Atlas of Digital Pathology: A Generalized Hierarchical Histological Tissue Type-Annotated Database for Deep Learning

Mahdi S. Hosseini[1,4], Lyndon Chan [1], Gabriel Tse [1], Michael Tang [1], Jun Deng [1], Sajad Norouzi [1]
Corwyn Rowsell [2,3], Konstantinos N. Plataniotis [1], and Savvas Damaskinos [4]
[1]The Edward S. Rogers Sr. Department of Electrical & Computer Engineering, University of Toronto
[2]Division of Pathology, St. Michaels Hospital, Toronto, ON, M4N 1X3, Canada
[3]Department of Laboratory Medicine and Pathobiology, University of Toronto
[4]Huron Digital Pathology, St. Jacobs, ON, N0B 2N0, Canada
mahdi.hosseini@mail.utoronto.ca

Abstract

In recent years, computer vision techniques have made large advances in image recognition and been applied to aid radiological diagnosis. Computational pathology aims to develop similar tools for aiding pathologists in diagnosing digitized histopathological slides, which would improve diagnostic accuracy and productivity amidst increasing workloads. However, there is a lack of publicly-available databases of (1) localized patch-level images annotated with (2) a large range of Histological Tissue Type (HTT). As a result, computational pathology research is constrained to diagnosing specific diseases or classifying tissues from specific organs, and cannot be readily generalized to handle unexpected diseases and organs.

In this paper, we propose a new digital pathology database, the "Atlas of Digital Pathology" (or ADP), which comprises of 17,668 patch images extracted from 100 slides annotated with up to 57 hierarchical HTTs. Our data is generalized to different tissue types across different organs and aims to provide training data for supervised multi-label learning of patch-level HTT in a digitized whole slide image. We demonstrate the quality of our image labels through pathologist consultation and by training three state-of-the-art neural networks on tissue type classification. Quantitative results support the visually consistency of our data and we demonstrate a tissue type-based visual attention aid as a sample tool that could be developed from our database.

1 Introduction

The field of computational pathology aims to develop computer algorithms to assist the diagnosis of disease from digital pathology images by pathologists [21]. Leveraging the impressive advances of computer vision in solving difficult general image recognition problems (e.g. recognizing dogs and cats) [9, 32, 16, 33, 12, 38], computational pathology promises to improve the diagnostic accuracy and productivity of pathologists, who face an increasingly heavy workload [27, 41, 21, 15]. In fact, the use of computer vision algorithms in computer-aided diagnosis (CADx) has already become widely accepted in the sister field of radiology [29, 40, 5]. But as there exist serious shortcomings in publicly-available digital pathology databases, and most state-of-the-art computer vision systems are trained by supervised learning, this constrains the development of useful and generalizable computational pathology tools.

Most databases are either annotated with textual descriptions or annotated at the slide level with minimal localized information provided by drawn outlines [21]. This is problematic because, unlike general image classification, digital pathology slide images are very large compared to the features of interest, where local patch-level (or even pixel-level) information is essential. Furthermore, existing databases are specially annotated with an incomplete set of possible tissue conditions (i.e. healthy, disease #1, disease #2, ..., disease #N) and hence cannot be generalized to classify unexpected diseases or unseen organs. Since only a limited number of possible tissue types are observable, it makes more sense to annotate tissue types. Some databases do just that, but use only a few tissue types, which again limits their application. This lack of (1) localized patch-level images annotated with (2) a large range of Histological Tissue Type (HTT) prevents computational pathology from properly developing generalized clinically useful tools.

Inspired by this shortcoming in publicly available digital pathology databases, we introduce the Atlas of Digital Pathology (ADP or Atlas, for short), a database of digital pathology image patches originating from a variety of organs acquired by a whole slide imaging (WSI) scanner, and each annotated with multiple HTTs (both morphological and functional) organized in a hierarchical taxonomy. To the best of our knowledge, there are no other similar databases of digital pathology originating from different organs and annotated at the patch level with multiple, hierarchical HTTs, and this makes our database singularly well-suited for training useful computational pathology tools.

Our database annotations are validated by an experienced pathologist which uniquely exploits *a priori* knowledge of the hierarchical relations between HTTs. As it is labeled with tissue type at the patch level, it can be used to develop a computational pathology system for assisting pathologists in their diagnostic evaluations of whole slide images with patch-level resolution. Furthermore, we adopt three existing CNN architectures - VGG16 [33], ResNet18 [12], and Inception-V3 [38] - and validate the quality of our labeling by training them for tissue type classification using multi-label hierarchical classification loss. The excellent predictive performance of these networks not only further validates the labeling quality, but also suggests the possibility of applying the data to develop computational pathology tools for solving more difficult related problems.

An obvious immediate application of the Atlas of Digital Pathology would be the development of a multi-label supervised learning algorithm for predicting patch-level tissue type, which could be used to localize tissue types in a whole slide image as an automated attention aid for pathologists. For instance, such a tool could highlight glandular tissue regions for cancer diagnosis or show regions of tumor-infiltrating lymphocytes for disease prognosis, thus helping pathologists streamline their visual search procedure. Our Atlas data could be feasibly extended to develop: (a) pixel-level interpretation of WSI via weakly-supervised semantic segmentation to obtain instance-level outlines of individual tissue structures which can assist pathologists in visually searching for relevant tissues, which is known to be time consuming and inconsistent [17, 28]; (b) tissue morphometric analysis to automatically measure the shapes of tissue structure, that is needed for diagnosing certain diseases [11]; (c) abnormality detection that would associate disease prognosis with the shape and arrangement of tissue structures [4]; and (d) image retrieval based on HTT encoding which would enable more semantically-relevant approaches than existing unsupervised Content-Based Image Retrieval (CBIR) approaches [37, 24, 23, 44, 14, 25, 45, 46].

1.1 Related Work

As mentioned above, there exist many publicly available databases of labeled digital pathology images, which may be broadly classified into three types: (1) government or academic repositories, (2) educational sites, and (3) grand challenges. The first type is released by governmental agencies or academic medical research departments for enabling research into the relationships between diagnostic outcomes and histological image features, including The Human Protein Atlas [31], The Cancer Genome Atlas [42], and the Stanford Tissue Microarray Database [26]. These databases are very large and cataloged with structured slide-level mixed-format annotations (e.g. patient data, survival outcomes, pathologist diagnosis). Chang *et al.* train three-class tissue classifiers with TCGA in [6] and Sirinukunwattana *et al.* train four-class nuclei classifiers with STMD in [35]. The second type is released by medical teaching departments for educating their students in histology, contain fewer examples, and are annotated with unstructured slide-level textual descriptions (e.g. labels pointing at diseased regions) - these are very difficult to apply to computational pathology. The third type is released by academic research departments and teaching hospitals for evaluating computational algorithms applied to a chosen clinically-relevant problem, known as Grand Challenges, and include the MICCAI Nuclei Segmentation [18], MICCAI Gland Segmentation [34], and CAMELYON [20] contests. These databases tend to be moderately large in size, focused on particular tissues or conditions, and cataloged with patch-level categorical or positional annotations relevant to the problem at hand (e.g. cancerous or non-cancerous, locations of mitotic figures). These databases are commonly used for computational pathology: Chen *et al.* won MICCAI Gland Segmentation 2015 with their DCAN [7], while Lee and Baeng won CAMELYON17 with their multi-stage framework [19].

1.2 What is Missing?

As stated above, the existing databases are focused on slides exhibiting particular organs (e.g. breast tissue) either annotated at the slide level with particular diagnostic conditions (e.g. cancer grades) or at the local patch level with particular tissue structures (e.g. glands) to the exclusion of other conditions and structures. Therefore, the current compilation format of annotated database limits the proper means of developing tissue recognition tasks for computation pathology in the first place. In fact, a meaningful recognition tool should be capable of identifying a wide range of tissue spectrum, so one can narrow down to certain tissue(s) for diagnostic analysis. Our proposed database, though, includes histological tissue slides from different organs and annotated at the patch level with different tissue structures.

2 Construction of Atlas Database

This section explains the construction of Atlas database from diverse HTTs originating from various organs. First, we describe the WSI scanning workflow to digitize the glass slides and divide them into patches. Then, we explain our proposed hierarchical taxonomy of HTT and patch labeling procedure. Finally, we analyze statistical patterns of the resultant multi-label categorical HTT data through Association Rule Learning and visualize the label co-occurrences as a graph network, and describe the validation of our tissue type label data by an experienced pathologist.

2.1 Whole slide imaging (WSI) workflow

A total of 100 glass slides were selected from a larger size of 500 anonymized glass slides (each sized 1”×3” and 1.0 mm in thickness) by observing the slides under a Nikon H550L brightfield microscope using the 20x objective lens and 0.75 numerical aperture. We selected those slides with (a) acceptably few focus variations caused by tissue specimen thickness [13, 10], (b) diverse spectrum of color variations of tissue stains [3, 8], (c) acceptably few preparation imperfections such as air bubbles and tissue folding/crushing/cracks, (d) different organs of origin, such as brain, kidney, breast, liver, and heart, and (e) different diagnoses (i.e. disease or non-disease related).

These 100 selected glass slides were then digitized using a Huron TissueScope LE1.2 WSI scanner at 40X magnification (0.25μm/pixel resolution, uncompressed TIFF file) and each digital slide was then divided into a randomized subset of recognizable non-background patches of size 1088 × 1088 pixels with an overlap of 32 pixels - 17,668 patches were collected in total. Background patches - those with more than 97.5% of pixels exceeding 85% intensity on all three RGB channels - were excluded. Patches without any recognizable tissue due to significant focus problems or non-tissue objects (e.g. dust specks) were also excluded. On average, each glass slide produced 177 patches (minimum: 12 patches, maximum: 280 patches); most appeared to be stained with Hematoxylin and Eosin (H&E).

2.2 Hierarchical Taxonomy of Histological Tissues

Given digital pathology slide patches containing visibly-recognizable tissues, we assume that each patch originates from an unknown organ and can be classified with one or more histological types. Here, we elaborate on the chosen taxonomy of histological tissue types and their organizing principles. In histology itself, there are two practical approaches: (1) Basic Histology, which studies tissue structure (or morphology), and (2) Systematic (or Functional) Histology, which studies tissue functionality and organization into organs. The Basic Histology approach is readily applicable to the slide patches because even a small visual field is sufficient to identify the tissue structure. By contrast, the Systematic Histology approach is generally not applicable for labeling the slide patches since the organ of origin is usually unknown and a larger visual field is needed to provide the spatial context for understanding the tissue functionality. Note that this excludes the cases of smaller functional structures such as glands and transport vessels. Hence, we consulted standard histology texts [43, 30] and selectively combined both approaches to tissue type classification to suit patch-level analysis in our proposed three-level tissue type taxonomy. The top level contains seven basic histological (or morphological) tissue types (a superset of the five basic types in [43] and four basic types in [30]) and two systematic histological (or functional) tissue types (“Glandular” and “Transport Vessel”). We organized the taxonomy to further divide each top-level tissue type into more specific sub-types, each of which corresponds to a visually identifiable tissue type and not, for example, to an abstract grouping or organ of origin such as epithelium/endothelium/mesothelium. For cases where a more specific child node type cannot be identified, we associated a “Undifferentiated” child node type to the parent but did not consider such nodes as belonging to that level (due to its insufficient specificity). Table 1 details the full hierarchical tissue type taxonomy, the associated letter codes of each type, and provides the number of occurrences of each tissue type label.

Table 1. *Hierarchical taxonomy of histological tissue type used for supervised labeling of the proposed Atlas database. The tissue hierarchy consists of three layers, starting from the least specific top level to the most specific bottom level. The bracketed numbers in the first row correspond to the number of non-“Undifferentiated” types falling in the corresponding level.*

#	level-1 (9)	level-2 (23)	level-3 (36)	#patch
1	Epithelial (E)	Simple Epithelial (E.M)	Simple Squamous Epithelial (E.M.S)	3341
2			Simple Cuboidal Epithelial (E.M.U)	5240
3			Simple Columnar Epithelial (E.M.O)	2533
4		Stratified Epithelial (E.T)	Stratified Squamous Epithelial (E.T.S)	355
5			Stratified Cuboidal Epithelial (E.T.U)	3662
6			Stratified Columnar Epithelial (E.T.O)	783
7			Stratified Epithelial Undifferentiated (E.T.X)	22
8		Pseudostratified Epithelial (E.P)		50
9	Connective Proper (C)	Dense Connective (C.D)	Dense Irregular Connective (C.D.I)	4481
10			Dense Regular Connective (C.D.R)	68
11		Loose Connective (C.L)		8768
12		Connective Proper Undifferentiated (C.X)		291
13	Blood (H)	Erythrocytes (H.E)		7504
14		Leukocytes (H.K)		1739
15		Lymphocytes (H.Y)		5232
16		Blood Undifferentiated (H.X)		126
17	Skeletal (S)	Mature Bone (S.M)	Compact Bone (S.M.C)	298
18			Spongy Bone (S.M.S)	233
19		Endochondral Bone (S.E)		38
20		Cartilage (S.C)	Hyaline Cartilage (S.C.H)	10
21			Cartilage Undifferentiated (S.C.X)	35
22		Marrow (S.R)		157
23	Adipose (A)	White Adipose (A.W)		536
24		Brown Adipose (A.B)		2
25		Marrow Adipose (A.M)		137
26	Muscular (M)	Smooth Muscle (M.M)		4213
27		Skeletal Muscle (M.K)		783
28	Nervous (N)	Neuropil (N.P)		2198
29		Neurons (N.R)	Nerve Cell Bodies (N.R.B)	1840
30			Nerve Axons (N.R.A)	59
31		Neuroglial Cells (N.G)	Microglial Cells (N.G.M)	593
32			Astrocytes (N.G.A)	N/A
33			Oligodendrocytes (N.G.O)	N/A
34			Ependymal Cells (N.G.E)	N/A
35			Radial Glial Cells (N.G.R)	N/A
36			Schwann Cells (N.G.W)	22
37			Satellite Cells (N.G.T)	N/A
38			Neuroglial Cells Undifferentiated (N.G.X)	1856
39	Glandular (G)	Exocrine Gland (G.O)		6976
40		Endocrine Gland (G.N)		1115
41		Gland Undifferentiated (G.X)		66
42	Transport Vessel (T)			6045
-	TOTAL			17668

2.3 Tissue Type Labeling Workflow

A total of five labelers were assigned to perform the tissue type labeling of the slide patches. All labelers were first trained to recognize the tissue types (as proposed above) using annotated exemplar images sourced from standard histology texts such as in [43, 30]. Furthermore, the following labeling criteria were agreed upon in order to ensure labeling consistency: (a) non-cellular labels (those with singular form names, e.g. "Loose Connective") are assigned whenever the tissue is confidently discerned to be present in any quantity anywhere in the patch; (b) cellular labels (those with plural form names, e.g. "Erythrocytes") are assigned whenever five or more cells are confidently discerned to be present anywhere in the patch; (c) each pixel in a patch can correspond to at most one morphological type, but might also correspond to an additional functional type and vice versa (e.g. "Simple Columnar Epithelial" and "Exocrine Gland"); (d) labels are to be assigned at the most specific level possible; and (e) if a patch has no recognizable tissue structures, it is to be excluded from the database.

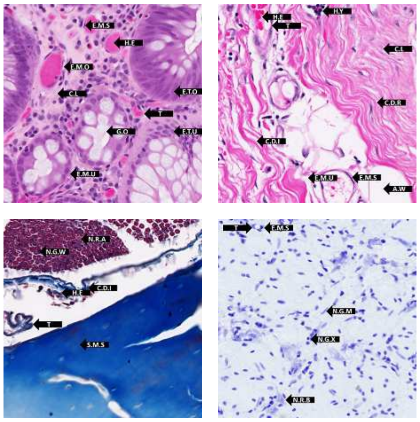


Figure 1. *Sample patches, annotated with arrows pointing to the regions displaying the respective tissue type labels - to interpret the tissue type letter codes, please refer to Table 1*

In Figure 1, four exemplar patches are indicated with their ground-truth labels overlaid as black arrows pointing to the patch regions displaying the corresponding tissue types (to be clear, each patch is assigned global labels without local position information and the arrow labeling indicated here is included for the purpose of demonstration only). Note that the tissue types tend to correspond to specific visual patterns and relative spatial relationships, which suggests that general visual pattern recognition algorithms sensitive to spatial context (such as Convolutional Neural Networks) can be readily applied for supervised learning. "Epithelial" appears as linearly-arranged nuclei-dense cells lining a surface, "Connective Proper" appears as long fibrous cells with scant nuclei between other tissues, "Blood" appears as circular/ovoid blobs which sometimes clump together and are often inside transport vessels[1], "Skeletal" appears as a woody material which sometimes contains layered rings and sometimes appears mottled with lighter regions, "Adipose" appears as clustered rings or bubbles, "Muscular" appears as parallel (longitudinal cut) or bundled (transverse cut) dense fibrous cells with oblong nuclei, "Nervous" appears as wispy strands connecting branching circular blobs, "Glandular" appears as epithelium-lined ovoid structures with or without inner openings, and "Transport Vessel" corresponds to string-like rings often containing blood.

2.4 Label Metadata

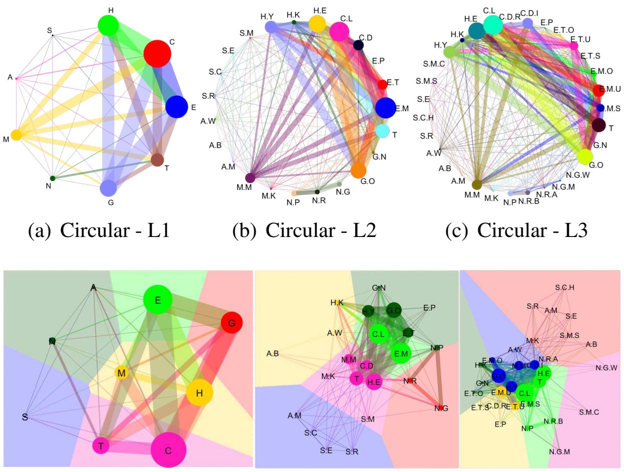
The tissue type labelers only assigned labels at the leaf nodes of the hierarchical taxonomy described in Table 1. However, we can also assign the non-leaf ancestor nodes tissue types based on their descendant nodes - this is done by assigning an ancestor node label if at least one descendant node label is present. For example, if $\mathbf{y} = [\cdots, y_{\text{C.D}}, y_{\text{C.D.I}}, \cdots]$ is the ground-truth tissue type label vector for a given patch and $y_{\text{C.D.I}} = 1$ (present), then its parent label $y_{\text{C.D}}$ is also set to 1. After augmenting the leaf-node labels with these originally un-labeled ancestor nodes, we associate each slide patch with a 57-dimensional binary label vector $\mathbf{y} \in \{0, 1\}^{57}$ (i.e. all the non-"Undifferentiated" types), which has at least one non-zero element. Our proposed database includes the patch image files and their associated augmented binary labels in a comma-separated file. See Table 1 for the class exemplar counts for the leaf-node types.

2.5 Tissue Type Label Statistics

In this section, we demonstrate some methods of understanding the statistics of HTT label data: (1) Co-occurrence Network Analysis, and (2) Association Rule Learning.

Firstly, we model the data as a graph (with the tissue type labels modeled as nodes and label co-occurrences modeled as edges weighted by their counts) and visualize the co-occurrence network. Figure 2 displays the co-occurrence networks at the three levels of the hierarchical taxonomy, visualized with the nodes in both circular (i.e. nodes positioned in a circle) and force-directed layouts (i.e. nodes are cluttered to equalize edge length and minimize edge crossings). For the force-directed layout plots 2(d), 2(e), and 2(f), the nodes are additionally colored by their k-means clusters ($k = 6$) and the cluster regions are displayed in the background. Note that the vast majority of the slide patches belong to a large central interconnected cluster consisting of epithelial, connective proper, blood, muscular, glandular, and transport vessel types (although significant sub-clusters exist for deeper levels of the hierarchy), while the less common skeletal, adipose, and nervous types tend to occur separately.

[1] The sub-type "Leukocytes" includes basophils, neutrophils, eosinophils, and macrophages, while "Lymphocytes" includes natural killer cells, T lymphocytes, and plasma cells



(a) Circular - L1 (b) Circular - L2 (c) Circular - L3

(d) Force-dir. - L1 (e) Force-dir. - L2 (f) Force-dir. - L3

Figure 2. *Tissue type co-occurrence networks, displayed for all three levels of the hierarchical taxonomy (from left to right) and in two layouts - circular (top) and force-directed (bottom) with k-means clustering. To interpret the tissue type letter codes, please refer to the level-appropriate columns of Table 1*

Secondly, we apply the Apriori Association Rule Learning algorithm [2] (with a support threshold of 0.01 and a confidence threshold of 0.5) and display the most significant rule for each unique consequent label where such a rule exists. Following Agrawal *et al.* in [1], we define a *rule* to be between an antecedent itemset and a consequent label. As may be seen from the results in Table 2, 12 out of the 36 possible labels have associated rules with confidence exceeding the confidence threshold. The most confident rule is "$\{\text{E.M.S}, \text{E.M.O}, \text{H.Y}\} \Rightarrow \text{T}$", which indicates that 99.643% of patches labeled with "Simple Squamous Epithelial", "Simple Columnar Epithelial", and "Lymphocytes" are also labeled with "Transport Vessel", which supports the observation about the relative spatial relationships between tissue types mentioned above.

Table 2. *Results of applying to the Apriori Association Rule Learning algorithm to the multi-label data, displaying only the most significant rule for each unique consequent label where such a rule exists.*

Antecedent Itemsets	⇒	Consequent Labels	Confidence
{C.D.I, H.E, H.Y, T}	⇒	E.M.S	0.78701
{G.N}	⇒	E.M.U	0.93274
{H.E, H.K, G.O, T}	⇒	E.M.O	0.50402
{M.K, T}	⇒	C.D.I	0.89457
{H.Y, E.T.U, E.T.O}	⇒	C.L	0.98404
{H.Y, G.O, T, E.M.S, E.T.U}	⇒	H.E	0.96296
{E.M.U, E.T.U, H.E, H.K}	⇒	H.Y	0.93299
{T, E.M.S, C.D.I}	⇒	M.M	0.68536
{N.G.M, T}	⇒	N.P	0.99576
{N.G.M}	⇒	N.R.B	0.8769
{E.M.U, E.M.O, H.Y}	⇒	G.O	0.98097
{E.M.S, E.M.O, H.Y}	⇒	T	0.99643

2.6 Pathologist Validation

In order to further ensure accuracy of labeling, a random set of 1000 tissue patches was reviewed by an experienced, board-certified pathologist. The pathologist was provided with the LabelViewer Graphical User Interface (GUI), where patches could be viewed one at a time in conjunction with the labels that were assigned by the original five labelers (see Section 2.3). The pathologist was provided with the same instructions and information regarding the tissue type hierarchical taxonomy as the original labelers. After comparing each image patch with its assigned labels, the pathologist was able to provide specific notes on each patch with respect to labels that should be added, removed, or modified.

There was excellent concordance between the original labelers and the Pathologist at the higher levels (1-2) of the tissue type hierarchical taxonomy, with most of the suggested modifications relating to Level 3 of the Epithelial branch (e.g. simple cuboidal epithelial vs. simple columnar epithelial) and occasionally at Level 2 (stratified vs. simple). This discordance is most likely attributable to the original labelers' inexperience with interpreting tissues that mimic other tissue types due to tangential sectioning, preparation artefacts, or suboptimal section thickness. Further detailed analysis of the pathologist's validation is provided in the supplementary materials.

2.7 How to Access ADP Database?

The detail information on accessing ADP database can be found in the website[2]. ADP is released as an early version V1.0 and will be updated as an ongoing research effort to provide useful computational pathology database for academic researchers and educators around the world.

3 Experiments

In this section, we explain our experiment to computationally evaluate the labeling quality of the slide patches with three state-of-the-art Convolutional Neural Network architectures: VGG16 [33], ResNet18 [12], and Inception-V3 [38]; we trained the networks separately on flat classification of the three levels of the tissue type hierarchical taxonomy. Furthermore, for the second and third levels, where flat classification assumes label independence, we tested exploiting prior knowledge about the hierarchical relations between tissue types with the *hierarchical binary relevance* method. In the rest of this section, we explain the hierarchical binary relevance method, the training setup, the experimental results in both patch level and slide level, and the failure modes of the best-performing neural network.

3.1 Hierarchical Binary Relevance Method

Hierarchical Binary Relevance (HBR) is a simple method proposed by Tsoumakas *et al.* in [39] for exploiting hierarchical label relations in a multi-label classifier. In our case, we implemented HBR by augmenting each level's tissue types during training with their ancestor types as a flat classification problem. At test time, we obtained optimal thresholds θ_i for all nodes through ROC analysis of

[2] http://www.dsp.utoronto.ca/projects/ADP/

the validation set, and then zeroed all node scores in the test set with any non-confident ancestor node predictions ($p_i < \theta_i$). For example, if the predicted score p_{C} was below θ_C, then its child node score $p_{\mathrm{C.D}}$ would be zeroed. In this way, HBR penalizes the network less during training for predicting a wrong tissue type hierarchically close to the target tissue type than one hierarchically distant; it also prevents the network from predicting tissue types with non-confident ancestor node predictions at test time.

Table 3. *Dataset configuration for training purpose.*

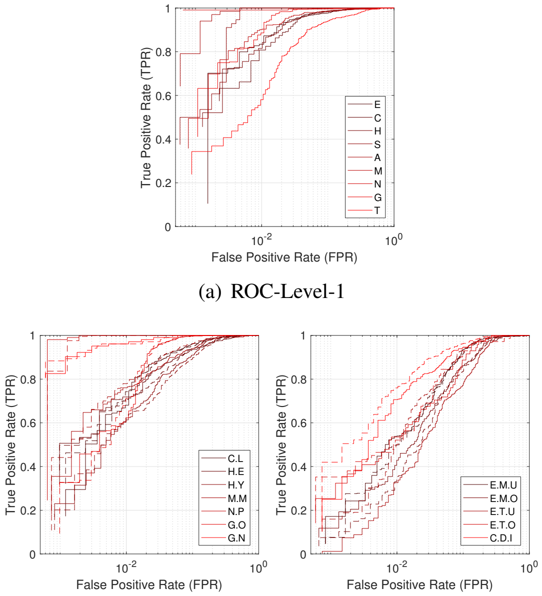
General Statistics	
Training Sample Size	14134
Validation Sample Size	1767
Test Sample Size	1767
Original Image Size	1088 × 1088
Level-1 classes	9
Level-2 classes	31
Level-3 classes	51

3.2 Training Setup

For the following experiments, training was conducted in Keras (TensorFlow backend) with an NVIDIA GTX 1080 Ti GPU. The following Keras implementations of the networks were used: VGG16[3], ResNet18[4], and Inception-V3[5]. Images were resized from original scan size (no pyramid image is used) to each networks' accepted input size (ranging from 224×224 to 229×229) using bilinear downsampling method; all networks were trained for 80 epochs with a batch size of 32, and ℓ_2 regularization weight decay of 0.0005 on all convolutional and dense layers. Moreover, we utilized a stochastic gradient descent optimizer with cyclical learning rate (resetting-triangular policy) [36] in order to adopt the optimum learning rate for our proposed database. Initial base learning rate was 0.001 and initial max learning rate was 0.02 for all the networks, step size was 4 epochs, and learning rates were halved every 20 epochs. The training/validation/test split was 80-10-10 (refer to Table 3 for more information), data augmentation (horizontal/vertical flip) was used, and class weighting was applied to the binary cross-entropy loss $\epsilon(\mathbf{y}, \mathbf{p}) = -\sum_{i=1}^{k} w_i \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right)$. Here, the class weight is defined by $w_i = N/n_i$, with N being the training set size and n_i being the class example count, target label vector $\mathbf{y} \in \{0, 1\}^k$, and predicted score vector $\mathbf{p} \in \mathbb{R}^k$ in k configuration-specific labels. For more information on the training progress please refer to supplementary materials. The Keras implementation of the training networks including all pre-trained models can be obtained from[6], [7].

3.3 Patch-Level Analysis

Figure 3 demonstrates the multi-label classification ROC curves in different hierarchical levels. The predicted scores here are obtained by evaluating the validation set using the VGG16-L3+HBP trained model, which was observed to perform the best at level 3. The ROC curves of level-2 and level-3 are overlaid with and without HBP augmented labeling. The vertical axis of each ROC curve corresponds to the true positive rate (TPR), i.e. sensitivity, and the horizontal axis corresponds to the false positive rate (FPR), i.e. 1 - specificity. Here, the horizontal axis is shown in logarithmic scale for better visual comparison between class labels. The area under the curve (AUC) of ROCs is reasonably high for all trained networks, where the average AUC for level-1, level-2, level-2+HBP, level-3, level-3+HBP are 0.992, 0.9885, 0.9813, 0.9867, and 0.9812, respectively. Overall, the sensitivity decreases in the lower hierarchical levels as a result of the corresponding increase in the number of classes. It is worth noting that we observe noticeable improvements in the sensitivity on level-3 (shown in Figure 3(c)) from using HBP, which suggests that exploiting *a priori* hierarchical knowledge is beneficial for predictive performance. Confusion matrices are also provided in the supplementary materials.



(a) ROC-Level-1

(b) ROC-Level-2 (+HBP)

(c) ROC-Level-3 (+HBP)

Figure 3. *ROC curves for all three hierarchical levels: (left) level-1 of all nine different labels, (middle) level-2 with and without HBP for select labels, and (right) level-3 with and without HBP for select labels. The HBP curves are shown as dashed lines and the non-HBP curves as solid lines for comparison.*

The quantitative performances of all three neural networks trained on five different configurations are listed in Table 4. Note that, for each combination of network and label configuration, we have obtained optimum thresh-

[3] https://github.com/geifmany/cifar-vgg

[4] https://github.com/keras-team/keras-contrib/blob/master/keras_contrib/applications/resnet.py

[5] https://github.com/keras-team/keras-applications/blob/master/keras_applications/inception_v3.py

[6] http://www.dsp.utoronto.ca/projects/ADP/

[7] https://github.com/mahdihosseini/ADP/

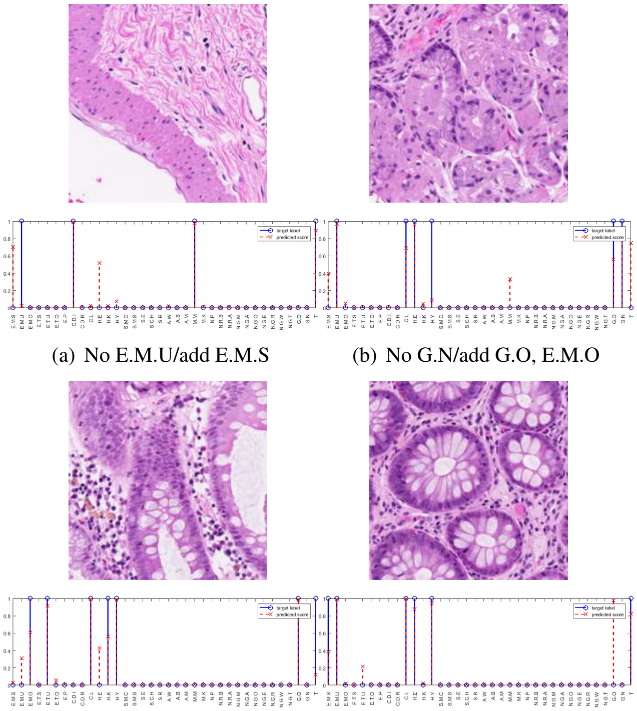
olds using the corresponding ROC analysis. The accuracy (ACC) results are consistently high (above 0.95) across all combinations. Note that, using the mean precision (F1-score) and missing rate (FNR) as evaluation metrics, adding HBP improves predictive performance in the lower levels and VGG16 out-performs out-performs ResNet18 and Inception-V3. We hypothesize that VGG16's superior performance is due to the reduced number of layers (16 instead of 18 and 48 respectively), which has been shown to correspond to a smaller effective receptive field for the final convolutional layer [22]. This is crucial because, as opposed to general image recognition databases such as MS-COCO and ImageNet, where better prediction of large objects is usually achieved by using deeper layers to improve viewpoint- and scale-invariance, the tissue structures to be recognized in our Atlas database seem to be small viewpoint- and scale-invariant textures, so using deeper layers will be redundant and may even promote overfitting.

Table 4. *Test set performance of VGG16, ResNet18, and Inception-V3 applied to tissue type classification on the proposed database, in three levels and five training configurations (with bracketed numbers of classes and best-performing architecture for each metric in boldface).*

		Tissue type configuration				
		L1 (9)	L2 (23)	L2+HBR (36)	L3 (31)	L3+HBR (51)
VGG16	TPR	**0.9505**	**0.9143**	**0.9099**	0.8597	**0.8948**
	FPR	0.0209	0.0219	0.0205	0.0197	0.0188
	TNR	0.9791	0.9781	0.9795	0.9803	0.9812
	FNR	**0.0495**	**0.0857**	**0.0901**	0.1403	**0.1052**
	ACC	**0.9689**	0.9674	0.9644	0.9671	**0.9676**
	F1	**0.9561**	0.9037	**0.9172**	0.8506	**0.8968**
ResNet18	TPR	0.9420	0.8896	0.8903	0.8507	0.8723
	FPR	0.0209	0.0284	0.0279	0.0260	0.0267
	TNR	0.9791	0.9716	0.9721	0.9740	0.9733
	FNR	0.0580	0.1104	0.1097	0.1493	0.1277
	ACC	0.9659	0.9579	0.9544	0.9606	0.9574
	F1	0.9516	0.8761	0.8943	0.8245	0.8657
Inception-V3	TPR	0.9351	0.8969	0.8987	**0.8636**	0.8800
	FPR	**0.0178**	**0.0162**	**0.0170**	**0.0143**	**0.0162**
	TNR	**0.9822**	**0.9838**	**0.9830**	**0.9857**	**0.9838**
	FNR	0.0649	0.1031	0.1013	**0.1364**	0.1200
	ACC	0.9654	**0.9693**	**0.9647**	**0.9725**	0.9675
	F1	0.9507	**0.9071**	0.9071	**0.8722**	0.8950

3.4 Failure Modes

From the quantitative performance results shown above, it is clear that the quality of the tissue type labeling is sufficient for consistently accurate prediction by deep neural networks. But what about the minority of patches that the neural networks fail to accurately predict? To answer this, we sorted the test set patches by unweighted binary cross-entropy loss and examined the class discordance between target labels and predicted scores (using the best-performing VGG16-level-3+HBR configuration). Overall, we observed that almost all failure mode patches were incorrectly annotated, that the neural network still predicts the correct labels, and these labeling mistakes corresponded to those observed by the validating pathologist. They can be grouped as: (1) fundamental mislabeling errors (i.e. tissue type consistently mistaken for another) and (2) human mislabeling errors (i.e. tissue type inconsistently omitted). In Figure 4, four characteristic patches demonstrating select failure modes (one human error, three fundamental errors) are shown with their discordance plots. In Figure 3.4, E.M.S (Simple Squamous) is mislabeled as E.M.U (Simple Cuboidal) and H.E (Erythrocytes) is omitted; in Figure 3.4, fundic gland G.O (Exocrine Gland) is mislabeled as G.N (Endocrine Gland); in Figure 3.4, tangentially cut gland G.O (Exocrine Gland) is mislabeled with E.T.U (Stratified Cuboidal); and in Figure 3.4, the glandular E.M.O (Simple Columnar) is mislabeled as E.M.U (Simple Cuboidal). Further statistical analysis on the failure modes are provided in the supplementary materials.



(a) No E.M.U/add E.M.S

(b) No G.N/add G.O, E.M.O

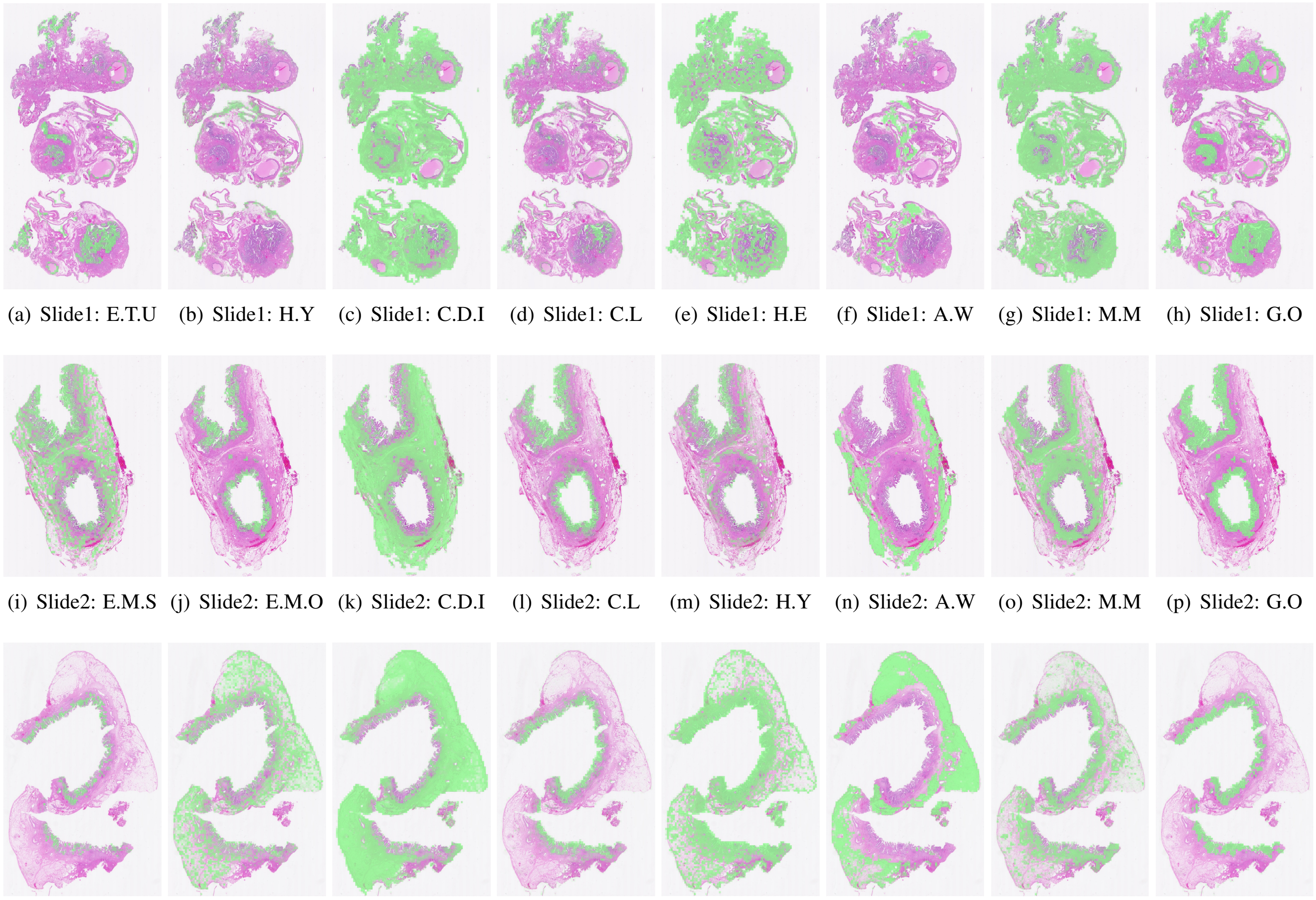
(c) No E.T.U

(d) No E.M.U/add E.M.O

Figure 4. *Failure modes: selected patch images at down-sampled 224×224 resolution on top, target label and predicted score discordance on bottom. Corresponding validating pathologist notes are shown in the sub-figure captions*

3.5 Slide-Level Analysis

In this section, we analyze the predictive performance of the best-performing neural network (VGG16-level-3+HBR) for the patch-level resolution prediction of tissue types in a whole slide image (WSI). As the image size of WSIs is very large (a digital slide of 1cm-by-1cm tissue scanned at 40X resolution is roughly 100K× 100K pixels in size), this demonstrates in a visual manner that the trained network in its present state is already a useful visual attention aid for pathologists in localizing the regions of tissue relevant to the diagnosis at hand, thus simplifying their work and enabling faster diagnosis. We have selected three differ-



(a) Slide1: E.T.U (b) Slide1: H.Y (c) Slide1: C.D.I (d) Slide1: C.L (e) Slide1: H.E (f) Slide1: A.W (g) Slide1: M.M (h) Slide1: G.O

(i) Slide2: E.M.S (j) Slide2: E.M.O (k) Slide2: C.D.I (l) Slide2: C.L (m) Slide2: H.Y (n) Slide2: A.W (o) Slide2: M.M (p) Slide2: G.O

(q) Slide3: E.T.U (r) Slide3: E.M.S (s) Slide3: C.D.I (t) Slide3: C.L (u) Slide3: H.E (v) Slide3: A.W (w) Slide3: M.M (x) Slide3: G.O

Figure 5. *Heatmap representation of confidence scores prediced from VGG16-level-3+HBP trained model on three different WSI images. Each WSI is divided into multiple image patches, their corresponding scores are predicted for all class labels, and represented as a heatmap of corresponding transparency overlaid on the original WSI image.*

ent WSIs known to originate from the Gastro-Intestinal (GI) tract and used them for our case study. Note that these slides are completely separate from the proposed Atlas database and have not been used for training, validating, or testing the neural networks. Each WSI is divided into multiple patches of 1088 × 1088 (excluding background patches), downsampled to 224 × 224, and fed into the trained CNN model (VGG16-level-3+HBR) explained in previous sections. The CNN outputs one confidence score for each predicted tissue type at a given patch, and the patch predictions are stitched together into a slide-level class confidence heatmap, which is overlaid on the whole slide image with transparency corresponding to the confidence level. In Figure 5, 24 different confidence score heatmaps are displayed for four different WSIs, each with eight tissue types. For instance, the White Adipose prediction confidence is visualized for all three WSIs in Figures 5(f), 5(n), and 5(v). Further results on the slide level analysis are provided in the supplementary materials.

4 Concluding Remarks

In this paper, we presented a new digital pathology database of slide patch images annotated with histological tissue types arranged in a hierarchical taxonomy. Given the lack of publicly-available databases of patch-level images annotated with a large range of histological tissue types, which constrains current computational pathology research to focus on particular diseases and organs, we propose that our database will enable research into generalized tissue type supervised learning. We demonstrated the quality of our patch annotations by consulting an expert pathologist and by training three state-of-the-art neural networks, both of which suggest that the data is of sufficiently good quality to be used for developing a useful computational pathology tool. As a proof of concept, we developed a slide-level tissue type-based visual attention aid and we hypothesize that the presented database could be readily applied in the future to solving related computational pathology tasks.

References

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993. 4325

[2] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994. 4325

[3] Pinky A Bautista, Noriaki Hashimoto, and Yukako Yagi. Color standardization in whole slide imaging using a color calibration slide. *Journal of pathology informatics*, 5, 2014. 4323

[4] Andrew H Beck, Ankur R Sangoi, Samuel Leung, Robert J Marinelli, Torsten O Nielsen, Marc J Van De Vijver, Robert B West, Matt Van De Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science translational medicine*, 3(108):108ra113–108ra113, 2011. 4322

[5] Matthew Brown, Patrick Browning, M Wasil Wahi-Anwar, Mitchell Murphy, Jayson Delgado, Hayit Greenspan, Fereidoun Abtin, Shahnaz Ghahremani, Nazanin Yaghmai, Irene da Costa, et al. Integration of chest ct cad into the clinical workflow and impact on radiologist efficiency. *Academic radiology*, 2018. 4321

[6] Hang Chang, Ju Han, Cheng Zhong, Antoine M Snijders, and Jian-Hua Mao. Unsupervised transfer learning via multiscale convolutional sparse coding for biomedical applications. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1182–1194, 2018. 4322

[7] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016. 4322

[8] Emily L Clarke and Darren Treanor. Colour in digital pathology: a review. *Histopathology*, 70(2):153–163, 2017. 4323

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 4321

[10] Filippo Fraggetta, Salvatore Garozzo, Gian Franco Zannoni, Liron Pantanowitz, and Esther Diana Rossi. Routine digital pathology workflow: the catania experience. *Journal of pathology informatics*, 8, 2017. 4323

[11] Metin N Gurcan, Laura Boucheron, Ali Can, Anant Madabhushi, Nasir Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147, 2009. 4322

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4321, 4322, 4325

[13] C Higgins. Applications and challenges of digital pathology and whole slide imaging. *Biotechnic & Histochemistry*, 90(5):341–347, 2015. 4323

[14] A Kallipolitis and Ilias Maglogiannis. Content based image retrieval in digital pathology using speeded up robust features. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 374–384. Springer, 2018. 4322

[15] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. 4321

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4321

[17] Elizabeth A Krupinski, Allison A Tillack, Lynne Richter, Jeffrey T Henderson, Achyut K Bhattacharyya, Katherine M Scott, Anna R Graham, Michael R Descour, John R Davis, and Ronald S Weinstein. Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience. *Human pathology*, 37(12):1543–1556, 2006. 4322

[18] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017. 4322

[19] Byungjae Lee and Kyunghyun Paeng. A robust and effective approach towards accurate metastasis detection and pn-stage classification in breast cancer. *arXiv preprint arXiv:1805.12067*, 2018. 4322

[20] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018. 4322

[21] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sanchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 4321

[22] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*, pages 4898–4906, 2016. 4327

[23] Yibing Ma, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Yushan Zheng, and Huaqiang Shi. Proposing regions from histopathological whole slide image for retrieval using selective search. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 156–159. IEEE, 2017. 4322

[24] Yibing Ma, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Yushan Zheng, Huaqiang Shi, and Yu Zhao. Breast histopathological image retrieval based on latent dirichlet allocation. *IEEE journal of biomedical and health informatics*, 21(4):1114–1123, 2017. 4322

[25] Yibing Ma, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Yushan Zheng, Huaqiang Shi, Yu Zhao, and Jun Shi.

Generating region proposals for histopathological whole slide image retrieval. *Computer methods and programs in biomedicine*, 159:1–10, 2018. 4322

[26] Robert J Marinelli, Kelli Montgomery, Chih Long Liu, Nigam H Shah, Wijan Prapong, Michael Nitzberg, Zachariah K Zachariah, Gavin J Sherlock, Yasodha Natkunam, Robert B West, et al. The stanford tissue microarray database. *Nucleic acids research*, 36(suppl_1):D871–D877, 2007. 4322

[27] GA Meijer, JJ Oudejans, JJM Koevoets, and CJLM Meijer. Activity-based differentiation of pathologists' workload in surgical pathology. *Virchows Archiv*, 454(6):623–628, 2009. 4321

[28] Jesper Molin, Morten Fjeld, Claudia Mello-Thoms, and Claes Lundstrom. Slide navigation patterns among pathologists with long experience of digital review. *Histopathology*, 67(2):185–192, 2015. 4322

[29] Robert M Nishikawa, Robert A Schmidt, Michael N Linver, Alexandra V Edwards, John Papaioannou, and Margaret A Stull. Clinically missed cancer: how effectively can radiologists use computer-aided detection? *American Journal of Roentgenology*, 198(3):708–716, 2012. 4321

[30] Wojciech Pawlina and Michael H Ross. *Histology: a text and atlas: with correlated cell and molecular biology*. Lippincott Wiliams & Wilkins Philadelphia, PA, 2006. 4323, 4324

[31] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. The human protein atlasa tool for pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 216(4):387–393, 2008. 4322

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4321

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4321, 4322, 4325

[34] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: the glas challenge contest. *Medical image analysis*, 35:489–502, 2017. 4322

[35] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016. 4322

[36] Leslie N Smith. Cyclical learning rates for training neural networks. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 464–472. IEEE, 2017. 4326

[37] Akshay Sridhar, Scott Doyle, and Anant Madabhushi. Content-based image retrieval of digitized histopathology in boosted spectrally embedded spaces. *Journal of pathology informatics*, 6, 2015. 4322

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4321, 4322, 4325

[39] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009. 4325

[40] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE, 2017. 4321

[41] Arne Warth, Albrecht Stenzinger, Mindaugas Andrulis, Werner Schlake, Gisela Kempny, Peter Schirmacher, and Wilko Weichert. Individualized medicine and demographic change as determining workload factors in pathology: quo vadis? *Virchows Archiv*, 468(1):101–108, 2016. 4321

[42] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013. 4322

[43] Barbara Young, Phillip Woodford, and Geraldine O'Dowd. *Wheater's functional histology: a text and colour atlas*. Elsevier Health Sciences, 2013. 4323, 4324

[44] Yushan Zheng, Zhiguo Jiang, Yibing Ma, Haopeng Zhang, Fengying Xie, Huaqiang Shi, and Yu Zhao. Content-based histopathological image retrieval for whole slide image database using binary codes. In *Medical Imaging 2017: Digital Pathology*, volume 10140, page 1014013. International Society for Optics and Photonics, 2017. 4322

[45] Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Yibing Ma, Huaqiang Shi, and Yu Zhao. Histopathological whole slide image analysis using context-based cbir. *IEEE Transactions on Medical Imaging*, 2018. 4322

[46] Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Yibing Ma, Huaqiang Shi, and Yu Zhao. Size-scalable content-based histopathological image retrieval from database that consists of wsis. *IEEE journal of biomedical and health informatics*, 22(4):1278–1287, 2018. 4322