal Cancer Grading Dataset (CCGD) [41]: The extended CRC dataset consists of 300 visual fields with an

TABLE I

PERFORMANCE COMPARISON OF TEACHER-STUDENT MODELS FOR TISSUE IMAGE CLASSIFICATION ON CRC-TP DATASET [19] USING WEIGHTED AVERAGE SCORE $\widehat{F}$ ON SEVEN DISTINCT CLASSES

Only Student Model	VGG-8 76.60%	VGG-13 77.88%	ShuffleNetV2 73.33%	ShuffleNetV2 73.33%	MobileNetV2 70.16%	ShuffleNetV1 72.56%	ResNet-18 80.11%	ShuffleNetV1 72.56%	MobileNetV2 70.16%	MobileNetV2 70.16%	VGG-8 76.60%
$KDTP_r$	78.10%	79.00%	74.60%	78.11%	73.20%	75.98%	82.33%	75.40%	73.70%	72.22%	78.55%
$KDTP_{1-1}$	82.40%	82.22%	77.20%	82.22%	77.20%	78.11%	84.66%	77.01%	77.33%	74.22%	79.91%
Proposed KDTP	84.30%	84.50%	79.10%	84.60%	82.29%	81.70%	86.10%	80.02%	83.33%	76.66%	80.22%
Only Teacher Model	ResNet-50 81.8%	ResNet-50 81.8%	VGG-13 77.88%	ResNet-50 81.8%	ResNet-18 80.11%	ResNet-50 81.8%	ResNet-50 81.8%	ResNet-18 80.11%	ResNet-50 81.8%	ShuffleNetV2 73.33%	VGG-13 77.88%

average size of 5000 × 7300 pixels [41]. This dataset is used for three class classification of tissue images into normal, low, and high-grade cancer. Similar to [41], we have also performed a three-fold cross-validation experiment. For each class in each fold, we extracted 25,000 patches each of size 224 × 224 from the visual fields. Each fold is once used for training, testing, and validation.

D. Performance Measures

The tissue image classification performance is evaluated using the weighted average $\widehat{F}$ score. For a particular class z, we compute F_z score as:

$$F_z = 2 \times \frac{Precision_z \times Recall_z}{Precision_z + Recall_z}, \quad \text{where}$$
$$Recall_z = \frac{TP_z}{TP_z + FN_z}, \quad Precision_z = \frac{TP_z}{TP_z + FP_z}, \tag{12}$$

where TP_z denotes the True Positives which are the number of tissue images belonging to class z and also predicted as class z, FN_z is the False Negatives which are the number of tissue images belonging to class z but predicted as some other class, FP_z are the False positives which are tissue images not belonging to class z but predicted as class z. The aim is to maximize F_z measure so that its value is close to one. The weighted $\widehat{F}$ measure is computed as a weighted average of F_z overall all classes as given below:

$$\widehat{F} = \sum_{z=1}^{c} p_z F_z, \tag{13}$$

where $p_z = n_z/n$ is the probability of the z-th class, n_z are the number of samples in that class, and n is the total number of tissue samples.

E. Variants of the Proposed Algorithm

In addition to the proposed KDTP algorithm, we have also evaluated the performance on two other variants including $KDTP_r$ which minimizes $\mathcal{L}_r$ given by (3). $KDTP_r$ minimizes the cross-entropy loss and KL divergence between the logits.

The second variant is $KDTP_{1-1}$ in which one layer of the student is projected to only one layer of the teacher model as in most of the existing methods. In the $KDTP_{1-1}$ variant, the later layers of the wider teacher model are not used. It minimizes cross-entropy loss, KL-divergence as in (3), and feature matching loss between only corresponding layers as in (4). It is because the number of layers in student models is less than that of the teacher models.

In the proposed KDTP algorithm, (6) is minimized which includes cross-entropy loss, KL-divergence, and feature matching loss with attention as given in (5).

F. Evaluation on CRC-TP Dataset

Table I shows the performance comparison of the proposed algorithm with other baseline methods. In all our experiments, we observe that the $KDTP_r$ variant is consistently better than the corresponding student model. The proposed KDTP algorithm has even performed better than $KDTP_r$. In most cases, the KDTP is even more accurate than the teacher model. For the case of ResNet-18/ResNet-50, the proposed KDTP has obtained 86.10% weighted average $\widehat{F}$ score which is 4.30% better than the corresponding teacher model.

G. Evaluation on Breast Cancer Dataset

Table II shows the comparative results of the proposed algorithm with other KD-based methods in terms of weighted average $\widehat{F}$ score. In all experiments, the proposed KDTP algorithm has remained more accurate than all variants including $KDTP_r$, $KDTP_{1-1}$, and the student model. In some of the cases such as ResNet-18 as student and ResNet-50 as teacher network, the proposed KDTP has obtained 83.10% weighted average $\widehat{F}$ score which is even higher than the only teacher model. A similar trend has been observed when student networks were VGG-8 and VGG-13 and the teacher network was ResNet-50. Compared to the teacher network, the maximum performance gained is 4.33% for the case of MobileNetV2 as a student and ShuffleNetV2 as a teacher. This demonstrates the effectiveness of our algorithm for histology image classification tasks.

H. Evaluation on Kather's Colon Cancer Dataset

Table III shows the performance comparison of our proposed algorithm in terms of weighted average $\widehat{F}$ score. The proposed algorithm variant $KDTP_r$ is more accurate than the only student model in all experiments. The second variant $KDTP_{1-1}$ which involves both feature-based and response-based knowledge distillation further improves the tissue image classification performance even beyond the teacher model. The final proposed KDTP algorithm has remained the most accurate among all variants. This is because of the multi-layer supervision obtained from the teacher model. The maximum performance gained by the KDTP algorithm from the student model is 13.12% for the case of MobileNetV2 as a student model and ResNet-18 as a teacher model. The performance of KDTP compared to the teacher

TABLE II
PERFORMANCE COMPARISON OF TEACHER-STUDENT MODELS FOR IDC VS NON-IDC CLASSIFICATION ON BREAST CANCER DATASET [8] USING WEIGHTED AVERAGE SCORE $\widehat{F}$

Only Student Model	VGG-8 72.22%	VGG-13 74.11%	ShuffleNetV2 70.11%	ShuffleNetV2 70.11%	MobileNetV2 68.55%	ShuffleNetV1 69.77%	ResNet-18 77.22%	ShuffleNetV1 69.77%	MobileNetV2 68.55%	MobileNetV2 68.55%	VGG-8 72.22%
$KDTP_r$	74.44%	75.11%	71.33%	72.33%	69.55%	71.20%	79.00%	71.22%	69.11%	70.00%	73.23%
$KDTP_{1-1}$	77.77%	79.10%	72.33%	75.55%	72.22%	74.44%	81.33%	72.66%	72.22%	71.11%	76.66%
Proposed KDTP	80.22%	82.30%	74.33%	78.56%	75.44%	78.22%	83.10%	75.81%	75.66%	74.44%	78.88%
Only Teacher Model	ResNet-50 79.11%	ResNet-50 79.11%	VGG-13 74.11%	ResNet-50 79.11%	ResNet-18 77.22%	ResNet-50 79.11%	ResNet-50 79.11%	ResNet-18 77.22%	ResNet-50 79.11%	ShuffleNetV2 70.11%	VGG-13 74.11%

TABLE III
PERFORMANCE COMPARISON OF TEACHER-STUDENT MODELS FOR TISSUE IMAGE CLASSIFICATION ON KATHER'S COLON CANCER DATASET [22] USING WEIGHTED AVERAGE SCORE $\widehat{F}$ ON NINE DISTINCT CLASSES

Only Student Model	VGG-8 82.84%	VGG-13 85.55%	ShuffleNetV2 83.22%	ShuffleNetV2 83.22%	MobileNetV2 76.10%	ShuffleNetV1 83.10%	ResNet-18 87.22%	ShuffleNetV1 81.32%	MobileNetV2 76.10%	MobileNetV2 76.10%	VGG-8 82.84%
$KDTP_r$	84.18%	87.00%	85.88%	86.11%	80.20 %	85.60%	89.33%	83.27%	78.91%	77.68%	83.20%
$KDTP_{1-1}$	92.11%	93.55%	87.65%	94.61%	84.20%	89.22%	92.22%	90.11%	88.10%	80.33%	85.22%
Proposed KDTP	94.80%	95.51%	90.11%	96.10%	89.22%	92.22%	95.51%	93.20%	92.11%	84.92%	86.10%
Only Teacher Model	ResNet-50 90.19%	ResNet-50 90.19%	VGG-13 85.55%	ResNet-50 90.19%	ResNet-18 87.22%	ResNet-50 90.19%	ResNet-50 90.19%	ResNet-18 87.22%	ResNet-50 90.19%	ShuffleNetV2 83.22%	VGG-13 85.55%

TABLE IV
PERFORMANCE COMPARISON OF TEACHER-STUDENT MODELS FOR MSS VS MSI TISSUE IMAGE CLASSIFICATION ON GASTROINTESTINAL CANCER DATASET [23] USING WEIGHTED AVERAGE SCORE $\widehat{F}$

Only Student Model	VGG-8 69.99%	VGG-13 72.22%	ShuffleNetV2 68.10%	ShuffleNetV2 68.10%	MobileNetV2 65.30%	ShuffleNetV1 67.11%	ResNet-18 73.33%	ShuffleNetV1 67.11%	MobileNetV2 65.30%	MobileNetV2 65.30%	VGG-8 69.99%
$KDTP_r$	72.10%	73.10%	70.20%	70.22%	67.33%	68.10%	75.22%	68.22%	68.22%	66.41%	72.20%
$KDTP_{1-1}$	75.51%	75.55%	71.88%	71.59%	69.11%	70.22%	78.22%	70.22%	71.21%	68.88%	74.55%
Proposed KDTP	77.77%	77.20%	74.22%	74.22%	71.71%	73.20%	80.22%	71.30%	74.44%	69.10%	76.77%
Only Teacher Model	ResNet-50 75.60%	ResNet-50 75.60%	VGG-13 72.22%	ResNet-50 75.60%	ResNet-18 73.33%	ResNet-50 75.60%	ResNet-50 75.60%	ResNet-18 73.33%	ResNet-50 75.60%	ShuffleNetV2 68.10%	VGG-13 72.22%

TABLE V
PERFORMANCE COMPARISON OF TEACHER-STUDENT MODELS ON COLORECTAL CANCER GRADING DATASET [41] USING WEIGHTED AVERAGE SCORE $\widehat{F}$ ON THREE DISTINCT CLASSES INCLUDING LOW, NORMAL, AND HIGH-GRADE CANCER

Only Student Model	VGG-8 80.22%	VGG-13 81.11%	ShuffleNetV2 72.55%	ShuffleNetV2 72.55%	MobileNetV2 70.33%	ShuffleNetV1 68.22%	ResNet-18 84.55%	ShuffleNetV1 68.22%	MobileNetV2 70.33%	MobileNetV2 70.33%	VGG-8 80.22%
$KDTP_r$	82.31%	83.22%	74.11%	73.90%	73.33%	71.11%	87.22%	72.88%	72.22%	73.33%	83.22%
$KDTP_{1-1}$	85.55%	85.11%	77.58%	76.55%	75.55%	74.44%	90.00%	75.22%	75.31%	75.09%	84.41%
Proposed KDTP	88.88%	87.90%	80.11%	80.22%	79.91%	78.88%	94.44%	78.11%	78.88%	78.88%	86.66%
Only Teacher Model	ResNet-50 87.20%	ResNet-50 87.20%	VGG-13 81.11%	ResNet-50 87.20%	ResNet-18 84.55%	ResNet-50 87.20%	ResNet-50 87.20%	ResNet-18 84.55%	ResNet-50 87.20%	ShuffleNetV2 72.55%	VGG-13 81.11%

model has improved up to 5.91% for the case of ShuffleNetV2 as a student and ResNet-50 as a teacher.

I. Evaluation on Gastrointestinal Cancer Classification Dataset

Table IV presents the comparative results of the proposed algorithm with other KD-based methods in terms of weighted average $\widehat{F}$ score. The proposed KDTP algorithm has consistently remained the best performer compared to other variants. Similar to other datasets, the teacher-student combination of ResNet-50 and ResNet-18 has obtained the best performance of 80.22% which is 4.62% better than the teacher network. The same trend has also been observed for other teacher-student combinations such as VGG-13 and VGG-8, and ShuffleNetV2 and MobileNetV2.

J. Evaluation on Colorectal Cancer Grading Dataset

Table V presents the comparative results of the proposed algorithm with existing KD-based methods and teacher-student models in terms of weighted average $\widehat{F}$ score. Similar to the aforementioned datasets, the proposed KDTP algorithm has maintained its superiority over the rest of the variants. Also, the best performer teacher-student pair is the ResNet-50 and ResNet-18 model which obtained 94.44% $\widehat{F}$ score higher than the rest of the other teacher-student combinations. It is also 7.24% better than the only-teacher model which obtained 87.20%. This shows the effectiveness of the proposed KDTP algorithm in performance improvement of a smaller network ResNet-18 to outperform a deeper network ResNet-50.

K. Comparison With SOTA Methods

We have also compared the proposed KDTP algorithm with existing State-of-the-Art (SOTA) methods including knowledge distillation methods proposed by Hinton et al. [17], Zagoruyko et al. [55], and Chen et al. [6]. For a fair comparison, we evaluated these methods using ResNet-18 as a student model and ResNet-50 as a teacher model. The source codes released by the original authors are used for our implementation. All methods are trained on CRC-TP, Kather's colon cancer, and colorectal cancer grading datasets similar to our proposed algorithm. The results of the trained student model are compared in Table VI. The proposed algorithm has consistently outperformed the compared methods on three datasets for the tissue image classification task.

TABLE VI
PERFORMANCE COMPARISON OF THE PROPOSED KDTP ALGORITHM WITH SOTA METHODS ON THE THREE DIFFERENT DATASETS FOR TISSUE CLASSIFICATION. RESULTS ARE REPORTED USING WEIGHTED AVERAGE SCORE $\widehat{F}$ ON SEVEN, NINE, AND THREE DISTINCT CLASSES OF CRC-TP, KCCD, AND CCGD. THE BEST TWO PERFORMANCES ARE SHOWN IN RED AND BLUE COLORS, RESPECTIVELY

Methods	CRC-TP	KCCD	CCGD
Hinton *et al.* [17]	78.22%	85.55%	82.71%
Zagoruyko *et al.* [55]	81.11%	90.66%	87.92%
Chen *et al.* [6]	82.33%	91.33%	88.22%
Proposed KDTP	86.10%	95.51%	94.44%

L. Computational Time Analysis

During testing, only the student model is employed for all teacher-student combinations. Therefore, the computational time will depend on the size of the student model. For the case of ResNet-18 as a student model, an average time of 1.31 seconds is observed for an image patch of 224 × 224 pixels. This demonstrates that the proposed KDTP algorithm provides significant performance gained despite the low computational time.

V. DISCUSSION AND CONCLUSION

In this paper, we proposed a KDTP algorithm for improving the performance of shallow networks for the task of tissue phenotyping. It is a fundamental clinical pathology task for analyzing the tumor micro-environment for better cancer grading and survival analysis. Automatic tissue phenotyping has been well investigated using deep neural networks. However, the practical implementation of these networks suffers from many clinical challenges such as the need for excessive memory and computational resources which may not be feasible in clinical settings. On the other hand, computationally less expensive neural networks have shown degraded performance for tissue phenotyping. In order to enable these low computationally complex neural networks to perform well in clinical applications, we propose the use of knowledge distillation. It has not been well explored in computational pathology. In this technique, supervision from deeper networks is utilized for better training of shallower networks. For this purpose, we have proposed the KDTP algorithm which is employed on many teacher-student combinations where the teacher is a deeper neural network and the student is a shallower network. The KDTP algorithm is evaluated on five different histology image classification datasets including CRC-TP, Breast cancer, Kather's colon cancer, Gastrointestinal cancer, and Colorectal cancer grading.

The trained shallow networks have performed significantly better than their previous versions as well as their teachers. For example, MobileNetV2 is trained under the supervision of ResNet50 as a teacher model. As a result, we observed significant performance improvements in MobileNetV2 on all datasets. For the CRC-TP dataset, it was originally obtained 70.16% weighted average $\widehat{F}$ score. Once, we retrained this network using the proposed KDTP algorithm its performance increased to 83.33% on the same dataset. Compared to the teacher network which obtained 81.80%, the shallow network has obtained even better scores.

The same teacher-student combination when employed in the breast cancer dataset has exhibited performance improvement from 68.55% to 79.11%. Similarly, on the Kather's colon cancer dataset the performance of MobileNetV2 improved from 76.10% to 92.11%. It is further evaluated on the gastrointestinal cancer dataset where the performance of MobileNetV2 is increased from 65.30% to 74.44%. In addition, we also evaluated this combination on the colorectal cancer grading dataset. The performance of MobileNetV2 is increased from 70.33% to 78.88%. These performance improvements are obtained without requiring any additional computational complexity at test time. Thus, our experiments demonstrate the significance of knowledge distillation algorithms in the field of computational pathology.

Considering another teacher-student combination of ResNet-50 and ResNet-18, the evaluations are performed on all five datasets. For the case of the CRC-TP dataset, the performance of ResNet-18 is improved from 80.11% to 86.10%. On the Breast cancer dataset, its performance is improved from 77.22% to 82.10%. On Kather's colon cancer dataset, its performance improved from 87.22% to 95.51%. For the case of the gastrointestinal cancer dataset, its performance is improved from 73.33% to 80.22%. Similarly, for the colorectal cancer grading dataset, its performance is improved from 84.55% to 94.44%. In all of these experiments, we observed that the shallower network, ResNet-18, has outperformed its teacher the deeper network, ResNet-50, by a significant margin. These experiments also demonstrated that the shallower networks can outperform deeper networks if trained properly using our proposed KDTP algorithm. Please note that the shallower networks are easy to deploy in clinical settings due to reduced resource requirements. Such a scheme can be potentially beneficial for the deployment of deep neural networks in resource-constrained hardware due to the reduction in computational and memory requirements.

In conclusion, we have presented a knowledge distillation network that can conduct histology image classification task in an automated and robust manner. The ability to automatically classify tissue images of various types has a direct bearing on the downstream analysis in pathology. It holds great potential not only for expediting the diagnostic process in clinics but also for extending our understanding of tissue/cellular characteristics, leading to an improved patient care and management. In the future, this technique may potentially be used for the discovery of low-cost cancer biomarkers.

REFERENCES

[1] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, "Fitnets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Representations*, 2015.

[2] S. Alinsaif and J. Lang, "Texture features in the shearlet domain for histopathological image classification," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 14, pp. 1–19, 2020.

[3] T. Araújo et al., "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, 2017, Art. no. e0177544.

[4] Y. Bai et al., "A scalable graph-based framework for multi-organ histology image classification," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5506–5517, Nov. 2022.

[5] S. Chaudhury, N. Shelke, K. Sau, B. Prasanalakshmi, and M. Shabaz, "A novel approach to classifying breast cancer histopathology biopsy images using bilateral knowledge distillation and label smoothing regularization," *Comput. Math. Methods Med.*, vol. 2021, 2021, Art. no. 4019358.

[6] D. Chen et al., “Cross-layer distillation with semantic calibration,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 8, pp. 7028–7036.

[7] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, “Learning efficient object detection models with knowledge distillation,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[8] A. Cruz-Roa et al., “Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks,” in *Proc. Med. Imag.: Digit. Pathol.*, vol. 9041, 2014, Art. no. 904103.

[9] M. Cui and D. Y. Zhang, “Artificial intelligence and computational pathology,” *Lab. Investigation*, vol. 101, no. 4, pp. 412–422, 2021.

[10] J. DiPalma, A. A. Suriawinata, L. J. Tafe, L. Torresani, and S. Hassanpour, “Resolution-based distillation for efficient histology image classification,” *Artif. Intell. Med.*, vol. 119, 2021, Art. no. 102136.

[11] T. N. N. Doan, B. Song, T. T. L. Vuong, K. Kim, and J. T. Kwak, “SONNET: A self-guided ordinal regression neural network for segmentation and classification of nuclei in large-scale multi-tissue histology images,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 3218–3228, Jul. 2022.

[12] B. Ehteshami Bejnordi et al., “Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies,” *Modern Pathol.*, vol. 31, no. 10, pp. 1502–1512, 2018.

[13] R. Feng, X. Liu, J. Chen, D. Z. Chen, H. Gao, and J. Wu, “A deep learning approach for colonoscopy pathology WSI analysis: Accurate segmentation and classification,” *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3700–3708, Oct. 2021.

[14] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.

[15] C. Han et al., “Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels,” *Med. Image Anal.*, 2022, Art. no. 102487.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[17] G. Hinton et al., “Distilling the knowledge in a neural network,” vol. 2, no. 7, 2015, *arXiv:1503.02531*.

[18] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” 2017, *arXiv:1704.04861*.

[19] S. Javed et al., “Cellular community detection for tissue phenotyping in colorectal cancer histology images,” *Med. Image Anal.*, vol. 63, 2020, Art. no. 101696.

[20] S. Javed, A. Mahmood, N. Werghi, K. Benes, and N. Rajpoot, “Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping,” *IEEE Trans. Image Process.*, vol. 29, pp. 9204–9219, 2020.

[21] X. Jin et al., “Knowledge distillation via route constrained optimization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1345–1354.

[22] J. N. Kather et al., “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study,” *PLoS Med.*, vol. 16, no. 1, 2019, Art. no. e1002730.

[23] J. N. Kather et al., “Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer,” *Nature Med.*, vol. 25, no. 7, pp. 1054–1056, 2019.

[24] J. N. Kather et al., “Multi-class texture analysis in colorectal cancer histology,” *Sci. Rep.*, vol. 6, no. 1, pp. 1–11, 2016.

[25] J. Ke, Y. Shen, J. D. Wright, N. Jing, X. Liang, and D. Shen, “Identifying patch-level MSI from histological images of colorectal cancer by a knowledge distillation model,” in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2020, pp. 1043–1046.

[26] J. Kim, S. Park, and N. Kwak, “Paraphrasing complex network: Network compression via factor transfer,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[28] A. Lahiani, I. Klaman, N. Navab, S. Albarqouni, and E. Klaiman, “Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 403–411, 2021.

[29] J. Lin et al., “PDBL: Improving histopathological tissue classification with plug-and-play pyramidal deep-broad learning,” *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2252–2262, Sep. 2022.

[30] X. Liu, X. Kang, X. Nie, J. Guo, S. Wang, and Y. Yin, “Learning binary semantic embedding for large-scale breast histology image analysis,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 3240–3250, Jul. 2022.

[31] Y. Liu, K. Wang, G. Li, and L. Lin, “Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition,” *IEEE Trans. Image Process.*, vol. 30, pp. 5573–5588, 2021.

[32] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, “Data-efficient and weakly supervised computational pathology on whole-slide images,” *Nature BME*, vol. 5, no. 6, pp. 555–570, 2021.

[33] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: Practical guidelines for efficient cnn architecture design,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.

[34] T. Mahmood, S. G. Kim, J. H. Koo, and K. R. Park, “Artificial intelligence-based tissue phenotyping in colorectal cancer histopathology using visual and semantic features aggregation,” *Mathematics*, vol. 10, no. 11, 2022, Art. no. 1909.

[35] K. Mariam et al., “On smart gaze based annotation of histopathology images for training of deep convolutional neural networks,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 3025–3036, Jul. 2022.

[36] N. Marini, S. Otálora, H. Müller, and M. Atzori, “Semi-supervised learning with a teacher-student paradigm for histopathology classification: A resource to face data heterogeneity and lack of local annotations,” in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 105–119.

[37] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, “Cancer diagnosis using deep learning: A bibliographic review,” *Cancers*, vol. 11, no. 9, 2019, Art. no. 1235.

[38] N. Passalis and A. Tefas, “Learning deep representations with probabilistic knowledge transfer,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 268–284.

[39] O. Russakovsky et al., “Imagenet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[40] R. Sarkar and S. T. Acton, “SDL: Saliency-based dictionary learning framework for image similarity,” *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 749–763, Feb. 2018.

[41] M. Shaban et al., “Context-aware convolutional neural network for grading of colorectal cancer histology images,” *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2395–2405, Jul. 2020.

[42] J. Shen, Y. Liu, X. Dong, X. Lu, F. S. Khan, and S. C. Hoi, “Distilled siamese networks for visual tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8896–8909, Dec. 2022.

[43] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

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

[45] C.L. Srinidhi, O. Ciga, and A. L. Martel, “Deep neural network models for computational histopathology: A survey,” *Media*, vol. 67, 2021, Art. no. 101813.

[46] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The cancer genome atlas (TCGA): An immeasurable source of knowledge,” *Contemporary Oncol.*, vol. 19, no. 1A, 2015, Art. no. A68.

[47] J. Van der Laak, G. Litjens, and F. Ciompi, “Deep learning in histopathology: The path to the clinic,” *Nature Med.*, vol. 27, no. 5, pp. 775–784, 2021.

[48] A. Vaswani et al., “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[49] T. T. L. Vuong, B. Song, K. Kim, Y. M. Cho, and J. T. Kwak, “Multi-scale binary pattern encoding network for cancer classification in pathology images,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 3, pp. 1152–1163, Mar. 2022.

[50] K. Wang, X. Gao, Y. Zhao, X. Li, D. Dou, and C.-Z. Xu, “Pay attention to features, transfer learn faster CNNs,” in *Proc. Int. Conf. Learn. Representations*, 2019.

[51] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[52] Y. Xu et al., “Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features,” *BMC Bioinf.*, vol. 18, no. 1, pp. 1–17, 2017.

[53] E. Yildirim and D. J. Foran, “Parallel versus distributed data access for gigapixel-resolution histology images: Challenges and opportunities,” *IEEE J. Biomed. Health Inform.*, vol. 21, no. 4, pp. 1049–1057, Jul. 2017.

[54] S.-Y. Yoo et al., “Whole-slide image analysis reveals quantitative landscape of tumor–immune microenvironment in colorectal cancers,” *Clin. Cancer Res.*, vol. 26, no. 4, pp. 870–881, 2020.

[55] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” 2016, *arXiv:1612.03928*.

[56] F. Zhang, X. Zhu, and M. Ye, “Fast human pose estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3517–3526.

[57] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep mutual learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4320–4328.