applied sciences

MDPI

Article

Detection of Ki67 Hot-Spots of Invasive Breast Cancer Based on Convolutional Neural Networks Applied to Mutual Information of H&E and Ki67 Whole Slide Images †

Zaneta Swiderska-Chadaj [1], Jaime Gallego [2] and Lucia Gonzalez-Lopez [3] and Gloria Bueno [2,*]

1 Faculty of Electrical Engineering, Warsaw University of Technology, 00-662 Warsaw, Poland; zaneta.swiderska@pw.edu.pl
2 VISILAB, University of Castilla-La Mancha, Av. Camilo José Cela s/n, 13071 Ciudad Real, Spain; jaime.gallego@uclm.es
3 Hospital General Universitario de Ciudad Real, 13005 Ciudad Real, Spain; lgonzalezl@sescam.jccm.es
* Correspondence: gloria.bueno@uclm.es
† The work was conducted during a secondment at VISILAB, Spain.

Received: 1 September 2020; Accepted: 27 October 2020; Published: 2 November 2020



Abstract: Ki67 hot-spot detection and its evaluation in invasive breast cancer regions play a significant role in routine medical practice. The quantification of cellular proliferation assessed by Ki67 immunohistochemistry is an established prognostic and predictive biomarker that determines the choice of therapeutic protocols. In this paper, we present three deep learning-based approaches to automatically detect and quantify Ki67 hot-spot areas by means of the Ki67 labeling index. To this end, a dataset composed of 100 whole slide images (WSIs) belonging to 50 breast cancer cases (Ki67 and H&E WSI pairs) was used. Three methods based on CNN classification were proposed and compared to create the tumor proliferation map. The best results were obtained by applying the CNN to the mutual information acquired from the color deconvolution of both the Ki67 marker and the H&E WSIs. The overall accuracy of this approach was 95%. The agreement between the automatic Ki67 scoring and the manual analysis is promising with a Spearman's ρ correlation of 0.92. The results illustrate the suitability of this CNN-based approach for detecting hot-spots areas of invasive breast cancer in WSI.

Keywords: Ki67 hot-spot detection; Ki67 labeling index; proliferation index; CNNs; color deconvolution; WSI; digital pathology

1. Introduction

The revolution of deep learning (DL) techniques, visible in many areas [1–3], has also reached medical diagnosis [4,5]. In histopathology, the advances in digital pathology, image processing, and artificial intelligence techniques can be applied to cancer diagnosis [6,7].

The quantification of cellular proliferation is one of the metrics used to study cell activity in oncology. In particular, the proliferative activity assessed by Ki67 immunohistochemistry (IHC) is an established prognostic and predictive biomarker that determines the choice of therapeutic protocols applied to breast cancer [8,9]. Tumor regions that exhibit high proliferating activity are called hot-spots. Several standardization issues for both the planning process and the related quantification protocols must be solved prior to the implementation of quantification in regular practice [9–12]. S. Robertson et al. [12] showed significant variations in Ki67 hot-spot scoring depending on the number

of included tumor cells and the hot-spot's size, shape and location. There is a need to accurately measure the proportion of the Ki67 labeling index (Ki67 LI) [13–15], and digital image analysis methods may help with this task [16].

Several methods to automatically quantify the Ki67 index or detect hot-spots have been reported in the literature (see Table 1). Extant approaches are based on two main strategies: (I) machine learning with clustering, adaptive thresholding, and edge-based segmentation [9,17–25] or (II) deep learning [26–29]. Moreover, two main goals can be distinguished—(I) labeling index (LI) estimation or (II) hot-spot detection. The KI67 LI and hot-spots have been detected for tumors occurring in various organs including the breasts, brain, adrenal glands, and pancreas. Due to the large size of a single WSI (2-4GB), in many research papers, the results are presented only for tissue micro arrays (TMAs) or small parts of tissue samples (sub-images).

Table 1 shows the success in producing outcomes and implementing more advanced approaches. The drawback of methods based on classical machine learning is that they are parameter-dependent and cannot be easily tuned for non-homogeneous staining, which can vary in a range of hot-spot density values. Moreover, the methods described in the literature do not distinguish tumors from stromal tissues and, therefore, do not exclude nonepithelial cells that potentially express the biomarker—that is, proliferating Ki67-positive lymphocytes. Only cells that express cytokeratin should be eligible for the positive or negative detection of the biomarker concerned. Another aspect not considered in works focused on TMA or sub-images is the artifacts that occur on WSIs and how to process the WS to obtain the final evaluation.

In this paper, we propose and compare three methods based on Deep Learning to detect and characterize Ki67 hot-spots in breast cancer WSIs. The proposed techniques solve the problem of handcrafted feature selection by using convolutional neural networks (CNNs) to automatically determine the best features to characterize the hot-spot regions. In the present work, we explore the possibility of using DL methods for extracting and combining information from KI67 and hematoxylin and eosin (H&E) stained slides.

Three methods have been developed, as mentioned above. Method 1 is the simplest one, where a single CNN is used to process samples extracted from KI67-stained slides. Method 2 uses both Ki67 proliferation and tumor localization obtained by H&E analysis, where KI67 and consecutive H&E slides are processed by two independent CNNs. Method 3 is the most advanced one, where we combine Ki67 and H&E WSIs by means of color deconvolution to use the mutual information of both biomarkers in one CNN. This method achieved the best performance, with average accuracy of 95%, sensitivity of 77% and specificity of 98%.

Table 1. Review of reference methods, where:Acc-accuracy, F1-F1-score, DL-deep learning, * after augmentation, det-detection, clas-classification, LI-labeling index, PPV-positive predictive value and rho-Spearman's rank correlation coefficient.

Type	Authors	Cancer Origin	Goal	Test Set	Results
Classical Machine Learning	Lopez et al. [19]	brain	hot-spot detection	4 WSI	Acc = 93.8%
	Niazi et al. [18]	digestive system	hot-spot and cell detection	55 images	Acc = 94.60%
	Lu et al. [17]	adrenal glands	hot-spot detection	50 WSI	rho $>$ 0.8.
	Xing et al. [20]	gastro-intestinal tract, pancreas	cell detection	46 cases	F1 = 90%
	Swiderska et al. [21]	brain	hot-spot detection and LI estimation	104 WSI	rho >0.95
	Laurinavicius et al. [9]	breast	hot-spot detection	152 WSI	good agreement
	Valous et al. [22]	pancreas	heterogeneity of proliferation	43 WSI	-
	Paulik et al. [23]	breast	detecting and count nuclear signals	25 WSI	PPV = 90.23%
	Pilutti et al. [24]	breast	hot-spot detection	50 cases, 3 experts	Acc = 82%
	Razavi et al. [30]	breast	LI estimation	30 TMA, 1 expert	Acc = 91%
	Gerard et al. [31]	breast	LI estimation	80 TMA	Acc = 93,5%
	Ko et al. [25]	breast	LI estimation	15 WSI	Acc = 89%
Deep Learning	Saha et al. [29]	breast	hot-spot detection and LI estimation	450 images, 5-fold cross validation	Acc = 97%
	Zhang et al. [27]	breast	cell detection	3000 samples *	Acc = 98%
	Lakshmi et al. [26]	bladder, breast	LI estimation	bladder: 60 images; breast: 40 images	bladder: F1 = 92%; breast: F1 = 74%
	Zhang et al. [28]	pancreas	nucleus recognition	20 annotated images	det.F1 = 81.3% clas.F1 = 62.3%
DL	Proposed method	breast	hot-spots detection and LI estimation	50 WSI, 2 experts	Acc = 95%, rho: 0.89

2. Materials

2.1. Database

This study involved digital slides from the AIDPATH breast cancer database [32], composed of breast tissue cohorts from four institutions and pathology labs around Europe, including Spain, Italy and Lithuania. The dataset includes:

- 50 WSI breast cancer specimens stained with Ki67/MIB-1 antibody (3,3'-diaminobenzidine tetrahydrochloride, DAB marker, brown color) and counter-stained with hematoxylin (blue).
- 50 WSI breast cancer specimens stained with H&E (violet).

H&E staining shows tumor localization, while Ki67 staining allows evaluation of tumor proliferation or the labeling index (Ki67 LI). An example of an H&E and Ki67 stained WSI is presented in Figure 1. It should be noted that there is significant biological and staining variability.

The slides were prepared serially using Formalin-Fixed Paraffin-Embedded (FFPE) tissue sections with 4 µm thickness. H&E and Ki67 stains were applied to consecutive pair sections to allow their posterior alignment for this study.

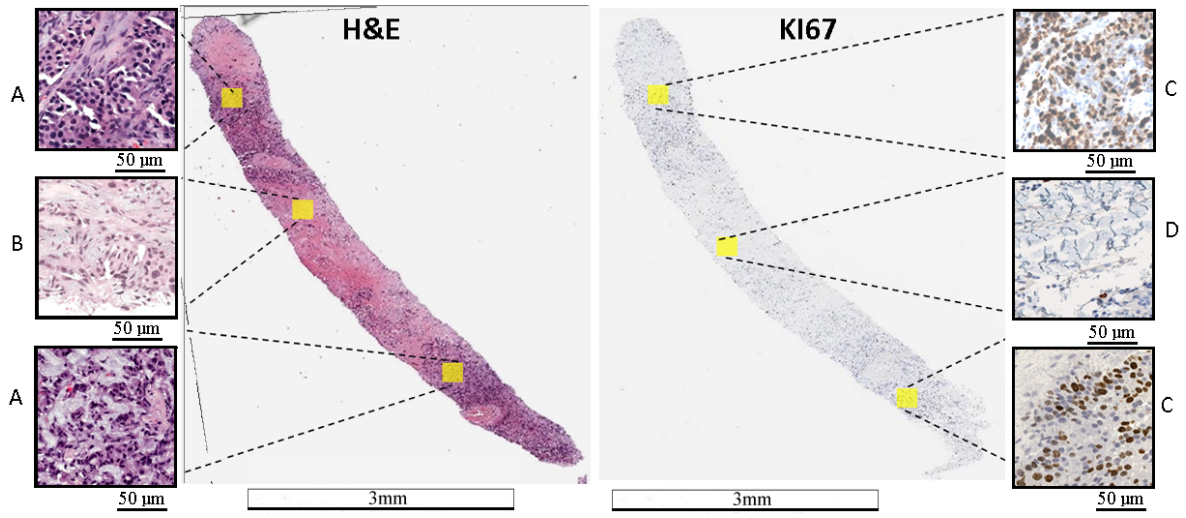


Figure 1. Example of WSI stained with H&E and Ki67. (**A**) area with a tumor on the H&E image, (**B**) area without a tumor on the H&E image, (**C**) area with a hot-spot on the Ki67 image, and (**D**)—area without a tumor on the Ki67 image. On the Ki67 image, immunopositive cells are marked by brown and immunonegative cells are marked by blue. The length bars shown in the WSIs are 3 mm and 50 µm in the zoomed areas.

2.2. Slide Digitization

The slides were digitized using an Aperio Scanscope CS (Leica Biosystems GmbH, Nussloch, Germany) whole slide scanner at 20× or 40× magnification. The area of tissue under a single field of view (FOV) at 40× was 0.037 mm^2, which resulted in a size of 1424 × 1064 pixels at 40× and 712 × 532 pixels at 20×. Depending on the size of tissue fixed on the slide, digital slides range between 0.5 GB and 2 GB. To reduce the time and effort required to manipulate the digital slides, create the ground truth annotations, and engage in computational processing, we worked at 20× magnification.

2.3. WSI Registration

To combine the H&E and Ki67 data to detect hot-spot areas, the H&E slides were registered to their Ki67 counterparts to achieve the correct alignment between serial sections. Thus, the H&E and Ki67 regions matched to achieve a correct inter-staining analysis. We used a WSI registration system based on intensity analysis to deal with multimodal images. This method compares the

intensity patterns in images via correlation metrics to achieve optimal transformation [33]. In our design, we applied the registration system proposed in Reference [34] , where the intensity analysis is applied to down-sampled slides (i.e., WSI at 2× magnification) to speed up the computational process. The final transformation obtained at 2× was transferred to the original 20× slides, thus obtaining the 20× registered WSIs.

2.4. Ground Truth

Each WSI was reviewed by pathologists participating in the AIDPATH project. These pathologists generated the ground truth annotations for tumor (H&E) and hot-spot (Ki67) areas (Figure 2) using the Aperio ImageScope software ver.12.3. Next, the pathologists evaluated the Ki67 LI.

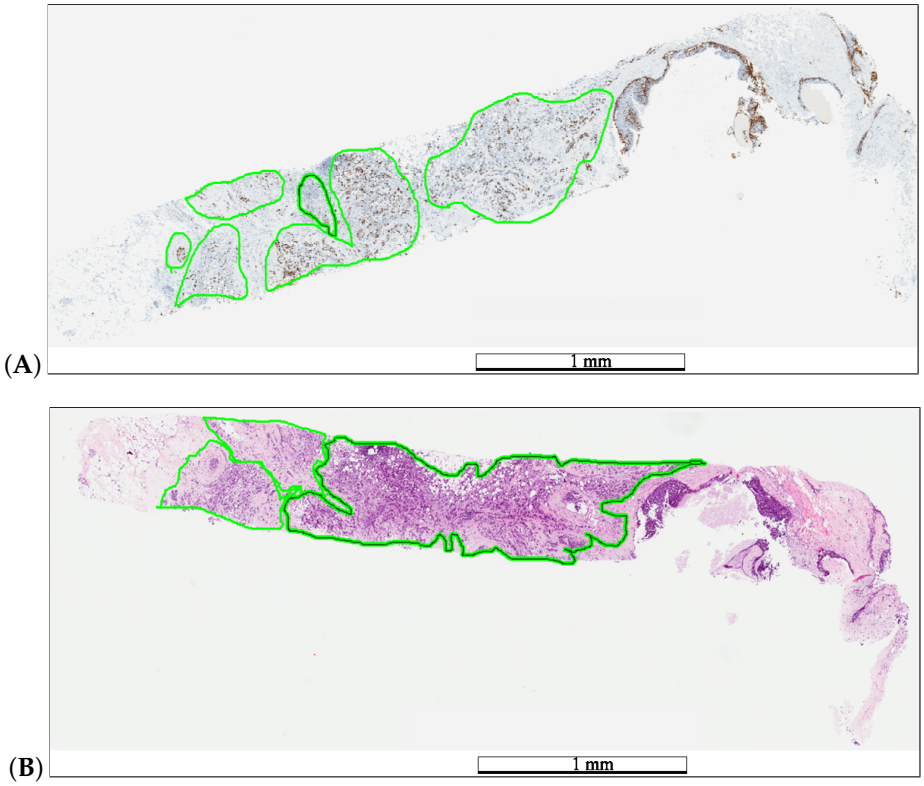


Figure 2. Example of medical annotations in a WSI pair before registration. The pathologist annotated hot-spot areas on the Ki67 specimen (**A**) and tumor areas on the H&E specimen (**B**) in a green color. The length bars shown in the WSIs are 1 mm.

2.5. Database Creation

The annotated areas were divided into 512 × 512 non-overlapping pixel image tiles. The H&E and Ki67 samples were utilized according to the information required:

- The H&E data were divided into two classes: tumor and non-tumor areas.
- The Ki67 data were divided into 11 classes: non-tumor areas, lymphocytes, and 9 classes with tumor areas (Figure 3). The lymphocyte class was added to this analysis to avoid the false positive regions generated by its high proliferation. The tumor areas were divided into 9 groups based on the tumor proliferation factor, which was calculated as the ratio between immunopositive cells (brown cells) and all tumor cells (brown and blue cells). The detection of all tumor cells was performed using the cell detector developed by Markiewicz et al. [35].

The dataset was randomly split into a training and test set at the whole slide level, where 35 cases were used for method development and 15 cases were used as a test set. The same test set, including

data from 15 cases, was used for the whole evaluation. Data augmentation was applied on the training set to increase the number of samples. The following methods were applied: color modification, 90° rotation, vertical flipping, and Gaussian blurring. Color modification was performed with Reinhard's method [36]. For each staining, we used four color patterns based on the various sites or institutions from which the databases were derived. The original dataset was significantly imbalanced due to biological variability. Some classes, such as lymphocytes, were significantly less represented than other classes (with up to 8 times less data). To achieve a balanced training database for classes with many fewer data, such as lymphocytes, we applied additional data augmentation, including 45° rotation, vertical flipping, and Gaussian blurring.

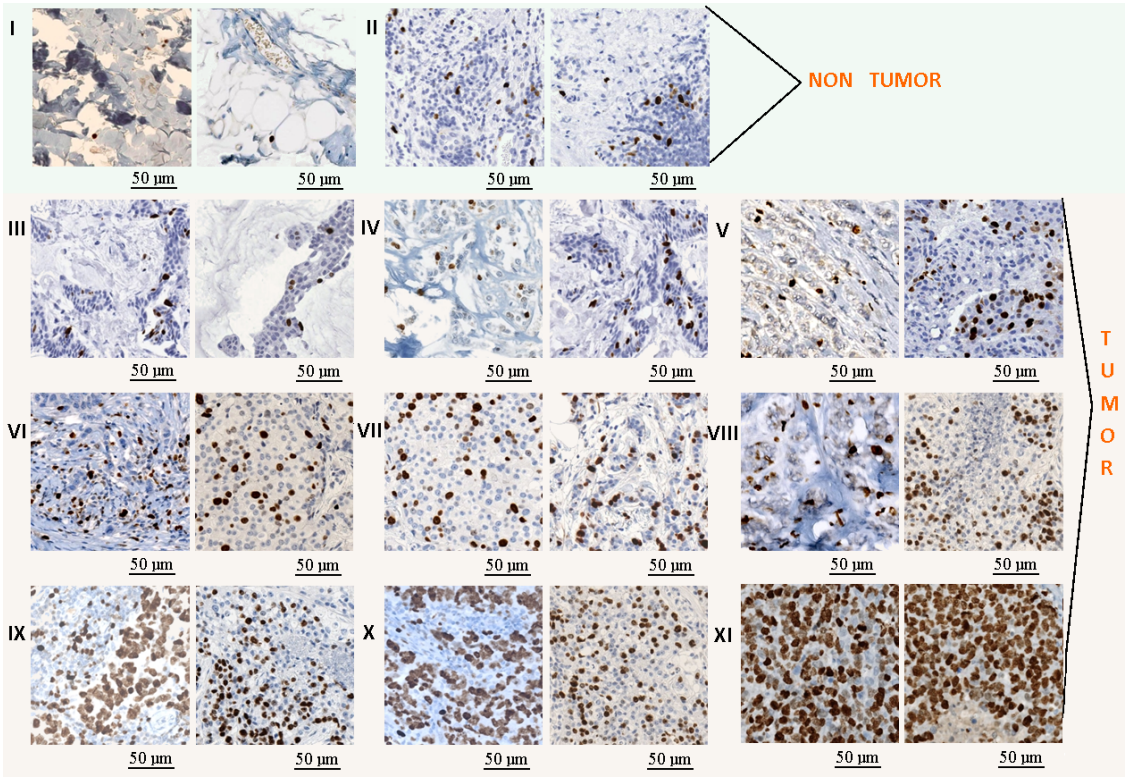


Figure 3. Example of tiles extracted from the KI-67 stained images: (**I**) non-tumor class, (**II**) lymphocytes, and (**III**–**XI**) tumor classes with several progressive tumor proliferations. The length bars shown in the tiles are 50 µm.

3. Methods

In our study, we developed and compared three strategies: the basic one (using only the Ki67 dataset) and two that use the H&E and Ki67 datasets. The idea of using the H&E and Ki67 datasets was inspired by the routine work of pathologists, who observe HE and KI67 stained slides to evaluate the specimens. In routine practice, pathologists have 2 separate slides—one stained with HE and one stained with KI67; based on this, the pathologists then perform an analysis. This approach is modeled as the second method. In the third developed method, we used the computational power and deconvolutional technique to merge HE and KI67 stained slides into one artificial "HEKI" image. All approaches are based on Deep Learning and were developed to explore the possibilities of combining information from several stains. The proposals follow a three-step workflow: (a) tissue area segmentation, (b) hot-spot detection, and (c) Ki67 LI evaluation. All methods were developed using Matlab [37], and the calculations were performed on a computer with an Nvidia Quadro 400 graphics card and Cuda version 8.0.

3.1. Tissue Area Segmentation

Tissue segmentation is the first step in WSI analysis. It allows one to limit the area under analysis, thus reducing the computational time. To detect the WSI tissue regions, background segmentation based on color histogram analysis was performed.

3.2. Hot-Spot Detection

Hot-spot detection is central to achieving the correct tumor proliferation index. Here, we present three approaches to detect hot-spot areas on the Ki67 WSIs. These approaches were classified according to the type of data used and how those data are utilized to achieve hot-spot detection:

- Approach I: Using Ki67 data with 1 CNN.
- Approach II: Using Ki67 and H&E data with 2 CNNs, one for each stain.
- Approach III: Combine Ki67 and H&E data by means of color deconvolution with 1 CNN.

These approaches are based on CNN classification, but the classification is implemented in various ways in order to integrate all the information from the slides. That is, only to the Ki67 slides (method 1), to both the H&E and the Ki67 slides but separately, in a sequential process (method 2) and to the combination of the H&E and the Ki67 slides (method 3). The final result in each case is a hot-spot area map. These methods are illustrated in Figure 4.

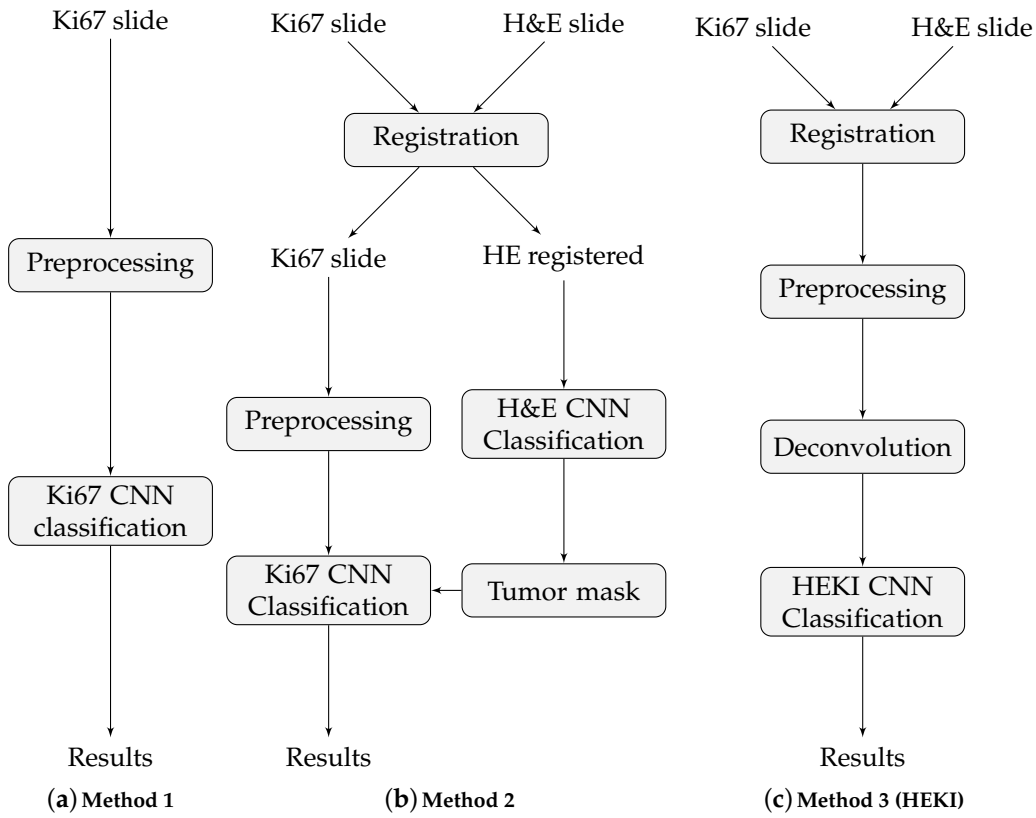


Figure 4. The main steps of the three proposed methods.

3.2.1. CNN Model Training

The CNNs were set up using the AlexNet model [38], a popular CNN model applied to image classification and digital pathology tasks [39]. This CNN has eight layers of neurons: five convolutional and three fully connected. We applied the full-training and fine-tuning methods to the training process. Fine-tuning is a transfer learning technique based on transferring the features and parameters of the

network from a broad domain to a specific one. This model was pretrained on the LSVRC-2010 ImageNet database with 1000 classes of natural objects and was fine-tuned using our database. The fine-tuning training took 20 epochs. The learning rate, gamma, and momentum for the stochastic gradient descent (SGD) optimizer were set to 0.01, 0.9, and 0.1, respectively. The retraining operation allows for optimization of the network to minimize the errors in another, more specific domain. In the case of fine-tuning, the CNN model was adjusted by modifying the network weights and trimming the number of outputs in the last layer. As a result, the CNN model returns a vector with class scores representing the affinity of the image to each of the predefined classes. The image tile classification is computed by the CNN by finding the maximum class score for each tile. With this information, we can construct the classification map of the WSIs.

3.2.2. Method 1

Under this approach, we used only Ki67 images to train the CNN (Figure 4a). The network classifies each tile into one of the 11 classes, defined in Section 2.5. A high value tumor class indicates a high tumor proliferation index, which defines the hot-spot areas. We trained the Alexnet CNN with the data extracted from the 35 WSIs, augmented as described in Section 2.5, which resulted in 210,312 training tiles. For testing, we used 25,264 tiles.

3.2.3. Method 2

This method uses information from registered H&E and Ki67 slides in two separated CNN classifiers, one for each stain, to combine the outputs in a final step (Figure 4b). Since we prepared consecutive WSI sections in a serial way, these sections can be registered to use the tumor area location and thereby enhance hot-spot detection. H&E staining allows for tumor area detection, while Ki67 staining can detect the hot-spot areas and evaluate the proliferation index. For each H&E and Ki67 registered slide, the process is as follows: (I) The H&E CNN is used to classify the registered H&E data into two groups: tumor area and non-tumor area; second (II), we analyze the Ki67 slide only in tumor areas detected in the H&E tissue. As a result, the H&E tumor areas are used to restrict the Ki67 regions to be classified by the Ki67 CNN.

In this proposal, each CNN is trained independently based on separate data cohorts. The first CNN checks whether a tile is a tumor area, while the second CNN shows the tumor proliferation degree. The H&E CNN is trained via fine-tuning with data from 10 H&E slides. After data augmentation, we achieved 7408 training tiles. The Ki67 CNN was trained using the full-training method with data from the 35 WSIs. After data augmentation, the number of training tiles was 210,312. The significant difference in the number of tiles in each of the datasets is related to the available data cohorts and the number of distinct classes defined in each CNN. The larger Ki67 dataset can be used for the full CNN training process. Since the number of H&E slides is limited, it is more appropriate to use the fine-tuning method to train the CNN. Finally, this approach was tested using 25,264 H&E and Ki67 tiles.

3.2.4. Method 3 (HEKI)

This last approach combines the information from registered H&E and Ki67 WSIs in one CNN (Figure 4c). We used registered H&E and Ki67 data to enhance hot-spot detection by combining the information of both stains (Figure 5). The color deconvolution operation [40] can convert an RGB image into a stain image. In other words, in each of the image layers we have separate stains. For example, for H&E images after color deconvolution, we obtained an image with hematoxylin staining and an image with eosin staining (Figure 5). As a result, the artificial HEKI image we created includes information about tumor regions and tumor proliferation. These data are the CNN classifier's input.

In this proposal, the AlexNet model was trained by fine-tuning with data from 70 WSIs (35 H&E slides and 35 KI67 slides). The image augmentation operation was performed according to the methods described previously. The number of training tiles was 61,163, while the number of testing tiles was 25,264 from 15 slides.

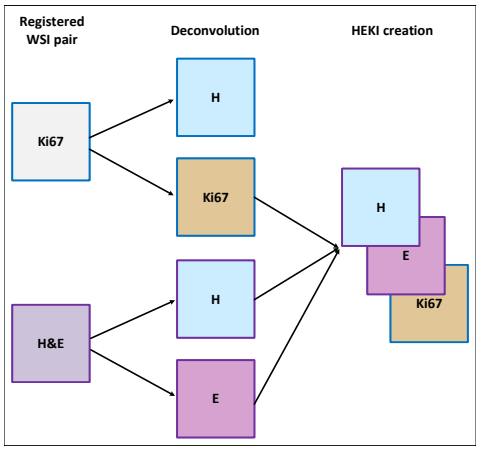


Figure 5. Overview of the HEKI image creation.

3.3. Ki67 Proliferation Index Evaluation

We evaluated the Ki67 proliferation factor in a field of view (FOV) according to the relation between the number of pixels belonging to immunopositive cells (brown color in Ki67 stain) and the number of pixels belonging to all cancer cells (brown and blue cells) inside the tumor area. We detected the brown cells using the cell detector method of Markiewicz et al. [35]. To detect the blue cells and the cancer region inside the FOV, we applied a grayscale histogram normalization and thresholding based on the three-class Otsu method [41]. Notably, it is important not to count other types of cells, such as lymphocytes, as cancer cells. To achieve this, we filtered cells smaller than the threshold size defined by the average size of the immunopositive cells (Figure 6).

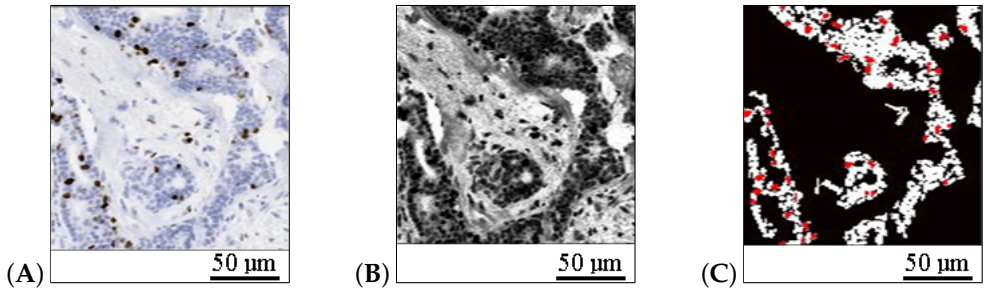


Figure 6. The detection of cancer regions and cells: (**A**) hot-spot area, (**B**) image in gray scale after histogram normalization, and (**C**) detected tumor area (white) and immunopositive cells (red). The length bars shown in the sample images are 50 µm.

The Ki67 LI (LI_{Ki67}) is calculated according to the following formula:

$$LI_{\mathrm{Ki67}} = \frac{\mathrm{pix}_{\mathrm{br}}}{(\mathrm{pix}_{\mathrm{br}} + \mathrm{pix}_{\mathrm{bl}})\,\alpha}, \tag{1}$$

where $\mathrm{pix}_{\mathrm{br}}$ and $\mathrm{pix}_{\mathrm{bl}}$ are respectively, the number of brown cell pixels and the number of blue cell pixels inside the tumor area, and α is the corrector factor selected experimentally to correct deviations in the tumor area detected, whose value ranges from 1 to 0.75. In our experiments, based on the analysis of 5 specimens, we determined that $\alpha = 0.9$ yields correct results.

The final value of Ki67 LI is the average value of I_{Ki67} for 10 hot-spot fields. The hot-spot fields were selected from the hot-spot map created in the first step from the FOV with the highest classification results. An example of automatically selected hot-spot fields is presented in Figure 7.

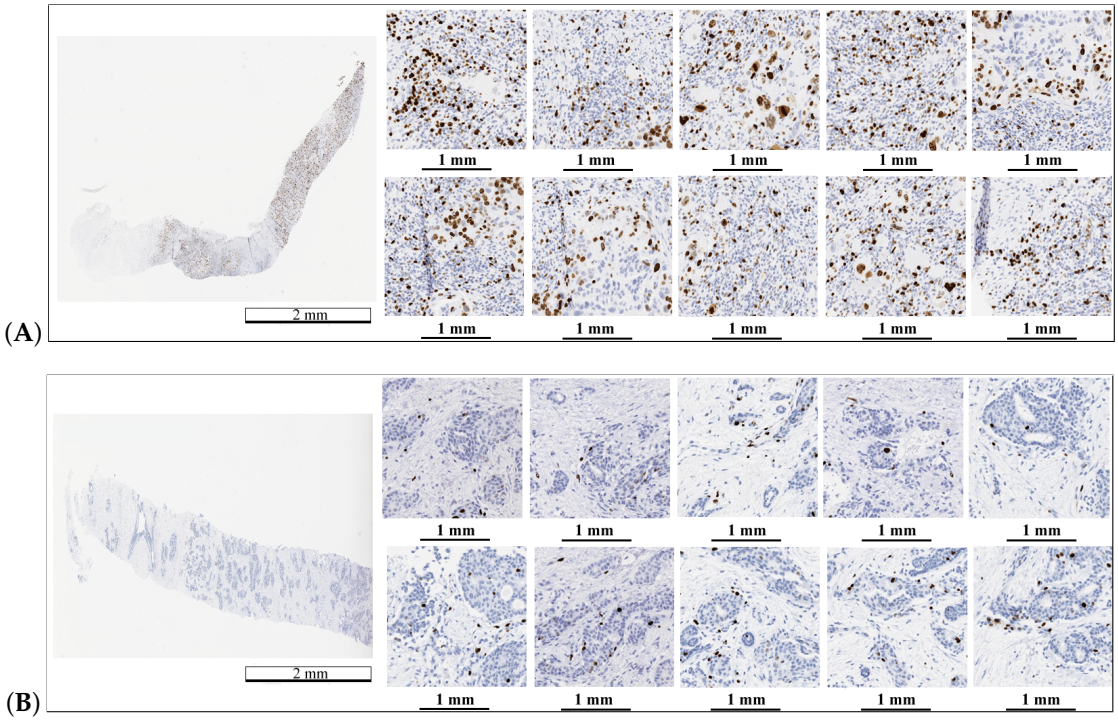


Figure 7. Automatic hot-spot area detection and selection for two WSIs: (**A**) a WSI with the Ki67 LI equal to 50% and (**B**) a WSI with the Ki67 LI equal to 5%. The length bars shown in the WSIs are 2 mm and 1 mm in the zoomed areas.

4. Results

We performed a complete evaluation of the proposed study for automatic WSI Ki67 LI diagnosis. The results are presented in two sections. First, we show the results of hot-spot detection and classification, while the second section is devoted to explaining the Ki67 LI evaluation results, including a comparison with a pathologist's evaluation.

4.1. Hot-Spot Detection

The classification metrics were calculated for a test set including 25,264 tiles using a one-shot experiment. The statistical results are shown in Table 2 and were calculated as follows:

$$Acc = \frac{TP + TN}{P + N} \tag{2}$$

$$F1score = 2\frac{PPV * TPR}{PPV + TPR} \tag{3}$$

where TP is true positive prediction, TN is true negative prediction, P is positive (ground true), N is negative (ground true), PPV is the positive prediction value (precision), and TPR is the true negative rate (sensitivity).

Table 2. Average values for the statistical results of tile classification (N = 25,264).

	Accuracy	Sensitivity (TPR)	Specificity	Precision (PPV)	F1-Score
Method 1	96%	47%	98%	47%	47%
Method 2	95%	39%	96%	44%	41%
Method 3 (HEKI)	95%	77%	98%	70%	73%

The results confirm the correct tile classification using each method. However, incorrect classifications between classes must also be considered since they are crucially important to the classification quality. A correct hot-spot classification must accomplish the following:

- Data from classes I (no tumor) and II (lymphocytes) must not be classified as tumor areas (hot-spots).
- Data with a low proliferation index must not be classified as data with high tumor proliferation.

Figure 8 shows the confusion matrices for each method of classification within various tumor groups (low, medium, and high tumor proliferation). The analysis of the confusion matrix for each method shows that Method 3 (HEKI) provides the best performance in discriminating classes. Ultimately, Method 3 achieved correct lymphocyte classification, while Methods 1 and 2 incorrectly classified lymphocytes as tumor areas with a high value of index proliferation. This is important because the presence of lymphocytes means that the patient is reacting against the tumor, which must not be incorrectly classified as a false positive hot-spot.

A	no tumor	tumor proliferation [%]								
no tumor	lymphocytes	low		medium		hight				
		0-10	10-20,	20-30	30-40	40-50	50-60	60-70	70-80	>80
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.04	0.00	0.16	0.00	0.00	0.00	0.00	0.60	0.20
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.02	0.00	0.03	0.34	0.54	0.06	0.01	0.00	0.00	0.00	0.00
0.02	0.00	0.01	0.05	0.55	0.20	0.10	0.01	0.01	0.01	0.03
0.02	0.00	0.00	0.00	0.14	0.39	0.17	0.18	0.06	0.01	0.03
0.01	0.00	0.00	0.00	0.05	0.11	0.33	0.34	0.06	0.05	0.05
0.00	0.00	0.00	0.00	0.01	0.02	0.13	0.51	0.20	0.07	0.05
0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.16	0.32	0.35	0.13
0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.10	0.34	0.51	0.01
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.10	0.88

B	no tumor	tumor proliferation [%]								
no tumor	lymphocytes	low		medium		hight				
		0-10	10-20,	20-30	30-40	40-50	50-60	60-70	70-80	>80
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.04	0.00	0.16	0.00	0.00	0.00	0.00	0.60	0.20
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.24	0.00	0.02	0.24	0.45	0.04	0.01	0.00	0.00	0.00	0.00
0.17	0.00	0.01	0.05	0.44	0.19	0.09	0.01	0.01	0.01	0.02
0.22	0.00	0.00	0.00	0.11	0.29	0.14	0.17	0.05	0.01	0.01
0.17	0.00	0.00	0.00	0.05	0.10	0.27	0.28	0.06	0.04	0.03
0.19	0.00	0.00	0.00	0.01	0.02	0.09	0.41	0.17	0.06	0.05
0.19	0.00	0.00	0.00	0.01	0.00	0.03	0.13	0.26	0.28	0.10
0.26	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.06	0.26	0.38
0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.08	0.73

C	no tumor	tumor proliferation [%]								
no tumor	lymphocytes	low		medium		hight				
		0-10	10-20,	20-30	30-40	40-50	50-60	60-70	70-80	>80
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.11	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.01	0.00	0.00	0.79	0.07	0.07	0.04	0.02	0.00	0.00	0.00
0.00	0.00	0.00	0.01	0.91	0.03	0.04	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.01	0.05	0.81	0.07	0.04	0.01	0.00	0.00
0.00	0.00	0.00	0.00	0.02	0.03	0.76	0.07	0.07	0.05	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.88	0.06	0.03	0.00
0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.04	0.83	0.10	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.96	0.00
0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.04	0.08	0.79	0.03

Figure 8. Confusion matrix for the tile classification (N = 25,264) under each method: (**A**) Method 1, (**B**) Method 2, and (**C**) Method 3 (HEKI).

Contrary to our initial assumptions, Method 2 provided a worse confusion matrix than Method 1, even though the system combined the Ki67 and H&E stain information in a more complex framework. In Method 2, we can effectively observe the incorrect classification of many hot-spot areas as non-tumor areas. The reason for this result was found after an analysis of the results, as small tumor proliferation appeared in these places, leading to incorrectly classified non-tumor areas, thereby avoiding analysis by the Ki67 CNN.

The goal of medical doctors is to select only a few hot-spot areas in WSIs, while automatic evaluation can detect a significantly higher number of such regions. The proposed methods allow one to create a proliferation index map. Figure 9 shows the results of automatic hot-spot detection.

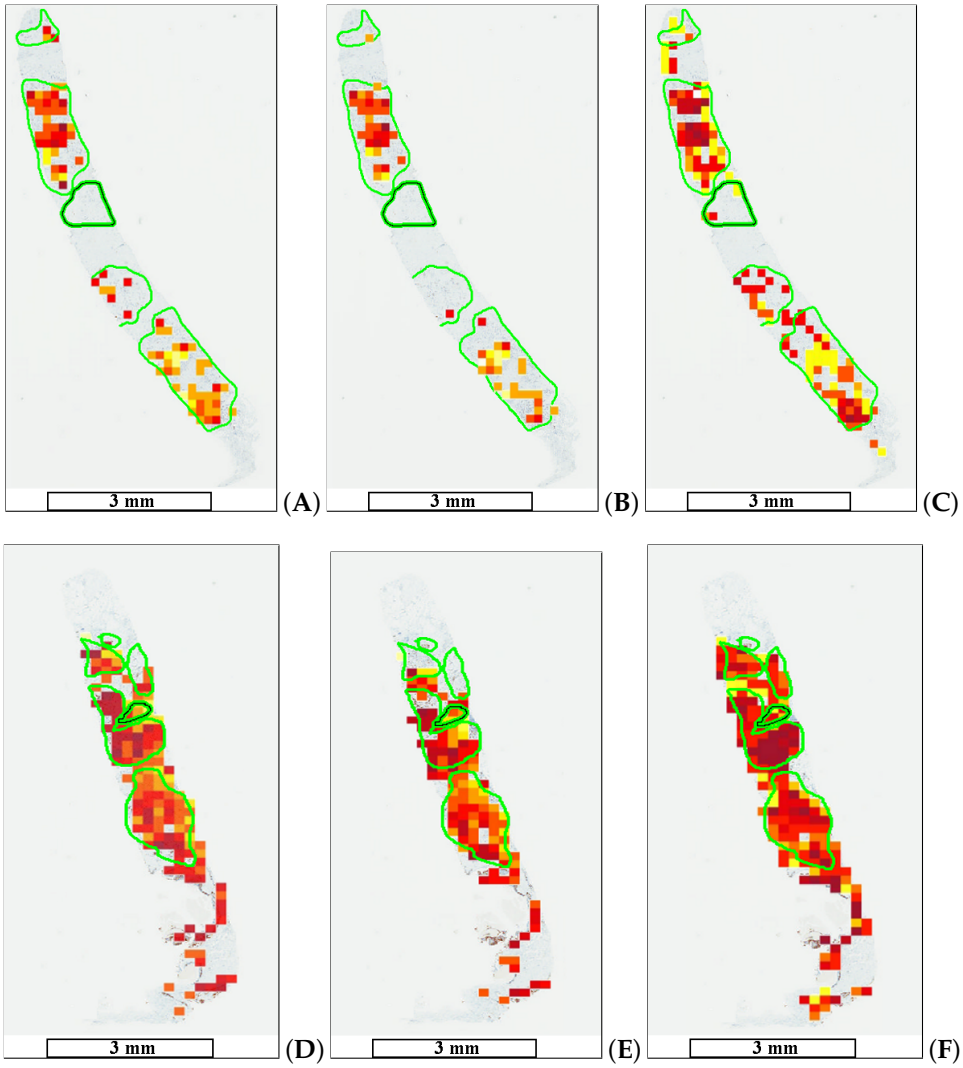


Figure 9. Automatic tumor proliferation map examples: (**A**,**D**) maps for Method 1, (**B**,**E**) maps for Method 2, and (**C**,**F**) maps for Method 3 (HEKI). High tumor proliferation areas are marked in red (dark red), whereas areas with low tumor proliferation are marked in yellow. The length bars shown in the WSIs are 3 mm.

The results presented in this section show that Method 3 (HEKI) provides both correct lymphocyte classification and the correct detection of a significant number of gold-standard hot-spot areas. It also performs better in terms of the confusion between classes. Hence, we focus on Method 3 (HEKI) to compute the quantitative results for the Ki67 proliferation index estimation.

4.2. Ki67 lAbeling Index Estimation

Hot-spot detection is a complex process since each WSI presents specific characteristics for the tumor, cells, and Ki67 hot-spot distribution. In some WSIs, tumor proliferation is observed homogeneously throughout the whole specimen, while some samples present tumor proliferation areas in specific regions, which sometimes have multiple proliferation values. Therefore, WSIs usually present few tumor proliferation areas that represent significant values of the proliferation index. This fact makes manual Ki67 LI evaluation difficult, which leads to observed differences between automatic and manual evaluation.

The results of the automatic (Auto) and manual (Manual) Ki67 LI evaluations are shown in Table 3 for 15 samples The manual evaluation was performed by two pathologists using two evaluation strategies. One pathologist performed the evaluation based on 10 FOV, while the second one performed the evaluation based on 1 FOV. Both methods are accepted as standard pathological procedures.

Table 3. Ki67 evaluation results. Number of samples: N = 15.

Case	Auto 10FOV	Auto 1FOV	Manual 10FOV	Manual 1FOV
1	15.70%	13.87%	22%	15%
2	18.07%	19.36%	10%	15%
3	36.24%	18.35%	40%	30%
4	4.67%	6.07%	8%	5%
5	20.47%	24.38%	17%	20%
6	29.77%	27.77%	30%	25%
7	48.40%	52.59%	50%	50%
8	30.36%	23.78%	20%	25%
9	67.60%	61.1%	90%	90%
10	10.42%	2.68%	15%	15%
11	31.44%	35.32%	30%	38%
12	3.74%	10.23%	5%	10%
13	24.62%	34.85%	25%	20%
14	30.85%	30.29%	15%	15%
15	9.34%	3.36%	8%	5%

As we can observe, the results achieved by our approach (Auto) provided a Ki67 LI that correctly estimated the manual values. There are several samples that deserve special comments:

- Five specimens presented higher results using our automatic diagnosis with 10 FOV compared to the manual evaluation. This is because the automatic system always detects hot-spot areas with the highest tumor proliferation, which can be difficult to perform by manual evaluation, thereby skipping some regions of interest.
- In three cases, we observed significantly smaller results than the medical evaluation for our automatic diagnosis of 10 FOV. We determined that the automatic system counts all cancer cells, while manual counting estimates the number of cells. This can lead to an imprecise estimation of the Ki67 proliferation factor.

To evaluate the variations between medical estimations and the automated Ki67 LI process, three statistical tests were performed: the Bland–Altman plot, one-sided t-test, and Spearman's ρ correlation. The Bland–Altman test allows one to compare two quantitative measures by setting agreement limits. These statistical limits are determined using the mean and standard deviation of the two measures.

This is a popular method applied in clinical measurement comparisons, where each of the methods provides some errors. The results are presented in Figure 10 as a comparison between the various doctors' results and between the automatic results and the doctors' results. From these final results, the following can be concluded:

- There is a difference between medical evaluation based on 1 or 10 FOV with a discrepancy equal to 0.47. This indicates a slight overestimation of Ki67 LI by the expert evaluating 1 FOV (Figure 10E).
- Automatic diagnosis for 10 FOV and manual evaluation using 1 FOV resulted in a small underestimation for our method, with a discrepancy value of −0.25 (Figure 10A).
- For the comparison between the automatic hot-spot evaluation for 1 FOV and the medical evaluation using 1 FOV, we observed a discrepancy of 0.93 (Figure 10C).
- For the results between the automatic hot-spot evaluation for 1 FOV and the medical evaluation using 10 FOV, we observed the highest discrepancy: 1.4 (Figure 10D).
- The comparison of the automatic evaluation for 10 FOV and the medical evaluation for 10 FOV presented the smallest difference between the two methods. The discrepancy value is 0.22 (Figure 10B).

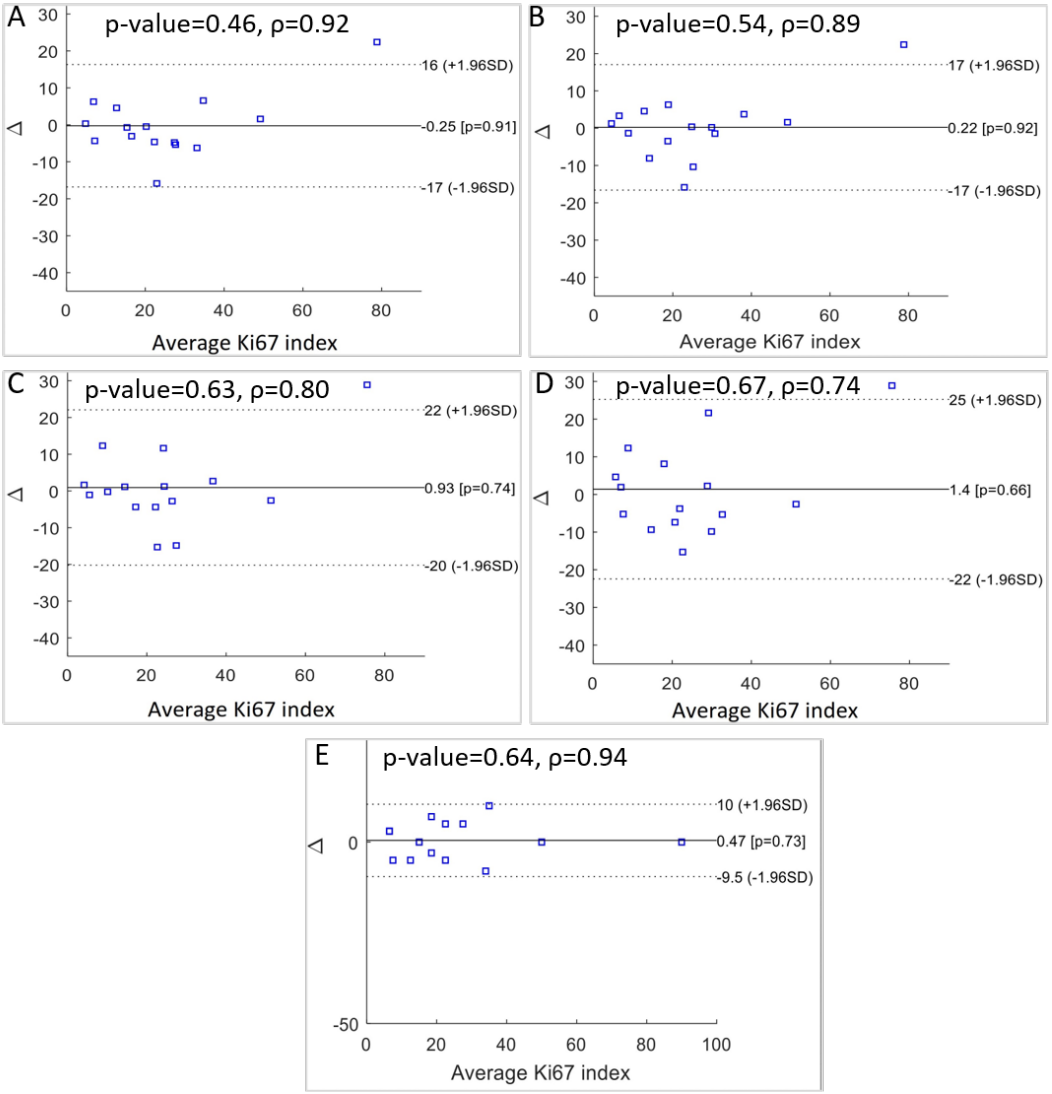


Figure 10. Statistical significance *p*-value, Spearman's ρ correlation coefficients, and Bland-Altman plots of the differences between: (**A**) the automatic method for 10 FOV and manual for 1 FOV, (**B**) automatic for 10 FOV and manual for 10 FOV, (**C**) automatic for 1 FOV and manual for 1 FOV, (**D**) automatic for 1 FOV and manual for 10 FOV, (**E**) and between the two manual scores for 1 FOV and 10 FOV.

For the one-sided *t*-test, no significant differences in Ki67 LI evaluation were observed using a one-sided t-test for the manual and automatic evaluations. The Spearman's ρ correlation coefficient was used for statistical confirmation of the results' consistency. The correlation is equal to 0.94 for

both the manual results (Figure 10E). Between the automatic evaluation for 10 FOV and the medical evaluation for 10 FOV, the correlation is equal to 0.89 (Figure 10B), and between the automatic evaluation for 10 FOV and the medical evaluation for 1 FOV, it is equal to 0.92 (Figure 10A). All statistical examinations confirm the good level of agreement between the automatic quantitative evaluations of the Ki67 LI for 10 FOV and the experts' results.

Regarding the computational cost, Method III computes the Ki67 analysis of the WSI, in 2–3 min, according to the area of tissue under evaluation.

5. Conclusions

In this paper, we presented three methods for hot-spot detection and Ki67 LI evaluation based on CNN classification. These methods were differentiated according to the data used, that is, Ki67 stain only or combination of Ki and H&E stains (HEKI). Our study shows that hot-spot areas can be detected by using the proposed methods, although some aspects make the HEKI Method 3 better suited for this purpose. Method 1 and Method 2 present a confusion matrix that provides worse results than the HEKI method since the confused values are more critical than those that appear in the former method. Moreover, Methods 1 and 2 cannot correctly differentiate lymphocyte regions, which is a key issue for avoiding false positive classifications.

Method 3 (HEKI), based on combining both Ki67 and H&E, provided better characteristics for Ki67 hot-spot detection and Ki67 LI calculation. It included WSI registration, color deconvolution, and the creation of artificial image structures (HEKI) used for CNN classification. The average value of CNN accuracy for tile classification was 95%, which means that the proposal correctly dealt with the different types of data and their combinations. Moreover, this approach allowed us to create a tumor proliferation map and compute both automatic hot-spot detections and Ki67 LI.

To establish a robust evaluation of the proposed methods, complete training and complex tests were performed. The database used contains 50 WSIs stained by Ki67 and 50 WSIs stained by H&E. The ground truth was created by four medical experts with extensive experience selecting hot-spot areas and evaluating the Ki67 LI. The Ki67 LI was evaluated using 1 FOV and 10 FOV, as there is neither a consensus among pathologists nor a standard protocol on how many FOVs to use.

Method 3's results were compared with the medical diagnosis ground truth. The Bland–Altman analysis and Spearman's ρ correlation coefficient confirmed the agreement between the medical and automatic diagnosis. In the Bland–Altman plot, only one point was observed to be out of range. The Bland–Altman analysis showed that the discrepancy between the automatic diagnosis and medical evaluation of Ki67 LI, based on 10 FOV, is only 0.25. Spearman's ρ correlation coefficient was 0.89 and 0.92 compared to the manual calculations for 10 FOV and 1 FOV, respectively. All statistical tests confirmed the good level of agreement between the automatic and medical evaluation of Ki67 LI.

Finally, we showed that automatic diagnosis based on CNN and data combinations from two types of staining can facilitate the development of an efficient tool to support Ki67 LI evaluations in breast cancer specimens.

Author Contributions: Conceptualization, G.B. and Z.S.-C.; methodology, G.B., Z.S.-C. and J.G.; investigation: G.B., Z.S.-C. and J.G., data curation: G.B. and L.G.-L., formal analysis: Z.S.-C. and G.B.; writing—original draft preparation: Z.S.-C., J.G., and G.B.; writing—review and editing: G.B., supervision: G.B. and L.G.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the EU FP7 Program, AIDPATH project, grant number 612471, and partially funded by National Science Center (Poland) by grant UMO-2016/23/N/ST6/02076 and Junta de Comunidades de Castilla-La Mancha (Spain), with grant SBPLY/19/180501/000273.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 637–653.
2. Muhammad, K.; Hussain, T.; Baik, S.W. Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognit. Lett.* **2020**, *130*, 370–375. [CrossRef]
3. Barra, S.; Carta, S.M.; Corriga, A.; Podda, A.S.; Recupero, D.R. Deep learning and time series-to-image encoding for financial forecasting. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 683–692. [CrossRef]
4. Kawauchi, K.; Furuya, S.; Hirata, K.; Katoh, C.; Manabe, O.; Kobayashi, K.; Watanabe, S.; Shiga, T. A convolutional neural network-based system to classify patients using FDG PET/CT examinations. *BMC Cancer* **2020**, *20*, 1–10. [CrossRef] [PubMed]
5. Kitao, T.; Shiga, T.; Hirata, K.; Sekizawa, M.; Takei, T.; Yamashiro, K.; Tamaki, N. Volume-based parameters on FDG PET may predict the proliferative potential of soft-tissue sarcomas. *Ann. Nucl. Med.* **2019**, *33*, 22–31. [CrossRef] [PubMed]
6. Bueno, G.; Fernandez-Carrobles, M.; Deniz, O.; Garcia Rojo, M. New Trends of Emerging Technologies in Digital Pathology. *Pathobiology* **2016**, *83*, 61–69. [CrossRef] [PubMed]
7. Litjens, G.; Sánchez, C.; Timofeeva, N.; Hermsen, M.; Nagtegaal, I.; Kovacs, I.; Hulsbergen-Van De Kaa, C.; Bult, P.; Van Ginneken, B.; Van Der Laak, J. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **2016**, *6*, 26286. [CrossRef] [PubMed]
8. Khoury, T.; Zirpoli, G.; Cohen, S.; Geradts, J.; Omilian, A.; Davis, W.; Bshara, W.; Miller, R.; Mathews, M.; Troester, M.; et al. Ki-67 Expression in Breast Cancer Tissue MicroarraysAssessing Tumor Heterogeneity, Concordance With Full Section, and Scoring Methods. *Am. J. Clin. Pathol.* **2017**, *148*, 108–118. [CrossRef] [PubMed]
9. Laurinavicius, A.; Plancoulaine, B.; Rasmusson, A.; Besusparis, J.; Augulis, R.; Meskauskas, R.; Herlin, P.; Laurinaviciene, A.; Muftah, A.; Miligy, I.; et al. Bimodality of intratumor Ki67 expression is an independent prognostic factor of overall survival in patients with invasive breast carcinoma. *Virchows Arch.* **2016**, *468*, 493–502. [CrossRef]
10. Elie, N.; Plancoulaine, B.; Signolle, J.; Herlin, P. A simple way of quantifying immunostained cell nuclei on the whole histologic section. *Cytom. Part A* **2003**, *56A*, 37–45. [CrossRef]
11. Gudlaugsson, E.; Skaland, I.; Janssen, E.; Smaaland, R.; Shao, Z.; Malpica, A.; Voorhorst, F.; Baak, J. Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology* **2012**, *61*, 1134–1144. [CrossRef]
12. Robertson, S.; Acs, B.; Lippert, M.; Hartman, J. Prognostic potential of automated Ki67 evaluation in breast cancer: Different hot spot definitions versus true global score. *Breast Cancer Res. Treat.* **2020**, *183*, 161–175. [CrossRef]
13. Besusparis, J.; Plancoulaine, B.; Rasmusson, A.; Augulis, R.; Green, A.R.; Ellis, I.O.; Laurinaviciene, A.; Herlin, P.; Laurinavicius, A. Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer. *Diagn. Pathol.* **2016**, *11*, 82. [CrossRef] [PubMed]
14. Kitson, S.; Sivalingam, V.; Bolton, J.; McVey, R.; Nickkho-Amiry, M.; Powell, M.; Leary, A.; Nijman, H.; Nout, R.; Bosse, T.; et al. Ki-67 in endometrial cancer: Scoring optimization and prognostic relevance for window studies. *Mod. Pathol.* **2017**, *23056*, 459–468. [CrossRef] [PubMed]
15. Penault-Llorca, F.; Radosevic-Robin, N. Ki67 assessment in breast cancer: An update. *Pathology* **2017**, *49*, 166–171. [CrossRef]
16. Stålhammar, G.; Martinez, N.F.; Lippert, M.; Tobin, N.P.; Mølholm, I.; Kis, L.; Rosin, G.; Rantalainen, M.; Pedersen, L.; Bergh, J.; et al. Digital image analysis outperforms manual biomarker assessment in breast cancer. *Mod. Pathol.* **2016**, *29*, 318–329. [CrossRef] [PubMed]
17. Lu, H.; Papathomas, T.; van Zessen, D.; Palli, I.; de Krijger, R.; van der Spek, P.; Dinjens, W.; Stubbs, A. Automated Selection of Hotspots (ASH): Enhanced automated segmentation and adaptive step finding for Ki67 hotspot detection in adrenal cortical cancer. *Diagn. Pathol.* **2014**, *9*, 216. [CrossRef] [PubMed]
18. Khan Niazi, M.; Yearsley, M.; Zhou, X.; Frankel, W.L.; Gurcan, M.N. Perceptual clustering for automatic hotspot detection from Ki-67-stained neuroendocrine tumour images. *J. Microsc.* **2014**, *256*, 213–225. [CrossRef]

19. Lopez, X.; Debeir, O.; Maris, C.; Rorive, S.; Roland, I.; Saerens, M.; Salmon, I.; Decaestecker, C. Clustering methods applied in the detection of Ki67 hot-spots in whole tumor slide images: An efficient way to characterize heterogeneous tissue-based biomarkers. *Cytom. Part A* **2012**, *81A*, 765–775. [CrossRef]
20. Xing, F.; Su, H.; Neltner, J.; Yang, L. Automatic Ki-67 Counting Using Robust Cell Detection and Online Dictionary Learning. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 859–870. [CrossRef]
21. 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. *Anal. Cell. Pathol.* **2015**, *2015*, 15. [CrossRef]
22. Valous, N.; Lahrmann, B.; Halama, N.; Bergmann, F.; Jäger, D.; Grabe, N. Spatial intratumoral heterogeneity of proliferation in immunohistochemical images of solid tumors. *Med. Phys.* **2016**, *43*, 2936–2947. [CrossRef] [PubMed]
23. Paulik, R.; Micsik, T.; Kiszler, G.; Kaszál, P.; Székely, J.; Paulik, N.; Várhalmi, E.; Prémusz, V.; Krenács, T.; Molnár, B. An optimized image analysis algorithm for detecting nuclear signals in digital whole slides for histopathology. *Cytom. Part A* **2017**, *91*, 595–608. [CrossRef]
24. Pilutti, D.; Della Mea, V.; Pegolo, E.; La Marra, F.; Antoniazzi, F.; Di Loreto, C. An adaptive positivity thresholding method for automated Ki67 hotspot detection (AKHoD) in breast cancer biopsies. *Comput. Med. Imaging Graph.* **2017**, *61*, 28–34. [CrossRef] [PubMed]
25. Ko, C.C.; Lin, C.H.; Chuang, C.H.; Chang, C.Y.; Chang, S.H.; Jiang, J.H. A Whole Slide Ki-67 Proliferation Analysis System for Breast Carcinoma. In Proceedings of the 2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media), Bali, Indonesia, 6–9 August 2019; pp. 210–213.
26. 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 Proceedings of the TENCON 2019—2019 IEEE Region 10 Conference (TENCON), Kochi, India, 17–20 October 2019; pp. 2310–2314.
27. Zhang, R.; Yang, J.; Chen, C. Tumor cell identification in ki-67 images on deep learning. *Mol. Cell. Biomech.* **2018**, *15*, 177.
28. Zhang, X.; Cornish, T.C.; Yang, L.; Bennett, T.D.; Ghosh, D.; Xing, F. Generative Adversarial Domain Adaptation for Nucleus Quantification in Images of Tissue Immunohistochemically Stained for Ki-67. *JCO Clin. Cancer Inform.* **2020**, *4*, 666–679. [CrossRef]
29. 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. *Sci. Rep.* **2017**, *7*, 1–14. [CrossRef]
30. Razavi, S.; Khameneh, F.D.; Serteli, E.A.; Cayir, S.; Cetin, S.B.; Hatipoglu, G.; Ayalti, S.; Kamasak, M. An Automated and Accurate Methodology to Assess Ki-67 Labeling Index of Immunohistochemical Staining Images of Breast Cancer Tissues. In Proceedings of the 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP), Maribor, Slovenia, 20–22 June 2018; pp. 1–5.
31. Geread, R.S.; Morreale, P.; Dony, R.D.; Brouwer, E.; Wood, G.A.; Androutsos, D.; Khademi, A. IHC Colour Histograms for Unsupervised Ki67 Proliferation Index Calculation. *Front. Bioeng. Biotechnol.* **2019**, *7*, 226. [CrossRef]
32. Database AIDPATH (Academia and Industry Collaboration for Digital Pathology). Available online: http://aidpath.eu/ (accessed on 1 November 2020).
33. Goshtasby, A.A. *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*; Wiley-Interscience: Hoboken, NJ, USA, 2005.
34. Deniz, O.; Toomey, D.; Conway, C.; Bueno, G. Multi-stained whole slide image alignment in digital pathology. In *SPIE Proceedings Medical Imaging 2015: Digital Pathology*; SPIE: Orlando, FL, USA, 2015; Volume 9420, p. 94200Z.
35. Markiewicz, T.; Osowski, S.; Patera, J.; Kozlowski, W. Image processing for accurate cell recognition and count on histologic slides. *Anal. Quant. Cytol. Histol.* **2006**, *28*, 281–291.
36. Reinhard, E.; Adhikhmin, M.; Gooch, B.; Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 34–41. [CrossRef]
37. *MATLAB (R2017a)*; Version 9.2.0; The MathWorks Inc.: Natick, MA, USA, 2010.

38. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc. (NIPS): Lake Tahoe, NV, USA, 2012; pp. 1097–1105.
39. Bueno, G.; Fernandez-Carrobles, M.M.; Gonzalez-Lopez, L.; Deniz, O. Glomerulosclerosis identification in whole slide images using semantic segmentation. *Comput. Methods Programs Biomed.* **2020**, *184*, 105273. [CrossRef] [PubMed]
40. Ruifrok, A.C.; Johnston, D.A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **2001**, *23*, 291–299.
41. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Syst. Man Cybern. Soc.* **1979**, *9*, 62–66. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).