Deep Learning for Necrosis Detection Using Canine Perivascular Wall Tumour Whole Slide Images

Taranpreet Rai (tr00300@surrey.ac.uk)
University of Surrey
Ambra Morisi
University of Surrey
Kevin Wells
University of Surrey
Miroslaw Bober
University of Surrey
Roberto La Ragione
University of Surrey
Barbara Bacci
University of Bologna
Nicholas J. Bacon
Fitzpatrick Referrals Oncology and Soft Tissue
Spencer Angus Thomas
National Physical Laboratory
Tawfik Aboellail
Colorado State University
Michael J. Dark
University of Florida

Article

Keywords:

Posted Date: March 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1447968/v1

Deep Learning for Necrosis Detection Using Canine Perivascular Wall Tumour Whole Slide Images

Taranpreet Rai[1,*], Ambra Morisi[2], Barbara Bacci[5], Nicholas J. Bacon[6], Michael J. Dark[7], Tawfik Aboellail[8], Spencer Angus Thomas[4], Miroslaw Bober[1], Roberto La Ragione[2,3], and Kevin Wells[1]

[1]University of Surrey, Centre for Vision, Speech and Signal Processing, Guildford, GU2 7XH, United Kingdom
[2]University of Surrey, School of Veterinary Medicine, Guildford, GU2 7AL, United Kingdom
[3]University of Surrey, School of Biosciences and Medicine, Guildford, GU2 7XH, United Kingdom
[4]National Physical Laboratory, London, TW11 0LW, United Kingdom
[5]University of Bologna, Department of Veterinary Medical Sciences, Bologna, 40126, Italy
[6]Fitzpatrick Referrals Oncology and Soft Tissue, Guildford, United Kingdom
[7]Department of Comparative, Diagnostic, and Population Medicine, College of Veterinary Medicine, University of Florida, Gainesville, FL, USA
[8]Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO, USA
[*]t.rai@surrey.ac.uk

ABSTRACT

Necrosis seen in histopathology Whole Slide Images is a major criterion that contributes towards scoring tumour grade which then determines treatment options. However conventional manual assessment suffers from inter-operator reproducibility impacting grading precision. To address this, automatic necrosis detection using AI may be used to assess necrosis for final scoring that contributes towards the final clinical grade. Using deep learning AI, we describe a novel approach for automating necrosis detection in Whole Slide Images, tested on a canine Soft Tissue Sarcoma data set consisting of canine Perivascular Wall Tumours (cPWTs). A patch-based deep learning approach was developed where different variations of training a DenseNet-161 Convolutional Neural Network architecture were investigated as well as a stacking ensemble. An optimised DenseNet-161 with post-processing produced a hold-out test F1-score of 0.708 demonstrating state-of-the-art performance. This represents a novel first-time automated necrosis detection method in the canine Soft Tissue Sarcoma domain as well specifically in detecting necrosis in cPWTs demonstrating a significant step forward in reproducible and reliable necrosis assessment for improving the precision of tumour grading.

Introduction

Canine Soft Tissue Sarcoma (cSTS) are a heterogeneous group of mesenchymal neoplasms that derive from tissues of mesenchymal origin[1,2,3,4]. The anatomical site of cSTS varies significantly, but mostly involve the cutaneous and subcutaneous tissues[5]. Canine Soft Tissue Sarcoma (cSTS) are a large group that can be broken down into several subtypes, but are grouped together nonetheless, due to the similarities of microscopic and clinical features for each subtype. The general treatment of choice for cSTS is to surgically remove cutaneous and subcutaneous sarcomas, where they have a low re-occurrence rate after surgical excision. However, it is higher-grade tumours that can prove to be problematic leading to poorer prognosis and outcomes. Histological grade is the most important prognostic factor in human STS, and is likely one of the most validated criteria to predict outcome following surgery in canine patients[6,7,8]. It is widely accepted that the histological grading system for cSTS is applied to all cSTS subtypes to adopt simplicity. However, there can also be an inconsistent naming of subtypes which can lead to a poor correlation between classification of tumours and their histogenesis (tissue of origin). This sometimes results in confusion for pathologists, therefore, highlighting a need for standardisation[5]. Due to poor agreement when identifying sarcoma subtypes, we are focusing on one common subtype found in canines: Perivascular Wall tumours (cPWT). Perivascular Wall tumours (PWT) arise from vascular mural cells and can be recognisable from their vascular growth patterns which include staghorn, placentoid, perivascular whorling, and bundles from tunica media[9].

The scoring for cSTS grading is broken down into three major criteria: differentiation, mitotic index and necrosis[5]. For the purposes of this paper, the study is focused on necrosis detection which is an important indicator of disease progression and its severity.

A sub-field of machine learning known as deep learning is used for necrosis detection in this work. Such deep learning algorithms are abundant in the medical imaging field and especially digital pathology, assisting in computer-aided diagnosis to classify images or automatically detect diseases. Deep learning has become increasingly ubiquitous and has proven to be very successful in recent image classification tasks in digital pathology[10111213]. The digitisation of histological slides into Whole Slide Images (WSI) has created the field of digital pathology. The field of digital pathology has allowed cellular pathology labs to move into digital workflows[14]. This has resulted in a change in working practices as clinical pathologists are no longer required to be present at the same location as pathology equipment. Potential benefits of this innovation includes remote working across borders, collaborative reporting, and the curation of large teaching databases. Nevertheless, several pathology tasks remain exposed to inter-observer variability, where two or more pathologists will differ in their assessment of a histological slide[12]. As a result, there is much interest in improving and automating pathology workflows whilst promoting standardisation for scoring certain criteria within grading, with greater reproducibility. Automatic necrosis detection in cSTS could decrease viewing times for the pathologists and reduce inter and intra observer variability, positively impacting accuracy in the tumour's diagnosis and prognosis.

The study presented here aimed to classify regions demonstrating necrosis against regions that do not, in canine Perivascular Wall Tumour (PWT) Whole Slide Images (WSIs), by using deep learning models such as pretrained Convolutional Neural Networks (DenseNet161). In the literature, relatively few authors have investigated necrosis detection using machine learning methods. As necrosis detection is typically an image classification task, depending on the image resolution and "field of view" (size of image), necrosis detection can be considered a texture detection problem. Earlier work in necrosis detection applied machine learning methods where texture features were used for SVM classification[15]. The same authors later published literature where deep learning was compared with traditional computer vision machine learning methods in digital pathology. For necrosis detection, their proposed deep learning Convolutional Neural Network (CNN) architecture performed best with an average test accuracy of 81.44%[16]. Another set of authors investigated necrosis detection comparing both an SVM machine learning model and deep learning for viable and necrotic tumour assessment in human osteosarcoma WSIs[17]. The aim was to label the regions of WSI into viable tumor, necrotic tumor, and non-tumor. For evaluation the Volume Under the Surface score (VUS) was computed for non-tumour versus viable tumour versus necrotic. Their models produced 0.922 and 0.959 VUS scores for SVM and deep learning models respectively. Nevertheless, these works do not investigate canine Soft Tissue Sarcoma (cSTS) and so this paper addresses whether such deep learning models can also positively impact necrosis detection in cSTS.

Several methods of training deep learning models were investigated in this work, such as a pretrained DenseNet161 (with and without augmentations), an extension of training this model via hard negative mining, to reduce False Positive (FP) predictions and a stacking ensemble model. To the best of our knowledge this is the first work in automated detection of necrosis in canine PWTs, as well as in cSTS and thus this methodology could be used for the necrosis scoring in an automated detection and grading system for cSTSs. To our best of knowledge, these results represent the highest F1-scores in regard to canine PWT necrosis detection to date.

Methods

Data Description and Patch Extraction Process

A set of canine soft tissue sarcoma histology slides obtained from the Department of Microbiology, Immunology and Pathology, Colorado State University were diagnosed by a veterinary pathologist. A senior pathologist at the University of Surrey confirmed the grade of each case (patient) and chose a representative histological slide for each patient. These slides were then digitised using a Hamamatsu NDP slide scanner (Hamamatsu Nanozoomer 2.0 HT) and viewed with the NDP.viewer platform. These slides were scanned at 40x magnification (0.23µm/pixel) with a scanning speed of approximately 150 seconds at 40x mode (15mm x 15mm) to create a digital Whole Slide Image (WSI).

Two pathologists independently annotated the WSIs for necrosis using the open-source Automated Slide Analysis Platform (ASAP) software, as contours around the necrotic regions. The pathologists used different magnifications (ranging from 5x to 40x) to analyse the necrotic regions before drawing contours. As a result, two class labels were created from these annotations: positive (necrosis) and negative, for subsequent analysis as a binary patch-based classification problem. In order to categorise a region as containing necrosis, both pathologist annotators needed to form an "agreement". Therefore, the intersection of the necrosis annotations were labelled as necrosis. Similarly, areas that are agreed to have no necrosis are labelled as negative. We used these annotations to create image masks for the patch extraction process and applied Otsu thresholding to remove non-tissue background from both classes creating tissue masks. A patch-based approach was applied due to the large nature of Whole Slide Images, which typically produce gigapixel images in the higher resolution layers of the pyramid format. Such large images cannot be directly fed into machine learning models and so patches of a smaller size are extracted from WSIs for further analysis. Using the aforementioned intersection of the annotator's necrosis binary maps, non-overlapping patches

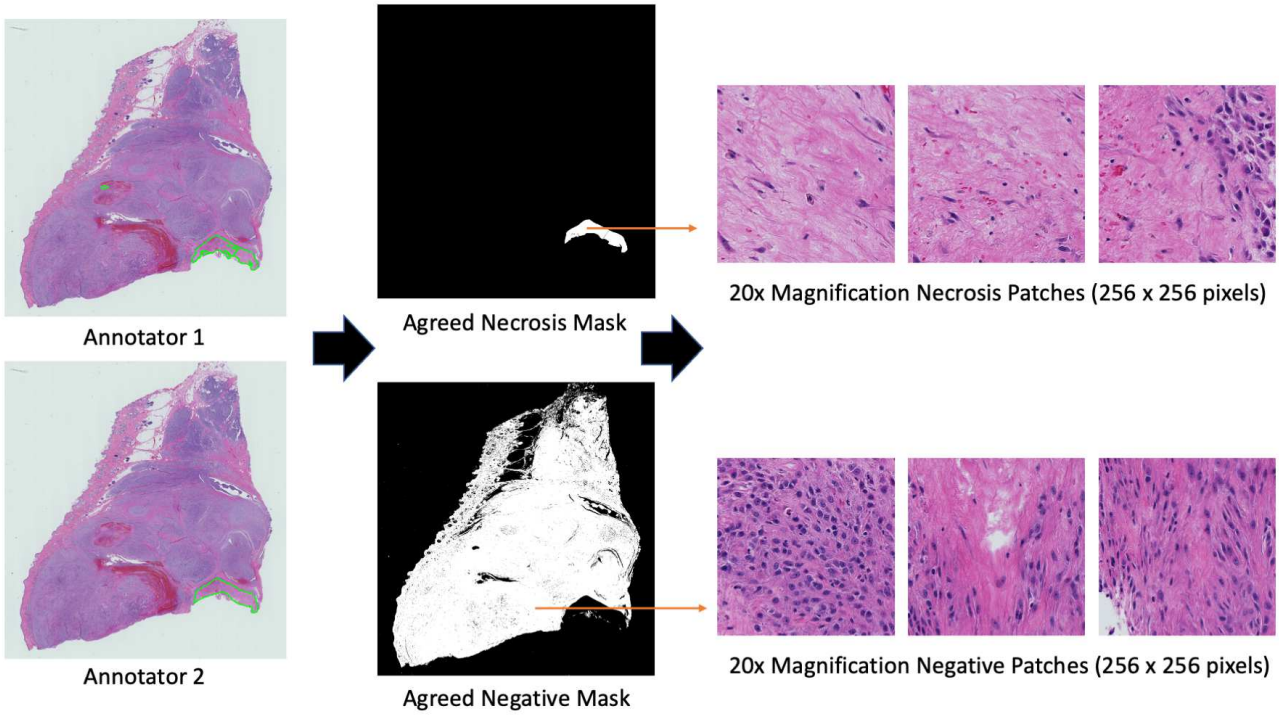

Figure 1 In **a)**, Annotations by "Annotator 1" and "Annotator 2" applied to the same Perivascular Wall Tumour (PWT) Whole Slide Image (WSI). For the patch extraction process, binary masks (or maps) are generated, (shown in **b)**. A necrosis mask is created, highlighting the intersection agreement between both annotators, when considering a region as necrotic. Any disagreement is dismissed from the necrosis and negative binary masks. From applying Otsu thresholding we dismissed any non tissue related regions and by using the intersection agreement for both annotators, we created a "negative mask", highlighting in white regions that do not contain necrosis. We used these masks to extract patches, as shown in **c)**. In this case we extract 20x magnification necrosis and negative patches of 256 x 256 pixels.

of size 256 x 256 pixels were extracted from both necrosis and negative regions (2 classes). A demonstration of the patch extraction process can be visualised in figure 1.

The study investigated 20x magnification resolutions for necrosis detection as suggested by the on-board pathologists. The pathologists chose 20x magnification over 5x, 10x and 40x as the ideal resolution for necrosis detection. Non-overlapping patches of 256 by 256 pixels were extracted from regions of both classes using a minimum decision threshold for percentage of necrosis present in a patch. For a magnification of 20X, 30% of the patch must have contained necrosis pixels (determined from the expert defined labels) in order for it to be labelled as necrosis. A threshold of 30% for necrosis pixels was chosen as it was needed to take into account boundary effects of the necrosis clusters in the images. Patches extracted from boundaries of the necrosis cluster would almost certainly contain non-necrotic tissue. However, a suitable amount (in this case 30%) of necrosis tissue is be required to sufficiently label a patch as necrosis for the effective training of deep learning models. A higher number would risk dismissing useful necrotic patches, whereas a lower number (thus more negative tissue in a patch) would likely cause confusion during the training of deep learning models.

Deep Learning Model and Experimental Set-Up

In order to evaluate the robustness and the veracity of our approach, we performed 3-fold cross validation, where a hold-out test set created to compare the models trained on the three different folds. In total we extracted patches from 32 patients (WSIs) to create our train, validation and test sets.

There were a total of 5784 necrosis patches from 20 slides for training/validation and 1151 necrosis patches from 12 slides for testing. Additionally, there were a total of 50975 negative patches for training/validation and 31351 negative patches for testing.

Class imbalance is apparent throughout the different folds of the datasets. We ran initial experiments with an imbalanced dataset, by using all necrosis and negative patches per WSI (further details on the experimental set up are provided in the next section). To address the large variation of the negative class with a relatively small presence of the necrosis class, we reduced

class imbalance by randomly extracting 800 negative patches per WSI and used these with all necrosis patches per WSI. This reduces the class imbalance to 1:4 for necrosis to negative, respectively. We opt to train with this mild imbalance ratio, as balancing the dataset to 50/50 risk throwing away useful (vital) information from the negative class.

The deep learning model implemented transfer learning bottleneck feature extraction. We investigated several pretrained networks including VGG and ResNet architectures as they have been shown to have a positive impact in digital pathology[18,19]. However, we chose DenseNet-161 due to its performance in preliminary experiments[20]. According to one study, DenseNet-161 can be used for fast and accurate classification of digital pathology images to assist pathologists in daily clinical tasks[21].Bottleneck features were extracted from DenseNet161, producing an output of 2208 features. These features were then fed into a classification layer, to classify "necrosis" or the "negative" class per patch. See part **a** figure 2.

For all experiments a batch size of 32 was input into each model and the loss function used was cross entropy loss. A grid search was previously implemented on several folds of preliminary experiments to determine optimal hyperparameters. The Adam optimiser initialised with a learning rate of 0.0001 was used, with a scheduler step of 20 and a scheduler gamma of 0.5 as "optimal" values[22]. We calculated the RGB mean and standard deviation values per fold for patch image normalisation. For every fold, each model was trained for 100 epochs, where the model from the epoch with the lowest validation loss was chosen as the best performing model. This selected model was then applied to both the validation and test sets for evaluation during training and final testing, respectively.

Other Types of Models Investigated

In this section we describe the different type of deep learning and ensemble models investigated for necrosis detection. Apart from the ensembles, all deep learning models used the same hyperparameters and training scheme as described in the previous section.

DenseNet-161 with Augmentations

Adding augmented patch images to a training set is a common strategy to mitigate the lack of variation and size of limited datasets[23]. Modifications were thus applied to an existing image to produce new images, using random horizontal/ vertical flips and colour jitter (random changes to the brightness, contrast, saturation and hue of a patch image). A change of upto/minus 40% is randomly applied for brightness, contrast and saturation.

Hard Negative Mining Model

Hard negative mining was performed in order to reduce the number of false positives. This is an approach that allowed us to train the model with additional "difficult" examples presented to the network[24]. Firstly, a "full" training set was created, where we extracted every single negative class patch from the training set. Secondly, we then applied the best model on the full training set to infer a new set of predictions. The model with the lowest validation loss was selected as the best model. Any prediction on a patch that was a false positive (and did not exist in the original dataset) was added to the sub-sampled dataset; thus creating a new dataset known as the "hard negative training set". Lastly, we then trained our 20x magnification model using this new dataset and evaluated as before. A flow diagram of the implementation of the hard negative experiments can be visualised in part **b** of figure 2.

Ensemble Model

A common approach to boost the performance of machine learning models is via the employment of ensemble models where the ensemble process is depicted in part **c** of figure 2. The basic concept of ensembles is to train multiple models (or base-member models) and combine their predictions into one single output[25]. Ensemble models typically outperform single models (individual base-member models) on their respective target datasets. This can be seen in recent machine learning based competitions such as on the Kaggle platform and MICCAI[26,27,28].

The ensemble model was trained on the sub-sampled training dataset. To make use of all the data from the training WSIs, we made inferences on the full training set patches (including the original training set patches). There are various methods and combinations to create ensemble models, however, preliminary experiments demonstrated that an ideal combination consisted of base-member models was the DenseNet-161 model, the DenseNet-161 model with augmentations and the hard negative mining model.

Preliminary experiments investigated combining ensemble predictions and training a logistic regression model (as a meta-model) on the full training set of patches to produce prediction probability outputs. We then tested this trained logistic regression model on validation and hold-out test sets. Pseudo code for the stacking ensemble is also represented in Algorithm 1.

Post Processing

It is important to note that although sensitivity is a vital measure in the medical domain, the number of false positives greatly influences the score for necrosis, thus impacting overall grading. For our problem, both the precision and sensitivity were considered to be equally as important and therefore the F1-score was used to determine the optimal thresholds for each folds

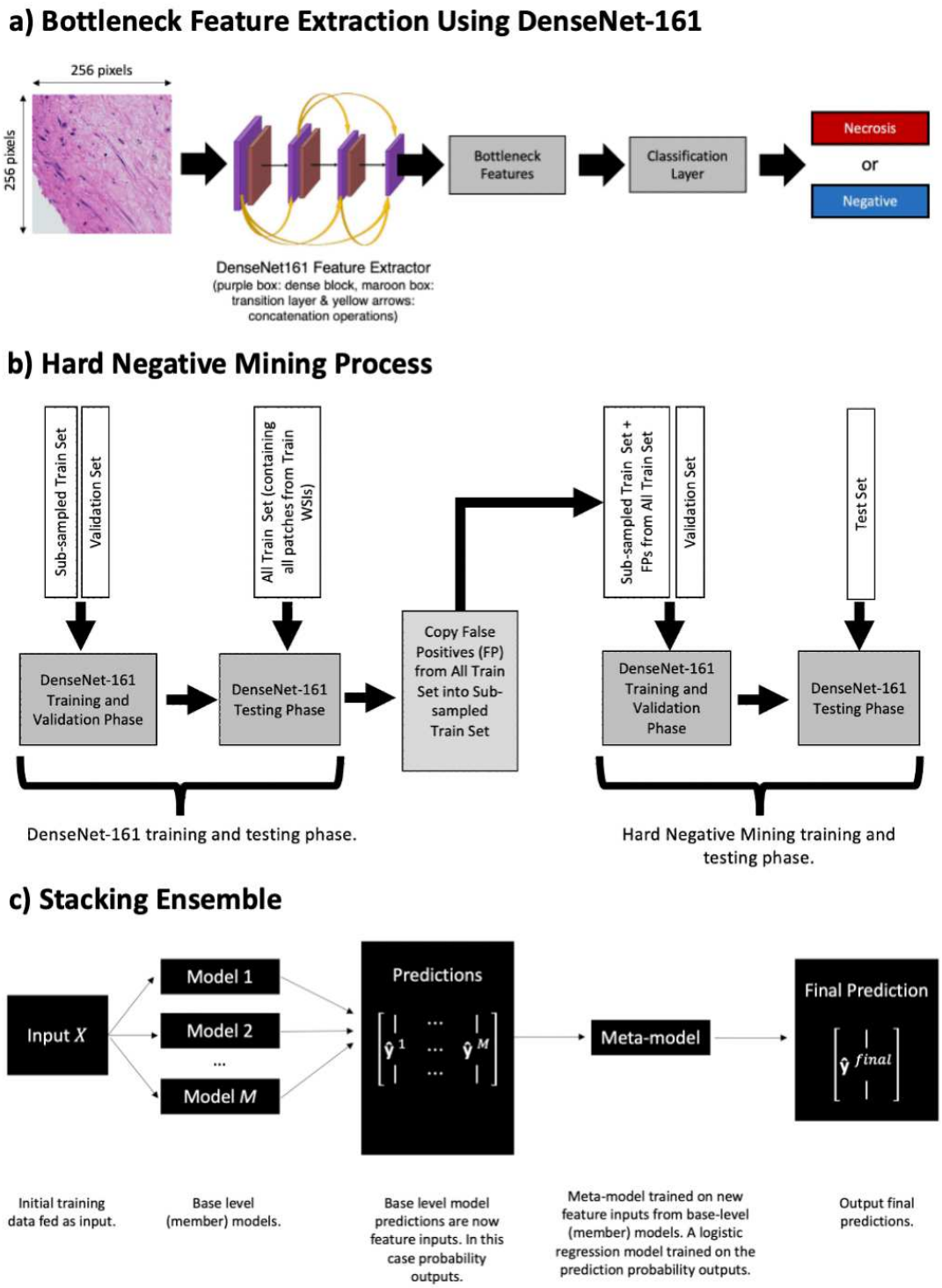

Figure 2 **a)** Bottleneck feature extraction using DenseNet-161. A patch size of 256 x 256 pixels is fed into a DenseNet-161 feature extractor, where bottleneck features are obtained. These features are then fed into a classification layer for further training and validation, classifying necrosis or negative patches. **b)** Hard negative mining approach to train the model with additional "difficult" examples presented to the network. **c)** Stacking ensemble. The input X is fed into M base-level member models: 20x baseline model, the 20x model with augmentations, the 20x hard negative mining model. The prediction outputs of these models $\hat{y}_M$ are combined and fed into a logistic regression meta-model as new feature inputs. New coefficients are learnt in this logistic regression model, before final predictions are output $\hat{y}^{final}$.

Algorithm 1 Stacking Ensemble

Input: training data $D = x_i, y_{i=1}^{m}$
Output: Ensemble logistic regression meta-model Z where, $Z(x) \approx y$
Step 1: Randomly sub-sample training data D to create T subsets of $[x_s, y_s]$
Step 2: Train base-level member models
for `t = 1 to` T **do**
    `learn` h_t `such that` $h_t(x_s) \approx y_s$
end for
Step 3: Aggregate predictions $x' = h_1(x), \ldots, h_T(x)$
Step 4: Train the meta-model Z
learn Z such that $Z(x') \approx y$
return Z

validation. These thresholds were then applied to our hold-out test set for each fold. Additionally, for comparison, the mean optimal threshold for the three folds was computed and applied this to our hold-out test set. The F1-score is the harmonic mean between the precision and sensitivity. It takes into account both the sensitivity and precision producing a weighted average of the two metrics. Both precision (formula 1) and sensitivity (formula 2) contribute equally to the F1 score (formula 4):

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

$$\mathrm{F1} = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision} \tag{4}$$

where, TP, FP and FN are true positives, false positives and false negatives, respectively.

Figure 3 depicts histograms showing true positives, true negatives, false positives and false negatives for the DenseNet-161 model. The y-axis is log-normalised to compress and better visualise the frequency of predictions as the number of true negatives significantly outweigh the number of true positives. The x-axis (probabilities) is split into 100 bins. The left side of this figure shows true negatives (TN) and false positive (FP) histograms for each validation fold, whereas on the right depicts histogram plots of false negatives (FN) and true positives (TP). The validation set was used to choose optimal thresholds. The hold-out test set was used as a data set purely for evaluation and not contribute towards any change in strategy (i.e. to prevent data/information leaks). These classification output combinations were chosen as they complement one other. For example, increasing the probability decision threshold would increase the number of TNs and reduce the number of FPs. However, this would subsequently increase the number of FNs thus reducing the number of TPs. For all folds, the TN and FP plots shows a wide range of prediction probabilities, with a heavy skew towards the lower probabilities. The FN and TP plot for fold 3 (validation) appears to shows slight sparsity among FN predictions in comparison to other folds.

The sensitivity, specificity and F1-scores were calculated for several probability thresholds for each validation fold, as shown in figures 4. For both the DenseNet-161 model and the ensemble model, it was apparent that high probability thresholds had an adverse effect on sensitivity. The F1-score was used as the metric to determine optimal thresholds.

The probability thresholds t ranged from 0.01 to 1 and so choosing the optimal validation threshold T for the F1-score $F1$ can be represented formally as:

$$T = \arg\max_{t} F1(t) \tag{5}$$

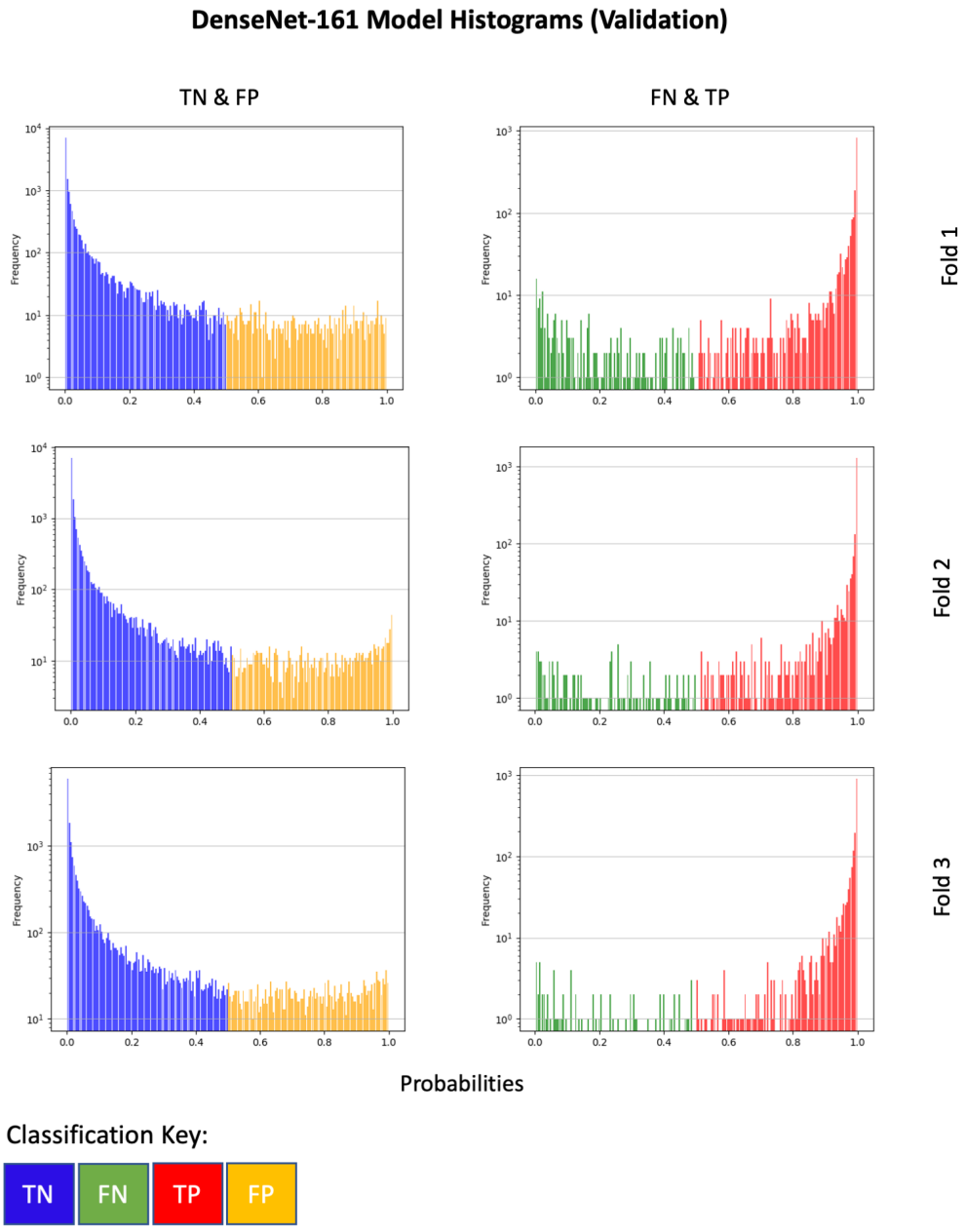

Figure 3 Histograms of the initial classification results based on a standard 0.5 (50%) probability decision threshold. Depicted are true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP) for the DenseNet-161 model. On the left side depicts histogram plots of TN and FP for each validation fold, whereas on the right side depicts histogram plots of FN and TP. These combinations were chosen for the plots as they complement each other. It can be seen that all three folds are characteristically similar in distribution. TN and TP predictions typically produce high probabilities, as can be seen by the frequency of such predicted probabilities. Increasing the probability threshold would increase the number of true negatives and reduce the number of false positives. However, this would subsequently increase the number of false negatives and reduce the number of true positives.

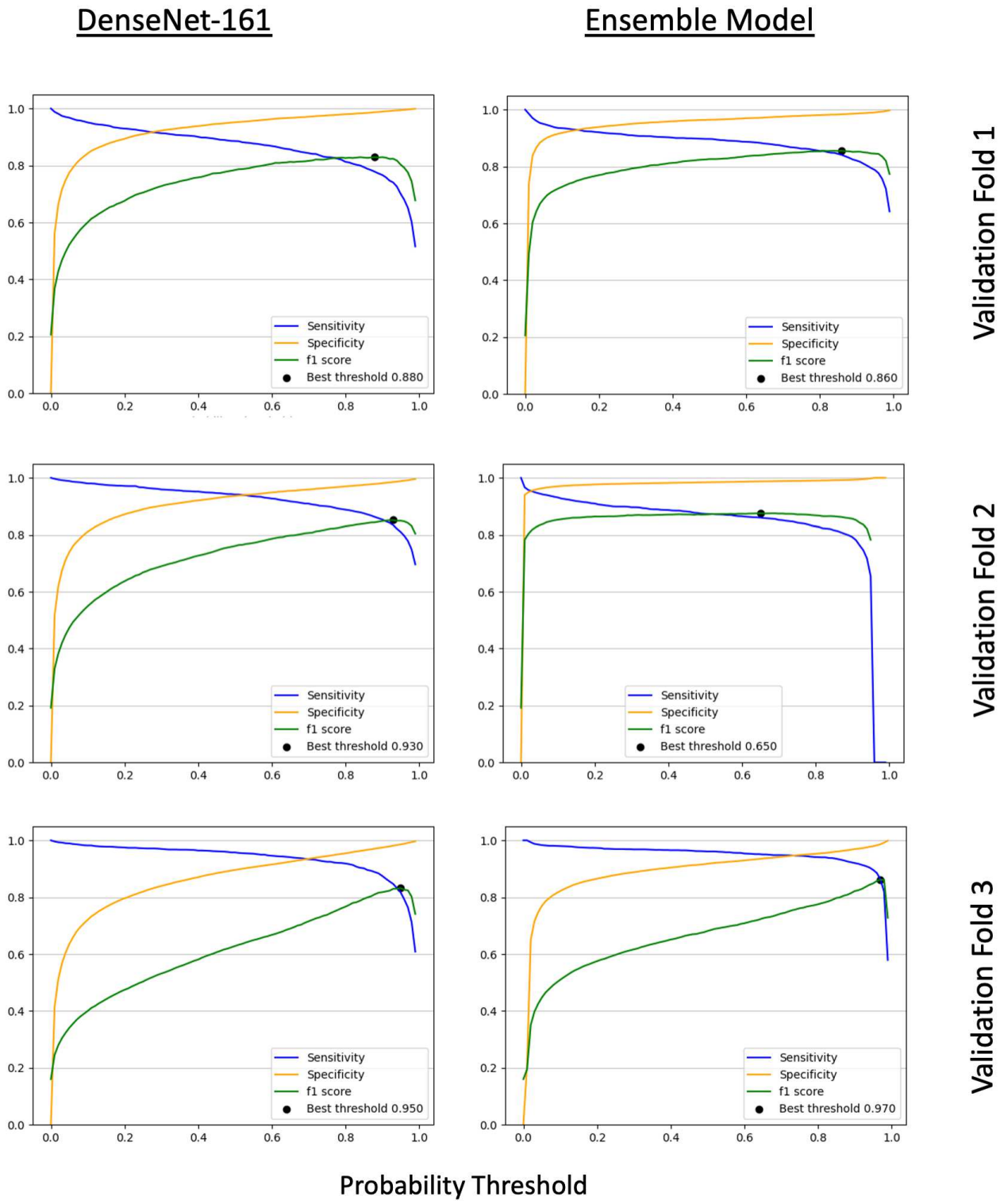

Figure 4 Line graphs that depict the sensitivity, specificity and weighted F1-score calculated for each probability threshold, for the three validation folds from the 20x ensemble "1" model. To determine the optimal probability threshold, we choose the threshold with the highest F1-score. In the above plots, these are denoted as "Best threshold". For validation fold 1, this threshold was 0.86, for fold 2 it was 0.65 and for fold 3 this was 0.97.

In general, the DenseNet-161 model followed a similar trend for all 3 folds where the optimal threshold was high (between 0.88 to 0.93). The ensemble model demonstrated similar results apart from fold 2 where the optimal probability decision threshold was found to be 0.65.

Necrosis detection in WSIs often displays sporadic false positive predictions. From domain knowledge and discussion with the on-board pathologists, it was determined that single tile (patch) necrosis predictions in a WSI would typically not be considered necrotic in most circumstances, if surrounded by non necrotic tissue. This is due to the size of the isolated region playing a part in quantifying on whether a region should be considered necrosis and scored. Of course, single patches could be necrotic, however, this would be analogues to outlier detection or statistical noise and fluctuations. As a result, a post-processing step was applied to remove these single tile necrosis predictions.

Results

Individual base-member model results

Results for the DenseNet-161 model, DenseNet-161 with augmentations model and the hard negative mining model are presented in table 1. It must be denoted that these results are patch-based and not WSI (or patient) based. Therefore, classification results are based on whether a sampled patch is considered necrotic or not. Furthermore, these results are based on a default "decision threshold" or simply "threshold" of 50%. This means that if a patch is predicted to have a probability confidence of more than 50%, then it is considered necrosis. Anything less than 50% is considered negative (not necrotic).

The addition of augmentations appeared to show a marginal improvement in sensitivity as shown with the validation (italicised) and hold-out test sets. The hard negative mining model demonstrated an improvement on the F1-score and specificity scores across the 3 folds, compared to the DenseNet-161 model for both the validation and test sets. However, sensitivity was adversely affected across the average of the validation and test sets.

Nevertheless, across all models, sensitivity scores were higher in the test set than validation. This is most likely due to the models finding unencountered tissue types to be suspicious during test, a consequence of training with limited datasets.

The highest validation and test specificity was produced by the ensemble model; the model trained on the combined probability outputs of the DenseNet-161 model, DenseNet-161 with augmentations model and the hard negative mining model. Consequently, the highest validation and test F1-scores are also from the ensemble model. As the F1-score is used as a basis for choosing the best performing model, we decided to continue with both the ensemble model and DenseNet-161 model as a comparison for further post-processing.

Table 1 3-fold averaged results for the DenseNet-161 model, the hard negative mining model and the DenseNet-161 with augmentations model, with reported mean sensitivity, specificity and F1-scores averaged across all three folds. The highest score for a metric is highlighted in bold for the test set, whereas this is italicised for the validation set. Plus/minus values shows the mean subtracted from the highest and lowest results from the 3-fold experiments, for each model.

Model	Set	Sensitivity		Specificity		F1 Score	
DenseNet-161	Validation	0.928	+0.029 -0.004	0.928	+0.024 -0.032	0.724	+0.060 -0.096
	Test	0.939	+0.003 -0.004	0.907	0.020 -0.019	0.404	+0.053 -0.049
Hard Negative Model	Validation	0.906	+0.033 -0.040	*0.943*	+0.021 -0.030	0.753	+0.060 -0.096
	Test	0.917	+0.013 -0.010	0.926	+0.011 -0.016	0.449	+0.053 -0.049
DenseNet-161 with Augmentations	Validation	*0.930*	+0.029 -0.052	0.922	+0.018 -0.022	0.710	+0.041 -0.075
	Test	**0.944**	+0.014 -0.008	0.900	+0.020 -0.017	0.389	+0.046 -0.041
Ensemble	Validation	0.910	+0.051 -0.037	*0.955*	+0.029 -0.038	*0.793*	+0.011 -0.009
	Test	0.924	+0.022 -0.039	**0.943**	+0.031 -0.031	**0.535**	+0.028 -0.025

After applying optimal thresholds and post-processing

The post-processing was applied to the results after obtaining optimal thresholds for each fold. In this case, the DenseNet-161 model and the ensemble model. Results are presented in table 2. From this table it is clear that post-processing has a significant impact, especially on specificity and F1-scores. The best sensitivities for both validation and test sets were found in the DenseNet-161 model results. However, the best performing specificity result was from DenseNet-161 (threshold per fold + post-processed), with 0.992 and 0.984 for validation and test, respectively. As a result, the highest performing test F1-score also came from this model (0.708).

From Table 1, it can be seen that the models on the test data produced sub-optimal F1 scores, suggesting that the models based on a 50% probability decision threshold, did not generalise at an optimal standard. However, Table 2 demonstrates after thresholding and post-processing, the F1 scores significantly improve, suggesting that higher decision thresholds may be required when applied to unseen data. This could be due to textures and different colours presented from the staining process residing in the test data. As a result, these unseen artefacts could lead to an increase in low confidence necrosis predictions, thus producing false positives. This also suggests that structures related to necrosis are learnt well using the training data as true positive necrosis predictions tend to produce high confidence predictions.

Additionally, accuracy and an average of the sensitivity and specificity (denoted as Sens./Spec. Avg.) are also introduced into the table 2. The Sens./Spec. Avg. can be directly compared to the results from[16] where the authors averaged their necrosis and non-necrotic classification results producing 81.44%. Comparatively, our 3-fold scores range from 89.5% t0 93.4%, thus producing the highest accuracies for this metric for necrosis vs. non necrotic (negative) classification.

A set of exemplar spatial confusion maps are shown in Figure 5 where we present the results of the tile-based classification overlaid onto the original WSIs created from the hold-out test set are shown in figure 5. In this figure, we depict results from the DenseNet-161 model and the ensemble model. The left side shows results with the standard 50% probability threshold applied, whereas the right shows optimal validation threshold applied to the fold 3 test set results, with single tile removal. It is apparent that there are far fewer FPs after the post-processing for both the DenseNet-161 and ensemble models.

The post-processing improved the DenseNet-161 results becoming the top performing model. This is attributed to spatially sparse FP predictions in slides. All models experienced a slight reduction in sensitivity after applying the optimal thresholds and post-processing. There is a slight increase in false negatives, especially around the borders of the necrosis clusters (TPs) in the images. However, the ensemble model spatial confusion matrices depict less FN predictions in the middle of the cluster, in comparison to DenseNet-161. This is important and allows us to understand the limitations of patch-based approaches and may in fact highlight disagreement between annotators. We are aware that boundary cases may exist around the borders of these clusters due to the annotation and patch extraction process. The deep learning models may reflect these uncertainties by producing less confident predictions in these areas. The ensemble model also demonstrates the power of combining multiple different models, mimicking the combination of different "teachers" or "experts", as there are fewer FNs in the middle of the necrosis clusters compared to the DenseNet-161 model.

Discussion

.A necrosis detection method was created after investigating a pretrained DenseNet-161 model, hard negative mining and ensemble models. We further investigated the application of optimal thresholds and further post-processing. This is the first known necrosis detection model for PWTs and in general cSTS. As a result, we also produce state-of-the-art performance metrics, especially regarding accuracy and sensitivity/ specificity averages for necrosis detection in PWTs.

The post-processing alongside applying optimal thresholds allowed the DenseNet-161 model to produce the best F1-scores: the key metric for evaluation for this work. This is most likely due to the DenseNet-161 model generating more "sparse" FP predictions in than the ensemble model, where there are more clustered FP predictions. Nevertheless, although producing the highest F1-scores, this difference was marginally higher than the Ensemble model with post-processing.

However, upon inspection of the spatial confusion matrix heatmaps, it was observed that both optimised models differed slightly especially in regards to false negative (FN) predictions, with slightly fewer FNs inside the true positive clusters for the Ensemble model. This further demonstrates the difference in learning between alternative types of machine learning models such as deep learning models and ensembles with logistic regression as their backbone. This paper demonstrates that deep learning models can be successfully used as a diagnostic support tool for grading canine PWT in cSTS. Necrosis detection should also be investigated with other cSTS subtypes.

References

1. Bostock, D. & Dye, M. Prognosis after surgical excision of canine fibrous connective tissue sarcomas. *Vet. Pathol.* **17**, 581–588 (1980).

Table 2 3-fold averaged results after applying optimal thresholds and the single tile removal post-processing to the DenseNet-161 and ensemble models, for the validation and test sets. Presented in this table are the mean sensitivity, specificity and f1-scores averaged across all three folds. "Threshold per fold + post-processed" is where optimal thresholds derived from three-way cross-validation followed by single tile removal were applied to all validation folds and the hold-out test set. "Mean threshold + post-processed" is where the optimal thresholds for each fold have been averaged and then applied to each folds validation and hold-out test set. The highest score for a certain metric is highlighted in bold for the test set, whereas it is italic for the validation set. Plus/minus values shows the mean subtracted from the highest and lowest results from the 3-fold experiments, for each model.

Model	Set	Sensitivity		Specificity		F1 Score		Accuracy		Sens./ Spec. Avg.	
DenseNet-161	Validation	*0.928*	+0.029 -0.004	0.928	+0.024 -0.032	0.724	+0.060 -0.096	0.927	+0.017 -0.026	0.928	+0.010 -0.009
	Test	**0.939**	+0.003 -0.004	0.907	+0.020 -0.019	0.404	+0.053 -0.049	0.908	+0.019 -0.019	0.923	+0.011 -0.012
DenseNet-161 (threshold per fold + post-processed)	Validation	0.808	+0.022 -0.032	*0.992*	+0.001 -0.001	0.860	+0.015 -0.011	*0.973*	+0.003 -0.005	0.900	+0.011 -0.015
	Test	0.807	+0.028 -0.032	**0.984**	+0.001 -0.001	**0.708**	+0.010 -0.011	**0.978**	+0.000 -0.000	0.896	+0.013 -0.016
DenseNet-161 (mean 3-fold threshold + post-processed)	Validation	0.813	+0.042 -0.069	0.991	+0.006 -0.005	0.855	+0.020 -0.016	0.972	+0.003 -0.005	0.902	+0.018 -0.032
	Test	0.807	+0.010 -0.011	**0.983**	+0.004 -0.004	**0.702**	+0.033 -0.033	**0.977**	+0.004 -0.004	0.895	+0.005 -0.003
Ensemble	Validation	0.910	+0.051 -0.037	0.955	+0.029 -0.038	0.793	+0.078 -0.112	0.950	+0.022 -0.029	*0.933*	+0.006 -0.004
	Test	0.924	+0.022 -0.039	0.943	+0.031 -0.031	0.535	+0.133 -0.120	0.943	+0.029 -0.029	**0.934**	+0.011 -0.007
Ensemble (threshold per fold + post-processed)	Validation	0.853	+0.009 -0.015	0.990	+0.001 -0.001	*0.878*	+0.011 -0.006	*0.976*	+0.002 -0.004	0.922	+0.004 -0.008
	Test	0.846	+0.050 -0.060	**0.981**	+0.031 -0.051	**0.704**	+0.026 -0.019	**0.977**	+0.003 -0.004	0.914	+0.022 -0.028
Ensemble (mean 3-fold threshold + post-processed)	Validation	0.868	+0.070 -0.048	0.982	+0.012 -0.018	0.851	+0.022 -0.043	0.969	+0.006 -0.008	0.925	+0.015 -0.022
	Test	0.871	+0.032 -0.059	0.974	+0.015 -0.014	0.673	+0.089 -0.085	0.971	+0.012 -0.012	0.923	+0.026 -0.018

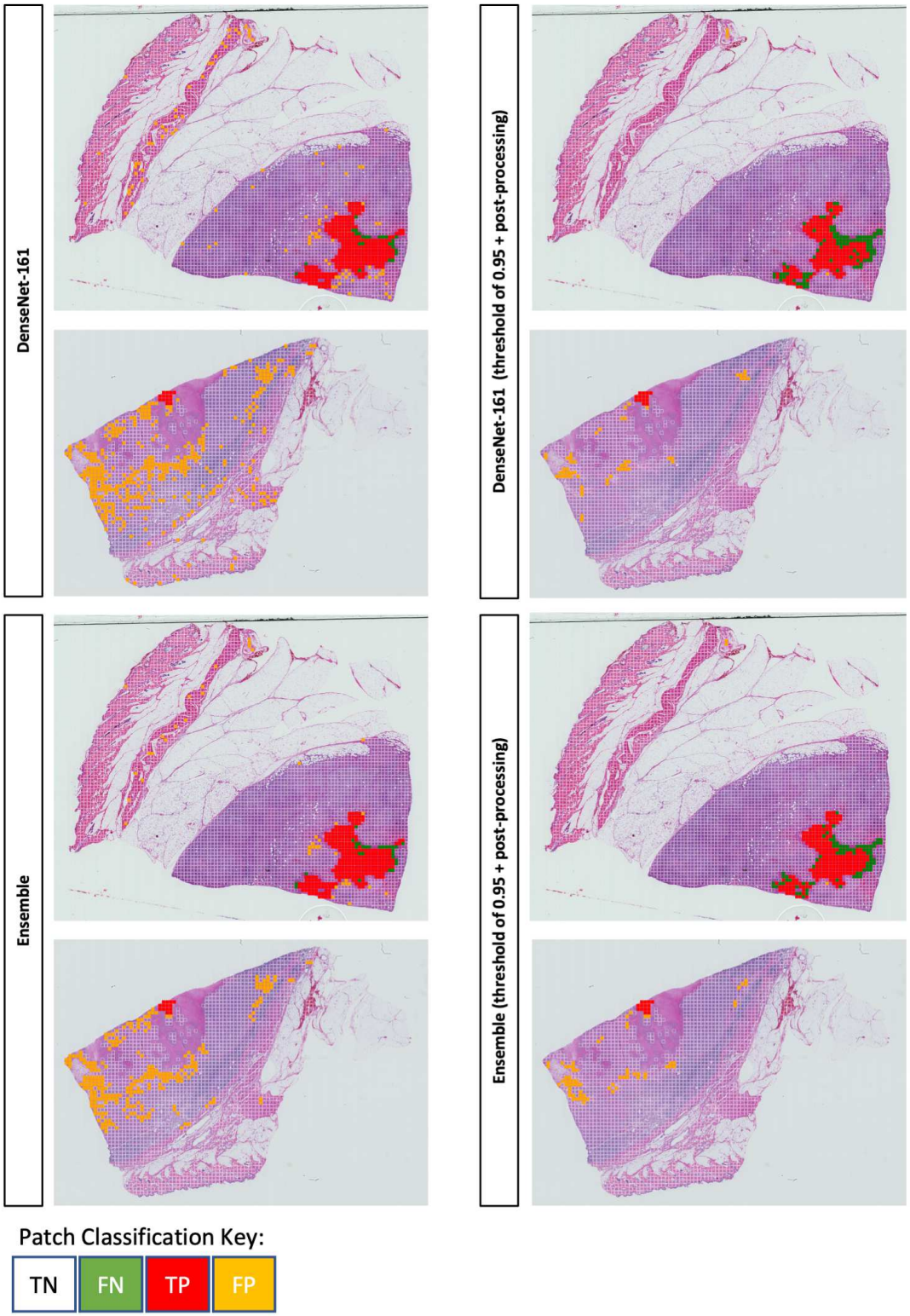

Figure 5 Sample Whole Slide Image (WSI) spatial confusion maps before and after applying optimal threshold (determined from the fold 3 validation set) and post processing; removing single tile predictions. The left side images shows predictions from the 20x baseline and ensemble models with the standard 50% probability decision thresholds. The right side shows predictions after applying the optimal threshold and post-processing. True positives (TP) are displayed in red, false negatives (FN) in green, false positives (FP) in yellow and true negatives (TN) in clear.

2. Dernell, W. S., Withrow, S. J., Kuntz, C. A. & Powers, B. E. Principles of treatment for soft tissue sarcoma. *Clin. techniques small animal practice* **13**, 59–64 (1998).
3. Ehrhart, N. Soft-tissue sarcomas in dogs: a review. *J. Am. Animal Hosp. Assoc.* **41**, 241–246 (2005).
4. Mayer, M. N. & LaRue, S. M. Soft tissue sarcomas in dogs. *The Can. Vet. J.* **46**, 1048 (2005).
5. Dennis, M. *et al.* Prognostic factors for cutaneous and subcutaneous soft tissue sarcomas in dogs. *Vet. Pathol.* **48**, 73–84 (2011).
6. Kuntz, C. *et al.* Prognostic factors for surgical treatment of soft-tissue sarcomas in dogs: 75 cases (1986-1996). *J. Am. Vet. Med. Assoc.* **211**, 1147–1151 (1997).
7. McSporran, K. Histologic grade predicts recurrence for marginally excised canine subcutaneous soft tissue sarcomas. *Vet. Pathol.* **46**, 928–933 (2009).
8. Bray, J. P., Polton, G. A., McSporran, K. D., Bridges, J. & Whitbread, T. M. Canine soft tissue sarcoma managed in first opinion practice: outcome in 350 cases. *Vet. Surg.* **43**, 774–782 (2014).
9. Avallone, G. *et al.* The spectrum of canine cutaneous perivascular wall tumors: morphologic, phenotypic and clinical characterization. *Vet. Pathol.* **44**, 607–620 (2007).
10. Xing, F., Xie, Y., Su, H., Liu, F. & Yang, L. Deep learning in microscopy image analysis: A survey. *IEEE Transactions on Neural Networks Learn. Syst.* **29**, 4550–4568 (2017).
11. Ing, N. *et al.* Semantic segmentation for prostate cancer grading by convolutional neural networks. In *Medical Imaging 2018: Digital Pathology*, vol. 10581, 105811B (International Society for Optics and Photonics, 2018).
12. Ertosun, M. G. & Rubin, D. L. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. In *AMIA Annual Symposium Proceedings*, vol. 2015, 1899 (American Medical Informatics Association, 2015).
13. Madabhushi, A. & Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities (2016).
14. Cross, S., Dennis, T. & Start, R. Telepathology: current status and future prospects in diagnostic histopathology. *Histopathology* **41**, 91–109 (2002).
15. Sharma, H. *et al.* Appearance-based necrosis detection using textural features and svm with discriminative thresholding in histopathological whole slide images. In *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*, 1–6 (IEEE, 2015).
16. Sharma, H., Zerbe, N., Klempert, I., Hellwich, O. & Hufnagl, P. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Comput. Med. Imaging Graph.* **61**, 2–13 (2017).
17. Arunachalam, H. B. *et al.* Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one* **14**, e0210706 (2019).
18. Rai, T. *et al.* Can imagenet feature maps be applied to small histopathological datasets for the classification of breast cancer metastatic tissue in whole slide images? In *Medical Imaging 2019: Digital Pathology*, vol. 10956, 109560V (International Society for Optics and Photonics, 2019).
19. Rai, T. *et al.* An investigation of aggregated transfer learning for classification in digital pathology. In *Medical Imaging 2019: Digital Pathology*, vol. 10956, 109560U (International Society for Optics and Photonics, 2019).
20. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
21. Talo, M. Automated classification of histopathology images using transfer learning. *Artif. Intell. Medicine* **101**, 101743 (2019).
22. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
23. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 1–48 (2019).
24. Li, M. *et al.* Deep instance-level hard negative mining model for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 514–522 (Springer, 2019).
25. Yang, Y. *Temporal Data Mining via Unsupervised Ensemble Learning* (Elsevier, 2016).
26. Kang, J. & Gwak, J. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access* **7**, 26440–26447 (2019).

27. Ataloglou, D., Dimou, A., Zarpalas, D. & Daras, P. Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning. *Neuroinformatics* **17**, 563–582 (2019).

28. Qummar, S. *et al.* A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access* **7**, 150530–150539 (2019).

Acknowledgements

We would like to thank the Doctoral College, University of Surrey (UK), National Physical Laboratory (UK) and Zoetis through the vHive initiative, for making this research possible.

Author contributions statement

T.R., conducted all experiments. A.M. and B.B. conducted the annotations process. T.R., M.B., R.L., K.W. and S.T. analysed the results. All authors contributed to manuscript preparation and reviewed the manuscript.

Additional information

Competing interests The authors declare no potential conflict of interest.