Data labeling impact on deep learning models in digital pathology: A breast cancer case study

K. Benaggoune, Z. Al Masry, C. Devalland, S. Valmary-degano, N. Zerhouni, L.H. Mouss

Abstract Image data labeling is a vital step for deep learning models training. Studies on data labeling have not considered its impact on model performance and only focused on problems such as the curse of big data labeling or labeling tools. Furthermore, it seems clear that errors in labeling have a significant impact and should be fixed. However, in the medical domain, it is hard to ensure proper data labeling. In general, trained engineers are asked to annotate histology images, which causes errors in labeling. The aim of this study is to highlight the impact of data labeling on deep learning models. For that purpose, deep learning models are trained on two different annotations with different levels of expertise. Results show the importance of including expertise in deep learning model development. The impact of data labeling is shown through a case study on the proliferation of biomarker Ki-67 labeling index scoring.

Key words: Digital pathology, Ki-67 index scoring, Data labeling, Deep learning.

K. Benaggoune
LAP, Batna 2 University, Batna, Algeria / FEMTO-ST institute,ENSMM, Besançon, France, e-mail: k.benaggoune@univ-batna2.dz

Z. Al Masry
FEMTO-ST institute,ENSMM, Besançon, France e-mail: zeina.almasry@femto-st.fr

C. Devalland
Service D'anatomie Et Cytologie Pathologiques, Hôpital Nord Franche-Comté, Trevenans, France, e-mail: christine.devalland@hnfc.fr

S. Valmary-degano
service d'Anatomie pathologique, CHU Grenobles-Alpes, 38043 GRENOBLE CEDEX 9, e-mail: svalmarydegano@chu-grenoble.fr

N. Zerhouni
FEMTO-ST institute,ENSMM, Besançon, France, e-mail: noureddine.zerhouni@femto-st.fr

L.H Mouss
LAP, Batna 2 University, Batna, Algeria, e-mail: h.mouss@univ-batna2.dz

1 Introduction

Over the past decade, advancements in algorithms have allowed machine learning techniques to set the cutting edge in many healthcare settings. [8], [25], [27]. Digital image processing methods process high magnification images of pathology slides, initially for research use, but more and more as a clinical tool [18]. Hence, computer scientists and pathologists met to use the most recent artificial intelligence techniques to address digital pathology problems as diagnosis, prognosis, prediction, and other clinically related goals.

The primary purpose of digital pathology is to assist pathologist to improve histological interpretation and to reduce the laborious work by applying machine learning algorithms. The digital pathology process with a deep learning model consists of slide preparation, digital imaging, image post-processing, cell annotation, cell identification models, and output results. However, most of the existing digital pathology studies focus on the development of accurate models for segmentation and classification and give less attention to data labeling. Data labeling in medicine, precisely in digital pathology, is unique in different ways. As an example, for data segmentation annotation, it is easy to highlight the borders of a cat, on the other side, delineating nuclei in a tissue slide is a tough task even for experts—the same thing as well for image classification. Therefore, the pathologist expertise (PE) is a critical factor that should be regarded when developing deep learning models. The aim of this paper is to highlight the impact of expertise through a case study of breast cancer. Up to our knowledge, the PE area is not so much explored when applying deep learning tools in digital pathology. It is related directly to the label quality in elaborating the dataset, selecting the regions of interest, annotating and classifying patches, and nuclei. Hence, different levels of expertise could head to different annotations, which affect the precision of models. Therefore, incorporating pathologist expertise into the design of advanced deep learning models for digital pathology could enhance their performance.

The remainder of the paper is organized as follows. Section 2 provides the background for the biomarker Ki-67 scoring. In Section 3 the framework of the proposed approach is described. Then the experimental results are presented in Section 4. Finally, a discussion and conclusion are drawn in Section 5.

2 Deep learning for Ki-67 scoring

The ki-67 labeling index (LI) is a reliable tumor proliferation marker. Its scoring has many roles in breast cancer and other cancers [16], both in standard clinical practice as a prognostic [11], and a predictive marker [6]. However, the usage of the Ki-67 LI in daily clinical practice is no easy task. The interpretation method is still a concern, with the manual estimation of Ki-67 being subjective, error-prone, and dependent on intra- and inter-observer uncertainties [12]. From now on, the automated scoring evaluation of Ki-67 LI is strongly recommended. Automated scoring will provide an increased flow and more consistent results than manual scoring. Ki-67 LI is computed as the ratio of the number of immunopositive nuclei to the total number of nuclei present in a region of interest saha2017advanced.

Classical machine learning techniques are widely used for Ki-67 LI scoring. In [1], a computer vision algorithm for Ki-67 scoring has been proposed in breast cancer tissue images. The proposed approach shows better performance compared to other techniques. In [21], smoothing decomposition and feature extraction are used with k-means clustering for Ki-67 quantification with 91.8 % segmentation accuracy. An automatic algorithm for selecting hotspots from the set of slide images is proposed in [22]. Color channel selection, feature selection, Otsu thresholding, and classification were used in this work.

However, the performance of these conventional techniques tends to depend on many thresholds that can be tricky to tune for users such as clinicians. Further, image problems in digital pathology, associated with tissue cuts on/or folds, uneven color cast, unspecific coloration, and varying intensity in background structures, misguide the image analysis. Recently, powerful deep learning techniques have been suggested to address these problems. In [20], the authors have proposed a deep learning model for automatic recognition of candidate hotspots. This work uses the Gamma mixture model for nuclei detection and patch selection together with deep learning for proliferation scoring. In [26], a combination of two models was used in the proposed method: the Single Shot Multibox Detector for nuclei detection and a Convolutional Neural Network for image classification. The proposed approach obtained 98% for classification and 90 % for segmentation, where Ki-67 quantification results are not reported. The authors of [15] used the MobileUnet model for nuclei segmentation and classification, and the connected component-based algorithm for Ki-67 index estimation. The results yield an average F-score of 0.92, a die score of 0.96, and it has a mean absolute error in the scored Ki-67 index of 2.1.

Accordingly, the use of deep learning models for Ki-67 LI scoring is highly dependent on the quality of data labeling. Therefore, one should take into consideration this problem in order to improve the Ki-67 LI scoring based on these techniques.

3 Methodology

As mentioned before, the impact of the PE in deep learning models development is here addressed. Since the annotation in the medical field presents different levels of uncertainties, the impact of different annotations is studied on the performance of deep learning models for Ki-67 LI scoring. For that purpose, an expert pathologist and a biomedical engineer (BE) annotated the same dataset for segmentation and classification tasks. Two deep learning models for segmentation and classification are then trained and evaluated on a third test set annotated by different pathologists for Ki-67 LI scoring (considered as a reference set). In Figure 1, the proposed methodology is exposed.

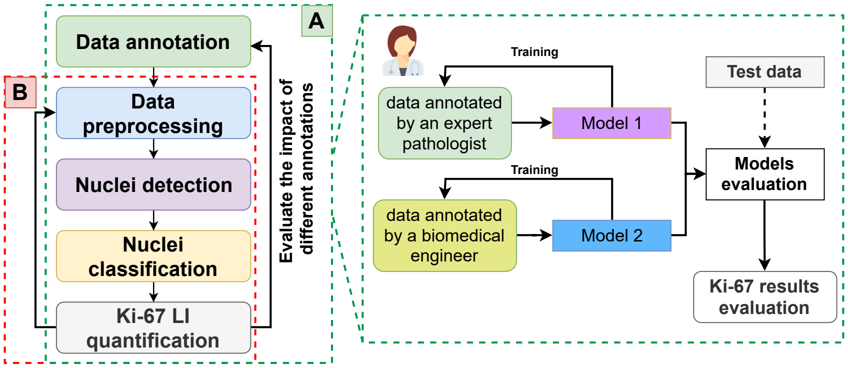

Fig. 1 Proposed approach: evaluate the impact of data annotation on Ki-67 LI scoring.

The first step consists of data preparation and annotation, then data preprocessing is applied before feeding the models and quantifying the Ki-67 LI scoring. The details of each step are explained as follows:

Data annotation:

The dataset collected from HNFC hospital [1] consists of breast cancer proliferation immunostaining DAB slides, stained and captured at 40x magnification with 40 slides scanned on a Hamamatsu scanner. Three sub-images are cropped of size 256x256 pixels from three regions of interest hot-spot, edges, and medium selected from each patient to guarantee data variation. The dataset is divided based on patients into 80%, 20% for train and test subsets, respectively. An expert pathologist and trained BE annotated train images separately. The test subset is annotated by the second expert pathologist considered as reference. The open-source software Qupath is used for annotation using the Brush tool on 25” monitor [2]. Annotators are asked to delineate all nuclei’s boundaries and indicate each nucleus class as positive or negative. Finally, three datasets are generated: (i) expert dataset for training, (ii) BE dataset for training, (iii) reference dataset for testing. The first two datasets are used to train the proposed segmentation and classification models. Figure 2 shows the difference between an expert and a BE annotation.

Data preprocessing:

Stain coloring and reactivity generate color variation in histopathology images. This variation could unfavorably affect the training of machine learning models. In this case, data needs to be preprocessed, therfore two preprocessing methods are adopted

[1] Hospital of Nord Franche-Comté in France.

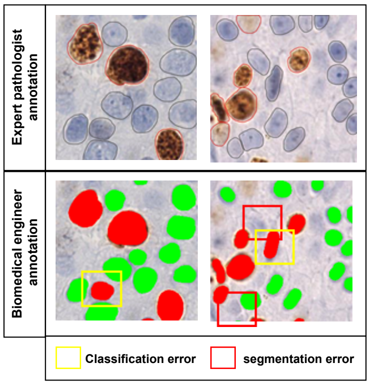

Fig. 2 Data annotation difference between an expert pathologist and a BE.

in this work: (i) color normalization with Structure preserving color normalization technique [23]. This method uses a simple statistical processing to move the color features from one image to another. (ii) Data augmentation is an essential part of the process of digital pathology since there is limited data to train the model for discriminative information learning. Albumentations library [5] is used to perform augmentations such as random rotation, flip, transpose, Gaussian noise, blur, optical distortion, grid distortion, and elastic transformation.

Data segmentation:

Two steps are used to perform an instance segmentation of nuclei. Firstly, the Unet algorithm is used for semantic segmentation. Unet is able to work with very few training images and produce more accurate segments. [19]. The general architecture of Unet (Figure 3) is made up of two paths, a contracting path on the left side and an expanding path on the right side. In this work, the contracting path is replaced by a Squeeze Excitation ResNet backbone ([9], [10]). The SE-ResNet makes use of the concept of residual mapping commonly used in computer vision. It is applied to create a basic residual learning block. Instead of using a reference layer to directly learn the correspondence between inputs and outputs as in a typical CNN, it uses some reference layers to learn the residual representation between input and output.

Second, the generated probability map from SE-ResNet Unet requires post-processing to separate nuclei. Thereby, the following steps for post-processing are respected: (1) Apply adaptive threshold on the probability map; (2) Fill small holes; (3) Apply binary opening;(4) Remove small objects. Once the binary mask is processed, a distance tranform watershed is applied for touched nuclei split [24].

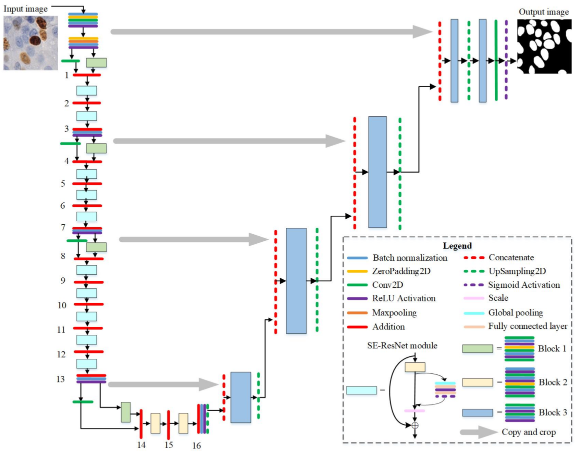

Fig. 3 Unet with Squeez-Excitation Resnet50 backbone for nuclei segmentation.

Data classification:

Complex textures and structural and morphological diversity make image classification of histopathological slides a difficult task. Recently, CNN-based approaches have achieved some success in image classification problems, of which medical image analysis is included [3]. CNNs explore the ability to learn features from data directly, sidestepping hand-crafted features [4]. Therefore, CNN is used to extract the characteristics of cell patches segmented in the previous step automatically and take full advantage of them for classification.

Transfer learning involves the application of pre-trained CNNs on large annotated image databases to images from different domains. Pre-trained CNNs may be further fine-tuned on medical image datasets, enabling a faster convergence of large networks and an ability to learn domain features. In this study, the well-known pre-trained CNN architectures on the ImageNet dataset, namely the deep residual convolutional network (ResNet50), is used for patch classification.

4 Experimental setup

All algorithms are developed in Python with Keras library and TensorFlow backend and trained with a Tesla K80 GPU free unit available in Google Colab. For model training, in both cases, BE and the expert pathologist, the segmentation and classification models are trained on the same training set. Then, the reference test set is used for models evaluation.

4.1 Segmentation and classification results

For the segmentation task, the model was initialized with SE-ResNet50 pre-trained backbone with ImageNet. Two categories of evaluation metrics are used to evaluate the performance of the model: (i) pixel level metrics such as accuracy (acc), mean intersection over union (MIU), and the frequency weighted MIU (FIU) [17]. (ii) object level metrics such as Dice2, Aggregated Jaccard Index (AJI), and Panoptic quality (Panoptic Q) ([14], [13]). Note that Dice2, which is an ensemble dice, measures the separation of all nuclei from the background, AJI captures the quality of the segmentation and Panoptic Q is a unified comparison score that sets the standard for measuring the performance of nuclear instance segmentation methods.

Trained models are evaluated on the reference test set, which is considered as a ground truth. Qualitative evaluation is depicted in Figure 4. As the figure shows, many false positive and false negative detection are noticed from the model trained with the BE compared to the model of the expert pathologist. These results show that the impact of the labeling on the final results of the detection is significant, which is clear because deep neural networks are strictly dependent on the quality of the labeled data. As well, the quantitative evaluation shows that models are close based on pixel-level metrics and the superiority of the pathologist model in terms of object-level metrics, which is most meaningful for nuclei counting (see Table 1).

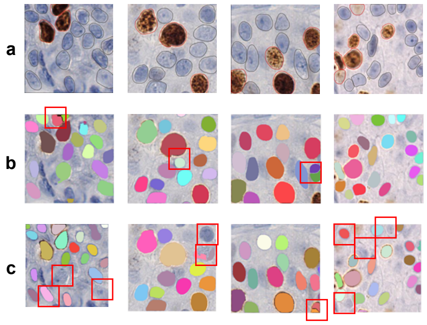

Fig. 4 Segmentation results: a- Ground truth, b- model trained on a dataset annotated by an expert. c- model trained on a dataset annotated by a BE. Red square indicates false positive and false negative detection made by the model.

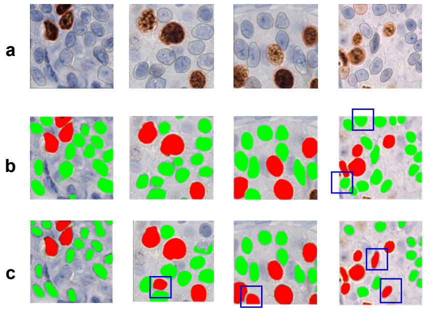

Fig. 5 Classification results: a- Ground truth, b- model trained on a dataset annotated by an expert. c- model trained on a dataset annotated by a BE. Blue square indicates false positive and false negative classifications.

For patches classification, the pre-trained ResNet50 is used to associate nuclei to one of the two classes immunopositive or immunonegative nuclei. The pathologist model gives an accuracy of 92 % higher than the BE model with 80 % accuracy. Qualitative test results are exposed in Figure 5 and quantitative results in Table 1.

Table 1 Segmentation and classification results.

Segmentation results							Classification results		
	Pixel-level metrics			Object level metrics					
Annotation	Acc	MIU	FMIU	Dice2	AJI	Panoptic Q	Acc	Precision	Recall
Pathologist	**0.93**	**0.84**	**0.87**	**0.79**	**0.71**	**0.64**	**0.92**	**0.94**	**0.95**
BE	0.91	0.79	0.83	0.65	0.61	0.48	0.80	0.83	0.88

4.2 Ki67 LI results

Ki67 labeling index (Ki67 LI) or proliferation rate is the percentage of stained nuclei to the total number of malignant nuclei counted [7]. It should be mentioned that the reference test set is annotated by different expert pathologists and eight patients with three images extracted from high, medium, and low proliferation regions for each patient.

For each model, the associated Bland-Altman plots of the difference between the automated and manual Ki67 LI are drawn against the mean of the two measurements (Figures 6 and 7). Additionally, a scatter plot of the Ki67 LI fit between the manual and automated methods for each model is provided. A high bias for the BE model can be explained by the model's error in accurately classifying the nuclei. In Figure 7, the R2 and Spearman correlation were calculated, for which the expert pathologist model had the highest Spearman correlation coefficient of 94%, and R2 of 88%. This shows a strong increasing monotonic relationship between the automatic and manually counted Ki67 LI by the pathologist, which is better than the BE model.

5 Discussion and conclusion

AI in medicine is rapidly going from research to application. This transition requires high verification of the credibility of AI algorithms. Data annotation in the medical field is entirely distinct and inflexible from other conventional domains. Hence, trained engineers are asked to annotate histology slides, causing errors and less generalization of trained models. Therefore, in this study, the impact of the pathologist expertise integration in the process of building deep learning models was highlighted. A pathologist and trained BE annotated the data to train two deep learning models for segmentation and classification tasks. The models are then used to estimate the Ki67 LI in a new test set annotated by another expert pathologist. Results showed the superiority of the expert pathologist model with a Pearson correlation coefficient of 94% and R2 = 88%. Hence, the BE model had difficulties defining the staining threshold to separate nuclei from the background or separate the two classes. These difficulties are directly related to the errors made at the annotation phase. Besides, pathologists also struggle in using annotation software, which induces small errors in data annotation.

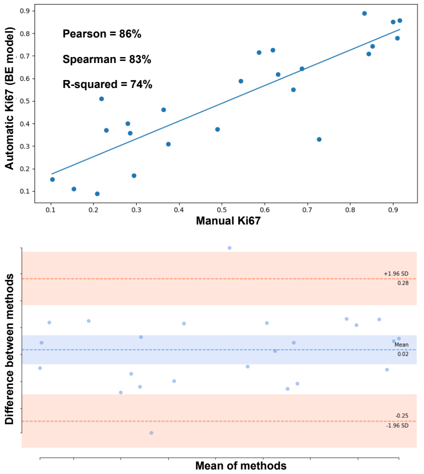

Fig. 6 Ki67 evaluation results of BE model. (i) Top: A scatter plot of the agreement in PI between manual and automated approaches. (ii) Bottom: The corresponding Bland-Altman plots based on the difference in Ki67 LI between automated and manual approaches

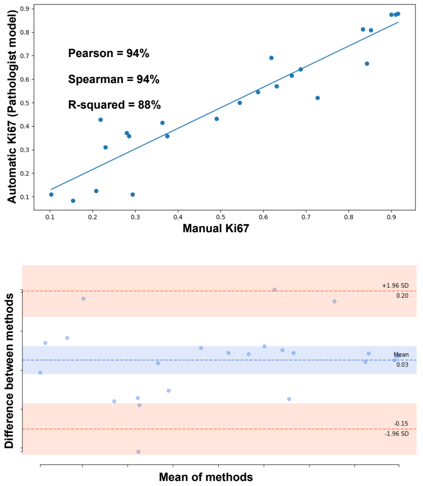

Fig. 7 Ki67 evaluation results of expert pathologist model. (i) Top: A scatter plot of the agreement in PI between manual and automated approaches. (ii) Bottom: The corresponding Bland-Altman plots based on the difference in Ki67 LI between automated and manual approaches

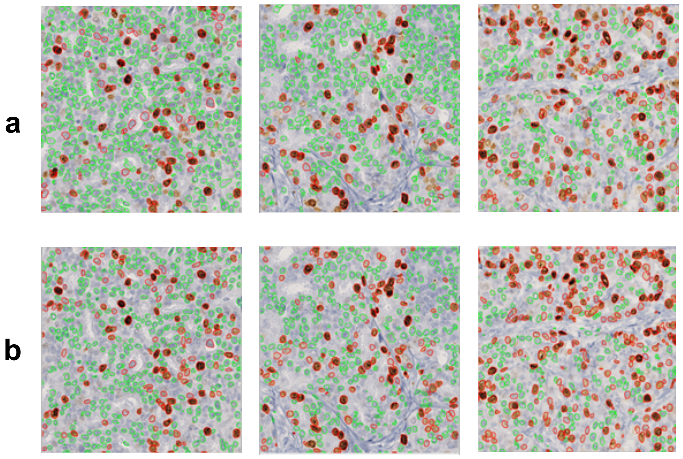

Fig. 8 Results on three different regions: a- model trained on a dataset annotated by an expert. b- model trained on a dataset annotated by a BE.

Data annotation is highly dependent on how images are visualized and the flexibility of annotation tools. Future studies will explore the impact of data normalization and subjective visualization on different expertise for Ki67 LI scoring. Also, the impact of expertise will be evaluated and compared based on different types of deep neural networks for both segmentation and classification tasks. Further, from our point of view, using healthcare data is not enough. To sum up, it is highly clear that the advancement of AI in healthcare is dependent on the integration of AI and clinician's expertise.

References

1. Abubakar, M., Howat, W.J., Daley, F., Zabaglo, L., McDuffus, L.A., Blows, F., Coulson, P., Raza Ali, H., Benitez, J., Milne, R., et al.: High-throughput automated scoring of ki67 in breast cancer tissue microarrays from the breast cancer association consortium. The Journal of Pathology: Clinical Research **2**(3), 138–153 (2016)
2. Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., et al.: Qupath: Open source software for digital pathology image analysis. Scientific reports **7**(1), 1–7 (2017)
3. Bayramoglu, N., Kannala, J., Heikkilä, J.: Deep learning for magnification independent breast cancer histopathology image classification. In: 2016 23rd International conference on pattern recognition (ICPR), pp. 2440–2445. IEEE (2016)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
5. Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. arXiv preprint arXiv:1809.06839 (2018)
6. Criscitiello, C., Disalvatore, D., De Laurentiis, M., Gelao, L., Fumagalli, L., Locatelli, M., Bagnardi, V., Rotmensz, N., Esposito, A., Minchella, I., et al.: High ki-67 score is indicative of a greater benefit from adjuvant chemotherapy when added to endocrine therapy in luminal b her2 negative and node-positive breast cancer. The Breast **23**(1), 69–75 (2014)
7. Dowsett, M., Nielsen, T.O., A'Hern, R., Bartlett, J., Coombes, R.C., Cuzick, J., Ellis, M., Henry, N.L., Hugh, J.C., Lively, T., et al.: Assessment of ki67 in breast cancer: recommendations from the international ki67 in breast cancer working group. Journal of the National cancer Institute **103**(22), 1656–1664 (2011)
8. Hamet, P., Tremblay, J.: Artificial intelligence in medicine. Metabolism **69**, S36–S40 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141 (2018)
11. Inwald, E., Klinkhammer-Schalke, M., Hofstädter, F., Zeman, F., Koller, M., Gerstenhauer, M., Ortmann, O.: Ki-67 is a prognostic parameter in breast cancer patients: results of a large population-based cohort of a cancer registry. Breast cancer research and treatment **139**(2), 539–552 (2013)
12. Jang, M.H., Kim, H.J., Chung, Y.R., Lee, Y., Park, S.Y.: A comparison of ki-67 counting methods in luminal breast cancer: the average method vs. the hot spot method. PLoS One **12**(2) (2017)
13. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9404–9413 (2019)
14. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal **16**, 34–42 (2018)
15. Lakshmi, S., Vijayasenan, D., Sumam, D.S., Sreeram, S., Suresh, P.K.: An integrated deep learning approach towards automatic evaluation of ki-67 labeling index. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp. 2310–2314. IEEE (2019)
16. Lei, Y., Li, Z., Qi, L., Tong, S., Li, B., He, W., Chen, M.: The prognostic role of ki-67/mib-1 in upper urinary-tract urothelial carcinomas: a systematic review and meta-analysis. Journal of endourology **29**(11), 1302–1308 (2015)

17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015)
18. Ma, J., Shang, P., Lu, C., Meraghni, S., Benaggoune, K., Zuluaga, J., Zerhouni, N., Devalland, C., Al Masry, Z.: A portable breast cancer detection system based on smartphone with infrared camera. Vibroengineering PROCEDIA **26**, 57–63 (2019)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015)
20. Saha, M., Chakraborty, C., Arun, I., Ahmed, R., Chatterjee, S.: An advanced deep learning approach for ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. Scientific reports **7**(1), 1–14 (2017)
21. Shi, P., Zhong, J., Hong, J., Huang, R., Wang, K., Chen, Y.: Automated ki-67 quantification of immunohistochemical staining image of human nasopharyngeal carcinoma xenografts. Scientific reports **6**, 32127 (2016)
22. Swiderska, Z., Korzynska, A., Markiewicz, T., Lorent, M., Zak, J., Wesolowska, A., Roszkowiak, L., Slodkowska, J., Grala, B.: Comparison of the manual, semiautomatic, and automatic selection and leveling of hot spots in whole slide images for ki-67 quantification in meningiomas. Analytical cellular pathology **2015** (2015)
23. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE transactions on medical imaging **35**(8), 1962–1971 (2016)
24. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis & Machine Intelligence (6), 583–598 (1991)
25. Zemouri, R., Omri, N., Morello, B., Devalland, C., Arnould, L., Zerhouni, N., Fnaiech, F.: Constructive deep neural network for breast cancer diagnosis. IFAC-PapersOnLine **51**(27), 98–103 (2018)
26. Zhang, R., Yang, J., Chen, C.: Tumor cell identification in ki-67 images on deep learning. Mol Cell Biomech **15**(3), 177–187 (2018)
27. Zuluaga-Gomez, J., Al Masry, Z., Benaggoune, K., Meraghni, S., Zerhouni, N.: A cnn-based methodology for breast cancer diagnosis using thermal images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization pp. 1–15 (2020)