scientific reports

Check for updates

OPEN

Scalable deep learning artificial intelligence histopathology slide analysis and validation

Colin Greeley[1], Lawrence Holder[1]✉, Eric E. Nilsson[2] & Michael K. Skinner[2]✉

Deep learning involves an artificial intelligence (AI) approach and has been shown to provide superior performance for automating image recognition tasks, as well as exceeding human capabilities in both time and accuracy. Histopathology diagnostics is one of the more popular challenges at the intersection of artificial intelligence, computer vision, and medicine. Developing methods to automatically detect and identify pathologies in digitized histology slides imposes unique challenges due to the large size of these images and the complexity of the features present in biological tissue. Most methods that are capable of human-level recognition in histopathology are tuned to a specific problem since the computational complexity exceeds that of traditional image classification problems. In the current study, a deep learning approach is developed and presented that can be trained to locate and accurately classify different types of pathologies in gigapixel digitized histology slides along with completing the binary disease classification for the entire image. The approach uses a novel pyramid tiling approach to take advantage of spatial awareness around the area to be classified, while maintaining efficiency and scalability for gigapixel images. The approach is trained and validated on a wide variety of tissue types (i.e., testis, ovary, prostate, kidney) and pathologies taken from an epigenetically altered histology study at Washington State University. The newly developed procedure was optimized and validated along with comparison and validation on public histology datasets. The current developed procedure was found to be optimal and more reproducible when compared to manual procedures, and optimal to previous protocols that used fragmented tissue or slide analysis. Observations demonstrate that the deep learning histopathology analysis is significantly more efficient and accurate than standard manual histopathology analysis.

Keywords Artificial Intelligence, AI, Deep learning, Histopathology, Pathology, Gigapixel, Digitated, Histology, Slides

Abbreviations

AI	Artificial Intelligence
CNNs	Convolutional Neural Networks
WSI	Whole Slide Image
PTO	Pyramid Tiling with Overlap
SA-HCNN	Spatially Aware Histology CNN
TIME	Tumor Immune Microenvironment
ROI	Region of Interest
SE	Squeeze and Excitation
MEL	Manually Expert Labeled
H&E	Hematoxylin and Eosin
FFPE	Formalin Fixed Paraffin-Embedded
DHMC	Dartmouth–Hitchcock Medical Center
PCam	PatchCamelyon
SA	Spatial Awareness
ICIAR	Image Analysis and Recognition

[1]School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752, USA. [2]Center for Reproductive Biology, School of Biological Sciences, Washington State University, Pullman, WA, USA. ✉email: holder@wsu.edu; skinner@wsu.edu

Convolutional neural networks (CNNs) have dominated image-processing related tasks since the 2010's[1], excelling in both whole image classification and subcomponent analysis. CNNs are designed to classify whole images but have been extended to classify subcomponents of images by breaking up large images into smaller images and patching the predictions together. This is the basic idea behind more complex methods such as object detection and semantic segmentation. In object detection, the goal is to predict bounding boxes paired with class labels for each object in an image. Object detection methods such as R-CNN[2], Faster R-CNN[3], and YOLO[4] have enabled real-time detection of objects within images. However, in biomedical imaging, this can result in severe cluttering due to numerous small objects existing in close proximity. Semantic segmentation techniques such as FCN[5], U-Net[6], Segformer[7], and PIDNet[8] are pixel-wise classifiers that are specialized for fine grained image segmentation. The idea behind semantic segmentation is to build a neural network whose output shape is equal to the input image size and be n channels deep where n is the number of classes being predicted. The most popular neural network architecture is the U-Net [6], which was originally built for biomedical image segmentation. A U-Net architecture allows the model to have many layers and an embedded understanding of the input image. The main issue with the U-Net structure is that it is not translatable to large images, since it predicts a pixel mask for each class, meaning that each pixel in the image gets n predictions. Additionally, the fine level of detail from pixel-wise classification techniques is unnecessary given that the task of this work is to find instances of pathologies, not to assign class values to every pixel in the image.

Scaling segmentation methods up to gigapixel images is impossible with modern computers due to memory constraints. A simple workaround to implement pathology detection for gigapixel images is to perform standard image classification on tiles created using a sliding window. This technique results in the model output being a super-pixel mask, where a super-pixel is a geometric grouping of pixels. While super-pixels are not as fine detailed as pixel-wise masks, they work well in histology domains since the images are so large; one super-pixel is generally smaller than the majority of objects in the tissue slide. While creating tiles with a sliding window, super-pixels can be reduced in size if the sliding window increment is smaller than the window size itself, since the tile predictions can be patched together to make finer super-pixels.

In most modern applied computer vision tasks, the largest constraint to work against is time. Computer vision technologies, no matter how accurate, are not easily implemented for near real-time applications using modern machinery due to long computation time for image processing through a deep convolutional neural network. In the case of deep histology and whole slide image (WSI) classification, near real-time computation is not a concern since finding pathological instances in gigapixel images is an exhaustive task for any human. It can take many hours for one person to annotate only one image. Consequently, annotating a large dataset of tissue images can easily take up to a year. On the other hand, trained deep learning networks can classify images in far less time than humans and can do so continuously, resulting in much faster, and typically more accurate, classification.

In the current study, a deep learning method is proposed to accurately and efficiently identify instances of disease in gigapixel histopathology slides. This method has two components: the data preparation, and the deep learning model. The data preparation method proposed is Pyramid Tiling with Overlap (PTO), where a sliding window approach is used to extract multiple resolution views of a subsection of an image. The model is Spatially Aware Histology CNN (SA-HCNN) which takes a tile pyramid as input and classifies the smallest tile in the pyramid. The segmentation ability is extended to be used in the process for classifying a currently existing and published Washington State University ifosfamide chemotherapy transgenerational study (WSIs) based on a statistical approach devised by the pathologists who are conducting the epigenetics experiments[9]. The goal of this research is to devise a computational method which can complete the same histopathology classification tasks as a trained human annotator in significantly less time, with the purpose of reducing tedious and time-consuming work for pathology researchers and increasing their research productivity.

Convolutional neural networks

Deep learning using CNNs dates to the LeNet proposed in 1998. There have been many key advancements in the subfield of computer vision that have allowed CNNs to become more complex by growing deeper and wider. Some of the most notable achievements are summarized in the following. **LeNet-5**[10]: One of the earliest CNNs created and used for handwritten digit recognition. It consists of several layers of convolutional and pooling operations, followed by fully connected layers. **AlexNet**[11]: A landmark CNN that achieved state-of-the-art performance on the ImageNet dataset. It introduced several key innovations, such as the use of rectified linear units (ReLU) as activation functions, data augmentation techniques, and dropout regularization. **VGG**[12]: A CNN architecture that uses a deeper network with smaller filter sizes. It improved performance on the ImageNet dataset, and its architecture has been widely used as a starting point for many subsequent works. **Inception**[13]: A family of CNN architectures that introduced the idea of using multiple parallel convolutional paths of different filter sizes to capture features at different scales. Inception has achieved excellent performance on a variety of benchmarks and has inspired many subsequent works. **ResNet**[14]: A family of CNN architectures that introduced the concept of residual connections to address the problem of vanishing gradients in deep networks. ResNet has achieved state-of-the-art performance on a variety of benchmarks and has inspired many subsequent works. **EfficientNet**[15]: A family of CNN architectures that use a novel compound scaling method to balance model depth, width, and resolution for better performance and efficiency. EfficientNet has achieved state-of-the-art performance on the ImageNet dataset and has become a popular architecture for many computer vision tasks. The family of CNN architectures that is **EfficientNetV2**[16] is an extension of the original idea, but the parameters of the architecture were optimized using neural architecture search. This family of models is the current best for image classification tasks. The proposed SA-HCNN approach is built using the latest best performing CNN backbone which at the time of writing is EfficientNetV2. Many papers in the field of deep learning histology

use VGG, Inception, or ResNet as the CNN backbone since these models have been widely popular and appear frequently in related works, thus resulting in recirculation of deep CNNs which are not state-of-the-art.

Biomedical image segmentation

Since the invention of the U-Net in 2015[6], there has been an explosion in research regarding biomedical image segmentation with deep learning. The auto-encoder model structure is currently the basis for all standard resolution image segmentation tasks as the problem has essentially been solved this way. Of course, state-of-the-art models are extensions on the U-Net structure using modern computer vision methods. The auto-encoder follows the structure of a standard CNN model which maps data in high dimensional space to low dimensional space (image $\rightarrow$ feature embedding). An auto-encoder takes a step further to map the low dimensional space back to high (feature embedding $\rightarrow$ image) to generate pixel-wise predictions to overlay on the original image. This works well for relatively small images, but the utility of this model comes at the cost of computational complexity. These models are more difficult to train since the gradients for the model are being computed per-pixel, which makes scaling up resolution a challenge.

For gigapixel biomedical image segmentation the standard method is to use the sliding window to predict tile patches[17–23]. Most research done in gigapixel biomedical image segmentation proposes some special deep learning architecture that is an extension on top of the standard tiling approach using a sliding window to cut out smaller patches of a larger image. One aspect of the research that often gets overlooked is the method of generating tiles. The papers published in this area often show that some arbitrary overlap value is used for making the decision on what class a tile belongs to. In other cases, the center pixel of a tile is used to make the determination of the class label. But this approach could be suboptimal given that a small sliding window increment is required to ensure that no small objects are skipped over.

The proposed SA-HCNN approach bypasses the need for the implementation of a U-Net for two reasons. First, the images are so large that using a U-Net to segment the pixels of tiles into groups would be unnecessary. Furthermore, a sliding window overlap is used during inference to decrease the super-pixel size of the tile classifications during inference. Second, the PTO tiling method results in highly accurate tile class assignment with the addition of contextual information through pyramid tiling.

Digital histopathology

Tomita et al., (2019)[22] used an attention-based deep learning approach as opposed to a tiling approach to break up a high-resolution image of cancerous tissue. The idea behind the attention-based model is that not all the pixels in the image are relevant so it focuses on areas of interest by using reinforcement learning and peripheral vision. In the tiling approach, each tile of the image is fed into a classifier resulting in most tiles being negatively classified and mostly irrelevant. The attention model uses a tile with a retina-like view to key in on centered features but have a large peripheral to be able to choose where to look next. This method results in only relevant features being sent to the classifier.

Nguyen et al., 2020[21] used a hybrid deep learning model that adds a location embedded element to the classification. The problem being solved in this paper is classification of tissue samples containing colorectal cancer. Similar to the last paper, the large images are broken up into (224×224) pixel tiles that are then input into a VGG-16 pretrained CNN for feature extraction. Before using the feature vector for prediction, another neural network called a capsule network is used to represent various properties such as position, orientation, skewness, scaling, translation and so on inside an image to define the probability of some entity existence. This is necessary since these features are removed in the max pooling layers of the VGG-16 network. The output of both networks is concatenated and used for prediction via a fully connected neural network layer. This means that the predictions are tile based, not the entire image.

Lee et al., 2021[24] discussed how the spatial organization of different cell types in the tumor immune microenvironment (TIME) can be used as biomarkers for predicting drug responses, prognosis and metastasis. It explores the use of deep learning approaches for digital histopathology images to diagnose and predict cancer. The article focuses on machine learning-based digital histopathology image analysis methods for characterizing the tumor ecosystem at three different scales: WSI-level, region of interest (ROI)-level, and cell-level. The authors also provide a perspective on generating cell-level training data sets using immunohistochemistry markers to "weakly-label" the cell types and discuss the limitations and future opportunities of integrating molecular omics data with digital histopathology images.

Hoefling et al., 2021[25] introduced HistoNet, a deep neural network trained to classify normal rat tissue. The network was trained on 1690 annotated slides from 6 preclinical toxicology studies, and small image patches were sampled at 6 different magnification levels. Three different models (VGG-16, ResNet-50, and Inception-v3) were trained on this data, with Inception-v3 and ResNet-50 outperforming VGG-16. Inception-v3 achieved an accuracy of up to 83.4% in identifying tissue from query images, with most misclassifications occurring between histologically similar tissues. The study also found that HistoNet's histological representation could be useful for other machine learning algorithms and data mining, as it identified subclusters corresponding to histologically meaningful structures that were not annotated or trained for. Finally, the models trained on rat tissues could be used on non-human primate and minipig tissues with minimal retraining.

Baxi et al., 2022[26] discussed how digital pathology-based approaches, including whole slide imaging and AI–based solutions, are changing the field of pathology and immuno-oncology. These approaches provide opportunities for the discovery of novel biomarkers and drug targets and support patient selection for diagnostic assays to identify the optimal treatment regimen. The paper highlights the challenges and limitations of implementing AI-based methods in biomarker discovery and patient selection. The authors also emphasize the potential of AI-powered analysis tools to enhance the role of pathologists in delivering accurate diagnoses or assessing biomarkers for companion diagnostics.

The current study proposed SA-HCNN approach builds on top of the current deep histology models by taking advantage of state-of-the-art deep learning practices and methods such as transfer learning, spatial attention, and memory efficient bootstrap-aggregating. Notably, tiling is the most prominent method that transpires across all recent work in gigapixel histology segmentation. The novel PTO tiling approach allows for an increased number of tiles for minority classes to reduce overfitting along with memory efficient training methods using tiles. In addition to innovations in tiling approaches, the most successful deep learning models have been small extensions on the most popular publicly available pre-trained models. HistoNet takes a similar approach but is using an outdated model, Inception-V3, whereas the SA-HCNN uses the EfficientNetV2 model family. Additionally, visual attention has become increasingly popular in biomedical image segmentation due to its feature localization ability. The SA-HCNN includes squeeze and excitation (SE) + spatial attention[27] for learning localization information about the pathologies classified.

Results

Epigenetics study

Researchers at the Skinner Laboratory at Washington State University conducted an epigenetics study[28] to determine if exposure to the chemotherapy drug ifosfamide during puberty in rats results in transgenerational disease in multiple generations of offspring. On the histology side of the study, there were a fixed number of pathologies that were observed and used to make the determination of the animal being diseased or not. However, the images are so large that it was only feasible to count the pathologies in one section of each image (usually 1/2 or 1/3 of the whole image), so the information used to make the final determination is a subset of all available information.

There were two groups used for the study, the control group and the experimental group, where the F0 generation of the experimental group was exposed to ifosfamide. For each generation, tissue samples were taken from the organs of interest (testis, prostate, and kidney for males and ovary and kidney for females) and the pathology instances were counted for each image. To determine whether any individual image would be considered diseased tissue, the number of pathology instances for any pathology had to exceed 1.5 standard deviations of the mean pathology counts of the same pathology in the control group. The procedure for counting pathology instances is as follows: for any section of an image, two analysts would submit their predictions about how many instances of each pathology were present. If there was a disagreement, then a third analyst would be brought in as the tie breaker. This method was used for all images in the final analysis for the study.

The primary data used for training the deep learning model is a subset of all the tissue images from the ifosfamide study described above. The tissue types included testis, prostate, female kidney, male kidney, and ovary. The resolution of the images in all groups ranged from ~(80,000-120,000)x(30,000–80,000) pixels. For the epigenetics study, F1 and F2 generations were used to make the determination of transgenerational disease, so only F1 and F2 generations were used for the deep learning study. In the F1 generation, there are a total of 102 images in the experimental group and 137 images in the control group. In the F2 generation, there are a total of 224 images in the experimental group and 242 images in the control group. Figure 1 shows a sample whole slide image (WSI) of Testis tissue. Sample WSIs for the other four tissue types are shown in Supplemental Fig S1 (Prostate), Fig S2 (Female Kidney), Fig S3 (Male Kidney), and Fig S4 (Ovary).

To digitize the tissue slides into high resolution images, the samples taken from the rats were sent to a 3rd party company that used digital microscopy to generate NDPI images, which could be loaded into a quantitative pathology and bioimage analysis program called QuPath (https://qupath.github.io/). The QuPath annotation procedure differed from the manual counting procedure in the way that geometric regions of the image samples were assigned to classes that were then used for training the deep learning model. For example (see Fig. 2), some pathologies were circled and tightly bounded, and some large regions of tissue were bounded by a rectangle whose class was "normal structure" or examples of the negative class that would exist close to the decision boundary. Additionally, there was an "ignore" class which represented areas of the image that were not tissue at all. All classes other than the pathologies of interest and the ignore class were grouped together to become the negative class which can be seen as normal tissue.

Multiple annotated images were provided for each tissue type: 10 testis slides, 9 prostate slides, 10 female kidney slides, 11 male kidney slides, and 10 ovary slides. Again, the provided images were not fully annotated, only a subsection of the whole images were annotated. Additionally, the subsections that were annotated only provided sparse annotations since the images are so large. Only annotated regions of the images were used to create training data since assumptions could not be made about the tissue outside of those regions. The ground truth data takes the form of pixel masks that were created from the QuPath annotations provided by the researchers. These pixel masks are converted to polygons in code that are further used to label tiles, which are the input to the classification model.

Deep learning approach

The deep learning approach to histopathology begins with tiling the image, then using the deep learning model to predict pathologies in each tile, and then using these predictions to make a whole-slide determination. Tiling is the process in which a sliding window is used to extract patches of a large image. This method is used to break up an image into a grid-like structure where the tiles can be classified individually and then patched back together. Without tiling, using deep learning for gigapixel image classification or segmentation is not possible since the computational overhead would be unachievable. Figure 3 illustrates the tiling process. One issue with tiling is determining the pathology label for the training tiles, i.e., how much of the tile must overlap a pathology for the tile to be labeled with that pathology. Figure 4 shows the performance of the SA-HCNN method on each

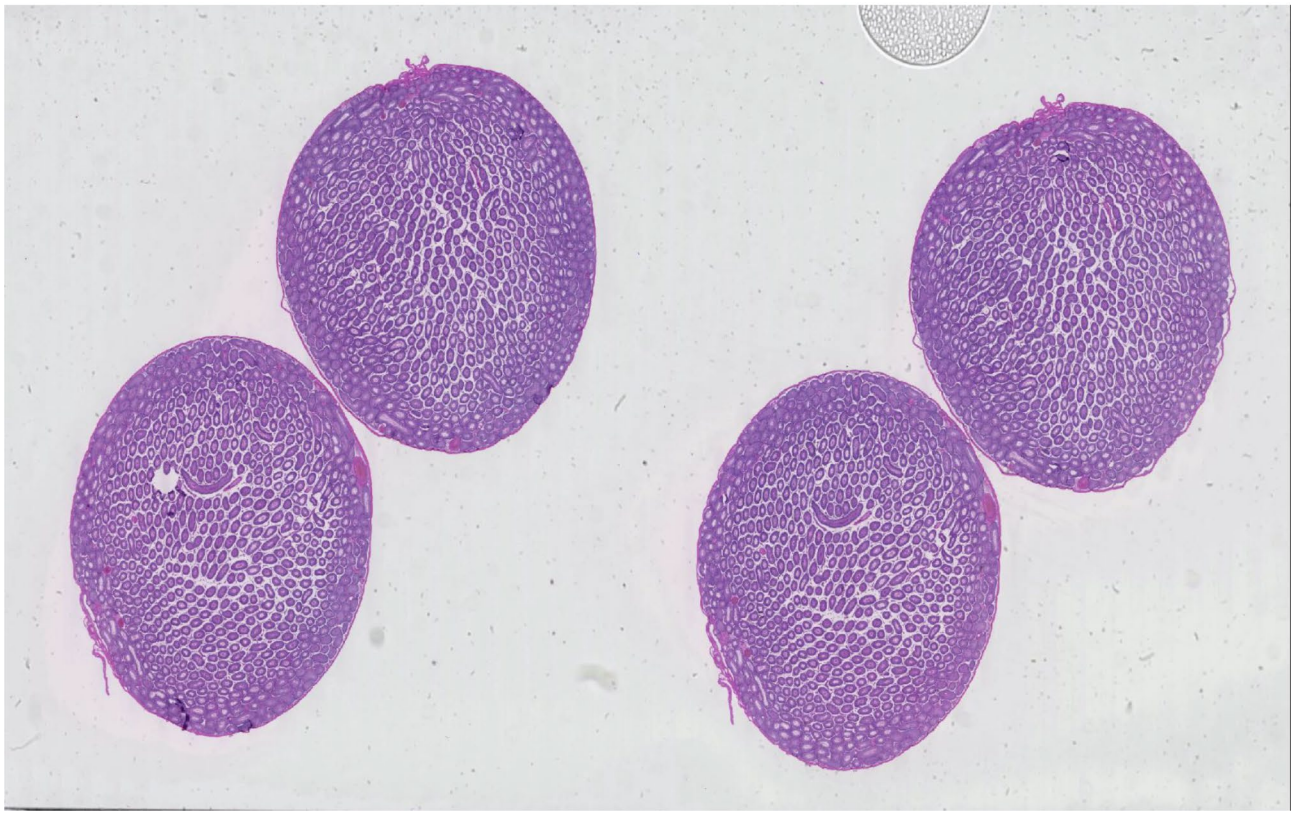

Fig. 1. Sample of testis WSI from the training set. In all testis images, there are two separate but similar looking cross sections of the testes tissue. The whole slides are composed of sequential cross sections of the same tissue. Additionally, pathologies are expected to appear mostly in the same regions across the sequential cross sections.

tissue type across several overlap percentages. An overlap value of 30% was found to be an optimal choice across all tissue types and pathologies, and this value is used for all experiments in this paper.

Another issue with tiling is that spatial information is lost since a deep leaning model would only process one tile at a time and neighboring pixels to the tile which could contain important information are cropped. One approach to this issue is to use center tiling, i.e., the model predicts the pathology of the pixel at the center of the tile, rather than the whole tile. The proposed approach, called Pyramid Tiling with Overlap (PTO), uses the model to predict pathology for the whole tile, but uses contextual information from larger tiles centered around the main tile, but scaled to the same size. Figure 5 depicts this process using two additional larger tiles to provide context, and the three tiles are input to the deep learning model to provide more context for the classification of the main tile. This approach provides the spatial awareness (SA) component to the SA-HCNN method. Table 2 shows the performance for each tissue type of SA-HCNN with PTO, SA-HCNN with Center tiling, and a state-of-the-art baseline EfficientNetV2 with Center tiling. Results show not only that HCNN is superior to the baseline EfficientNetV2, but also that PTO tiling is superior to Center tiling. Most alternative deep histology methods use models that pre-date EfficientNetV2; therefore, the EfficientNetV2 performance is representative of results using alternative methods on the epigenetics study data. The approach is also compared to a modern image segmentation approach called SegFormer[7], which is built using the transformer architecture and is pretrained to perform pixel-wise image segmentation. SegFormer is trained and tested on the same epigenetic histology data as SA-HCNN, and the results show that the image segmentation approach performs poorly for the task of gigapixel histopathology detection within tiles. Thus, the SA-HCNN approach is put forth as a state-of-the-art approach to histology.

The model used for tile classification (see Fig. 6) was built using the latest standard practices for high precision image classification. Google's EfficientNetV2B2 model was implemented for the convolutional backbone to the neural network. There are seven total convolutional blocks in the model. For the output of the model, N classifier blocks are used for prediction. The classifier blocks take the output of the CNN backbone where attention is used, followed by global average pooling and finally a fully connected layer used for prediction.

During inference on a new image, a sliding window is used to extract tiles from the image, and the model is used to predict the pathologies in each tile. Contiguous tile predictions are combined to portray an auto-annotated pathology region in the original image. Figure 7 shows two examples of pathology predictions overlayed on the input image. The images shown are snippets from an annotated testis image. Additional sample

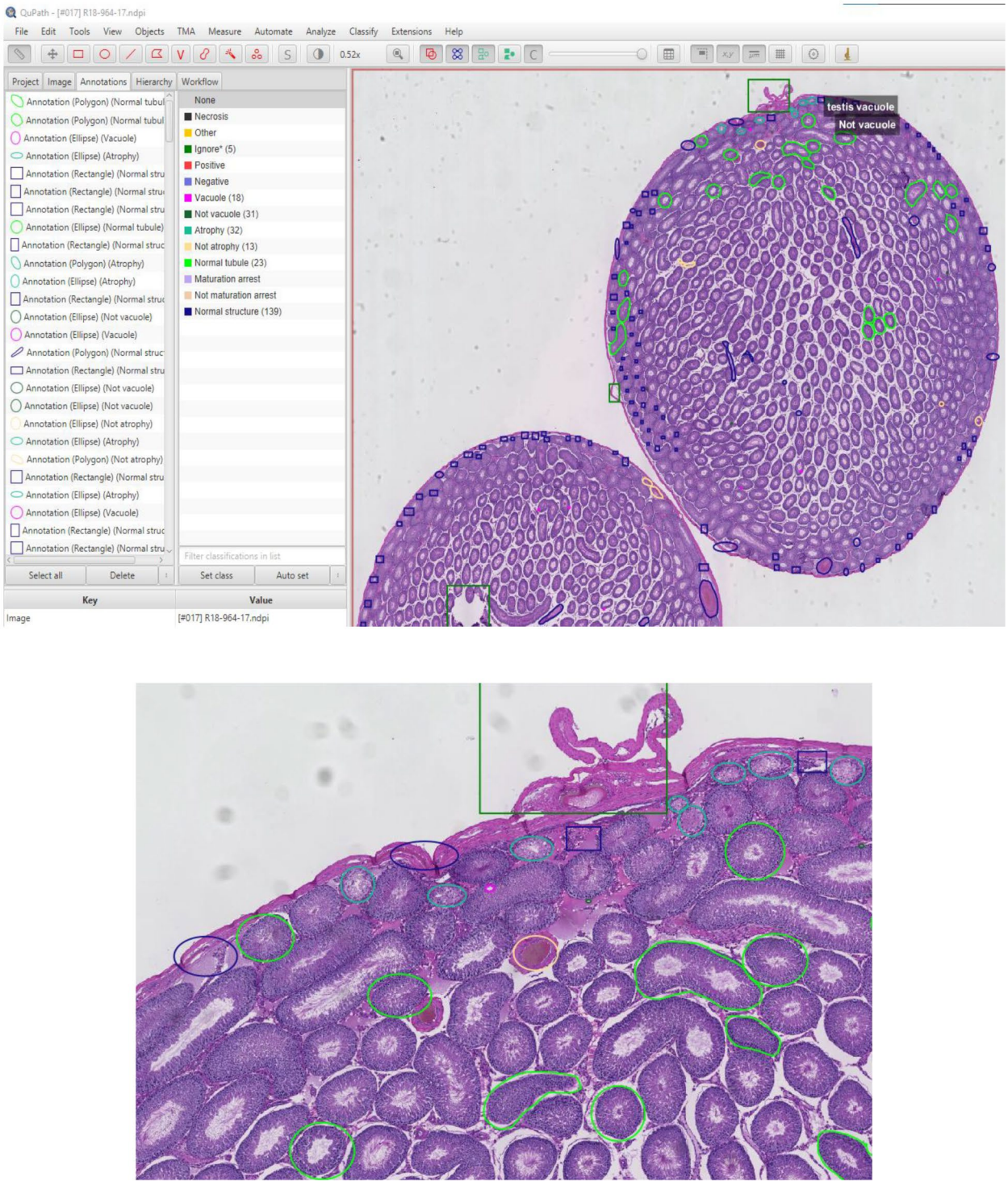

Fig. 2. Example of a testis tissue image viewed in the QuPath system. Various regions of pathology have been human annotated in the image. The bottom image shows a zoomed in section from the top image that more clearly shows the annotations. For example, the light green annotations identify normal tubules, and the mid annotations identify atrophy. The large dark green rectangle annotates an "ignore" region.

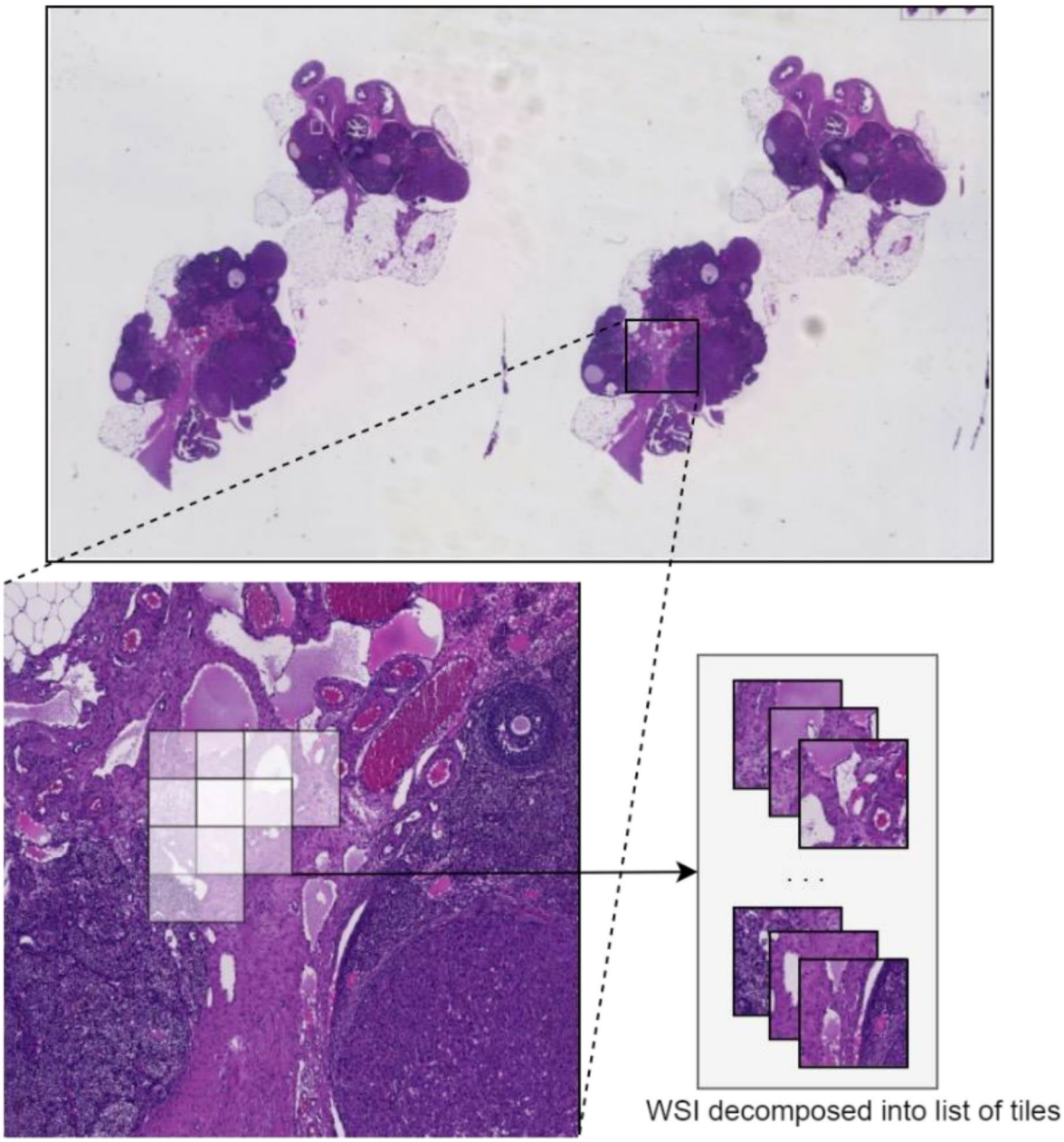

Fig. 3. Illustration for tiling procedure on an ovary image. Tiles are extracted from the WSI using the sliding window approach with an increment of 128 pixels. To minimize computation time and memory usage, tiles which are mostly white are excluded from the tiled dataset.

pathology predictions are shown in supplementary Fig S5 (Testis), Fig S6 (Prostate), Fig S7 (Female Kidney), Fig S8 (Male Kidney), and Fig S9 (Ovary). Overall, the model predictions show good alignment with manual annotations.

More details about the overlap tradeoffs, spatial awareness PTO tiling method, the deep neural network model used in SA-HCNN, and the inference procedure are discussed in the Methods section.

Deep learning training and testing

Table 1 shows the results of the SA-HCNN deep learning approach for each tissue type and for each pathology within a tissue type, along with the Ignore class which represents non-tissue regions in the image. Performance is measured according to the entropy loss, F-score, and the confusion matrix (true positives and negatives, false positives and negatives). The entropy loss of the model is the average cross-entropy which is a measure of the total uncertainty of the model.

Figure 4

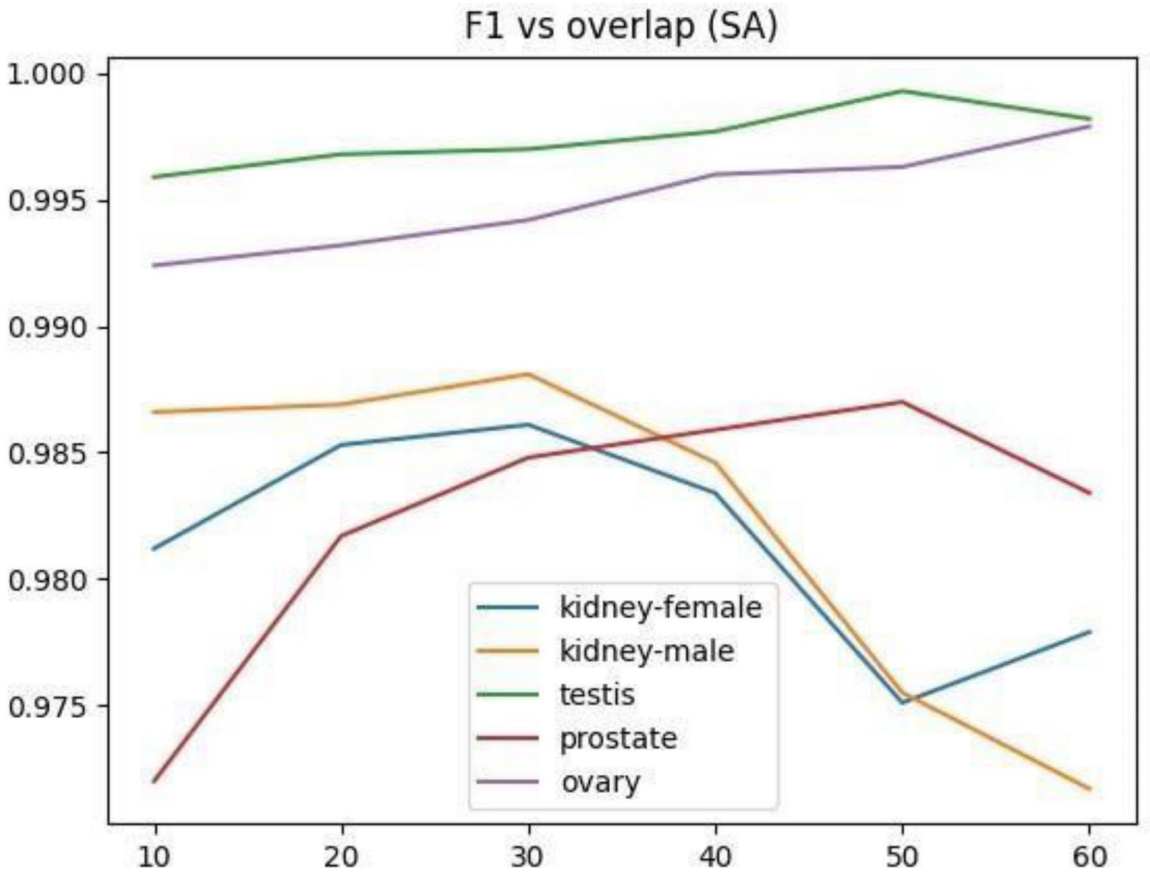

Fig. 4. The graph shows macro F-score vs. overlap φ from averaged test runs across all data. After an overlap of 30%, the model's predictions can become unstable due to overfitting the minority classes. $\varphi = 0.3$ provides the best stability generally while maximizing the F-score.

$$L = -\sum_i p(x_i) \log(q(x_i))$$

In the cross-entropy calculation above, $p \in [0, 1]$ is the ground truth probability distribution and $q \in [0.0, 1.0)$ is the model's predicted probability distribution. The entropy loss is arguably the best raw metric for evaluating a deep learning model's performance since it is a measure of distance between the set of ground truth labels and predicted labels. The F-score shown below is the harmonic mean between the recall and precision of a model.

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

F-score is a better metric than the accuracy of a model when working with imbalanced data. In the case of pathology data, some classes have an imbalance ratio up to 1:100, because healthy tissue areas typically outnumber pathology areas. For example, Table 1 shows a ratio of 178:18171 for testis vacuole tiles compared to non-vacuole tiles. The F-score considers the support for both the positive and negative class, so a high F-score ensures a low false-positive rate.

The values seen in Table 1 are the averaged results of the evaluation on the test data for three runs where 80% of the whole dataset is randomly sampled to create the training dataset and the remaining 20% becomes the test set. Models are trained individually with respect to the tissue type, meaning that tiles taken from two separate tissue types will never appear in the same training dataset. For the epigenetics data, five models are trained in total, one for each tissue type (testis, prostate, female kidney, male kidney, and ovary). The results in Table 1 indicate that the SA-HCNN deep learning model successfully learned the classification tasks, with entropy losses as low as 0.0 (testis maturation arrest and testis antral follicle) and no higher than 0.06 (prostate atrophy) across all tissue types and pathologies. F-scores were all higher than 0.95 with most higher than 0.99, indicating that the model predictions generalize well even in the presence of significant class imbalance. Table 1 also includes

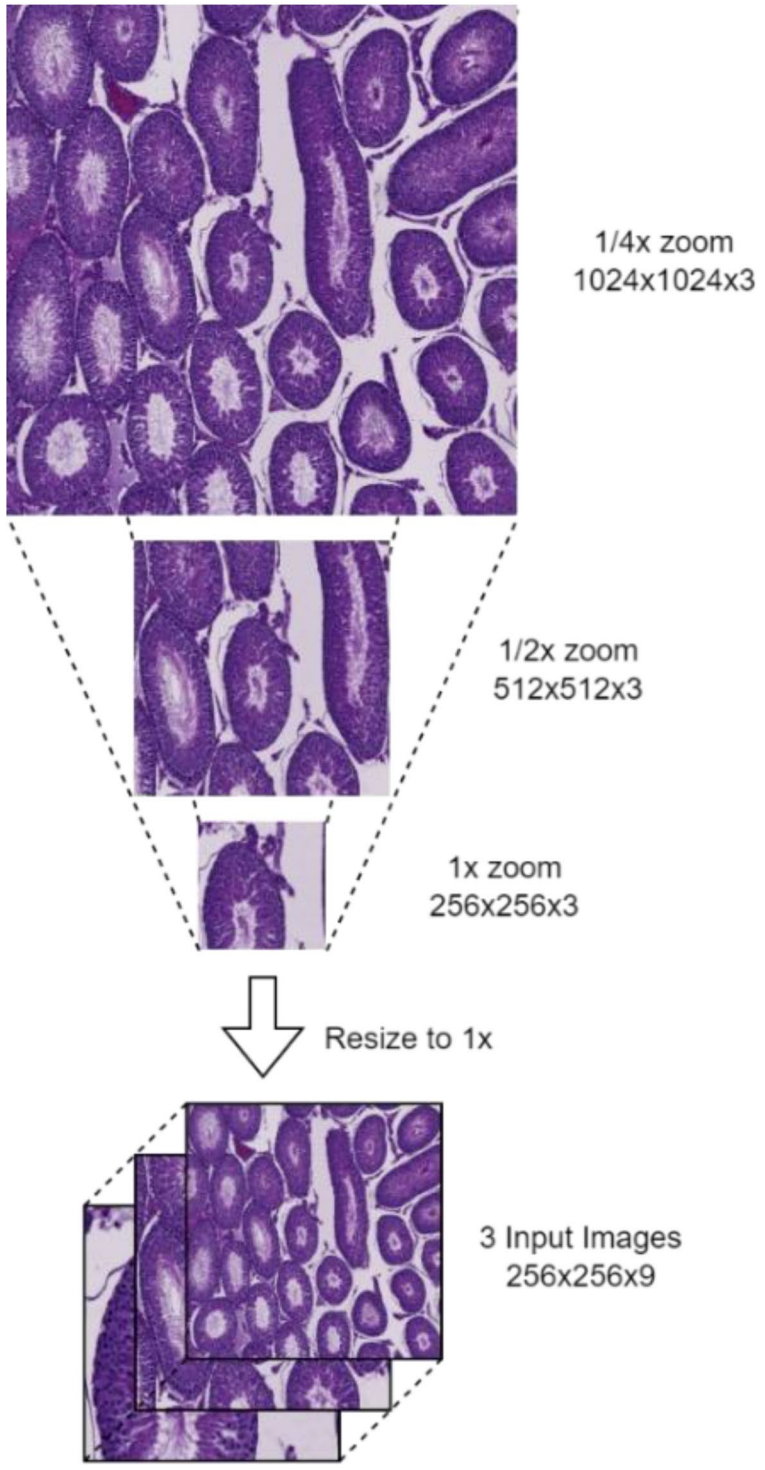

Fig. 5. Example of tile pyramid with a depth of 3. Lowest level tile is the true tile in which the classification is being made for. The larger tiles allow the model to understand its surroundings, thus being spatially aware.

the 95% confidence interval for each entropy loss and F-score value. The confidence intervals are small, which indicates that the means over the three runs are representative of true performance with high confidence.

Comparing manual predictions to deep learning predictions

The SA-HCNN deep learning results were computed for all images in the F1 and F2 generation ifosfamide datasets and compared with the analysis using the manual counts method. Figures 8 and 9 show the frequency of

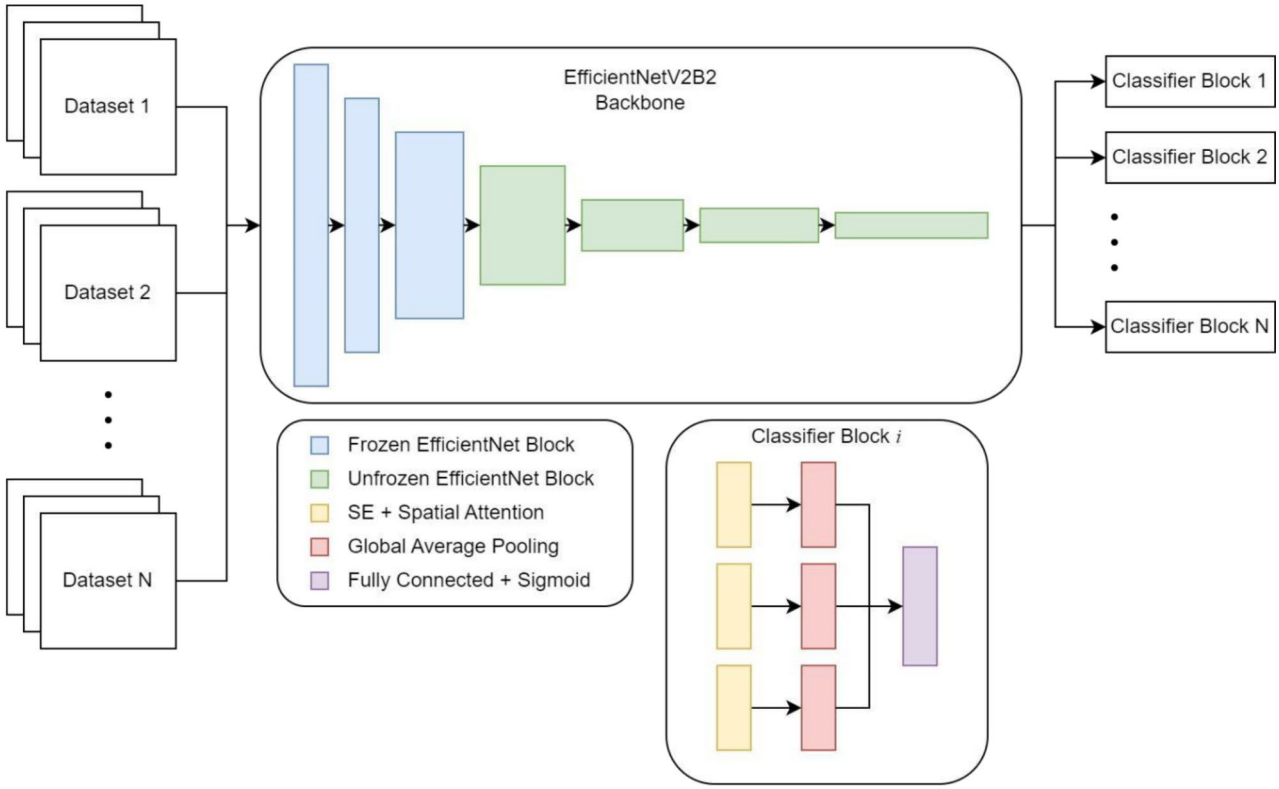

Fig. 6. Illustration of the architecture for the deep learning model with spatial awareness. The diagram represents the model architecture during training since bagging is used to train a model ensemble. There are N datasets as the input to the model along with N classifier blocks that are independently trained in parallel on the data ensembles. During inference, the model only has one input that is processed by each classifier block, and the prediction of each block is averaged to create the final prediction. Each classifier block has 3 inputs, 1 for each of the CNN outputs from the input image pyramid with a depth of 3.

organ disease for each group of rats. Pathology analysis from manual counts are shown at the top of each figure, and deep learning area ratios are shown at the bottom of each figure. The charts for the manual counts analysis were taken directly from the published paper (Thompson, et al., 2022) [28]. Each bar graph shows the frequency of disease for each of the examined organs in the animals. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p<0.05$, $(**)p<0.01$, and $(***)p<0.001$ are presented above the bar experimental group bars for both manual and deep learning predictions. For the deep learning predictions, there are some instances where the total number of tissue images for a specific organ is less than that of the manual counts. This is because a few images were ruined in the transfer process, but the frequency does not differ much since the sample size is sufficiently large (Table 2).

Figure 8 compares the results from the manual analysis (top) and SA-HCNN analysis (bottom) for testis disease, prostate disease, and male kidney disease. Both approaches identify significantly lower frequency of testis disease in the F2 generation. Both approaches also identify an increased frequency of male kidney disease in the F1 generation, but only the SA-HCNN approach finds this significant according to the Fisher exact test at the $p<0.05$ level (*). Figure 9 compares the manual and DL analyses for ovary disease and female kidney disease. Again, the general trends in disease frequency are highly correlated. In particular, both approaches identify a significant increase in female kidney disease in the F1 generation. The general conclusion made by the epigenetics researchers is that there is a direct correlation between ifosfamide exposure and transgenerational diseases. To elaborate, the exposure had little to no effect on the F0 generation of rats but resulted in significant changes in pathology frequency through generations F1 and F2.

Since the SA-HCNN method predicts the presence of each individual pathology, there is the opportunity to perform a more detailed pathology specific analysis of the tissue. Figures 10, 11, 12, 13 and 14 compares the frequency of each individual pathology for the control versus experimental animals. This allows consideration of significantly higher or lower frequencies for an individual pathology. For example, for the F2 prostate results in Fig. 11, the experimental group had a significantly higher presence of vacuoles than the control group. A similar result is seen for cysts in F1 male kidney (Fig. 12), and cysts and thickened Bowman's capsules in F1

Figure 7

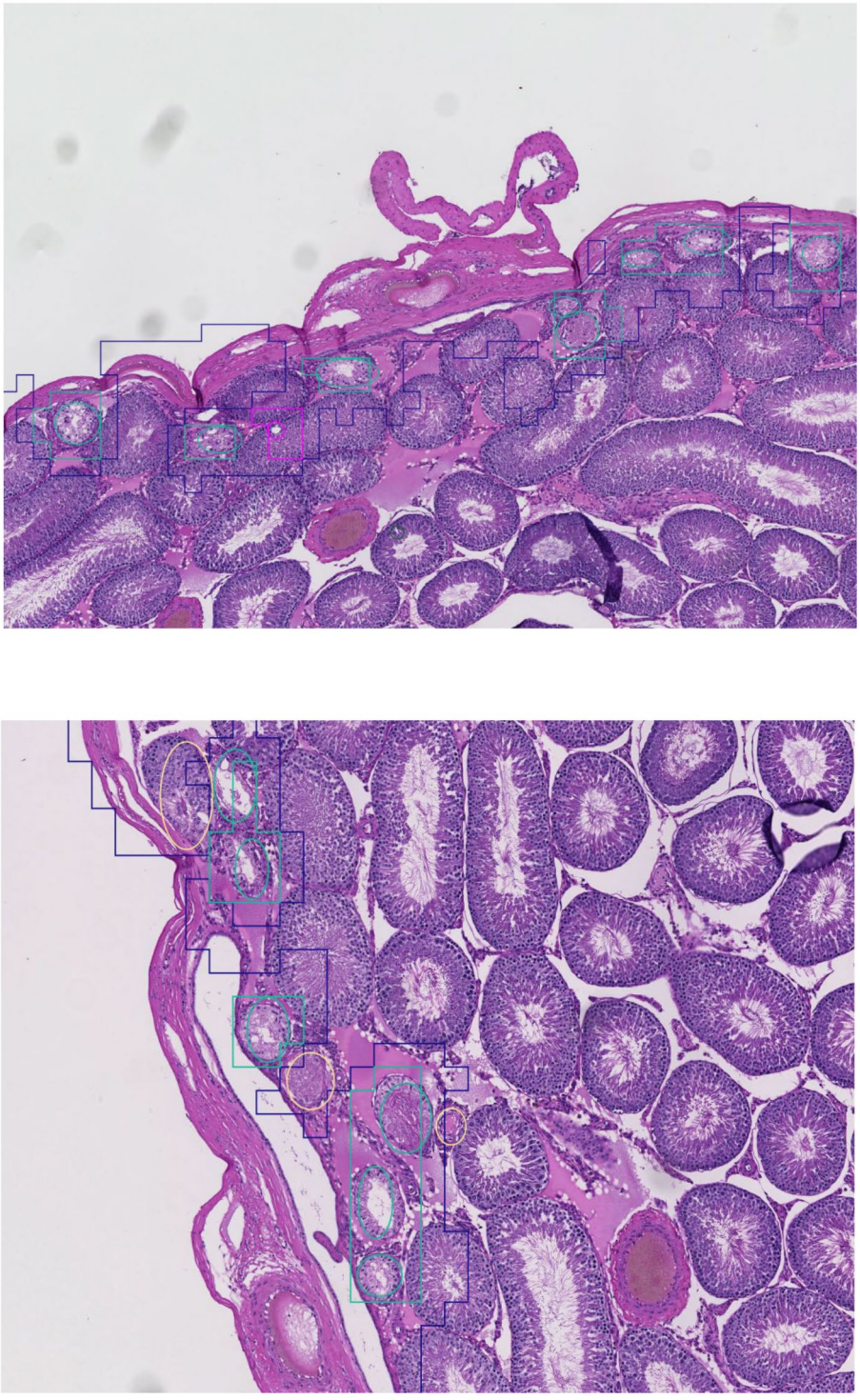

Fig. 7. The images shown are snippets from an annotated testis image. The circular annotations are the perimeter of the ground truth whereas the jagged/boxy annotations are the perimeter of the model prediction super-pixel masks. There is no prediction for the yellowish annotation since that annotation is a part of the negative class.

female kidney (Fig. 13). This more detailed analysis capability of the SA-HCNN deep learning approach provides further insights into the impact of the exposure on susceptibility to specific pathologies.

Using a machine with an Intel Xeon 3.6 GHz 8-core CPU, 2 Nvidia RTX 3090 GPUs and 376GB of RAM, processing one tissue image during inference takes 3.3 min on average. There are a total of 705 images in both the

Testis	Entropy Loss	F-score	TP	FN	FP	TN
Ignore	0.0069 ± 0.001	0.9993 ± 0.000	2,760	1	2	15,587
Atrophy	0.0383 ± 0.007	0.9921 ± 0.001	1,281	9	11	17,049
Maturation Arrest	0.0 ± 0.000	1.0 ± 0.000	312	0	0	18,038
Vacuole	0.0038 ± 0.003	0.9945 ± 0.004	177	1	0	18,171
Prostate	Entropy Loss	F-score	TP	FN	FP	TN
Ignore	0.0094 ± 0.001	0.9997 ± 0.000	68,621	25	10	62,249
Atrophy	0.0600 ± 0.005	0.9830 ± 0.001	6,562	132	95	124,116
Collapsed Prost	0.0026 ± 0.000	0.9988 ± 0.000	4,132	5	4	126,763
Hyperplasia	0.0035 ± 0.000	0.9798 ± 0.001	323	7	5	130,569
Vacuole	0.0078 ± 0.001	0.9509 ± 0.004	288	20	9	130,588
Female Kidney	Entropy Loss	F-score	TP	FN	FP	TN
Ignore	0.0100 ± 0.005	0.9996 ± 0.000	20,072	9	6	34,254
Cyst	0.0112 ± 0.009	0.9946 ± 0.003	1,610	8	9	52,714
Reduced Glomeruli	0.0159 ± 0.003	0.9675 ± 0.007	373	15	9	53,944
Thickened Bowmans	0.0097 ± 0.003	0.9695 ± 0.009	244	7	7	54,083
Male Kidney	Entropy Loss	F-score	TP	FN	FP	TN
Ignore	0.0322 ± 0.001	0.9986 ± 0.000	26,369	38	35	52,490
Cyst	0.0455 ± 0.004	0.9943 ± 0.001	9,030	49	54	69,798
Reduced Glomeruli	0.0308 ± 0.004	0.9597 ± 0.004	834	35	35	78,028
Thickened Bowmans	0.0352 ± 0.003	0.9913 ± 0.001	4,602	42	38	74,250
Ovary	Entropy Loss	F-score	TP	FN	FP	TN
Ignore	0.0010 ± 0.001	1.0 ± 0.000	11,767	0	0	11,449
Antral Follicle	0.0 ± 0.000	1.0 ± 0.000	82	0	0	23,134
Antral Follicle wo-oocyte	0.0198 ± 0.005	0.9988 ± 0.000	5,606	8	5	17,597
Large Cyst	0.0139 ± 0.003	0.9975 ± 0.000	1,881	4	5	21,326
Preantral Follicle ov	0.0069 ± 0.005	0.9750 ± 0.019	89	3	1	23,122
Primordial Follicle	0.0074 ± 0.004	0.9917 ± 0.005	300	0	4	22,912
Small Cyst	0.0149 ± 0.003	0.9939 ± 0.001	812	5	4	22,395

Table 1. Tracked metrics during model testing for each pathology of each tissue type. The first column shows the tissue type and pathology. Columns 2 and 3 are the averaged loss and F1 score for all training runs along with 95% confidence intervals. The remaining columns show the average confusion matrix entries over all training runs.

F1 and F2 datasets, resulting in the deep learning model fully annotating all images along with WSI predictions for each image in ~ 35 h. The same task took human annotators over a year to complete while needing multiple workers and only annotating portions of each image.

Analysis of other datasets

In order to evaluate the generalizability of the approach, SA-HCNN was applied to five other datasets. Researchers at the Pattern Recognition Lab, Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, created a completely annotated WSI dataset of canine breast cancer[29]. The dataset consists of tiles extracted from WSI's using coordinate annotations where each coordinate represented an instance of mitosis. In the paper, the researchers present three datasets, the first manually expert labeled (MEL) dataset is used to train the deep histology model and compare results. The second and third datasets in the paper are highly specialized and are partially generated using their own machine learning models, so there is inherent bias in the datasets. In total, there are 21 WSI's that contribute to the tiles gathered for the MEL dataset, 7 of which are used as the holdout set for evaluation.

TUPAC16 auxiliary dataset (https://tupac.grand-challenge.org/) consists of 73 weakly annotated WSI's for mitosis in human breast tissue. A recent study was done for testing the performance of many popular deep learning algorithms on this dataset, as well as state of the art deep neural networks[30].

MHIST: A Minimalist Histopathology Image Analysis Dataset, is a small-scale histology tile dataset proposed to be a new baseline in the field of deep histology[31]. This dataset comprises 3,152 hematoxylin and eosin (H&E)-stained Formalin Fixed Paraffin-Embedded (FFPE) fixed-size images (224 × 224 pixels) of colorectal polyps from the Department of Pathology and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (DHMC).

PatchCamelyon (PCam)[32] is another tile dataset extracted from the Camelyon16 dataset (https://camelyon16.grand-challenge.org/). The PCam dataset consists of 327,680 tile images (96 × 96 pixels) extracted from histopathologic scans of lymph node sections. Each image is annotated with a binary label indicating presence of metastatic tissue.

Figure 8

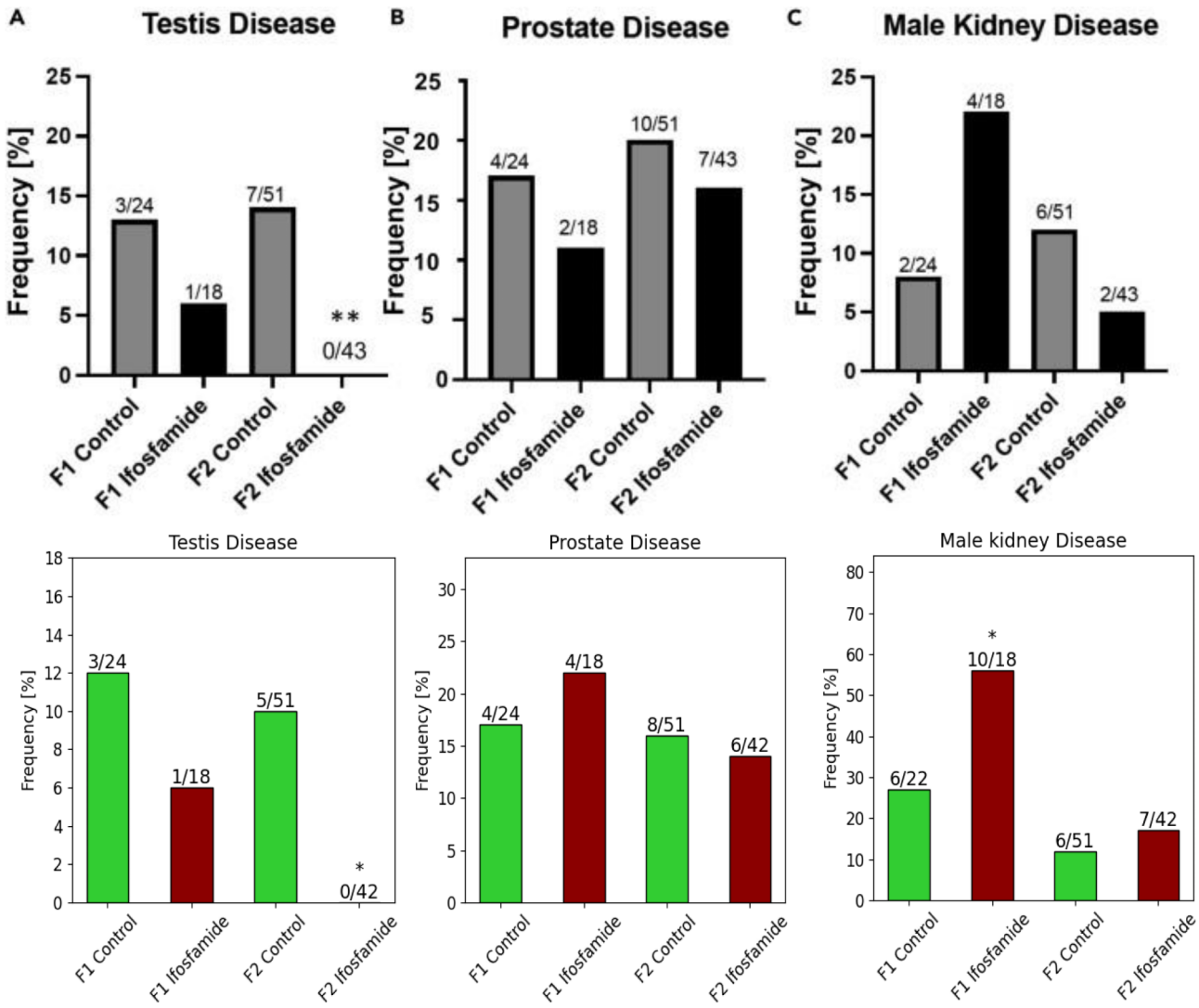

Fig. 8. Pathology analysis from manual counts (top) and deep learning area ratios (bottom) for male organs. The charts for the manual counts analysis were taken directly from the published paper (Thompson et al., 2022). Pathology analysis for F1 and F2 generation control and Ifosfamide lineage 1-year-old rats. Each bar graph shows the frequency of disease for each of the examined organs in the animals. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p < 0.05$, $(**)p < 0.01$, and $(***)p < 0.001$ are presented above the bar experimental group bars for both manual and deep learning predictions.

The results of applying SA-HCNN on the above four datasets are shown in Table 3, along with the best performance reported in the literature. The above four datasets consist of pre-generated tiles, so the spatial awareness (SA) component of SA-HCNN was not utilized; therefore, the method is listed as only HCNN in Table 3. The results show that the HCNN approach outperforms the previously best reported measure in all four datasets (F-score for MEL and TUPAC16, AUC for MHIST and PCam).

In regards to full WSI datasets, BACH is one of the most popular open-source histology datasets hosted by the International Conference on Image Analysis and Recognition (ICIAR)[33]. BACH is an acronym for breast cancer histology and the dataset consists of 40 tissue slide images with resolutions (39,980 – 62,952) x (27,972 – 44,889). The ground truth for the tissue images are pixel masks similar to those used in the training dataset. Again, the pixel masks are converted to polygons which are then used to further label tiles created in the tiled dataset. There are 3 classes in the BACH dataset: Benign, InSitu, and Invasive.

The procedure for training the SA-HCNN model on the BACH dataset is the same as that of the primary dataset. There were no changes to the model since the purpose of testing on external data is to ensure that the model and tiling method are general purpose and can perform well on all histology data. Table 4 shows the accuracy results of applying SA-HCNN to the BACH dataset along with the reported accuracy of expert

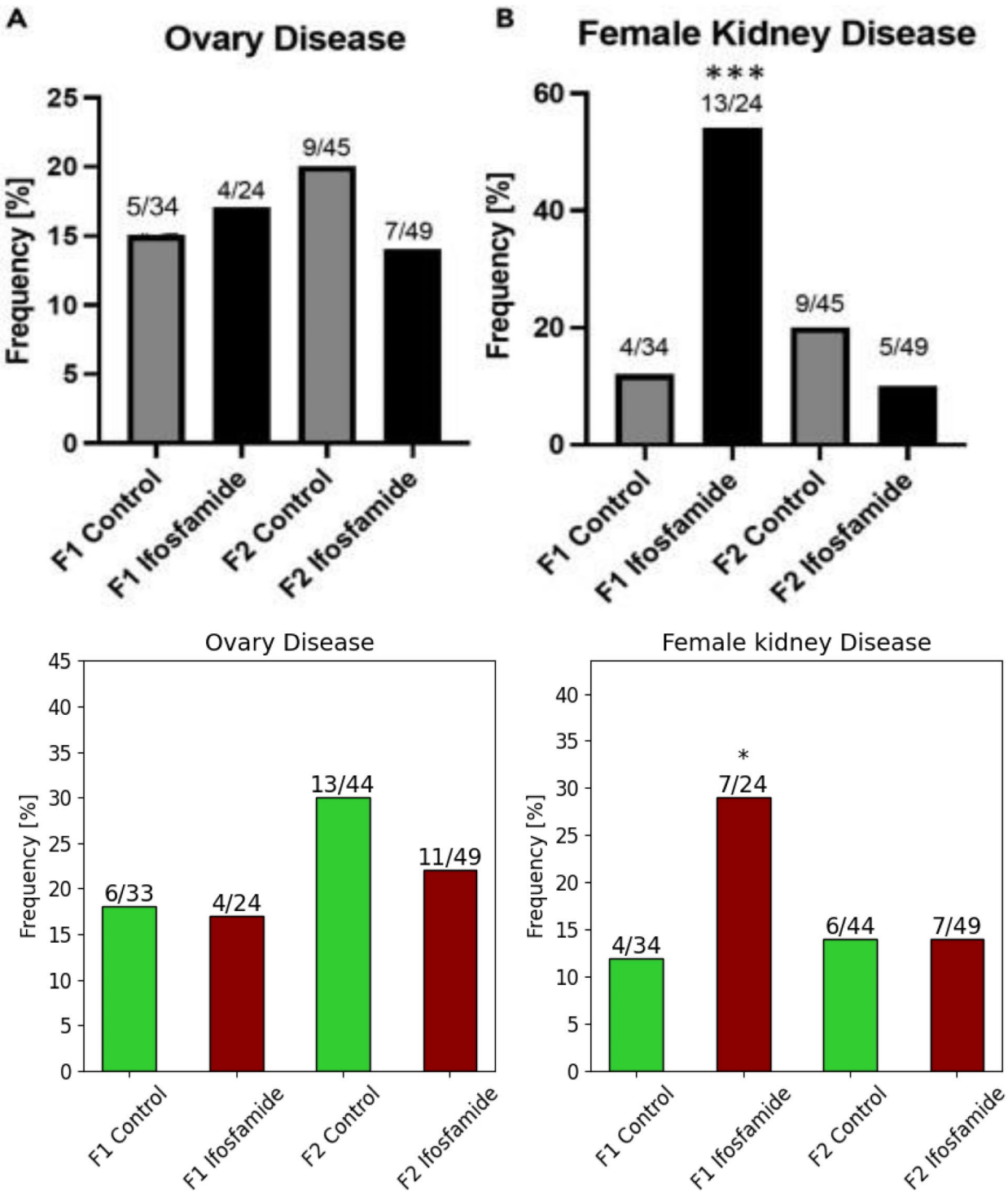

Fig. 9. Pathology analysis from manual counts (top) and deep learning area ratios (bottom) for female organs. The charts for the manual counts analysis were taken directly from the published paper (Thompson et al., 2022). Pathology analysis for F1 and F2 generation control and Ifosfamide lineage 1-year-old rats. Each bar graph shows the frequency of disease for each of the examined organs in the animals. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p < 0.05$, $(**)p < 0.01$, and $(***)p < 0.001$ are presented above the bar experimental group bars for both manual and deep learning predictions.

Macro F-Score	Testis	Prostate	Female Kidney	Male Kidney	Ovary
SA-HCNN + PTO Tiling Method	0.9970	0.9848	0.9861	0.9881	0.9942
SA-HCNN + Center Tiling Method	0.9847	0.9062	0.9779	0.9894	0.9095
Baseline EfficientNetV2 + PTO Tiling Method	0.9360	0.7789	0.8453	0.8519	0.8297
SegFormer + Center Tiling Method	0.4148	0.2477	0.5235	0.3548	0.2769
SegFormer + PTO Tiling Method	0.3265	0.2489	0.2394	0.3028	0.2633

Table 2. Macro F-score averaged over each tissue type for the SA-HCNN with PTO tiling, the SA-HCNN with Center tiling, the current best baseline CNN (EfficientNetV2) with PTO tiling, SegFormer with center tiling, and Segformer with PTO tiling. Results show that the SA-HCNN approach significantly outperforms the baseline EfficientNetV2 and image segmentation approach, and the PTO tiling provides a small additional performance improvement over Center tiling.

predictions and the best DL approach. The results show that SA-HCNN outperforms both the expert and DL methods. The superior performance of SA-HCNN is primarily due to the PTO tiling method, which allowed the model to train on a much larger tile dataset while boosting the size of the minority class which helps balance the data.

Discussion

Deep learning evaluation

As seen in Tables 1 and 2, the HCNN model architecture and training method are sufficiently effective for achieving high F-scores on the training set for all tissue types given. From the confusion matrix values, we also show that the data imbalance has not caused the model to be biased or overfit towards data of the majority classes. The generalization strength of the model comes from the ensemble learning that uses majority voting from models trained on bootstrapped subsets of the training data. The use of sigmoid activations is also key to performing well on the minority classes, as severe overfitting would occur otherwise. The averaged F-scores for SegFormer are significantly lower than the other tests because for most of the class cases, the model was insufficient to appropriately handle the severe class imbalance, and collapsed to becoming a majority classifier. This outcome can be expected given that transformers tend to scale best with extremely large datasets (more than 10^6 samples). This is why they are the model of choice for modern self-supervised learning tasks like next-token prediction and masked auto-encoders. In the case of the epigenetic histology data, there are less than 10^3 positive class samples trained against a more than 10^5 negative class samples for many classes. Overall, the evaluation of the SA-HCNN deep learning approach shows efficient and effective classification of multiple pathologies across multiple tissue types. The generality of the method is confirmed using other publicly available histology datasets. Therefore, the SA-HCNN is a general-purpose state-of-the-art histology classification method that can be applied to a variety of gigapixel histology image datasets given that some amount of manually-annotated pathologies are provided.

Manual versus deep learning histology analysis

The analysis from the manual counts and the area ratios in Figs. 8 and 9 show that the entire deep learning pipeline arrived at a similar conclusion about the datasets. The most obvious trend is the significantly increased frequency of kidney disease in the experimental group for F1 generation, which was also predicted by the deep learning method. Additionally, significant differences in pathology frequencies given by the Fishers exact test are similar across manual and deep learning predictions. While the frequencies are not the same across the manual and automated methods, the deep learning method resulted in a higher frequency of disease in the experimental group overall which is anticipated. Recall, the final WSI predictions for the manual method were determined by the sum of pathology counts in only subsections of all the images. The deep learning model looks at the entire image, so it is reasonable to assume that the deep learning method is still accurate even though the final predictions are not the same. Additionally, the deep learning predictions take the area of pathologies into account, whereas the manual counts predictions are determined only by the instance counts. This indicated that there could also be some correlation between the size of the pathologies and the probability of an organ being diseased.

The results from the Fisher exact test show similarities in significant statistical differences across both prediction methods. For the manual counts method, the control group was found to have a significantly higher frequency of testis pathology than the F2 generation ifosfamide lineage animals ($p < 0.05$), indicating a decrease in testis pathology frequency in F2 generation ifosfamide males[28]. This statistical significance was also predicted by the deep learning method. Furthermore, there was a significant increase in pathology frequency in the F1 generation ifosfamide lineage animals relative to the controls ($p < 0.05$) in female kidney[28] found by using both manual and deep learning methods. The only outlier from the Fisher exact test is male kidney, which the deep learning method found a significant increase in pathology frequency in the F1 generation ifosfamide lineage animals relative to the controls, whereas the manual counts did not. Although, there is still a visible increase in

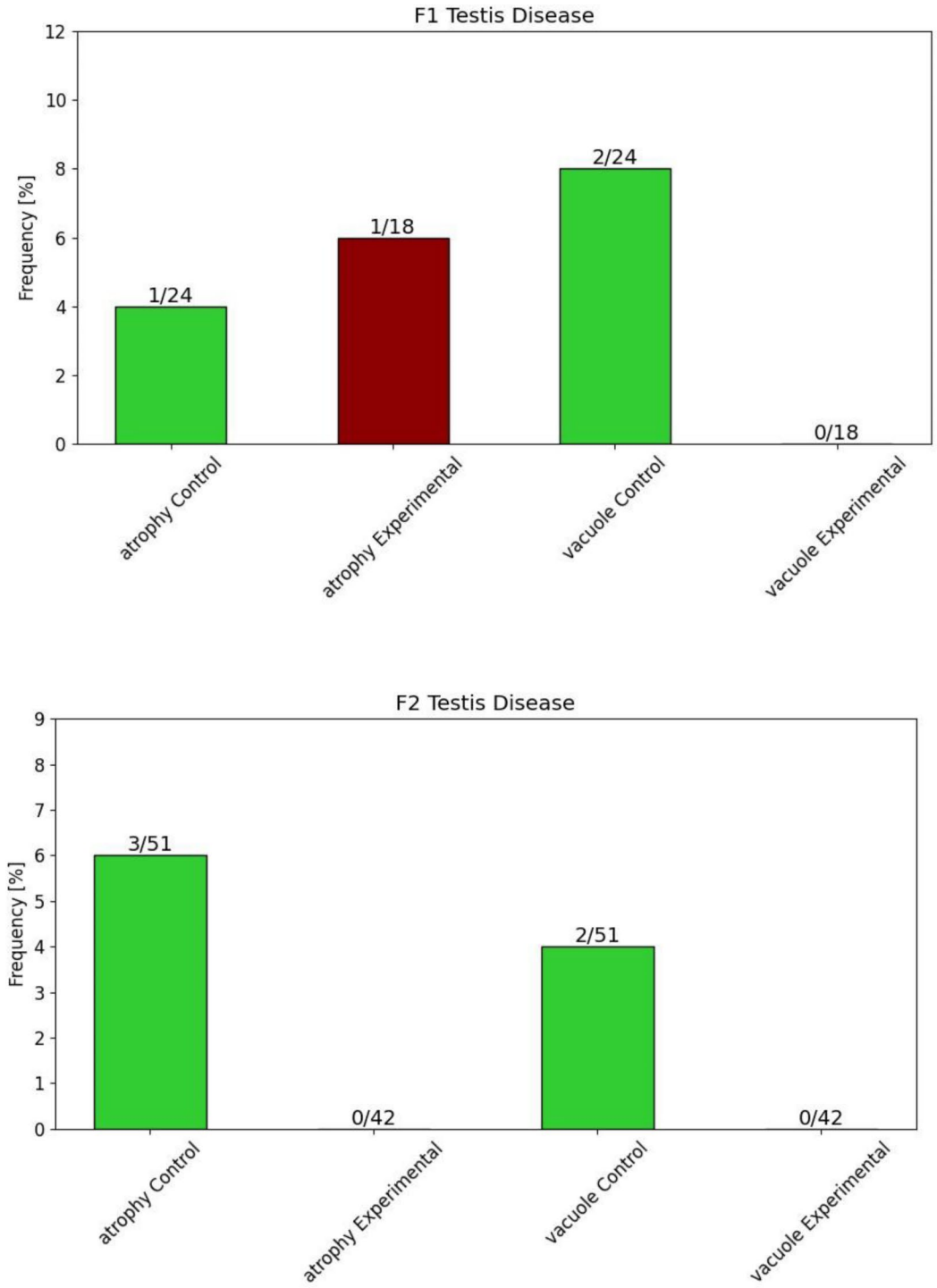

Fig. 10. Deep learning pathology independent analysis for F1 (top) and F2 (bottom) groups for testis tissue. Each bar graph shows the frequency of each pathology per image for each of the examined organs in the animals. The pathology/group pair is shown as text below each bar, the total number of pathologies is specific to the tissue type. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p < 0.05$, $(**)p < 0.01$, and $(***)p < 0.001$ are presented above the experimental group bars.

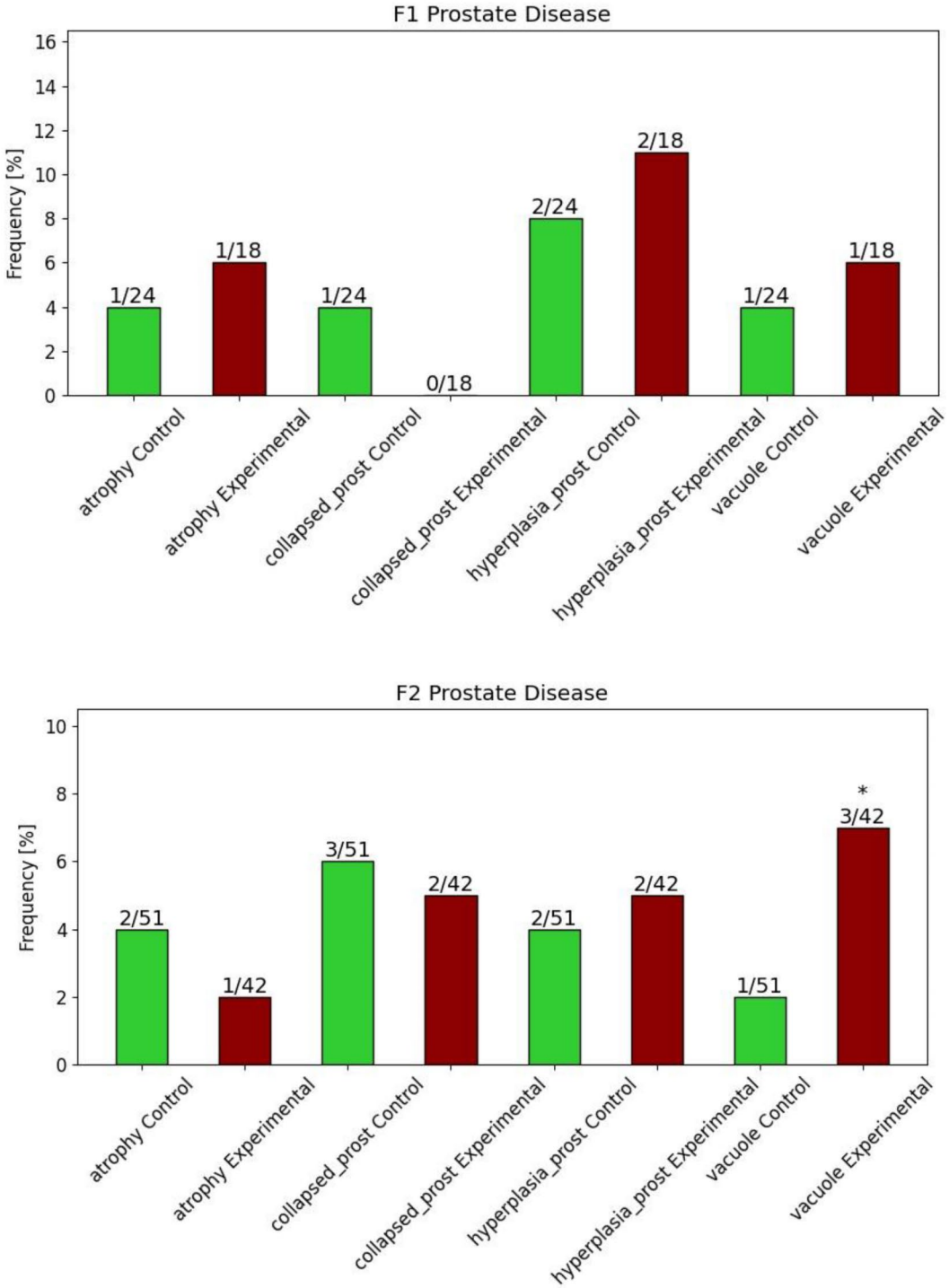

Fig. 11. Deep learning pathology independent analysis for F1 (top) and F2 (bottom) groups for prostate tissue. Each bar graph shows the frequency of each pathology per image for each of the examined organs in the animals. The pathology/group pair is shown as text below each bar, the total number of pathologies is specific to the tissue type. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p < 0.05$, $(**)p < 0.01$, and $(***)p < 0.001$ are presented above the experimental group bars.

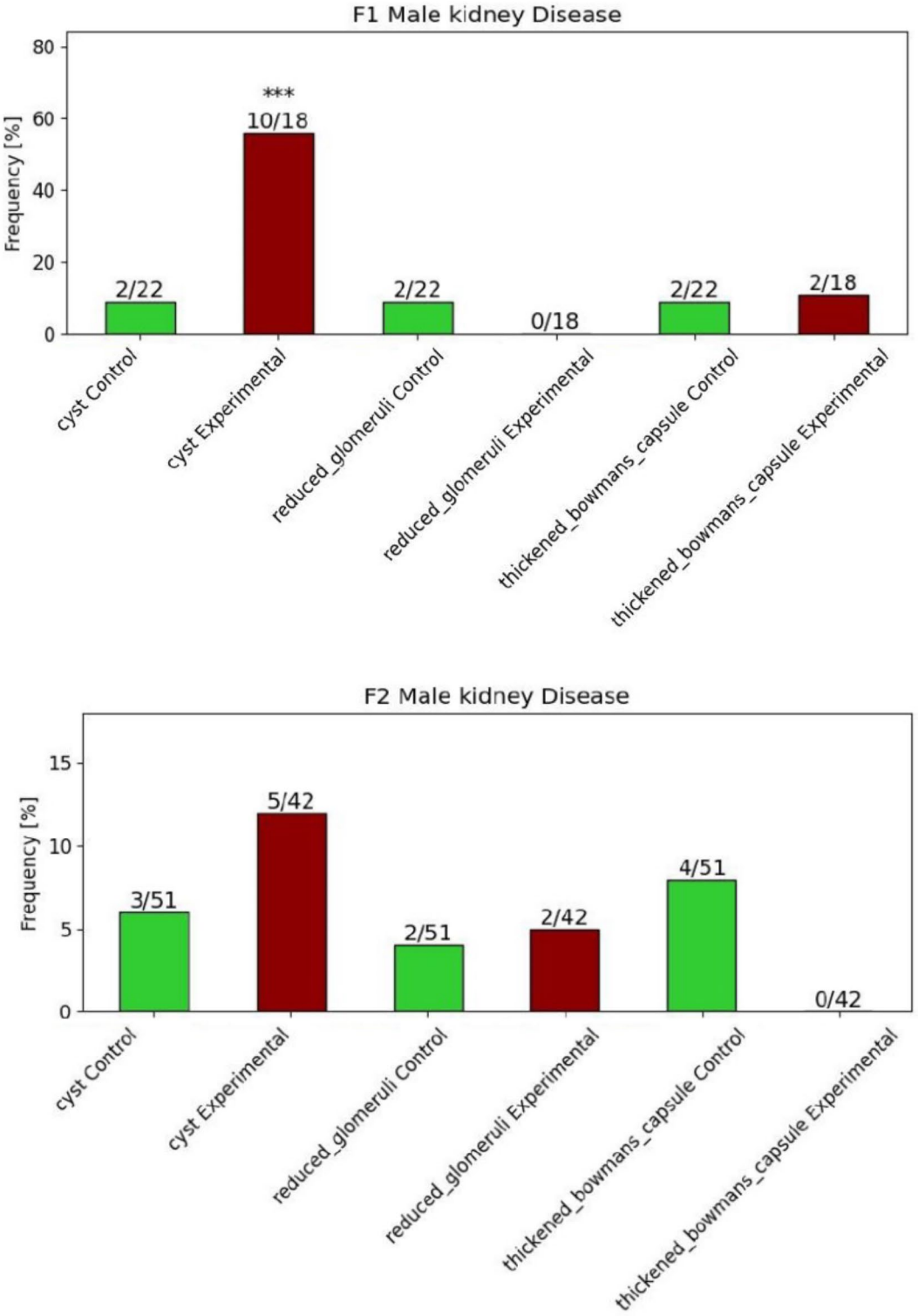

Fig. 12. Deep learning pathology independent analysis for F1 (top) and F2 (bottom) groups for male kidney tissue. Each bar graph shows the frequency of each pathology per image for each of the examined organs in the animals. The pathology/group pair is shown as text below each bar, the total number of pathologies is specific to the tissue type. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p < 0.05$, $(**)p < 0.01$, and $(***)p < 0.001$ are presented above the experimental group bars.

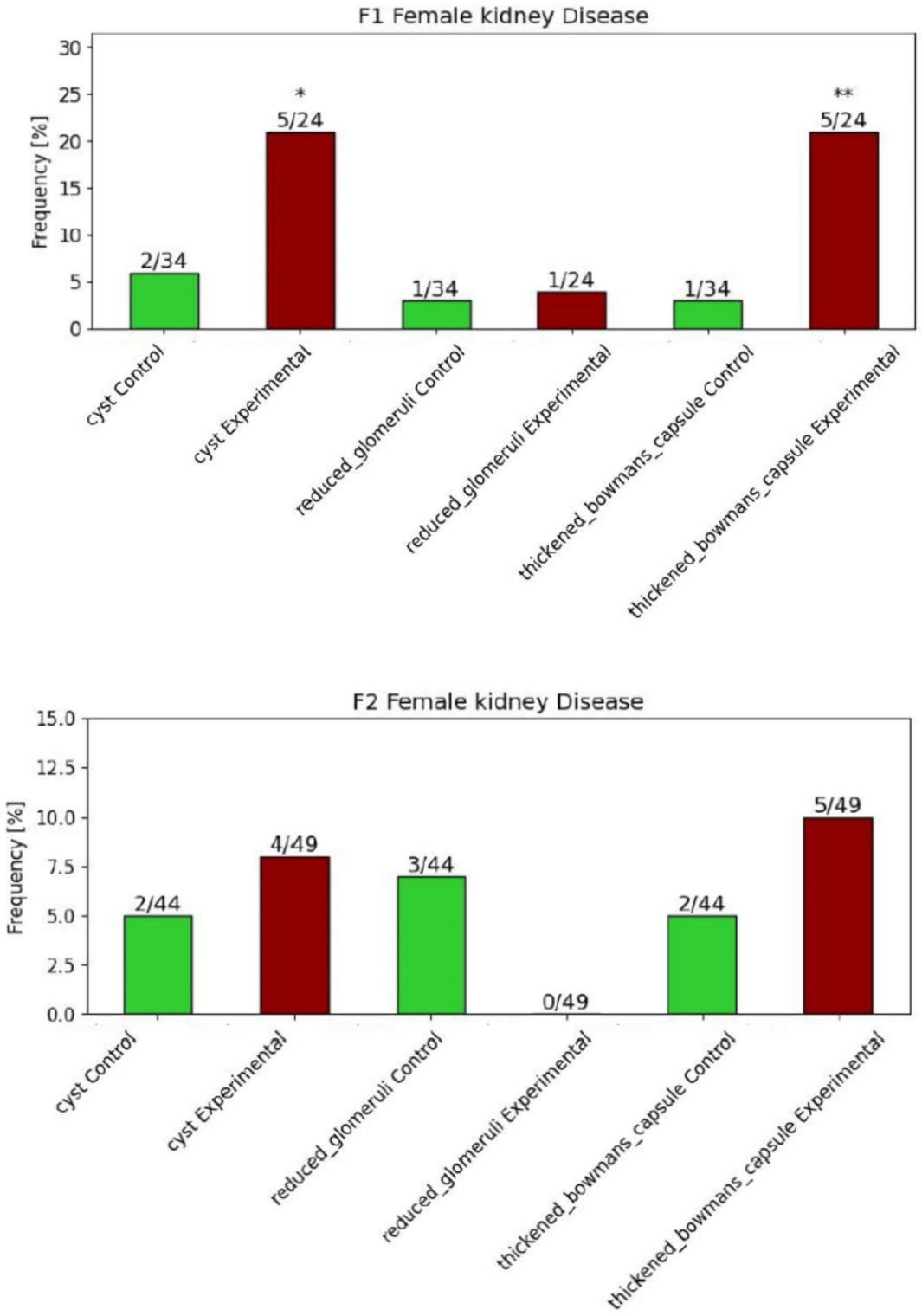

Fig. 13. Deep learning pathology independent analysis for F1 (top) and F2 (bottom) groups for female kidney tissue. Each bar graph shows the frequency of each pathology per image for each of the examined organs in the animals. The pathology/group pair is shown as text below each bar, the total number of pathologies is specific to the tissue type. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p < 0.05$, $(**)p < 0.01$, and $(***)p < 0.001$ are presented above the experimental group bars.

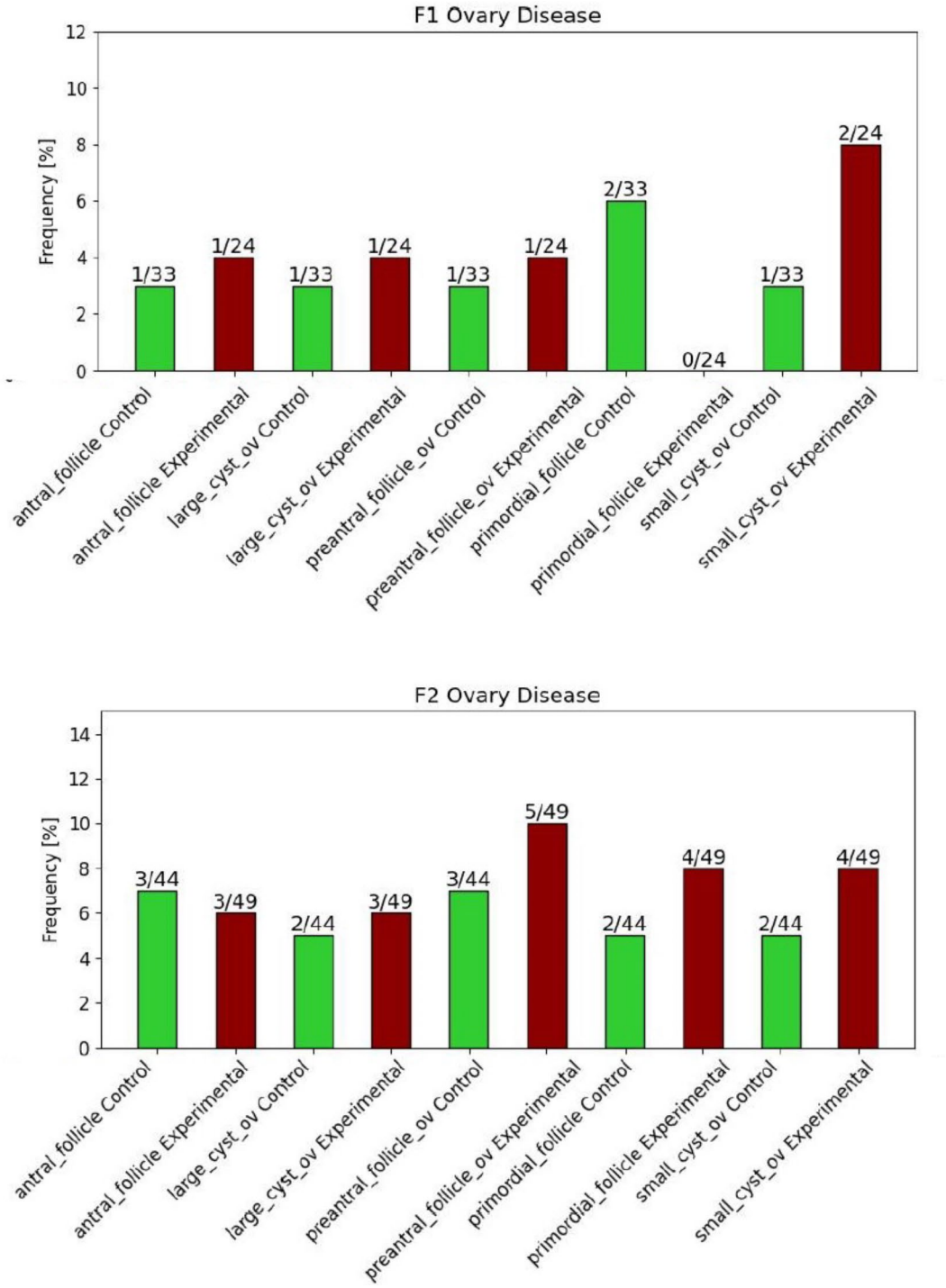

Fig. 14. Deep learning pathology independent analysis for F1 (top) and F2 (bottom) groups for ovary tissue. Each bar graph shows the frequency of each pathology per image for each of the examined organs in the animals. The pathology/group pair is shown as text below each bar, the total number of pathologies is specific to the tissue type. The total number in each bar is the total number of tissue slides for each group/tissue pair. Asterisks indicating a statistical difference as determined by Fisher exact test $(*)p < 0.05$, $(**)p < 0.01$, and $(***)p < 0.001$ are presented above the experimental group bars.

pathology frequency in F1 generation ifosfamide lineage animals for male kidney manual counts. The similarities in significant statistical differences across methodologies show that the PTO + SA-HCNN method is capable of drawing valid conclusions about the pathology found in histopathology images. Moreover, the deep learning method perceived about 50–66% more data per image to determine pathology frequencies, which supports a more robust conclusion for the images being classified.

Comparative results on additional datasets	Method	Comparative best result from paper
MEL	(Aubreville et al., 2020) DL	F-score: 0.707 ± 0.013
	HCNN	F-score: **0.832**
TUPAC16	(Sohail et al., 2021) DL	F-score: 0.72 ± 0.026
	HCNN	F-score: **0.986**
MHIST	(Wei et al., 2021) DL	AUC: 0.841 ± 0.011
	HCNN	AUC: **0.927**
PCam	(Veeling et al., 2018) DL	AUC: 0.963
	HCNN	AUC: **0.994**

Table 3. Comparison of best metrics shown in related work compared to those generated from the proposed deep learning model in this paper. Since these are tile datasets, spatial awareness via pyramid tiles was not used for the proposed model. This variant of the proposed model is referred to as HCNN rather than SA-HCNN.

BACH Tumor Classification	(Aresta et al., 2019) Expert predictions	(Aresta et al., 2019) Best DL approach	SA-HCNN approach
Accuracy	0.8849	0.9396	**0.9873**

Table 4. Comparison of performance on the BACH WSI dataset between the SA-HCNN deep learning approach including the PTO tiling method against the best reported deep learning method and the expert predictions.

The manual analysis of the ifosfamide data took five people and about a year to complete. The finalized PTO + SA-HCNN method took about two weeks to complete model training, and about two and a half days to complete full WSI pathology segmentation and image classification for the entire ifosfamide dataset. The model predictions have been verified at a pathological level by experts in the field, implying that the WSI classifications are also more accurate, given that the entirety of all slides are observed in detail by the SA-HCNN model. Finally, the deep learning method allows for pathology specific analyses shown in Figs. 10, 11, 12, 13 and 14. The pathology specific analysis is calculated the same way as the WSI disease analysis, except the pathologies are kept separate to show which pathologies were directly correlated to the WSI disease classification. This analysis is additionally helpful to the pathology researchers since it allows conclusions to be drawn about the exact type of disease that is correlated with certain exposures.

Conclusions

The proposed SA-HCNN deep learning approach efficiently and accurately classifies pathologies in gigapixel digital histology slides. This approach can process pathologies of any size while also being able to learn how to filter out irrelevant pixel data. High F-scores were achieved on all classes across all tissue types, even when the positive class was severely underrepresented. By utilizing a highly efficient training generator, the neural network can query only the necessary number of samples from the majority class, resulting in a minimized training time and improved performance on datasets with extreme class imbalance.

The approach also includes a super-pixel annotation generator which takes trained models and a tissue slide as input to construct super-pixel predictions that can be viewed in QuPath in just a matter of minutes. Using a statistical approach inspired by the researchers at Skinner labs, the deep learning method also produces WSI predictions at the level of a human pathologist, with the exception of the negligible false-positive rate (can be seen in negative class predictions of supplementary material). Given the human-level segmentation abilities of the deep learning model and the speed of the inference procedure and super-pixel generator, the proposed approach can be very helpful for biomedical researchers using histopathological slides to make a diagnosis. Not only would using this approach save significant amounts of time and money, it would deliver a consistent level of precision that tends to diminish over periods of time with human annotators, as annotating gigapixel tissue images is painstaking work that takes a long time. Moreover, the SA-HCNN model can complete the same task orders of magnitude faster than humans and at higher accuracy.

In conclusion, the use of deep learning in histopathology diagnostics holds significant promise for the advancement of medicine and biotechnology. With the speed and accuracy that deep learning provides, researchers will be able to better understand the environmental effects on biological systems, including humans, and develop technologies and medicines to combat the negative effects of these interactions. The findings from this research will contribute to the fight against cancer and other gene-related diseases. The potential impact of this technology is immense and presents an exciting opportunity for researchers to push the boundaries of medical research and improve human health.

Methods

Tiling procedure

Tiling the gigapixel tissue slides was done by using a sliding window of size 256×256 pixels. The motivation for this tile size stems from the standard input size of pretrained models (traditionally 244×244 in older models)

and from previous studies of deep histology where the authors either use a tile size of 128×128 or 256×256. This tile size is the nearest power of 2 that is optimal for minimizing computational overhead and maximizing information perception. The increment size, or the amount by which the tile window is moved, differs between training and testing. For training, a sliding window increment of 1/4 window size (64px) was used to build a large dataset. For inference, a sliding window increment of 1/2 window size (128px) is used for a 4x speedup in processing time. A smaller increment size can be used to build the dataset used for training since runtime is less of a concern during training, whereas a 1/2 overlap is all that is needed to make accurate super pixel predictions during inference. Additionally, less computation is needed during inference as a result of this.

For building the training dataset, a piecewise condition is used to determine which class a tile belongs to:

$$L_{i,j} = \begin{cases} 1 \;\; if \;\; Intersection\,(p_j,\, T_i) > \varphi \, * \, Area\,(T_i) \\ 1 \;\; if \;\; Intersection\,(p_j,\, T_i) = Area\,(p_j) \\ 0 \;\; otherwise \end{cases}$$

In the piecewise condition shown above, $L_{i,j}$ is the class label associated with the tile T at index i and pathology instance p at index j. p represents any pathology instance which overlaps T. $i \in [0,\, I]$ where I is the total number of tiles in a tissue slide (could be tens of thousands of tiles in one image) and $j \in [0, J]$ where J is the total number of pathologies (classes) that a classifier is trained to predict. $Intersection$ is a function that returns the area in pixels computed from the intersection between pathology instance p_j and tile T_i. $Area$ is a function that simply returns the area in pixels of a polygon. Note, there can be multiple instances of any given pathology in one image, that is why each tile is given its own label. Finally, φ is a scalar that is used to control how much overlap is needed for a tile to be considered a positive example for any given pathology where $\varphi \in [0,\, 1.0)$. The second constraint in the piecewise function ensures that a positive label will be assigned if the entire instance of a pathology is present inside of a tile. This constraint is necessary for small pathologies that would result in $Intersection\,(p_j,\, T_i) \ll \varphi \, * \, Area\,(T_i)$.

In order to find an optimal value for the overlap parameter φ, several values were tried ranging from 0.1 to 0.6. Figure 4 shows the F-score for each overlap value for each tissue type. The results indicated that an overlap of $\varphi = 0.3$ yields good overall performance across all tissue types. An overlap of 30% is large enough that the model would be able to see a reasonable amount of a pathology to make an accurate classification but small enough to be able to make a larger training dataset. An overlap value of $\varphi = 0.3$ is used in all results presented in this paper, including results on other datasets.

A smaller overlap constraint will result in a larger dataset for training given that more examples from each class will be present since there is less of a constraint for positive class examples. On the contrary, for a large overlap parameter, say $\varphi = 0.9$, a tile will only be taken into consideration as a positive example if a pathology is overlapping close to the entire tile. In more traditional sliding window semantic segmentation algorithms, a class label for any given tile would be decided by the center pixel of the tile, i.e., the class label for a tile would be equal to the class that the center tile overlapped. This method works well but is costly given that it generally only works with a small sliding window increment size, i.e., 1/8 or 1/16 of the tile size. The reasoning behind this is in the case of small class instances. Assuming that the tile size being used is 256×256 pixels and there exists a pathology that is ~ 10 pixels in diameter on average, the model is unlikely to see many instances of this pathology if a sliding window increment of 1/1 or 1/2 was used for inference given the center pixel condition.

Deep learning model with spatial awareness

The main issue with classifying an entire tile as opposed to the center pixel method for tile labeling is that spatial information could be lost since the class-relevant information is likely close to the edge of the tile or outside the tile boundary. For instance, when a model is being trained to classify what class the center pixel of a tile belongs to, the model can see 128 pixels in all directions to help understand what to predict. In the case of the overlap method for tile labeling, some of that important information could be just outside of the tile that is being classified, especially if the overlap φ is small. This is where introducing additional spatial information could become useful. Inspired by the glimpse-net[34], spatial awareness is implemented in the deep learning model through pyramid tiling. This is the process of zooming out from the original tile by powers of 2 and inputting all images into the model to be processed in parallel. All zoomed out images are resized to be equal to the model's input size, in this case 256×256. Pyramid tiling can be seen in Fig. 5. Pyramid tiling combined with the tile overlap approach constitutes the Pyramid Tiling with Overlap (PTO) approach used in SA-HCNN.

Table 2 shows the macro F-score for each tissue type for SA-HCNN with PTO tiling, SA-HCNN with Center tiling, and the baseline network EfficientNetV2 with Center tiling. The results for the SegFormer with PTO tiling and Center tiling are also included for comparison. Results show that PTO is superior to the Center tiling approach overall and that the HCNN model is superior to the baseline EfficientNetV2 model. Furthermore, the SegFormer method performed worse than all HCNN. Thus, SA-HCNN with PTO tiling is the recommended approach for histology in gigapixel images.

The model used for tile classification (see Fig. 6) was built using the latest standard practices for high precision image classification. Google's EfficientNetV2 models[15,16] were implemented for the convolutional backbone to the neural network. The EfficientNetV2 architecture provides the best performance while also minimizing the number of trainable parameters. The exact model used is EfficientNetV2B2 since the input size is 260×260 which is the closest to the tile size of 256. To utilize the effectiveness of transfer learning, we use the pretrained ImageNet [35] weights instead of training from scratch, as that has been shown to significantly improve training time and classification performance. Additionally, there are seven total convolutional blocks in EfficientNetV2B2. The first three blocks are left frozen while the last four layers are re-trained so that the convolutional filters learn

to be more representative of the tissue data that is being passed through the model. For the output of the model, N classifier blocks are used for prediction. The classifier blocks take the output of the CNN backbone where attention is used, followed by global average pooling and finally a fully connected layer used for prediction.

Sigmoid activations are used for the classification rather than softmax since multiple classes can be present in one tile making this a multilabel problem, not a multiclass problem. Additionally, we have found empirically that using sigmoid activations is the best way to handle training a model on heavily imbalanced data. Since the class predictions are independent of each other, training in this fashion effectively creates a set of one-vs-rest classifiers, which are prone to overfitting due to a data imbalance since they are binary classifiers. Furthermore, we use binary cross entropy to compute gradients for the model rather than categorical cross entropy. Binary cross entropy allows each classifier to learn independently, meaning that minority class learners will not be over saturated with gradients from the classes with high support. The traditional method of using categorical cross entropy is problematic for learning models on highly imbalanced data. Models trained using SoftMax activations paired with categorical cross entropy tended to be unstable during training, which could lead to model collapse where the model would only predict a single class. In most cases, this was the majority class.

Training procedure

The ensemble method used for training is Bootstrap aggregation, also known as bagging. The purpose behind using bagging for training is to be able to use the entire training data to train the model while simultaneously having large validation sets to accurately monitor the model performance and further use dynamic learning rates for training. Additionally, bagging is a standard machine learning technique which has proven to improve accuracy with more stable predictions[36]. During training, N datasets are created by randomly sampling from the whole dataset with replacement to create the training data for each set. The number of samples taken for each set is equal to the number of images in the whole dataset. By randomly sampling with replacement, the number of unique images in each bagged dataset is approximately $1 - \frac{1}{e} \approx 63\%$ of the original dataset, meaning that each classifier block has $\sim 37\%$ of the whole dataset to be used for validation which is better than the standard 10% traditionally used. By combining what each classifier block learns independently, the data seen by the entire classifier is $1 - \frac{1}{e^N}$ where N is the ensemble number. If $N = 7$ for example, then 99.91% of the whole dataset is used for training across all classifier blocks while still maintaining $\sim 37\%$ of the data for validation on each block. This method turns out to be incredibly beneficial for training stable models which avoid overfitting and makes the use of dynamic training parameters more accurate due to the large validation size. For this reason, learning rate decay is used during training by starting out the model at a learning rate of 10^{-4} and decaying by a factor of 0.1 when the validation loss converges.

During each epoch of training, a randomly sampled batch of images is sampled from each dataset 100 times before computing validation statistics and moving on to the next epoch. Batches are generated by randomly sampling tiles in such a way that the number of samples per class is equal. For example, if there are 4 classes and the batch size is 128, 32 random tiles are sampled from each class to allow the model to train on balanced data batches which leads to more stable training. This method is only applied during training, batches are sampled completely at random during validation. Since we do not move through the training data linearly, i.e., random batches are trained on each step, the number of steps in each epoch is effectively arbitrary. The only effect it has on training is how often the validation statistics are computed which are used to adjust the dynamic training parameters. Adam optimizer is used along with the loss being computed via binary cross entropy, since the model is multi-label.

Part of the novelty of the SA-HCNN method is within the algorithmic implementation of the training procedure. The training data generator is implemented such that memory is minimized without any increase in processing time. For instance, all tile batch samples are collected in real time via array slicing (live tiling) which adds negligible processing time to the training procedure. This means that the tile for training via the tiling procedure can be created using any sliding window increment less than the tile size. For example, if the sliding window increment is 1/4 of the tile size, then the total number of tiles will take ~ 16 times as much memory than one WSI, since the increment is inversely proportional to the quadradic increase in tiles generated from an image. Using live pyramid tiling, an arbitrarily large virtual training dataset can be used while only needing as much memory as required to load the WSIs into memory. This results in the ability to train SA-HCNN on a set of gigapixel images using any high-end desktop rather than a computing cluster. Additionally, this can be used to quadratically increase the number of unique tiles in each class, resulting in low probability of overfitting to the minority classes. Lastly, since the only information about the tiles stored in memory during training are the coordinate locations, the memory efficiency of live tiling is extended to the bagging method by creating N bagged datasets of tile indices, not tiles. So, the training algorithm uses constant memory with respect to the sliding window increment of the tiling procedure and the ensemble size without any significant increase in training time.

Inference procedure

During inference on a new image, a sliding window increment of 1/2 (128-pixel stride) is used. A set of blank canvases of all zeros with the same shape of the input image is created on which to overlay the tile prediction. The number of blank canvases in the set is equal to the number of classes that the model predicts. The prediction value for each tile/class pairing is added to all the corresponding pixels on the respective canvas for all tiles in an image. Recall that the value of a prediction is a probability meaning that it is a scalar between 0 and 1. The pixel values of the canvases are then normalized since the maximum pixel value is equal to the inverse of the sliding window increment squared. For example, with an increment of 1/2 the maximum value of any given pixel is 4 given the quadratic increase in tile overlap. Finally, all pixel values are rounded to the nearest whole number, resulting in pixel groupings with values of 1 and zeros everywhere else. This is how pixel masks are created for prediction.

The predicted pixel masks are jagged since they are created from patched tile predictions, but the images are so large that the mask predictions are sufficient for the segmentation task. Figure 7 shows two examples of pixel mask predictions overlayed on the input image. The images shown are snippets from an annotated testis image. The circular annotations are the perimeter of the ground truth. The polygonal annotations are the perimeter of the model prediction super-pixel masks. The green and light pink annotated regions show the manual and predicted annotations for two different pathologies. There is no prediction for the yellowish annotation since that annotation is a part of the negative class. The long dark blue line denotes the predicted boundary of the tissue. Overall, the model predictions show good alignment with manual annotations, and the pixel mask images can be used to visually confirm predictions and direct the human pathologist to the locations of different pathologies. Additional pixel mask predictions are shown in Supplementary Fig S5 (Testis), Fig S6 (Prostate), Fig S7 (Female Kidney), Fig S8 (Male Kidney), and Fig S9 (Ovary).

Evaluation and WSI classification

Given the ability to accurately classify individual tiles in a whole tissue slide, a method analogous to the manual instance counting was devised to allow us to determine if an animal is diseased or not. Since the model prediction results allows counting the number of predicted tiles for each class, the ratio of the total area of each pathology to the total area of tissue can be computed. This is better than simply using the tile counts to compute statistics because the size of the images may vary, so this is the normalization process to keep track of the relative amount of pathology in an image. For a given pathology p, the area ratio R_p represents the normalized area which can be used to compute group statistics. In the manual counting approach, the determining factor for an animal being diseased depends solely on the mean of any amount of pathology being 1.5 standard deviations higher than the mean of the same pathology count in the control group. The same procedure can be replicated using the area ratios of the pathologies rather than the instance counts. The mean μ and standard deviation σ of the area ratios of the control group are computed for each pathology. The area ratios of the experimental group can then be standardized by subtracting each area ratio by μ and then dividing by σ. Now that the experimental values are standardized relative to the control group, the classification can be made to determine if any whole slide tissue sample taken from the experimental group belongs to a diseased animal as follows:

$$Diseased\,(P) = \begin{cases} True, & \frac{(R_p - \mu_p)}{\sigma_p} \geq \delta\,,\ \exists\, p \in P \\ False, & otherwise \end{cases}$$

In the equation above, P represents the set of pathologies that belong to any one image. δ is the threshold on the number of standard deviations which must be met or exceeded by standardized area ratio for the statement to become true. In other words, if the area ratio R_p for any pathology p in the given image is δ standard deviations above the mean of the control group, then the entire image belongs to the diseased class. In this case, $\delta = 2$ resulted in final predictions that aligned more closely to the manual counts that used $\delta = 1.5$. This is likely due to noise in the deep learning predictions due to a constant but low false-positive rate. For this reason, the higher WSI threshold is justified.

Data availability

Computational script at: https://github.com/holderlb/deep-histology.

Received: 4 April 2024; Accepted: 16 October 2024
Published online: 05 November 2024

References

1. O'Shea, K. & Nash, R. *An Introduction to Convolutional Neural Networks*. arXiv, 1–11 (2015).
2. Girshick, R., Donahue, J., Darrell, T. & Malik, J. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 580–587 (IEEE, Columbus, OH, USA, 2014).
3. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach .Intell.***39**, 1137–1149 https://doi.org/10.1109/TPAMI.2016.2577031 (2017).
4. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 779–788 (IEEE, Las Vegas, NV, 2016).
5. Long, J., Shelhamer, E. & Darrell, T. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 3431–3440 (IEEE, Boston, MA, 2015).
6. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes Comput. Sci.***9351**, 1–8 (2015).
7. Xie, E. *et al.* SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* NeurIPS (2021).
8. Xu, W. *et al.* PIDNet: A real-time semantic segmentation network inspired from PID controller. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5752–5761 (2023).
9. Thompson, R. P., Nilsson, E. & Skinner, M. K. Environmental epigenetics and epigenetic inheritance in domestic farm animals. *Anim. Reprod. Sci.***220**, 106316 https://doi.org/10.1016/j.anireprosci.2020.106316 (2020).
10. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE***86**, 2278–2324 (1998).
11. Krizhevsky, A. *et al.* ImageNet Classification with Deep Convolutional Neural Networks Alex. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)***3**, 1–9 (2012).
12. Simonyan, K. & Zisserman, A. *3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings* 1–14 (Oxford, Oxford, UK, 2015).
13. Szegedy, C. *et al.Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1–9 (IEEE, 2015).

14. He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
15. Tan, M. & Le, Q. V. *36th International Conference on Machine Learning, ICML 2019 10691–10700* 1–11 (Long Beach, 2019).
16. Tan, M. & Le, Q. V. EfficientNetV2: Smaller Models and Faster Training. *CoRR, abs/2104.00298*, 1–11 (2021).
17. Bychkov, D. *et al.* Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.***8**, 3395 https://doi.org/10.1038/s41598-018-21758-3 (2018).
18. Kraus, O. Z., Ba, J. L. & Frey, B. J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics***32**, i52-i59 https://doi.org/10.1093/bioinformatics/btw252 (2016).
19. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.***6**, 26286 https://doi.org/10.1038/srep26286 (2016).
20. Liu, Y. *et al.* Detecting cancer metastases on gigapixel pathology images. *arXiv:1703.02442 [cs.CV]*, 1–13 (2017).
21. Nguyen, H. G., Blank, A., Lugli, A. & Zlobec, I. An effective deep learning architecture combination for tissue microarray spots classification of HE stained colorectal images. In *Proceedings - International Symposium on Biomedical Imaging* 2020-April, 1271–1274 (2020).
22. Tomita, N. *et al.* Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA Netw .Open.***2**, e1914645 https://doi.org/10.1001/jamanetworkopen.2019.14645 (2019).
23. Wang, H. *et al.* Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *J. Med. Imaging (Bellingham)***1**, 034003 https://doi.org/10.1117/1.JMI.1.3.034003 (2014).
24. Lee, K. *et al.* Deep learning of histopathology images at the single cell level. *Front. Artif. Intell.***4**, 754641 https://doi.org/10.3389/frai.2021.754641 (2021).
25. Hoefling, H. *et al.* HistoNet: A deep learning-based model of normal histology. *Toxicol .Pathol.* **49**, 784–797 https://doi.org/10.1177/0192623321993425 (2021).
26. Baxi, V., Edwards, R., Montalto, M. & Saha, S. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Mod .Pathol.* **35**, 23–32 https://doi.org/10.1038/s41379-021-00919-2 (2022).
27. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.***128**, 336–359 (2020).
28. Thompson, R. P. *et al.* Examination of generational impacts of adolescent chemotherapy: Ifosfamide and potential for epigenetic transgenerational inheritance. *iScience***25**, 105570 https://doi.org/10.1016/j.isci.2022.105570 (2022).
29. Aubreville, M. *et al.* A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Sci. Data***7**, 417 https://doi.org/10.1038/s41597-020-00756-z (2020).
30. 30. Sohail, A., Khan, A., Wahab, N., Zameer, A. & Khan, S. A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images. *Sci. Rep.***11**, 6215 https://doi.org/10.1038/s41598-021-85652-1 (2021).
31. Wei, J. *et al.* A petri dish for histopathology image analysis. *Lecture Notes Comput. Sci.***June**, 11–24 (2021).
32. Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. & Welling, M. Rotation equivariant CNNs for digital pathology. *Med. Image Comput. Comput. Assist. Intervent. Lecture Notes Comput. Sci.***11071**, 210–218 (2018).
33. Aresta, G. *et al.* BACH: Grand challenge on breast cancer histology images. *Med. Image Anal.* **56**, 122–139 https://doi.org/10.1016/j.media.2019.05.010 (2019).
34. Mnih, V., Heess, N., Graves, A. & Kavukcuoglu, K. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.***January**, 1–9 (2014).
35. Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.***115**, 211–252 (2015).
36. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 https://doi.org/10.1007/BF00058655 (1996).

Acknowledgements

We acknowledge Dr. Millissia Ben Maamar, and Dr. Jennifer L.M. Thorson for critically reviewing the manuscript. We acknowledge Ms. Heather Johnson for assistance in preparation of the manuscript. This study was supported by John Templeton Foundation (50183 and 61174) (https://templeton.org/) grants to MKS and NIH (ES012974) (https://www.nih.gov/) grant to MKS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

CG Conceptualization, formal analysis, investigation, validation, wrote original draft, reviewed, edited manuscript. LH Conceptualization, formal analysis, investigation, supervision, validation, writing, reviewed and edited manuscript.EN Tissue histology, image analysis, validation, writing, reviewed and edited manuscript.MKS Conceptualization, formal analysis, funding acquisition, investigation, supervision, validation, writing, reviewed and edited manuscript.

Funding

This study was supported by John Templeton Foundation (50183 and 61174) (https://templeton.org/) grants to MKS and NIH (ES012974) (https://www.nih.gov/) grant to MKS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1

0.1038/s41598-024-76807-x.

Correspondence and requests for materials should be addressed to L.H. or M.K.S.