A Framework for Automatic Analysis of WSI using Deep Learning

Peter Bokor *
Supervised by: Lukáš Hudec †

Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava

Abstract

Early diagnosis of breast cancer using microscopic analysis of tissue is a prerequisite for its treatment. This process is time-consuming and difficult due to the inter-intraclass variance of histopathology data, the vast size of histopathology slides, and environmental influences on pathologists. A robust, unbiased diagnostic assistance tool, resistant to outside influences would be of utmost help for pathologists performing the diagnosis.

We propose an end-to-end framework for processing Whole Slide Images (WSI) and their analysis using methods of computer vision and deep learning.

Multiple deep learning models are used to apply their knowledge to new WSIs, creating complex analyses, which may be used to assist with diagnosis. Our system is fast enough for everyday use and may speed up the workflow and ease up the workload of medical experts performing the analysis of tissue.

Keywords: Computer Vision, Deep Learning, Deep Convolutional Networks, Classification

1 Introduction

Cancer is a collection of related diseases that occur inside the human body. Breast cancer is the most common form of cancer among women. Early diagnosis is a prerequisite for its treatment. Microscopic analysis of tissue is required for its definitive diagnosis. This process is time-consuming and difficult due to the inter-intraclass variance of histopathology data, the vast size of histopathology slides, and environmental influences on pathologists. All of these reasons lead to a need for an unbiased decision support histopathology tissue analysis system, free of environmental influence, capable, reliable, and robust enough to assist pathologists with their analysis.

We propose a framework for automatic processing of Whole Slide Images and subsequent use of the WSIs to train *deep learning models*, with a focus on a specific task, that, in the context of our framework, we call *Experts*. An overview of our proposed framework is shown in Figure 1.

Whole Slide Images are processed by a module we design, called *WSI Analyzer*, that manages matching annotation to the WSI, and hierarchically organizing the data. WSIs processed by *WSI Analyzer* are subsequently used by another module we design - *WSI Generator* to produce large sets of data based on multiple parameters.

WSI Analyzer may also be used to produce a completely new analysis, using an *Expert*. The *Expert* is able to apply their knowledge to a new WSI, creating an analysis. The analysis may be used by doctors as a helping tool, or a standalone diagnostic.

We also propose several *Experts* using binary and multi-class classification, which when combined create such a diagnostic and an *Expert* for enhancing the multi-class classification results by bringing attention to different levels of zoom for different inputs - tumor types.

We evaluate our approach against state-of-the-art methods in the domain, quantitatively, but also against each other, with a focus on the enhancements, and empirically. Our contributions include: 1) An automatic universal framework for WSI processing, and data generation, able to handle large datasets - in the form of multiple WSIs and help train deep learning models to achieve state-of-the-art results. 2) Complex analysis of WSIs using multiple deep learning models, taking into account contextual and detailed information, mimicking analysis performed by a histopathologist. 3) A deep learning model with a special architecture capable of enhancing multi-class classification prediction results.

2 Related Work

Classification of breast cancer histopathology visual data and creation of automatic systems for breast cancer diagnosis is the focus of many recent works. As soon as it was possible to scan and load medical images into a computer, researchers have built systems for automated analysis [5]. Over time, the systems became more sophisticated and lately the most popular are the ones using *Deep learning*.

Spanhol et al. [6] create a large dataset called BreakHis consisting of 7909 images, containing 8 breast tumor types and also provide baseline results using traditional computer vision methods for feature extractions and traditional machine learning methods for classification. What Span-

*bokorpet98@gmail.com
†lukas.hudec@stuba.sk

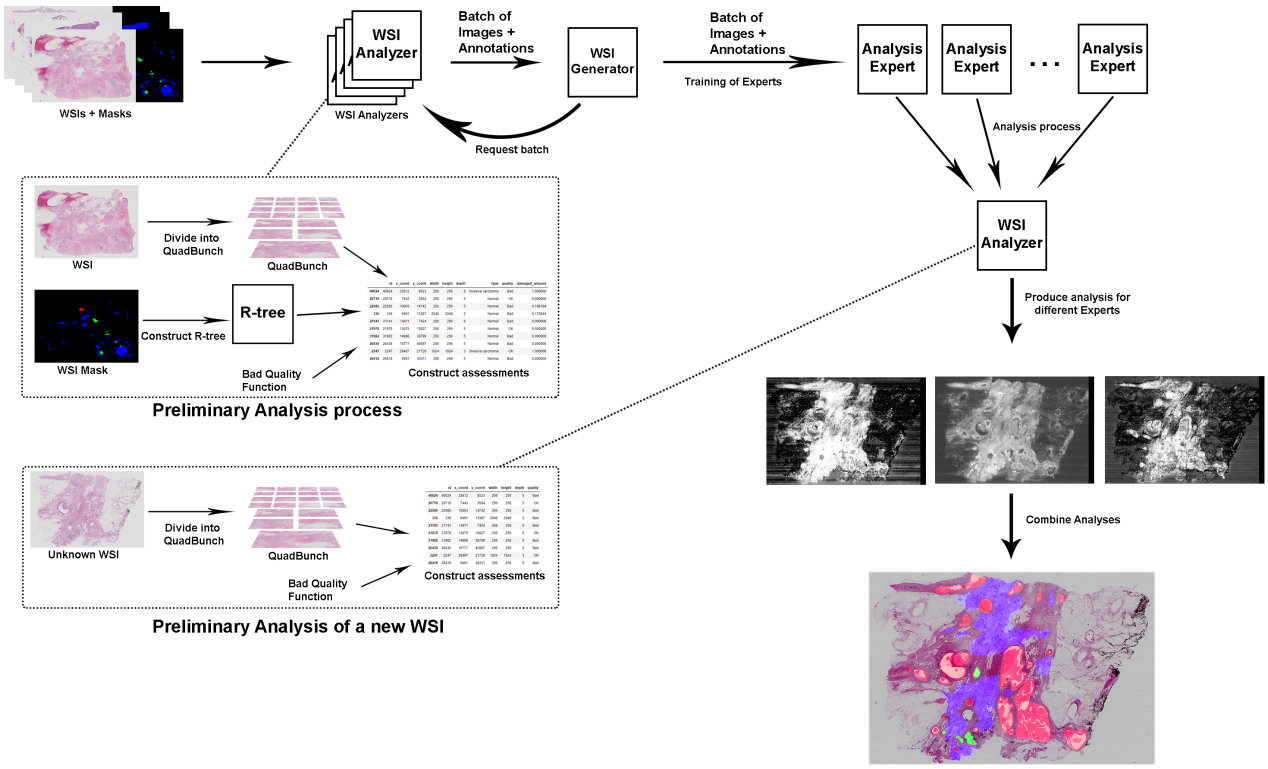

Figure 1: Overview of our system - Horizontal flow = training, vertical flow = analysis of new WSIs. The WSIs and their annotations are processed using WSI Analyzer. Annotation regions are matched to WSI, bad quality regions are assessed and windows of different zooms are selected from the WSI. WSI Analyzers are used to provide images with annotations to WSI Generators, which are used to train *Experts*. Leveraging the flexibility and customizability of the Analyzers and Generators, an *Expert* may be trained for numerous amount of tasks. The *Experts* are then used to produce complex analyses of unknown WSIs by combining simpler analyses.

hol stresses is the complex nature of histopathology data due to the large visual disunity of structures within tumor types, thus making the classification task to be even more difficult.

Bayramoglu et al. [2] focus their research on magnification invariant classification of breast histopathology images using a modern approach leveraging the power of deep learning to replace the whole traditional pipeline. Concluding their work, the authors remark that CNNs provide more promising results in breast cancer histopathology image classification than handcrafted features.

Han et al. [3] use an end-to-end deep learning recognition method using a newly proposed model capable of not only binary but also multi-class classification. Their motivation and primary problems identified with histopathology visual data are consistent with other works - differences in multiple classes due to the broad variability of high-resolution image appearances, high coherency of cancerous cells, and inhomogeneity of color distribution. They propose a class structure-based deep convolutional neural network (CSDCNN) capable of overcoming the problems mentioned above by optimizing the distance of different classes' feature space in the training stage. They achieve state-of-the-art results both in binary and multi-class classification. Concluding their work they lay emphasis on their work being the first time that automated multi-class classification for breast cancer is investigated in histopathological images. They also point out that CS-DCNN is capable of classifying whole slide images by preserving fully global information contained within the images and therefore avoids the limitations of patch extraction methods.

Li et al. [4] present a framework for epithelial cell detection and Gleason grading based on histological images using a two-stage model. Their work focuses on analyzing histopathology images of prostate cancer which includes stroma (tissue) and glands with stroma holding the glands together. They propose a novel model that can automatically diagnose prostate cancer and perform Gleason grading based on histological whole slide images. The model consists of the *Image parser* - ResNet used to generate feature maps, the *Region Proposal Network* - generates region proposals, the *Grading Network Head* - used for predicting class, box offset, and a binary mask for each region, the *Epithelial Network Head* - detects the presence of epithelial cells in the image, and a post-processing step

based on a conditional random field which is applied to the prediction. They note that the model is inspired by Mask R-CNN. Their approach achieves state-of-the-art performance in both epithelial cells detection, with the accuracy of 99.07%, as well as Gleason grading, with a mean IoU of 0.7956

Being able to meaningfully assist doctors with the analysis would help speed up their workflow. Analyzing only zoomed-in and detailed windows of tissue is not enough help for the doctors, as often they are only presented with an unexplained result of classification, produced by a deep learning model, which is a black-box in its nature. The task is also difficult for the network, as it is presented solely with zoomed-in detailed images, and does not see the context, which could provide useful additional information. Using our framework, we aim to ease up the task for the network by introducing contextual information for every zoomed-in window, thus creating an analysis of the whole WSI, which in turn may be used by the doctors. Furthermore, we aim to employ a method to solve the inter-intraclass variance, leveraging different levels of zoom.

3 Framework for WSI Analysis

Due to its vast size and multiple levels of zoom, the WSI contains a gigantic amount of data. Accessing this data, binding annotation to regions, and using it efficiently (generating a batch for training a model must not take too long) is a challenging task. To address this challenge, we propose a two-stage universal framework for complex WSI analysis consisting of *Data Processing* and *Data Analysis* pipelines.

The data processing part consists of two main units - the *WSI Analyzer* and the *WSI Generator*. WSI Analyzer is used to manage WSI and its annotation, and, upon request, produce images with annotations. These images may subsequently be used by a WSI Generator to train a machine learning model.

The data analysis part produces analyses for new, previously unseen WSIs. *WSI Analyzer* is used for this task as well, however, since the ground truth annotation is not input (since it does not exist for a completely new WSI), a machine learning model capable of analyzing images must be used instead. Using this model, the WSI Analyzer is able to infer annotations for different regions of the WSI using multi-level analysis. Multiple models may be used, producing complex combined analysis.

Analyzing the whole WSI using multiple zoom levels yields a complex multi-level analysis capturing both high-zoom details and also their lower-zoom context, and therefore is comparable to analysis performed by a human.

3.1 WSI Analyzer

The main function of WSI Analyzer is enabling easy access to multiple levels of data contained within a WSI, along with their annotations. When using the WSI Analyzer to process the WSI, it is split into smaller windows using a *complete Quadtree* (T-pyramid), in order to assess each of the windows' quality and an in-memory data structure containing information about the positions, sizes, types, and quality of each window is created for future access to the images.

To properly utilize the whole of the WSI and extract as much information as possible, images across all zoom levels are considered for analysis. Splitting for multiple levels is visualized in Figure 2. We employ *T-pyramid* with the aim to manage the images while keeping the pyramid structure of different zoom levels. Using the T-pyramid structure, we can easily combine specific information - acquired from higher zoom level images, with contextual information - from lower zoom level images.

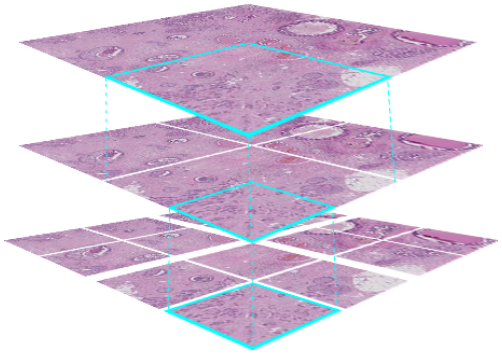

Figure 2: Visualization of first three levels of the T-pyramid.

3.1.1 Annotation Using the R-Tree

With WSI and its annotation as an input, the WSI Analyzer constructs an R-tree of the regions contained in the annotation. The R-tree is a search tree over axis-aligned rectangular regions, mainly used for spatial indexing and accessing geographical data, which are not dissimilar to our annotated regions. Using such a structure provides a convenient way of performing spatial queries on images, in the form of square-shaped regions. Spatially intersecting these square regions with the R-tree index, where annotation information about regions of the WSI is stored (as shown in Figure 3), we achieve a fast way to assess whether an image belongs to at least one tumor region, and if so, how much area of the image belongs to each region. Having access to such data, we can create actual annotations for each image within the T-pyramid by assigning each Quad a class.

3.1.2 Producing Images and Annotation

WSI Analyzers can select quads along with their assessments to create a batch of images with their corresponding annotations, based on multiple parameters: *classes* to use, *zoom levels*, *quality of images produced* (we can exclude

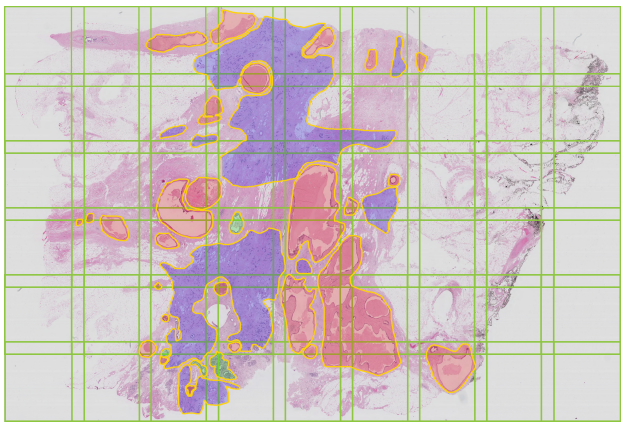

Figure 3: Visualization of first-level quads intersected with annotation regions. Green windows represent quads, different colors of regions mean different tumor type.

images that we consider bad quality - i.e. regions that are too white as they contain only background or fat), *batch size*, *custom preprocess function applied to each image*, *size of generated images*. To achieve the use of different parameters, a *Parametrized Batch Producer* is created for each class contained within the WSI. These producers then manage the creation of parts of the batches by providing images and annotations of their assigned class, which when combined produce a whole, perfectly balanced batch containing samples from each class. Parametrized batch producers may also be used with different, already mentioned, parameters, fine-tuning the batch production to all our needs.

3.2 WSI Generator

We use WSI Generators to produce training data in a format accepted by the models. WSI Generators may use multiple WSI Analyzers to request data from, and thus the final model is allowed to train on a vast set of different images and therefore generalize well. When the Generator unit is created, all classes from across all WSI Analyzers used are considered and unified, ensuring that each WSI Analyzer has the same classes present, which is needed for consistent one-hot encoding of the labels.

The Generator unit is designed with different types of machine learning tasks in mind - multiclass classification, binary classification, and even segmentation, and is able to produce data with different labels used by different models.

4 Deep Learning Experts

The WSI Analyzer unit may also be used to analyze WSIs with unknown annotation. In this case, at least one *Expert* has to be provided when creating the WSI Analyzer. The steps taken are the same as in the case annotation is present, except the R-tree is not constructed. Instead, along with assessing the quality of images, their type is also assessed using the analyzing model or models.

After the WSI is analyzed, an analysis containing the findings can be produced using results of the analysis recorded within the T-pyramid *Quads*. When creating the analysis, information from lower zoom nodes is combined with higher zoom nodes. If *multiple Experts* are used, multiple assessments are contained inside each node. The final analysis is, depending on the *Experts* used, a complex analysis of the slide. Analyses may be also generated for every *Expert* separately.

4.1 Dataset

ICIAR 2018 [1] dataset was used to train the *Experts*. The dataset consists of 10 annotated WSIs, 9 of which were used for training of our *Experts* and 1 was used for evaluation. The annotations associated with the WSIs contain 3 classes of damaged tissue - Benign, In situ, and Invasive carcinomas. Visualization of the annotation may be seen in Figure 10, right, with each of the RGB channels representing a single class. The annotation may also be used to divide the WSI into damaged and non-damaged regions, producing a binary analysis (shown in Figure 9, right).

4.2 Binary Classification Expert

The first *Expert* we design produces binary classification of the WSI into regions that should be analyzed further and ones that should not be. Inspired by the inner workings of Mask R-CNN, specifically its region proposal part, described and used by Li et al.[4] in a similar domain, we use the *T-pyramid* with a simple *binary classifier* to divide the WSI into two parts - one that contains diagnostically important structures, and one that does not (this part would then contain only healthy tissue and fat).

We train and use this first *Expert* to distinguish between healthy and damaged regions. This alone theoretically gives enough pointers to the doctor while deciding on a diagnosis, but what is more important, it may serve us to propose regions with suspected damage for further analysis using other, more specialized *Experts*. Furthermore, the total area which needs to be analyzed gets reduced, which in turn saves time.

4.2.1 Binary Classification Expert Design

After initial experiments using a simple convolutional network, we decided to rely on a network with already existing and well-established architecture *InceptionV3* with 1 *Dense(512)* layer and 1 *Dropout(0.3)* layer added at the end of the network. Scheme of the architecture is shown in Figure 4. Out of multiple experiments using different parameters and configurations, the best validation accuracy achieved was 0.8662.

The optimizer used with the best run was *Adam* with learning rate of $1e^{-6}$, the *batch size* used was 32, and the

model trained for 29 epochs before *early stopping*.

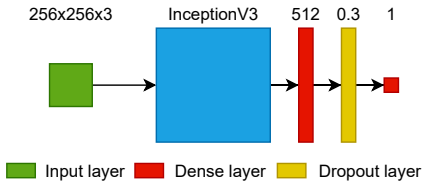

Figure 4: Architecture of the *Binary Expert*.

4.3 Multi-Class Classification Expert

After filtering out redundant structures using *Binary Expert*, we can focus purely on regions marked as damaged and analyze them further, more into details. For this purpose, we design a *Multi-Class Classification Expert*. The *Expert* is able to distinguish between 3 different tumor types contained withing the dataset we use - *Benign*, *Carcinoma in Situ* and *Invasive Carcinoma*.

4.3.1 Multi-Class Classification Expert Design

Considering the problems present within histopathology data of breast, mentioned by both Spanhol [6] and Han [3], but also other authors, we once more opt for a robust architecture - *InceptionResnetV2*. This architecture provides enough capacity to work out the differences and similarities between, but also within the classes and perform well enough multi-class classification. Following the same pattern as in the previous experiment, we added 1 *Dense(512)* layer and 1 *Dropout(0.3)* layer at the end of the network. The scheme of the architecture is shown in Figure 5.

The optimizer used was *Adam* with the learning rate of $1e^{-7}$. The *batch size* used was 18, the model trained for 167 epochs before *early stopping*, and achieved the best validation accuracy of 0.62.

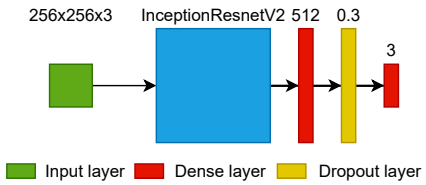

Figure 5: Architecture of the *Multi-Class Expert*.

4.4 Weighing Predictions Expert

Features that can be used to determine the tumor type are best visible using different zoom levels for different tumor types. For some tumor types, contextual information is more relevant, while in other cases, more detailed information from higher zoom levels is the most useful. To leverage the use of multiple zoom levels to the full extent, we can assign different weights to different zooms, which in turn puts more emphasis on different structures for different tumor types.

As we do not know which tumor type is the best comprehensible on which zoom level, we once again use the power of deep learning to solve this problem. We propose a *Weighing Expert*, which can learn to assign different weights to different zoom levels based on the structures present.

4.4.1 Weighing Predictions Expert Design

The *Expert* consists of 6 separate networks (*networks do not share weights*), and a single concatenate layer to join the final values produced by each network. The simplified schema of the architecture may be seen in Figure 6. Each of the networks has its input, therefore each network operates on a single level of T-pyramids created within the WSI Analyzer. The output is $6d$ vector - a weight for each of the levels. A *multiclass label* of ground truth annotation along with 6 *softmax labels*, containing predictions of the previous *Expert*, are used to calculate loss.

The loss for N classes (3 classes are used in our work) is calculated as follows: first of all, the weights are repeated N times, to achieve the shape of $6 \times N$. The predictions of previous *Expert* are multiplied by weights, producing $6 \times N$ matrix of weighted predictions -

$$W_{pred} = weights \odot predictions$$

The predictions are then summed to produce a Nd vector representing weighted multi-class label - W_{sum}, with each class represented in this vector having value from the range $< 0, \infty)$. Softmax is then applied to these labels, producing $W_{Softmax}$, and finally the loss is calculated as a categorical crossentropy between the gt labels and $W_{Softmax}$ values.

$$Loss = -\sum_{i=1}^{N} gt_i.log(W_{Softmax_i})$$

The optimizer used was *Adam* with the learning rate of $1e^{-4}$. The *batch size* used was 128, the model trained for 55 epochs before *early stopping*, and achieved the best loss of 0.5814, and the best evaluation loss of 1.154.

5 Results

In this section, we evaluate the *Experts* trained. First, we perform *Expert* level evaluation, focusing purely on whether the *Expert* performs the task it is assigned to well. In the second part, we perform WSI level evaluation, taking the context into account. All the metrics used to evaluate the different *Experts* are shown in Table 1

5.1 Expert Level Results

We first present the results for the network that represents each *Expert*. We evaluate the *Expert* regardless of the

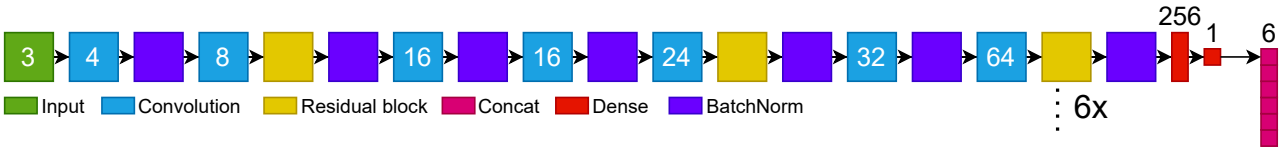

Figure 6: Simplified architecture of the *Weighing Expert*. The network shown is repeated 5 more times to produce the final 6d vector of weights. The number within Conv layers represent number of filters contained. Conv layers have the stride of 2, therefore the image is halved in size with each convolution.

Table 1: The best results achieved by our *Experts*. Note: IoU scores for *Multi-Class and Weighing Experts* are macro-averaged.

Expert	Classes	Accuracy	Precision	Recall	IoU
Binary	2	88.3%	91%	96%	0.74
Multi-Class	3	62.5%	63%	62%	0.44
Weighing	3	75.8%	53%	74%	0.46

"context", to assess how well the *Experts* handle the tasks they were made to perform. All of the results, compared to other, state-of-the-art approaches are shown in Table 2.

Table 2: The best achieved results of analyzed methods for classification. Last row are the best results achieved using our training strategy with state-of-the-art nets.

Authors	Approach	Number of classes	Accuracy
Spanhol et al.[6]	PFTAS descriptor + Traditional classifiers	2	83.33%
Spanhol et al.[7]	DCNN	2	87.28%
Bayramoglu et al.[2]	DCNN, magnification independent	2	82.1%
	DCNN, magnification specific	2	80.66%
Han et al.[3]	CSDCNN	2	96.25%
	CSDCNN + Aug	8	93,88%
Our Solution	InceptionV3	2	88.3%
	InceptionResnetV2	3	62.5%
	InceptionResnetV2 + WeighingNet	3	75.8%

5.1.1 Binary Classification Expert Results

The *Binary Expert* achieved an accuracy of 88.3%, the precision of 91%, and recall of 86% when evaluated using 20480 previously unseen images balanced across both classes. Judging from these metrics, we can see that the *Expert* is doing well in the task of distinguishing between healthy and damaged tissue. These results are satisfactory enough for the *Expert* to be used to filter out redundant and diagnostically not important structures.

5.1.2 Multi-class Classification Expert Results

Evaluating the *Expert* using 20480 previously unused images, evenly distributed among all classes, accuracy of 62.5% was reached, along with 63% precision and 62% recall. Figure 7 shows confusion matrix for different classes this model is trained to predict. These results show room for improvement, but as previously stated, there is a limit given by the difficult nature of breast histopathology data, to the performance of any deep learning model. Contrary to Han [3], who uses special architecture and loss function, we try to deal with this limit using a *Weighing Expert*.

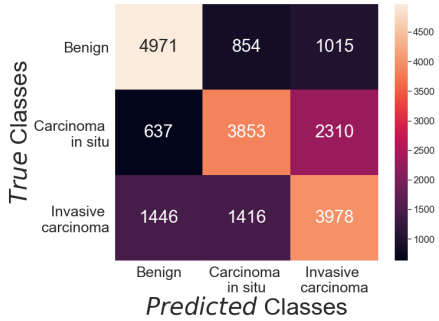

Figure 7: Confusion matrix for the *Multi-Class Expert*.

5.1.3 Weighing Predictions Expert Results

In order to evaluate whether the *Weighing Expert* is working correctly, we define the action of changing one class to another as a *flip*, and the action of not changing the class as a *no flip*. Using the non-weighted predictions, the weighted predictions, and the ground truth annotations, we may then count *correct flips* - the predicted class changes from an incorrect one to a correct one (equivalent to true positive), *incorrect flips* - the predicted class changes from a correct one to an incorrect one (false positive), *correct no flips* - the predicted class does not change and stays correct (true negative), and finally, *incorrect no flips* - the predicted class does not change and stays incorrect, or the predicted class changes, but from an incorrect one to an incorrect one (false negative). We can therefore calculate all the metrics, as usual, using true positives, true negatives, false positives, and false negatives.

We performed the evaluation using 12123 samples from previously unseen WSI. The *Expert* achieved the accuracy of 76%, the precision of 77%, and the recall of 36% in the task of *correcting the output of the previous Expert*.

Focusing not on the *flip* itself, but on the end result, therefore comparing the weighted predictions to the ground truth labels, the net achieved the accuracy of 75.8%, the precision of 53%, and the recall of 74%.

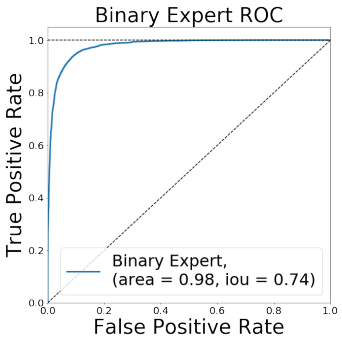

Figure 8: ROC of the Binary Expert achieved on the Evaluation slide. The dotted line = random guessing.

5.2 WSI Level Results

In this part of the evaluation of our solution, we focus on the whole of the solution and evaluate the predictions combined using our framework, for each *Expert*. The metric we chose as best suited for this evaluation is Intersection over Union, calculated as $IoU = \frac{gt \cap predicted}{gt \cup predicted}$, but also try to empirically evaluate each analysis result from the point of view of a human expert using the analysis as assistance.

5.2.1 Binary Expert WSI Level Results

The Figure 8 shows the ROC of the *Binary Expert*. Judging from the ROC, with an area of 0.98, and IoU score of 0.74, but also performing empirical evaluation using the GT annotation, and the results of *Binary Expert* analysis, shown in Figure 9, we can say that the *Expert* is able to distinguish between normal and damaged tissue, which serves both as a filtering step - parts of stroma not containing any significant structures are removed, but at the same time as a region proposal for more specific *Experts*.

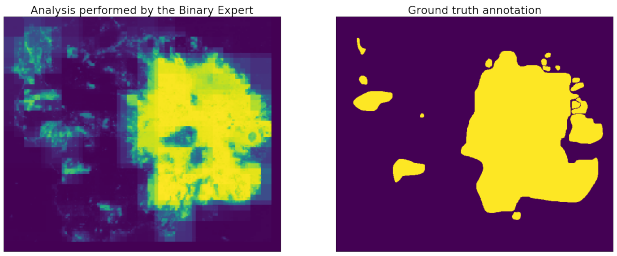

Figure 9: Left: Analysis performed by the *Binary Expert*. Right: GT annotations of the Slide. Higher intensity means higher confidence for the *Expert*.

5.2.2 Multi-Class Expert WSI Level Results

Figure 10 shows the analysis produced by the *Multi-class Expert* along with the ground truth annotations used to evaluate the *Expert*. IoU scores of the whole slide are shown in Table 3, distinguished as "Not Weighted". Only regions marked as damaged by the *Binary Expert* with confidence >0.3 were analyzed by the *Multi-Class Expert*. We can clearly see that the *Binary Expert* region proposal works and is precise enough, helping make the area analysed by subsequent *Experts* smaller, thus saving time and resources. Furthermore, we can see that the *Expert* is able to pick up the differences between tumor types and create a complex analysis.

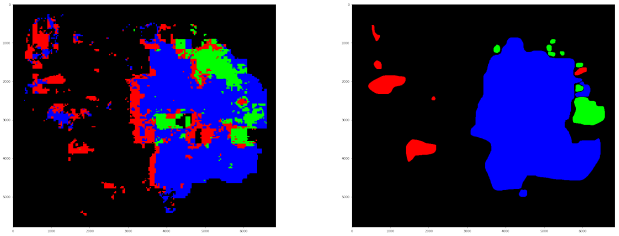

Figure 10: Left: Analysis produced by the *Multi-Class Expert*. Right: the GT analysis that should be matched. Red channel = *Benign* tumor type, Green channel = *In Situ* tumor type, Blue channel = *Invasive* tumor type.

Table 3: IoU scores achieved by the *Multi-Class Expert* when evaluated using a previously unseen WSI, along with IoU scores of the *Weighing Expert* applied to the *Multi-Class Experts'* predictions.

		Micro	Macro	Binary
All classes	Not Weighted	0.85	0.44	-
	Weighted	0.87	0.46	-
Benign	Not Weighted	-	-	0.11
	Weighted	-	-	0.13
In Situ	Not Weighted	-	-	0.15
	Weighted	-	-	0.18
Invasive	Not Weighted	-	-	0.57
	Weighted	-	-	0.61

Similarly to the results of *Multi-Class Expert* evaluation (Section 5.1.2), we see that although the *Expert* is often able to correctly analyze regions, confusion of the classes, caused by all of the mentioned problems is still present, even on WSI level.

5.2.3 Weighing Expert WSI Level Results

Figure 11 shows the comparison of the analysis produced by the *Multi-Class Expert* and the same analysis but weighted by the *Weighing Expert*. The change of the IoU may also be seen in Table 3, comparing the rows marked as "Weighted" with the ones marked as "Not Weighted". We see that the metric - IoU, is slightly enhanced by the *Weighing Expert*.

Empirically evaluating the performance of the *Weighing Expert*, and therefore comparing the *weighted analysis* (Figure 11, right) with the *non-weighted* one (Figure 11, left), but also with the *ground truth annotations* (Figure 10, right), we see that the *Weighing Expert* corrects the

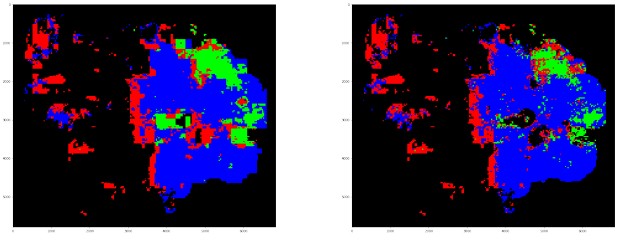

Figure 11: Analysis produced by the *Multi-Class Expert* and the same analysis weighted by the *Weighing Expert*. Red channel represents *Benign* tumor type, Green channel is *In Situ* tumor type and Blue channel is *Invasive* tumor type.

class in the right areas - specifically - in the middle, large area of the *In situ* type (green), and *Benign* (red) is correctly changed to *Invasive* type (blue), but also some minor changes may be seen in the top-left areas, upon closer inspection. Such results suggest, that the *Expert* is working correctly, but the metrics of automatic evaluation tell us that there is still room for improvement.

6 Conclusion

In this paper we presented a universal framework for complex WSI analysis, using various methods of deep learning in the form of *Experts*. We also presented multiple *Experts*: the *Binary Expert* for distinguishing between healthy and damaged parts of the WSI and proposing damaged regions to the other *Experts* for further analysis, the *Multi-Class Classification Expert* for classifying regions as different types of tumor and the *Weighing Expert* to make the analysis more precise. As a result of using the *Experts* combined, we achieve a complex and unbiased analysis of the whole WSI, which may serve as a reference, or a helping tool to pathologists.

The binary classification results we achieved using our method for data preparation and generation, and an already existing architecture - InceptionV3, are comparable to other state-of-the-art approaches presented by Spanhol [7], [6], and Bayramoglu [2].

Our multi-class classification results are not yet on par with the likes of Han [3], but with a more thorough training and parameter tuning of the InceptionResnetV2 we use for this task, but also with further exploring the capabilities of the *Weighing Expert*, and enhancing its capacity for learning by increasing the width of each individual network the *Expert* contains, we are on the right track to achieve state-of-the-art results in this domain as well in the future.

The universality of the framework allows for creating any number of *Experts*, each analyzing a specific part of the WSI, leading to even more complex and precise analyses. An *Expert* for segmentation of the tissue, distinguishing even better between the classes, especially around their borders would be valuable, although hard to train due to the nature of histopathology images of the breast. Another *Expert* of value would be one for counting mitoses on the WSI level, as this task is often required to grade the stage of cancer.

References

[1] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, Gerardo Fernandez, Jack Zeineh, Matthias Kohl, Christoph Walz, Florian Ludwig, Stefan Braunewell, Maximilian Baust, Quoc Dang Vu, Minh Nguyen Nhat To, Eal Kim, Jin Tae Kwak, Sameh Galal, Veronica Sanchez-Freire, Nadia Brancati, Maria Frucci, Daniel Riccio, Yaqi Wang, Lingling Sun, Kaiqiang Ma, Jiannan Fang, Ismael Kone, Lahsen Boulmane, Aurélio Campilho, Catarina Eloy, António Polónia, and Paulo Aguiar. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019.

[2] Neslihan Bayramoglu. Deep Learning for Magnification Independent Breast Cancer Histopathology Image Classification. pages 2440–2445, 2016.

[3] Zhongyi Han, Benzheng Wei, Yuanjie Zheng, Yilong Yin, Kejian Li, and Shuo Li. Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Scientific Reports*, 7(1):4172, 2017.

[4] Wenyuan Li, Jiayun Li, Karthik V. Sarma, King Chung Ho, Shiwen Shen, Beatrice S. Knudsen, Arkadiusz Gertych, and Corey W. Arnold. Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images. *IEEE Transactions on Medical Imaging*, 38(4):945–954, 2019.

[5] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(December 2012):60–88, 2017.

[6] Fabio A Spanhol, Paulo R Cavalin, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A Dataset for Breast Cancer Histopathological Image Classificatio. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[7] Fabio Alexandre Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using Convolutional Neural Networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2560–2567. IEEE, 07 2016.