HHS Public Access

Author manuscript

Comput Med Imaging Graph. Author manuscript; available in PMC 2022 May 17.

Published in final edited form as:

Comput Med Imaging Graph. 2018 April ; 65: 142–151. doi:10.1016/j.compmedimag.2017.09.001.

Towards Machine Learned Quality Control: A Benchmark for Sharpness Quantification in Digital Pathology

Gabriele Campanella[a,b], **Arjun R. Rajanna**[b], **Lorraine Corsale**[c], **Peter J. Schüffler**[b], **Yukako Yagi**[c], **Thomas J. Fuchs**[a,b,c,*]

[a]Weill Cornell Medicine, New York, USA

[b]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, USA

[c]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, USA

Abstract

Pathology is on the verge of a profound change from an analog and qualitative to a digital and quantitative discipline. This change is mostly driven by the high-throughput scanning of microscope slides in modern pathology departments, reaching tens of thousands of digital slides per month. The resulting vast digital archives form the basis of clinical use in digital pathology and allow large scale machine learning in computational pathology.

One of the most crucial bottlenecks of high-throughput scanning is quality control (QC). Currently, digital slides are screened manually to detected out-of-focus regions, to compensate for the limitations of scanner software.

We present a solution to this problem by introducing a benchmark dataset for blur detection, an in-depth comparison of state-of-the art sharpness descriptors and their prediction performance within a random forest framework. Furthermore, we show that convolution neural networks, like residual networks, can be used to train blur detectors from scratch. We thoroughly evaluate the accuracy of feature based and deep learning based approaches for sharpness classification (99.74% accuracy) and regression (MSE 0.004) and additionally compare them to domain experts in a comprehensive human perception study. Our pipeline outputs spacial heatmaps enabling to quantify and localize blurred areas on a slide. Finally, we tested the proposed framework in the clinical setting and demonstrate superior performance over the state-of-the-art QC pipeline comprising commercial software and human expert inspection by reducing the error rate from 17% to 4.7%.

Keywords

Computational Pathology; Digital Pathology; Quality Control; Machine Learning; Deep Learning; Quantitative Blur Detection

*Corresponding author.

Conflict of Interest Statement

The authors do not have any conflict of interest.

1. Introduction

Pathology has traditionally a labor intensive manual work-flow which includes tissue excision, slide preparation and staining, and the pathologist's diagnosis based on these slides. Unlike other medical imaging domains such as radiology, which have seen a complete transformation to a digital work-flow, pathology has yet to take this leap in the US. A large difference to radiology, where the images are always digitally generated and stored, is that the digitization of pathology requires this additional step of scanning the analog slides to the work-flow [12]. In addition, the food and drug administration requires pathologists to base their diagnostic report on the analog microscopic images instead of digital slides [12]. As a result, there has not been a strong push from the clinical side to digitize the pathology work-flow in the past.

In recent years, however, with the advent of high-definition, high-throughput scanners and important advances in the fields of computer vision and machine learning, computational pathology has emerged with the intent to create a unified framework for pathology image analysis to aid the pathologist work [11]. These efforts are adding strength in favor of digitizing pathology, with clinics all around the world steadily increasing their scanning efforts. The amount of scanned slides can exceed the tens of thousands per month, as e.g. at the Memorial Sloan Kettering Cancer Center (MSKCC). This necessarily automated digitizing process results frequently in undesirable artifacts such as out-of-focus scans or blurred regions due to tissue folds, varying tissue thickness, air bubbles, pen markers, dust, scratches or others. Figure 1 examples a partially blurred slide: some areas are blurred since the focus points were not set correctly by the machine. Blurred slides have to be rescanned for downstream processing.

Further, in the context of computational pathology, automated analysis pipelines rely on sharp, artifact-free images. Therefore, it becomes absolutely essential to have a robust quality control (QC) system in order to obtain significant results from the analysis.

Indeed, most currently available scanners have built-in quality control mechanisms [21], but according to the experienced staff at the MSKCC's pathology department, these seem to under-perform commonly. Others have reported similar difficulties [12, 3]. Frequent human intervention is therefore necessary to verify the quality of the scanned images, a very time consuming and repetitive task. The QC of digital slides at MSKCC as an example includes (i) automatic choice of focus points by the scanner, (ii) manual check of the focus points on every slide (and relocation if needed), (iii) automatic quality assessment after high resolution scanning by the scanner, (iv) manually inspection of very fourth to tenth slide. This last manual check is performed because of the poor confidence in the scanner's built-in QC system. There is no question that an automatic and reliable artifact detection would benefit both the clinic by reducing the time needed for manual assessment and the research by generating high quality datasets for any computational pipeline.

Out-of-focus detection has been an important topic of research and vastly applied in autofocusing mechanisms especially in digital cameras and microscopy [18, 9, 7]. An

exhaustive review of algorithms aiming to quantify the sharpness of an image is given in Ferzli *et al.*[9]. These metrics generally tend to perform well when describing blurriness in images with the same content (i.e. blurred versions of the same sharp image), but fail in recognizing blurriness in natural images with different content [9, 13, 8]. For pathology images, whose content can be very diverse, the combination of sharpness metrics has been suggested. For example, Lahrman *et al.*[15] have developed a complete pipeline for the scanning of cervical cytology slides where they asses quality of a slide in terms of sharpness and automatically re-scan the slide if the quality measure is below a set threshold. Their sharpness detector is based on a support vector machine classifier using five sharpness metrics as features (number of edges, gradient score, difference to sharpened, difference to smoothed, difference to blurred). While their method seem to perform very well, its scope is restricted to liquid-based cytology slides. Similarly, Ameisen *et al.*[3] developed a tool to assess the quality of virtual slides using a blur detector algorithm patented by the group.

In this work, we detect and quantify regions of different blur levels for different image sizes and for different tissues. We compare the engineered features approach using a random forest model [4, 6] with feature learning using a state-of-the-art convolutional neural network, called residual network (ResNet), that has been shown to have better convergence and accuracy [14]. In addition, we implemented a standalone blur detection application in python for usage in the clinic and for research. An overview of the pipeline is presented in figure 2. We tested the application in a clinical real-world scenario and it outperformed the current QC standards.

The remainder of the paper is structured as follows: Section 2 characterizes the newly created datasets along with details about the machine learning approaches used; Section 3.1 details the results of our benchmark on state-of-the-art engineered features; Section 3.2 includes our comparisons between feature engineering and feature learning; finally Section 3.3 describes the implementation of the blur detector software framework, its validation and the tests performed in the clinic.

2. Material and Methods

2.1. Benchmark of Engineered Features for Blur Detection

2.1.1. Dataset—In order to facilitate the quantitative evaluation of different machine learning algorithms, we generated a novel dataset specifically tailored to sharpness quantification. 30 tissue microarray (TMA) spots from clear cell renal cell carcinoma (kidney cancer) patients, and 159 whole slide images (WSIs) of prostate cancer were retrieved from MSKCC's pathology department. 18 hippocampus WSIs associated with hippocampal sclerosis were granted by the University Hospital Zurich. The prostate and kidney slides were scanned on an Aperio AT2 whole slide scanner (Leica Biosystems), whereas the hippocampus slides were scanned on a Nanozoomer C9600 virtual slide light microscope scanner (HAMAMATSU). All slides were subsequently anonymized to protect patient privacy. To guarantee broad applicability of the final prediction models not only in terms of instrumentation (e.g. using different scanners), the three sets were processed with different immunohistochemical stainings: Prostate was stained with H&E, hippocampus with SDF-1 and kidney with TOM20. All the slides were manually inspected

to be completely free of blurred regions. We extracted squared gray scaled patches of 64, 128, 256 and 512 pixels without overlap from those slides. The kidney samples originating from TMAs included wide areas of white background. Therefore, a simple thresholding approach (t=230 on gray scaled patches) was used to exclude patches with too much background. Finally, we artificially blurred the sharp patches using a Gaussian filter simulating out-of-focus blur. By increasing the standard deviation parameter σ of the Gaussian filter, we obtained increasing levels of blurriness: 0 (sharp), 0.8, 1.2, 1.6, 2, and 2.4. Table 1 summarizes the extracted patches: the final dataset contained 2880 prostate, 3816 hippocampus, and 4240 kidney patches of different sizes and blur levels. Due to the large number of patches and the arbitrary tissue orientation on the slides, no further data augmentation was deemed necessary.

2.1.2. Feature Selection—Many sharpness metrics for blur detection have been described in literature [18, 9, 7]. According to the approach used to estimate sharpness, they can be divided in 4 categories: pixel intensity based, gradient based, transform based and perceptual. We implemented 13 sharpness metrics (5 pixel intensity based, 3 gradient based, 3 transform based and 2 perceptual) from their respective papers, taking care to include the widest range of methods and approaches. For an exhaustive description of the metrics used, see appendix AppendixA. It is important to note that all features rely on gray-scale images.

2.1.3. Random Forest—A random forest [4, 6] consisting of 1000 trees was trained with 13 features using standard parameters. Both a classification task and a regression task were tested. For the classification task, all the patches with blur level equal to 0 were considered sharp and all the others blurred. In the regression task the hypothetical standard deviation of a Gaussian filter applied to the image is the value to be predicted. A leave-one-image-out cross-validation (CV) was employed where all patches extracted from a single TMA or slide were taken out of the training set and used as a test set. Iterating through all images, we obtained a prediction for every patch. Since the Random Forest is based on randomness, prediction error variability was then estimated by repeating the CV analysis 30 times. Finally the importance of features was obtained using a greedy search approach by adding at each step the metric that minimizes the CV prediction error.

2.2. Feature Engineering and Feature Learning Comparison

2.2.1. Datasets—We generated two new datasets, a prostate only dataset and a mixed tissues dataset, for comparing the performance of a feature engineering approach and a feature learning approach. From the prostate slides used before we generated 22896 gray-scale square patches of size 256 pixels. These were divided in 3 sets: a training set of 16028 patches, a validation set of 3434 patches and a test set of 3434 test set. The patches were artificially blurred as previously explained resulting in a final training set of 96168 patches for training and 20604 for each validation and test. From the same pool of slides used before plus 69 slides from skin biopsies, we generated a mixed dataset consisting of 7280 square patches of size 256 pixels per tissue type (prostate, hippocampus, kidney, skin) for training, and 1560 per tissue type for each validation and test. The patches were artificially blurred as previously explained, resulting in a final training set of 174720 patches for training and 37440 for each validation and test.

2.2.2. Logistic Regression and Random Forest—The 13 sharpness metrics were extracted from the datasets. Training was performed on the compound of training and validation sets, errors reported are calculated from predictions on the test set. A logistic regression and a random forest of 1000 trees were trained for a 6 class classification task. Similarly, a random forest was trained for a regression task. Using a model with all features and with the reduced set of 10 features was also investigated.

2.2.3. Residual Neural Network Training—We trained an 18 layer deep ResNet model[1] which takes as input 3-channel images of 224×224 pixels. We modified the architecture to accept as input gray-scale images. This was because, firstly, it allows for a better comparison with the engineered features approach which relies on gray-scale images, and, in addition, it can help the ResNet learn color independent features, especially since staining varies widely between tissues. Finally, during training, center crops are taken from our 256 pixels patches to accommodate the network's 224 pixels input requirement. A 6-class classification task was performed, where each blur level was considered a class with no ordinal information, with cross-entropy as cost function. A regression task was performed, where the blur level is the target, with MSE as cost function. Training from scratch was performed in parallel on four Nvidia TitanX GPUs for 300 epochs with hyper-parameters set as follows: batch size=1024, learning rate=0.1(multiplied by 1/10 every 30 epochs), momentum=0.1. For each epoch the training was done on the training set and validation error calculated for the validation set. After 300 epochs, the best performing model on the validation set was chosen. The test set error was then measured for the best model. Convergence plots are shown in appendix AppendixE.

2.3. Blur Detector

2.3.1. Human vs detector agreement—In the field of digital image quality control, testing a sharpness metric is most frequently done by showing the test set to individuals who give a score usually from 1 to 5. Then, the mean score for each image, referred to as MOS, is correlated to the proposed metric [9]. Similarly, we tested our entire pipeline on real pathology cases at MSKCC. The blur detector (previously trained as described in 2.2.2) was run on a set of 10 slides from the pathology department at MSKCC that were found to have out-of-focus regions by the technicians. It is important to note that these slides were not filtered to contain only tissue types that were used for training (kidney, prostate, hippocampus), and hence also include other tissue types. We extracted a balanced pool of images where the space of the regressed target was binned in 6 regions, obtaining in total 2345 patches of 512 pixels. The images were presented to experts in the field via a simple web application accessible only within MSKCC. Each user was asked to score the level of blurriness of a single image at a time out of 6 possible levels 0 through 5, with 0 being sharp and 5 very blurred. Scores and scoring time for each image were saved. Screen-shots of the web application are shown in appendix AppendixD. In the end we obtained scores for 1391 images from 10 experts: 2 experts that are in charge of the quality control of scanned slides, 4 pathologists, and 4 other scientists working at MSKCC. Spearman correlation is then used to measure the agreement between raw expert and detector scores. It is important to note that users were not trained or preconditioned. Prior to the task a brief explanation was given and

6 example images were shown with a different scoring system (3 classes) than the one used during the task. Screen-shots taken from the scoring are presented in appendix AppendixD.

2.3.2. Tests in the clinic—The blur detector was run on 193 whole slides. A slide was considered positive for blur if more than 5% of the tissue patches had a predicted score higher than 0.8.

3. Results and Discussion

3.1. Benchmark of Engineered Features for Blur Detection

Here we discuss the performance of several state-of-the-art features in detecting blurred regions in pathology slides. We investigated the performance of these features alone and using a random forest model, their importance and the relationships between performance, image size and tissue type.

3.1.1. Single Feature Performance—In the first instance, we tested the implemented sharpness metrics on a subset of "natural images" taken from the UT Austin LIVE Quality Assessment Database [23, 25, 2]. This database is commonly used to benchmark sharpness metrics: it contains sharp images and artificially blurred versions of them (a Gaussian filter with increasing standard deviation (0.8, 1.2, 1.6, 2, 2.4) is used to simulate out-of-focus blur). Results are extensively presented in appendix AppendixB. Briefly, most of the metrics were able to recapitulate the increase of blurriness for the same image, showing a monotonic increase or decrease. The same level of Gaussian blurriness on different images, gave very different responses for the different metrics. Perceptual metrics, which take into consideration the perception of blur by the human visual system, performed better than the other metrics but still failed to capture unambiguously the level of blurriness in natural images.

The UT Austin LIVE Quality Assessment Database contains only natural images, and the performance of the features could be different for images coming from pathology. We then tested the same metrics on pathology images, in particular tissue microarrays (TMAs) from clear cell renal cell carcinoma patients with the mitochondria staining TOM20. Patches of 512×512 pixels were randomly extracted from a set of six sharp TMAs and increasingly blurred versions were generated as previously explained. We show in appendix AppendixC that even with these apparently similar images, their content is variable enough that no metric correlates across patches with the blur level. In conclusion, single features were insufficient to detect blurriness across different images. In the next sections we discuss how the ensemble of features is descriptive enough to detect blurred regions across pathology images with different content.

3.1.2. Classification Experiments—Classification experiments were performed on single tissue datasets (e.g. prostate, hippocampus, kidney, as described in section 2.1.1), expecting prediction to be very accurate since the image content is fairly similar within these datasets. Indeed, the prediction accuracy was very high for all datasets with errors of 2%, 1.5% and 0.3% for prostate, kidney and hippocampus datasets respectively for patches of 64×64 pixels and 0.5%, 0.2% and 0.5% for patches of 512×512 pixels. It was observed

that the size of the patches influences accuracy with prediction on bigger patches being more accurate than on smaller patches as it can be seen in figure 3. The feature importance analysis determined that, in the case of intra-set tasks, only 2 or 3 metrics would be enough to reach minimum prediction error, with perceptual metrics and cosine transform metric being the most important. Going one step further, all datasets were combined to allow for a more general classifier able to predict blurriness independently from the tissue type, at least within the tissues in our dataset. Kidney samples with mitochondria staining were harder to predict with an error of around 3% mostly due to false positives (sharp patches that were predicted blurred), while prostate and hippocampus had instead errors of 0.5% and 0.2% respectively, as shown in figure 4. The feature importance analysis (figure 4) underlined how, in this case, more metrics are necessary to discern blurriness when image content has higher variance.

3.1.3. Regression Experiments—A regressor was then trained on all 3 datasets to also predict the level of blurriness of the patches. Results were encouraging with an RMSD close to 0.012 and a Spearman correlation coefficient larger than 0.98. In figure 5, the dispersion of the predictions is shown and a very flat distribution centered around the expected values can be observed, underlying the accuracy of the regression. The feature selection (figure 5) was performed minimizing the MSE and it was in accord with the results of the classification task.

3.2. Feature Engineering and Feature Learning Comparison

We compared the performance of two machine learning approaches, logistic regression and random forest, which rely on manually engineered features, with residual neural networks as state-of-the-art convolutional neural networks that are able to learn from scratch the features important for classification.

3.2.1. Classification Experiments—We started our experiments using the prostate only dataset. The ResNet converged to 0.03% error on the validation set after approximately 50 epochs (supplemental figure AppendixE.1). The best model (epoch 93) showed 99.95% accuracy on the test set across all classes. The result was comparable to the random forest approach, which achieved an accuracy of 99.39%. Interestingly, the logistic regression showed an accuracy of 94%, pointing to the fact that the non linearity introduced by the random forest or the neural network, is important for this task. In addition, by using a reduced set of features we lose only 0.29% and 0.66% accuracy for logistic regression and random forest respectively. Having obtained good results with one tissue, we then moved on to the mixed tissues dataset. For the ResNet, the validation error converged to 0.29% after roughly 100 epochs. Taking the best performing model (epoch 193) we had an accuracy of 99.74%, outperforming our random forest approach with accuracy 97.43%. Again, the logistic regression performed considerably worse (accuracy of 87%). Using the reduced set of features, the accuracy drops by 4.18% and 1.08% for the logistic regression and random forest respectively. A summary of these results is also presented in table 2. Even though the ResNet seems to outperform the random forest, in practice, they would perform similarly if the task was a binary classification: sharp and blurred. This can be seen in the confusion matrices presented in figures 6 and 7. In the case of mixed tissues, for example, the random

forest would yield 25 false positives and 14 false negatives, while the ResNet would give 14 false positives and 19 false negatives.

3.2.2. Regression Experiments—We then tested the performance of a regression task to predict the level of blur present in a patch. As it can be seen in figure 8, all the models were able to discern the different levels of blur with fair accuracy. The Resnet converged after roughly 30 epochs (supplemental figure AppendixE.2). Prediction on the test set using the best model gave an MSE of 0.018, whereas the two random forest approaches resulted in 0.004 and 0.005 when using all features or a reduced set respectively. The various methods showed high level of prediction agreement (supplemental figure AppendixF.1)

3.3. Blur Detector

3.3.1. Implementation—The results from previous experiments encouraged us to implement a blur detection application that could be used as part of a computational pipeline and in the clinic to more efficiently determine the quality of digital slides. Scanning of such slides results in image files of several gigabytes. Any computation performed on such images is intensive and particular attention had to be put in maximizing efficiency and computation speed. We decided to move forward with a random forest based regression as core module of our software. The random forest performance was comparable to that of the ResNet. In addition, the random forest pipeline could be more easily ported to the computers of the pathology department that are connected to the scanners.

The pipeline and metrics were implemented in python and performance was tested in terms of feature extraction time vs. prediction accuracy. This is because the bottleneck of this approach in terms of speed is actually the feature extraction, as opposed to the actual random forest prediction, which is very fast. The number of trees of the random forest was set to 19 in the implementation. In figure AppendixG.1 the time needed for feature extraction from a square patch of 512 pixels is plotted against the accuracy of the random forest with that particular set of features. The most time consuming features were removed one by one from the model and the increase in prediction MSE was measured. As expected, as the number of features used decreases, the computation time also decreases with the accuracy of the prediction and an increase in MSE. The final regressor used in the application uses a reduced set of 10 metrics. This set of features keeps the processing time to a minimum while maximizing the accuracy. The above described regressor was trained on the mixed dataset described in section 2.2.1 and it is the main component of the algorithm that we implemented to detect blurred regions in digital slides. The algorithm takes full advantage of the *OpenSlide*[22] API for python. *OpenSlide* is a C library that allows to read whole-slides for most of the commercially available scanners. Our pipeline is designed to run in parallel on all processors of a machine to increase efficiency: for a test slide 67,448 pixels wide and 27,817 pixels high, with tissue covering 20% of the slide area, the single process execution takes on average 309 seconds, while a parallel execution on 32 CPUs brings the execution down to 17 seconds. Further performance analysis is presented in appendix AppendixG.2. Briefly, a sliding window approach is used to scan through the slide's dimensions to find tissue regions. If a patch is detected as tissue, the metrics are extracted and blur level

predicted. In the end, a map of blurred and sharp regions is obtained. An overview of the algorithm is presented in figure 2.

3.3.2. Human vs detector agreement—The overall humans' score was highly correlated to the detector's score with a Spearman rank correlation of $\rho = 0.66$ (95% CI: 0.63–0.69), even without expert training. Figure 9 stratifies the users into "quality control personnel", "pathologists" and "others" revealing the highest correlation for people routinely engaged in quality control of virtual slides ($\rho = 0.786$), followed by pathologists ($\rho = 0.731$) and other scientists ($\rho = 0.581$). Focusing on the false negatives (images considered sharp by the blur detector but not so by the humans), we can understand where to focus our attention to improve our detector. Around one third of the false negatives were very homogeneous stroma images without any edges or structural elements hindering sharpness determination. A few other cases were interesting because they were mostly sharp but had small regions clearly out of focus. The detector seems to weight more the presence of the sharp regions, while the experts payed more attention to the blurred regions. Examples of the false negatives are shown in appendix AppendixH.

3.3.3. Clinical evaluation—Finally we evaluated the proposed framework in the clinical setting at Memorial Sloan Kettering. The current clinical pipeline consists of commercial quality control by the software of the scanner vendor followed by manual quality control by a team of QC technicians. To quantify the difference between the state-of-the-art and our machine learning approach, we analyzed all 196 slides scanned with a single scanner during one day of clinical operation. Three of these slides were loaded in reverse and correctly flagged by the scanner's QC application while all other 193 slides passed the commercial quality control checks. The subsequent routine manual assessment consisted of spot checking every fourth slide by the quality control technicians. The inspection of these 48 slides resulted in two additional detections and the slides were submitted for rescanning.

For comparison we automatically analyzed all 193 slides with the proposed framework. Besides the two slides that were found by the technicians, our system predicted 42 extra slides containing blurred regions. An expert QC technician was then tasked to inspect the detected slides. 33 slides were found to be blurred while 9 were considered false positives. It should be noted, that in clinical practice, 9 of the blurred slides would not have been rescanned due to blurred artifacts, that would not have improved with repeated scanning. We hypothesize that the false positives slides could have been classified correctly by optimizing the blur score decision threshold, which has been kept fix on this independent test set. An example of a positive detection is shown in figure 10. Meanwhile, examples of false positives are shown in appendix AppendixI.

In summary our algorithm detected 33 blurred slides the joint commercial and human pipeline missed, thus reducing the error rate from 17% to 4.7% (table 3).

4. Conclusions

Modern pathology departments, like the one of the MSKCC, digitize tens of thousands of whole slides per month, which makes high precision quality control a paramount element of

every high-throughput digital pathology pipeline. In addition, the fine-grained assessment of quality not only on slide level but within a slide itself is an indispensable prerequisite for real-world computational pathology pipelines.

In this paper we focused in depth on the aspect of blur detection. The contributions of this work are the following: (i) We created a comprehensive benchmark dataset consisting of 10936 patches from digital slides of prostate cancer, renal cancer and hippocampal sclerosis patients. (ii) We implemented 13 state-of-the art sharpness metrics and (iii) conducted a comprehensive comparison of their performance and extraction speed within a random forest framework. To compare feature engineering vs. feature learning (iv) we trained a residual network from scratch and compared its prediction accuracy to random forests and logistic regression. In addition to exhaustive quantitative evaluation we conducted a qualitative study with human domain experts from MSKCC's quality control group, practicing pathologists and additional scientists. To this end we (vi) implemented a web-application for collecting and visualizing human expert estimations. Furthermore, (vii) we implemented a parallel blur detection software package which can take advantage of modern multi-core systems and hence offers high speed quality control for high-throughput digital pathology workflows. The software produces fine-grained sharpness assessment maps for every slide, thus enabling computational pathology at scale on real-world data. An example blur map is presented in figure 11. Finally, (viii) we independently tested the proposed system in the clinical setting and compared it to the state-of-the-art joint QC pipeline of commercial scanner software and human QC experts, resulting in a reduction of detection error from 17% to 4.7%.

We expect that the use of automated methods for blur detection will substantially enhance the digital pathology work-flow in addition to overall greater virtual slide quality. On the computational pathology side, it has still to be shown quantitatively to what extent the introduction of a quality control step will enhance the accuracy of computational pipelines beyond the empirical evidence. We strongly believe that an effort has to be put into the standardization of the quality of virtual slides for achieving more reproducible and better performing computational approaches that will undoubtedly lead to better care of patients. In this paper, the crucial problem of detecting out-of-focus regions was addressed, but other quality related problems, such as color standardization and tissue fold-detection, should be studied to compile a comprehensive quality control pipeline for virtual slides in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank Joe Sirintrapun from pathology informatics of MSKCC for his insightful comments and all the members of the MSK community who helped in the validation of our blur detector.

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748.

References

[1]. facebook/fb.resnet.torch https://github.com/facebook/fb.resnet.torch.

[2]. Laboratory for image and video engineering - the university of texas at austin. http://live.ece.utexas.edu/research/quality/subjective.htm.

[3]. Ameisen David, Deroulers Christophe, Perrier Valerie, Bouhidel Fatiha, Battistella Maxime, Legres Luc, Janin Anne, Bertheau Philippe, and Yunes Jean-Baptiste. Automatic image quality assessment in digital pathology: from idea to implementation. In IWBBIO, pages 148–157, 2014.

[4]. Amit Yali and Geman Donald. Shape quantization and recognition with randomized trees. Neural Computation, 9:1545–1588, 1997.

[5]. Batten Christopher F.. Autofocusing and astigmatism correction in the scanning electron microscope. PhD thesis, University of Cambridge, 2000.

[6]. Breiman Leo. Random forests. Machine learning, 45(1):5–32, 2001.

[7]. Chern N. Ng Kuang, Neow Poo Aun, and Ang Marcelo H.. Practical issues in pixel-based autofocusing for machine vision. In Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on, volume 3, pages 2791–2796. IEEE, 2001.

[8]. De Kanjar and Masilamani V. Image sharpness measure for blurred images in frequency domain. Procedia Engineering, 64:149–158, 2013.

[9]. Ferzli R and Karam LJ. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). IEEE Transactions on Image Processing, 18(4):717–728, 2009. [PubMed: 19278916]

[10]. Firestone Lawrence, Cook Kitty, Culp Kevin, Talsania Neil, and Preston Kendall. Comparison of autofocus methods for automated microscopy. Cytometry, 12(3):195–206, 1991. [PubMed: 2036914]

[11]. Fuchs Thomas J. and Buhmann Joachim M.. Computational Pathology: Challenges and Promises for Tissue Analysis. Computerized Medical Imaging and Graphics, 35(7–8):515–530, 2011. [PubMed: 21481567]

[12]. Ghaznavi Farzad, Evans Andrew, Madabhushi Anant, and Feldman Michael. Digital imaging in pathology: Whole-slide imaging and beyond. Annual Review of Pathology: Mechanisms of Disease, 8(1):331–359, 2013.

[13]. Hassen R, Wang Zhou, and Salama MMA. Image sharpness assessment based on local phase coherence. IEEE Transactions on Image Processing, 22(7):2798–2810, 2013. [PubMed: 23481852]

[14]. He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[15]. Lahrmann Bernd, Valous Nektarios A., Eisenmann Urs, Wentzensen Nicolas, and Grabe Niels. Semantic focusing allows fully automated single-layer slide scanning of cervical cytology slides. PLoS ONE, 8(4):e61441, 2013.

[16]. Marichal Xavier, Ma Wei-Ying, and Zhang HongJiang. Blur determination in the compressed domain using DCT information. In Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on, volume 2, pages 386–390. IEEE, 1999.

[17]. Marziliano Pina, Dufaux Frederic, Winkler Stefan, and Ebrahimi Touradj. A no-reference perceptual blur metric. In Image Processing. 2002. Proceedings. 2002 International Conference on, volume 3, pages III–57. IEEE, 2002.

[18]. Lopez Xavier Moles, D'Andrea Etienne, Barbot Paul, Bridoux Anne-Sophie, Rorive Sandrine, Salmon Isabelle, Debeir Olivier, and Decaestecker Christine. An automated blur detection method for histological whole slide imaging. PLoS ONE, 8(12):e82710, 2013.

[19]. Moreno Pol and Calderero Felipe. Evaluation of sharpness measures and proposal of a stop criterion for reverse diffusion in the context of image deblurring. In VISAPP (1), pages 69–77, 2013.

[20]. Narvekar ND and Karam LJ. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). IEEE Transactions on Image Processing, 20(9):2678–2683, 2011. [PubMed: 21447451]

[21]. Pantanowitz Liron, Farahani Navid, and Parwani Anil. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathology and Laboratory Medicine International, page 23, 2015–06.

[22]. Satyanarayanan Mahadev, Goode Adam, Gilbert Benjamin, Harkes Jan, and Jukic Drazen. OpenSlide: A vendor-neutral software foundation for digital pathology. Journal of Pathology Informatics, 4(1):27, 2013. [PubMed: 24244884]

[23]. Sheikh HR, Sabir MF, and Bovik AC. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing, 15(11):3440–3451.

[24]. Tong Hanghang, Li Mingjing, Zhang Hongjiang, and Zhang Changshui. Blur detection for digital images using wavelet transform. In Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on, volume 1, pages 17–20. IEEE, 2004.

[25]. Wang Z, Bovik AC, Sheikh HR, and Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004. [PubMed: 15376593]

***Highlights**

- **i.** A comprehensive benchmark dataset for blur detection was created.
- **ii.** A comprehensive performance comparison of 13 sharpness metrics was obtained.
- **iii.** Feature engineering was compared to deep feature learning for blur detection.
- **iv.** A blur detection software was implemented for usage in the clinic.
- **v.** The blur detector was validated on 3 datasets, and against human experts.
- **vi.** The blur detector was tested in the clinical setting and compared it to the state-of-the-art joint QC pipeline of commercial scanner software and human QC experts.

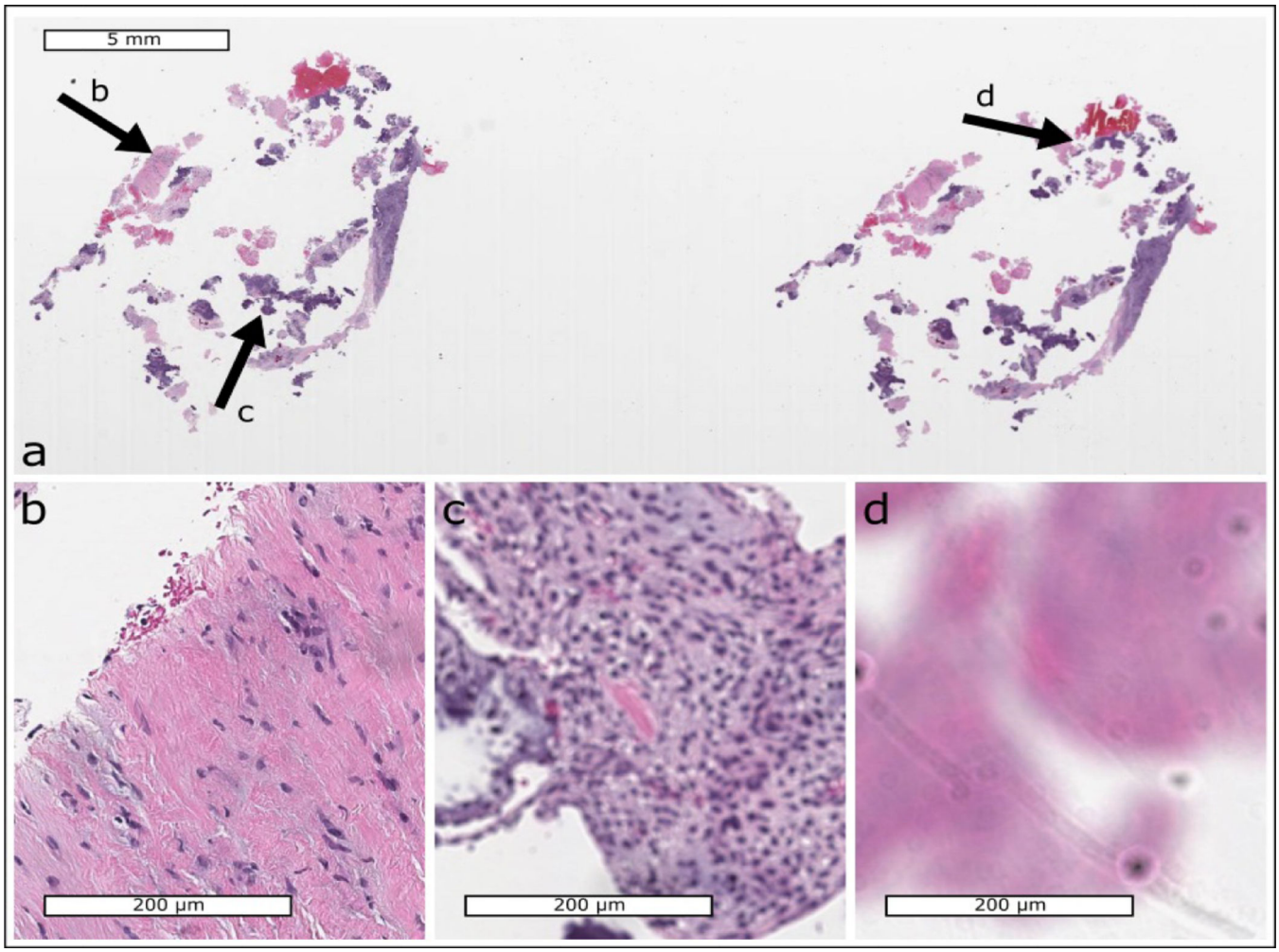

Figure 1:

Example of blur artifacts in a H&E slide of our dataset. **a:** Whole slide thumbnail. **b,c,d:** Comparison between sharp, slightly blurred, and very blurred regions at maximum magnification (20×), presented as arrows in **a**.

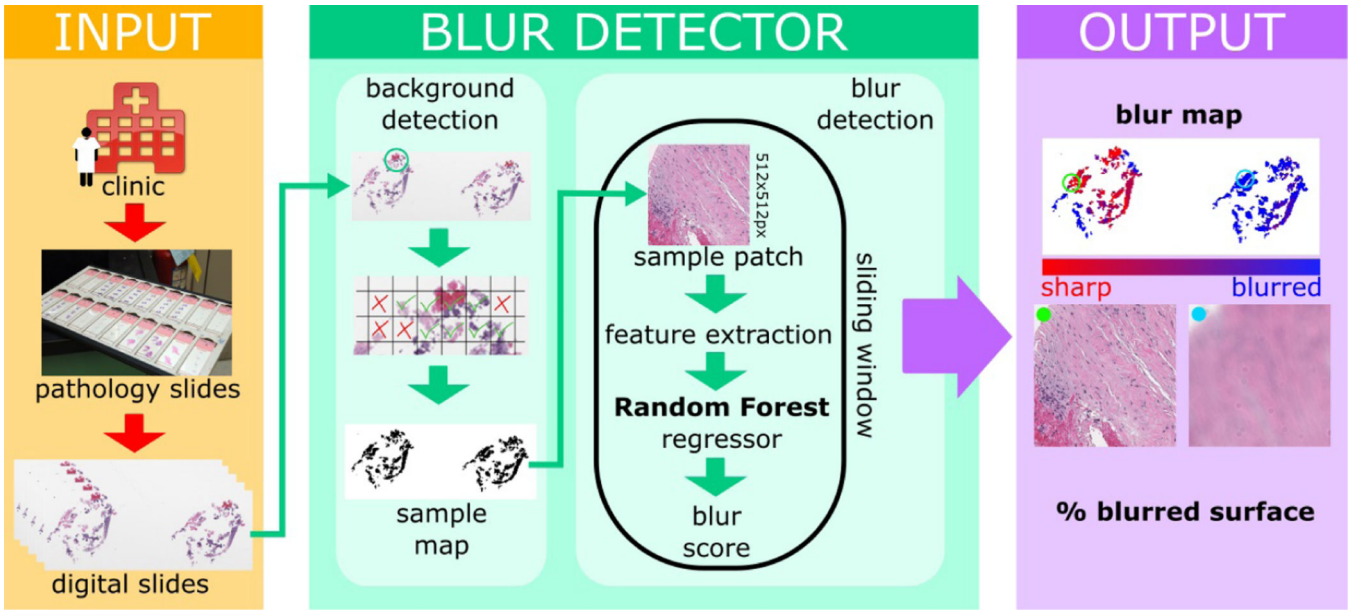

Figure 2:

Overview of our automatic blur quantification pipeline. After digitizing pathology slides (left) they are processed by our blur regression algorithm (middle): First, the tissue is separated from the white background. Then, foreground patches are extracted in a sliding window approach. Blurriness features from each patch are extracted and processed by a random forest regression algorithm assigning a blur score to the corresponding patch. The output of the system (right) is a visual representation and a quality score proportional to the amount of blurred regions present.

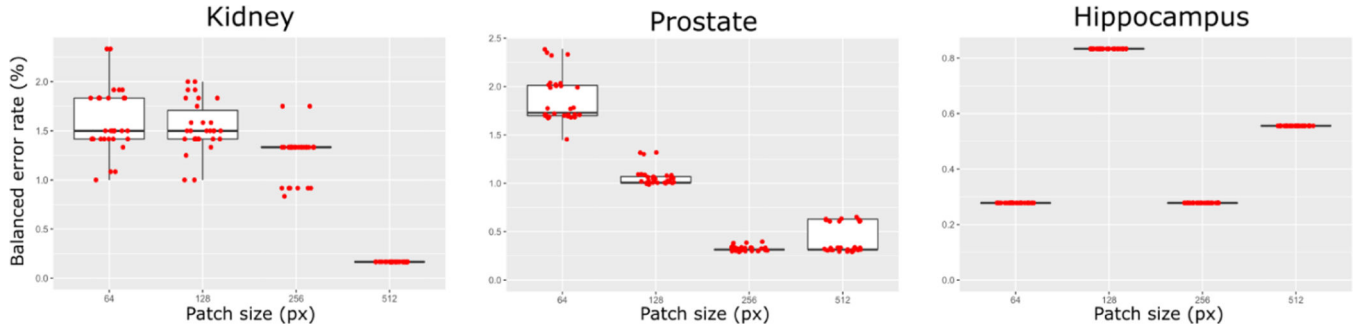

Figure 3:

Prediction error variability for the classification task trained on the tissue types separately. Hippocampus classifier showed best accuracy, followed by prostate and kidney. The relationship between patch size and classification error can be observed: bigger patches seem to have lower error rates.

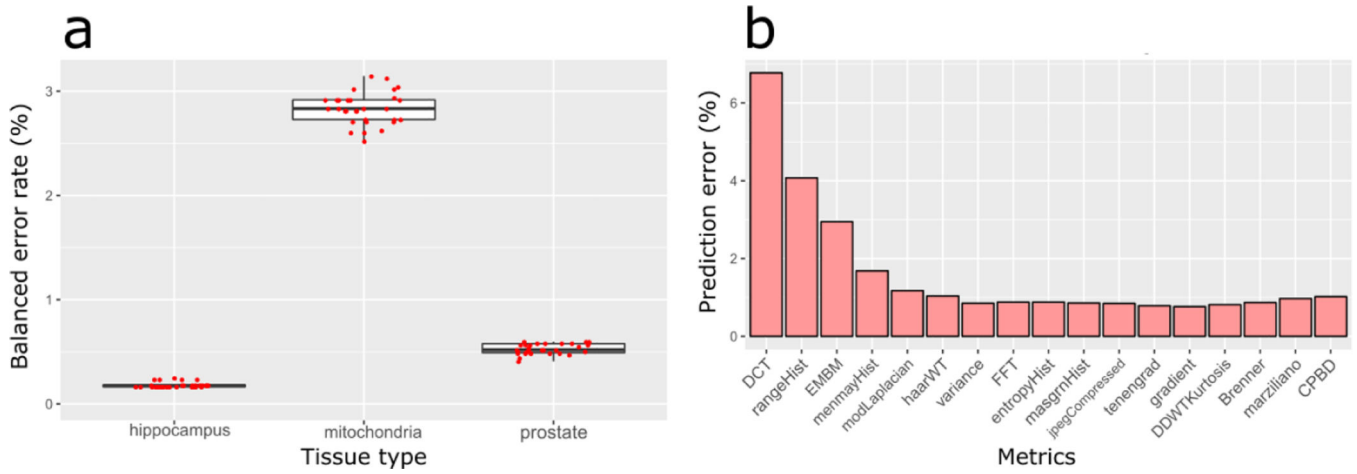

Figure 4:

Classification task performance for a classifier trained with all tissues together. **a)** Prediction error variability separated by tissue type. **b)** Feature importance analysis. From left to right the feature is successively added to the classifier which reduces the overall prediction error most (Greedy approach).

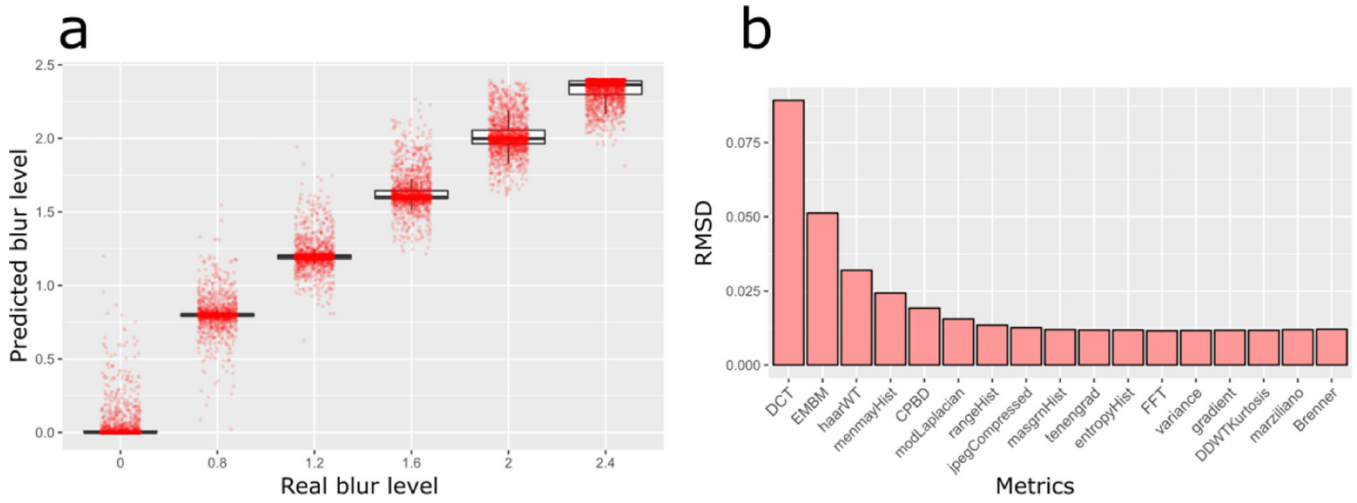

Figure 5:

Regression task performance for a classifier trained with all tissues together. **a)** Predicted blur level for each real blur level. Each red dot is a single patch. **b)** Feature importance analysis. From left to right the feature is successively added to the classifier which reduces the overall RMSD most (Greedy approach).

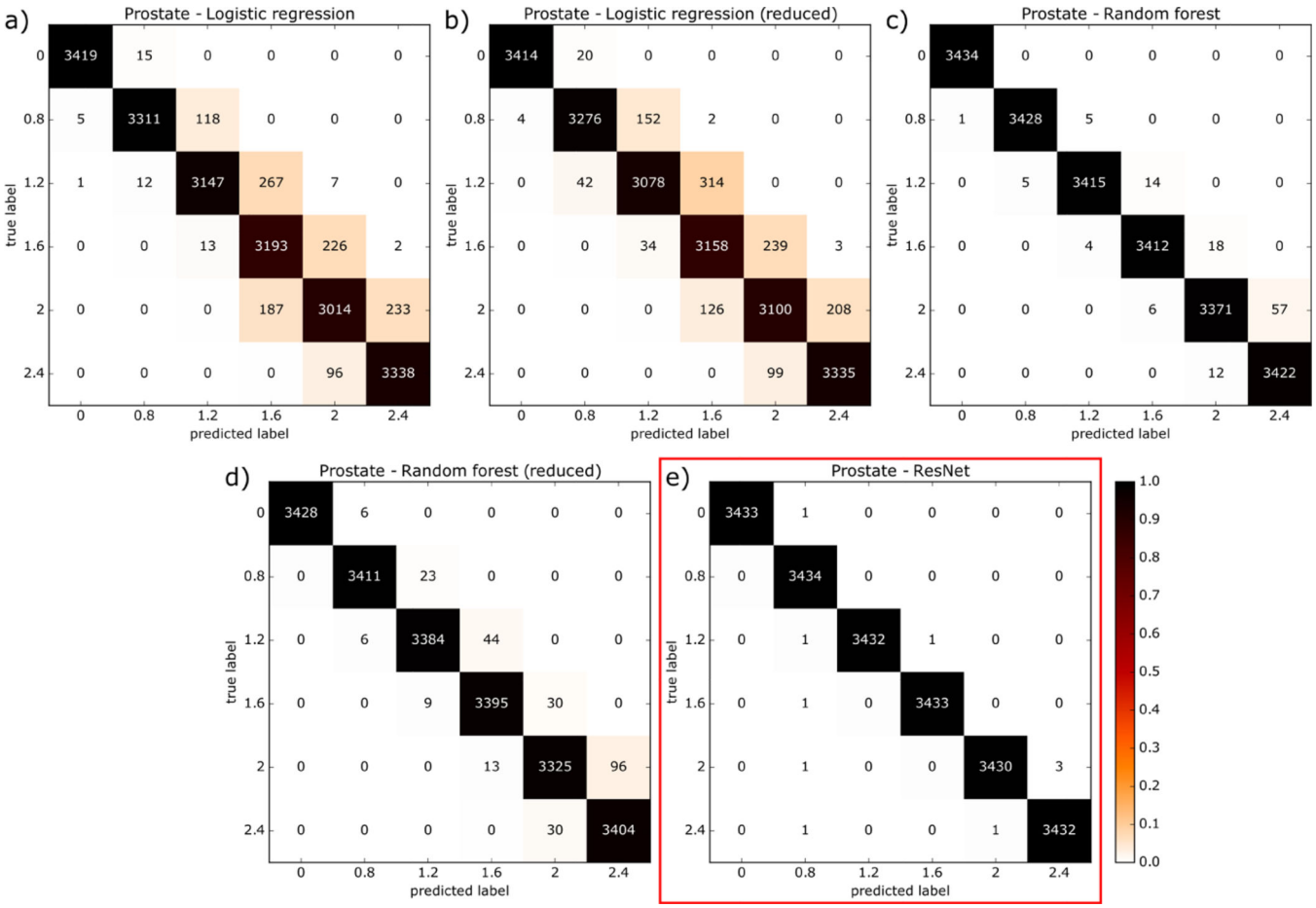

Figure 6:

Classification experiments on the prostate dataset. Comparison between models: **a)** Logistic regression on all features, **b)** Logistic regression on the reduced set, **c)** Random forest on all features, **d)** Random forest on the reduced set, **e)** 18 layer ResNet. ResNet (red) gives the best overall accuracy.

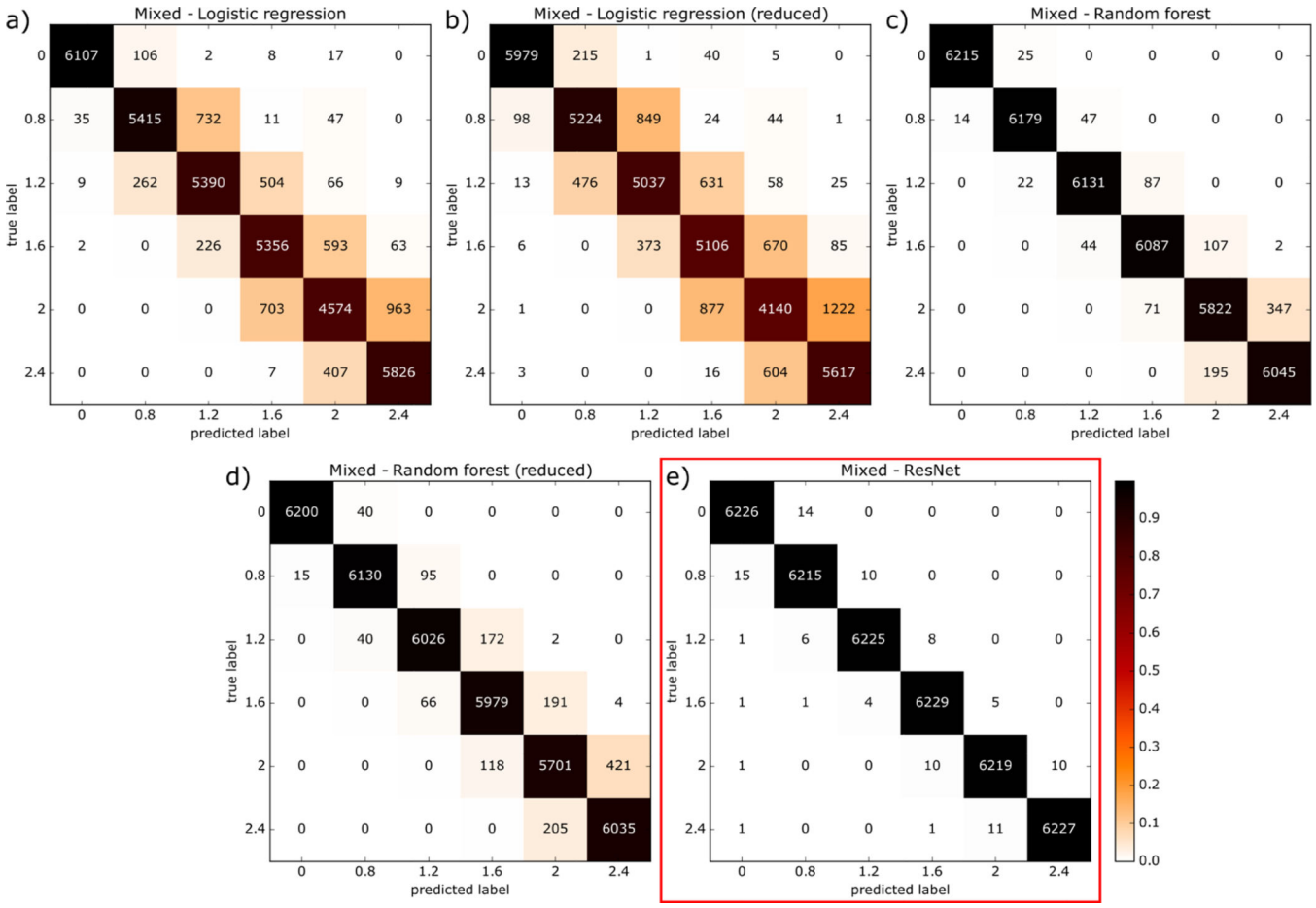

Figure 7:

Classification experiments on the mixed tissues dataset. Comparison between models: **a)** Logistic regression on all features, **b)** Logistic regression on the reduced set, **c)** Random forest on all features, **d)** Random forest on the reduced set, **e)** 18-layer ResNet. ResNet (red) gives the best overall accuracy.

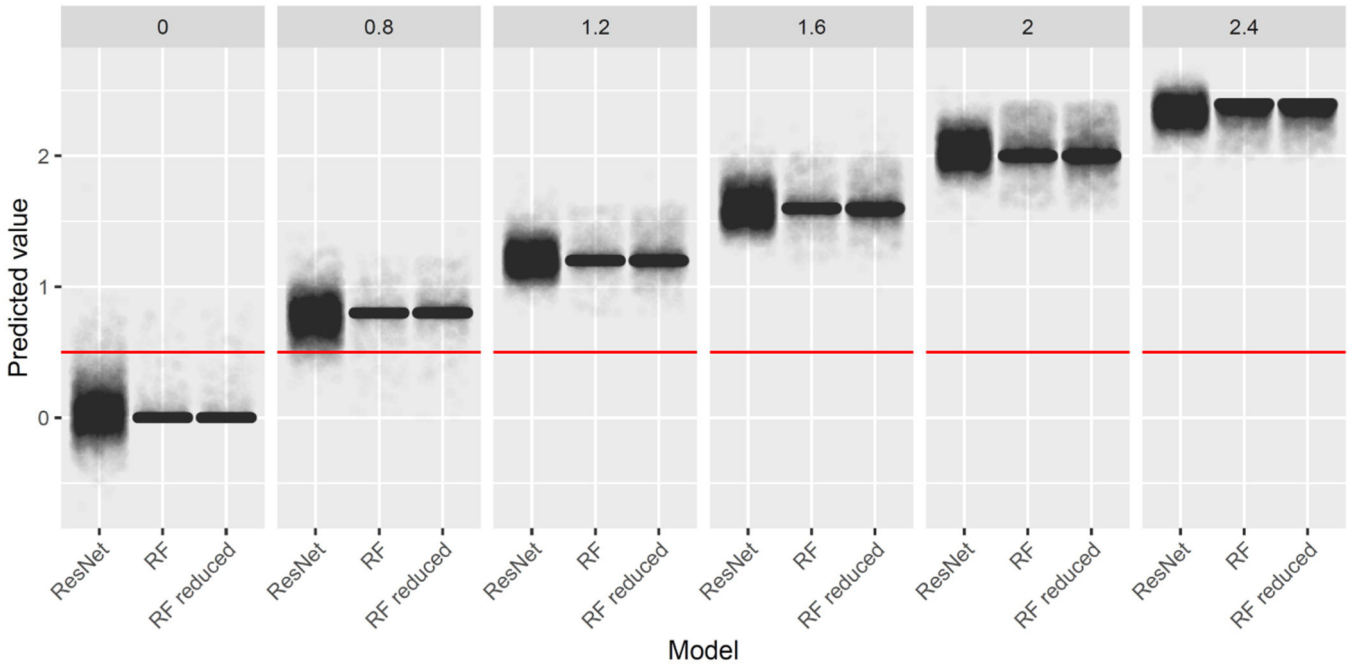

Figure 8:

Comparison of model performances for the regression task. Predicted values for each model are plotted side by side for each Gaussian σ. The distribution of predicted values of the ResNet is much wider spread than that of the random forest approaches, underlying a higher MSE. The red line indicates the decision boundary between ”sharp” and ”blurred”.

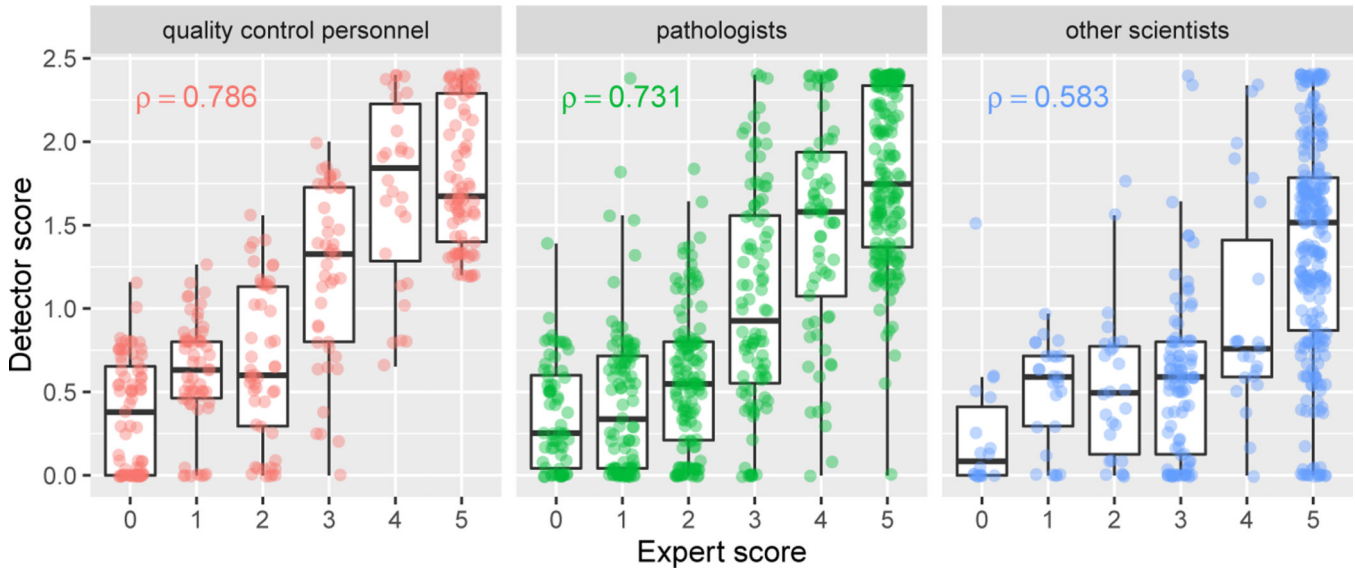

Figure 9:

Blur detector validation results: boxplot and scatter plot (artificially jittered horizontally for better visualization) showing the correlation between human score (x-axis) and detector score (y-axis), stratified by observer group. The Spearman rank correlation coefficient is highest for ”quality control personnel” ($\rho = 0.786$, left), followed by ”pathologists” ($\rho = 0.731$, middle) and ”other scientists”($\rho = 0.581$, right).

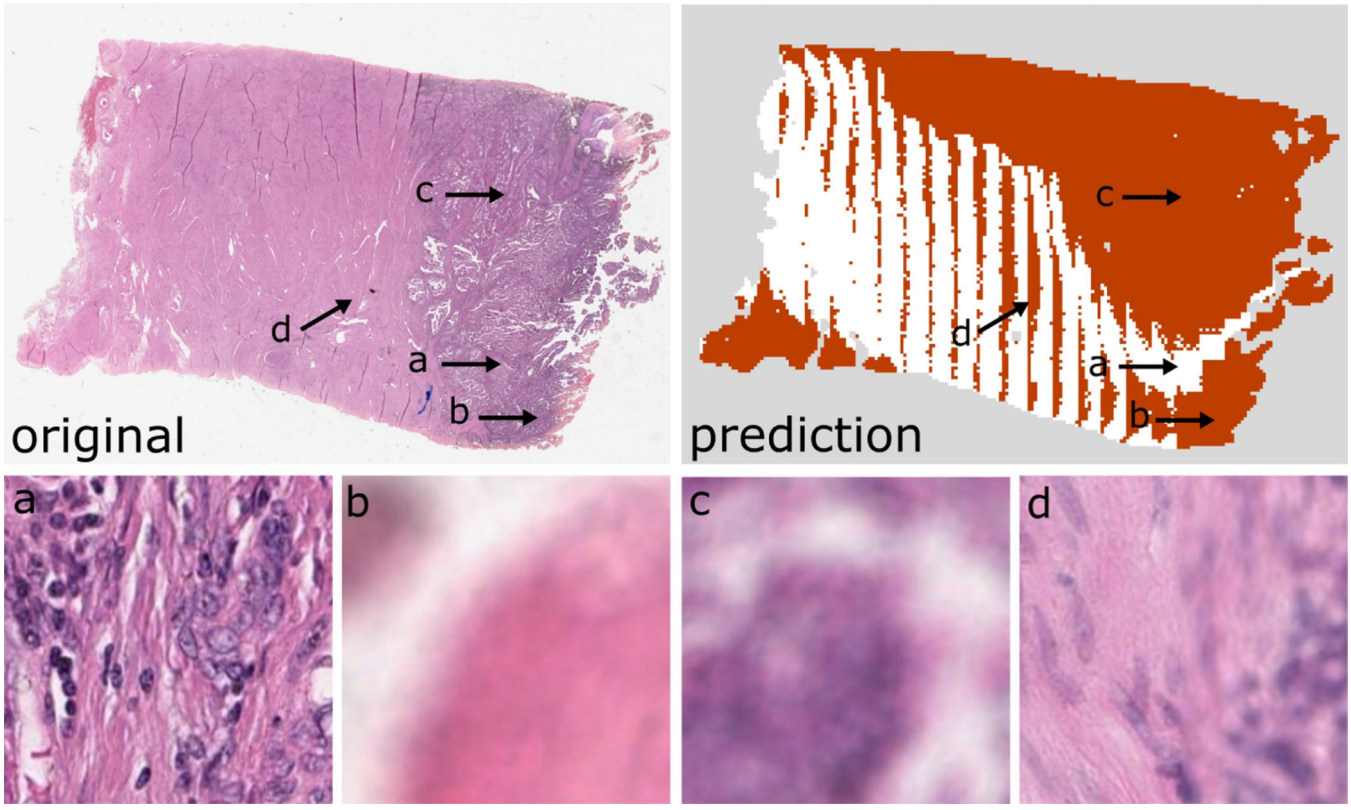

Figure 10:
An example slide that passed all quality control checkpoints in the clinic but was detected by our application. The original slide (left) is shown next to the blur prediction mask (right) where white means sharp and red indicates blurred. a-d show enlarged patches (a: sharp; b,c: blurred; d: mixed).

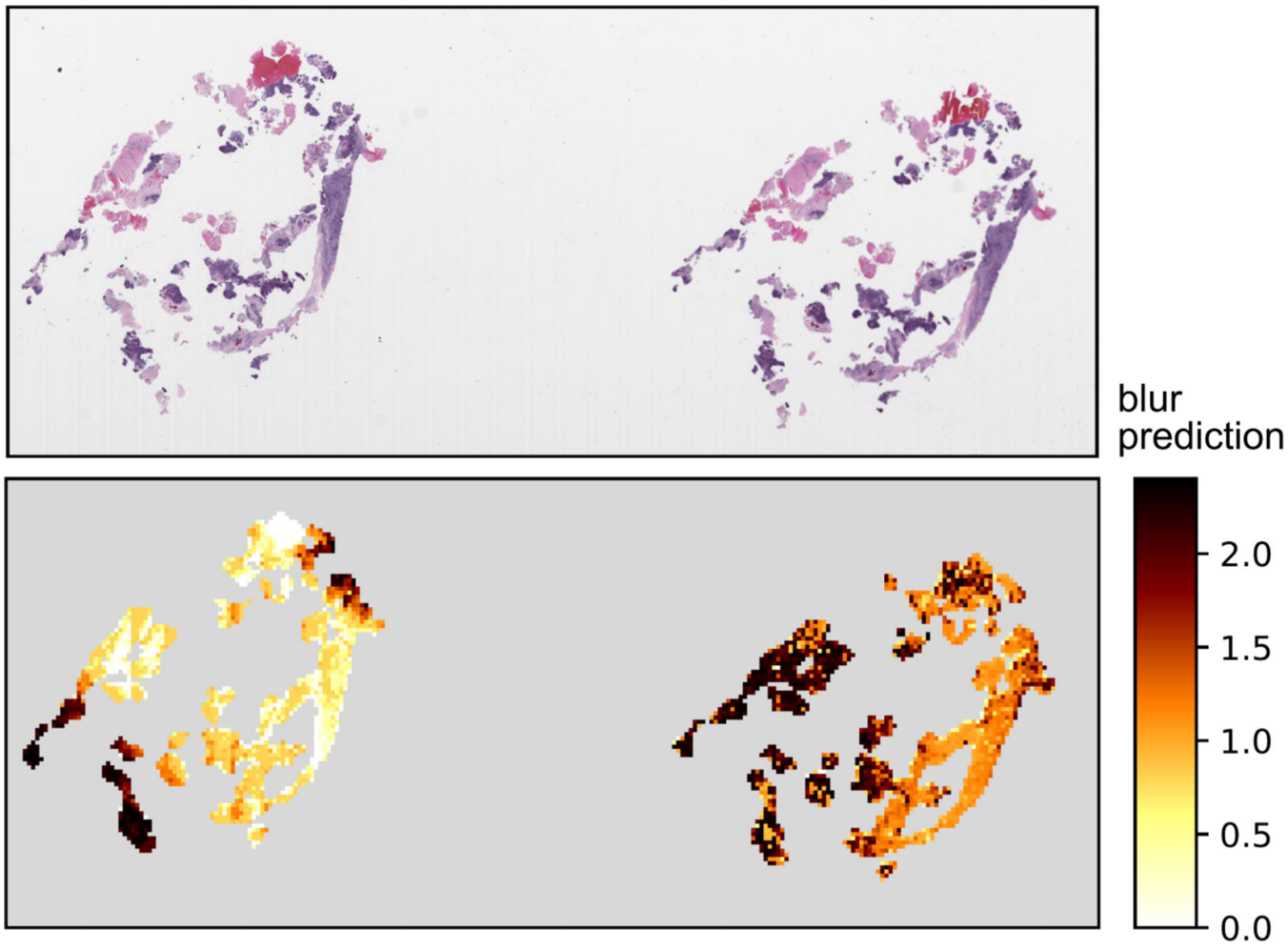

Figure 11:
Fine-grained blur prediction map. Each color on the colormap represents a blur level: white is sharp and black is very blurred.

Table 1:

Summary of patch extraction. 18 hippocampus sharp virtual slides were used. 10 patches were extracted from each slide for each of the 4 sizes. Subsequently, each sharp patch was blurred with a Gaussian filter with 5 levels of intensity yielding a total of 4240 hippocampal patches of varying size and blurriness. Similarly for prostate and kidney sample.

dataset	original sharp slides	patches per slide	patch sizes	blur levels	total # patches
kidney	30	4	4	6	2880
prostate	159	1	4	6	3816
hippocampus	18	10	4	6	4240

Table 2:

Classification accuracy of ResNet compared to different supervised classifiers on the test set. For logistic regression and random forest: *all* means training was done with all 13 metrics, *red.* means training was done on the subset of 10 metrics that maximized accuracy over extraction time.

	Logit		Random forest		
dataset	**all**	**red.**	**all**	**red.**	**resnet**
prostate	94.26%	93.97%	99.41%	98.75%	**99.95%**
mixed	87.25%	83.07%	97.43%	96.34%	**99.74%**

Table 3:

Results of the test in the clinic: comparison between current quality control (QC) procedures (right) and our blur detector (left). P is positive for blur, N is negative for blur.

Detector performance			**Current QC performance**		
	ground-truth			**ground-truth**	
detector	**P**	**N**	**current QC**	**P**	**N**
P	2+33	9	P	2	0
N	0	149	N	33	158
Error: 4.7%			Error: 17.1%		