

Article

Development of an Artificial-Intelligence-Based Tool for Automated Assessment of Cellularity in Bone Marrow Biopsies in Ph-Negative Myeloproliferative Neoplasms

Giuseppe D'Abbronzio ¹, Antonio D'Antonio ², Annarosaria De Chiara ³, Luigi Panico ⁴, Lucianna Sparano ⁵, Anna Diluvio ¹, Antonello Sica ⁶ , Gino Svanera ⁷, Renato Franco ^{1,8,*}  and Andrea Ronchi ^{1,8}

- ¹ Department of Mental and Physical Health and Preventive Medicine, Università degli Studi della Campania “Luigi Vanvitelli”, 80138 Naples, Italy; dabbronzogiuseppe@gmail.com (G.D.); anna.diluvio@gmail.com (A.D.); andrea.ronchi@unicampania.it (A.R.)
- ² Pathology Unit, Hospital “Ospedale del Mare”, 80147 Naples, Italy; ada66@inwind.it
- ³ Histopathology of lymphomas and Sarcoma SSD, Istituto Nazionale dei Tumori I.R.C.C.S. Fondazione “Pascale”, 80131 Naples, Italy; a.dechiara@istitutotumori.na.it
- ⁴ Pathology Unit, Hospital “Monaldi”, 80131 Naples, Italy; luigi.panico@ospedaledicolli.it
- ⁵ Pathology Unit, Hospital “Andrea Tortora”, 82100 Pagani, Italy; luciannasparano@gmail.com
- ⁶ Haematology and Oncology Unit, Vanvitelli Hospital, 80131 Naples, Italy; antonello.sica@fastwebnet.it
- ⁷ Haematology Unit, ASL Na2 North, 80014 Giugliano, Italy; gsvanemato@inwind.it
- ⁸ Pathology Unit, Vanvitelli Hospital, 80138 Naples, Italy
- * Correspondence: renato.franco@unicampania.it; Tel.: +39-0815664084



Citation: D'Abbronzio, G.; D'Antonio, A.; De Chiara, A.; Panico, L.; Sparano, L.; Diluvio, A.; Sica, A.; Svanera, G.; Franco, R.; Ronchi, A. Development of an Artificial-Intelligence-Based Tool for Automated Assessment of Cellularity in Bone Marrow Biopsies in Ph-Negative Myeloproliferative Neoplasms. *Cancers* **2024**, *16*, 1687. <https://doi.org/10.3390/cancers16091687>

Academic Editors:

Katja Pinker-Domenig and
Nandita deSouza

Received: 14 February 2024

Revised: 22 April 2024

Accepted: 24 April 2024

Published: 26 April 2024



Copyright: © 2024 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 (<https://creativecommons.org/licenses/by/4.0/>).

Simple Summary: In recent years, artificial intelligence has been used increasingly in medical practice. Technological progress has made it possible to digitize the slides of histological preparations, allowing the use of image processing and AI technologies in surgical pathology. This potentially reduces variability and improves the uniformity of some histological evaluations, such as cellularity assessment in bone marrow biopsies, which is traditionally performed visually by expert human observers, resulting in inter-observer and intra-observer variability. In this work, we developed an accurate AI-based tool for the automated quantification of cellularity in BMB histology and compared its performances with cellularity estimates from five expert hematopathologists. Our results showed the robustness of our model across users and two different scanners for digitized image generation.

Abstract: The cellularity assessment in bone marrow biopsies (BMBs) for the diagnosis of Philadelphia chromosome (Ph)-negative myeloproliferative neoplasms (MPNs) is a key diagnostic feature and is usually performed by the human eyes through an optical microscope with consequent inter-observer and intra-observer variability. Thus, the use of an automated tool may reduce variability, improving the uniformity of the evaluation. The aim of this work is to develop an accurate AI-based tool for the automated quantification of cellularity in BMB histology. A total of 55 BMB histological slides, diagnosed as Ph- MPN between January 2018 and June 2023 from the archives of the Pathology Unit of University “Luigi Vanvitelli” in Naples (Italy), were scanned on Ventana DP200 or Eprexia P1000 and exported as whole-slide images (WSIs). Fifteen BMBs were randomly selected to obtain a training set of AI-based tools. An expert pathologist and a trained resident performed annotations of hematopoietic tissue and adipose tissue, and annotations were exported as .tiff images and .png labels with two colors (black for hematopoietic tissue and yellow for adipose tissue). Subsequently, we developed a semantic segmentation model for hematopoietic tissue and adipose tissue. The remaining 40 BMBs were used for model verification. The performance of our model was compared with an evaluation of the cellularity of five expert hematopathologists and three trainees; we obtained an optimal concordance between our model and the expert pathologists' evaluation, with poorer concordance for trainees. There were no significant differences in cellularity assessments between two different scanners.

Keywords: digital pathology; artificial intelligence; cellularity; myeloproliferative neoplasms; bone marrow biopsy

1. Introduction

Artificial intelligence (AI) is increasingly being used in clinical practice and translational medical research. Various AI systems have been employed to improve the diagnosis and treatment of diseases such as diabetes [1] and cancer [2]. In surgical pathology, AI technologies have been facilitated by the ability to digitize slides. The first guideline for the production of whole-slide image (WSI) systems for clinical practice [3] was developed in 2016 by the Food and Drug Administration (FDA), who set out technical standards that ensured the production of acceptable images (varying color intensity, different cutting thickness) even in the presence of uneven tissue or tissue from different laboratories. Additionally, the WSI quality obtained from an approved scanner was required to have appropriate focus quality, color reproducibility, and spatial resolution to accurately reflect the quality of the original histology slides. Several commercial WSI and digital pathology systems have been approved by the FDA for diagnostic use and have received CE marks for in vitro diagnosis [4]. Digitized slides that meet these standards can be used to evaluate specific histological features such as the number of mitoses [5], biomarkers in breast tumor settings [6], or histological activity in ulcerative colitis [7]. Recently, an AI-supported diagnostic tool for prostate cancer based on digitized pathology slides (Paige Prostate) was approved by the FDA [8]. Despite these regulatory processes, variability in scanner performance has been shown to influence AI-based assessments [9,10].

Myeloproliferative neoplasms (MPNs) are a heterogeneous group of bone marrow disorders characterized by the abnormal proliferation of one or more myeloid cell lines, and include conditions like polycythemia vera (PV), essential thrombocythemia (ET), and primary myelofibrosis (PM) [11]. These diseases, collectively referred to as Philadelphia chromosome-negative (Ph-) MPNs, often require a complex diagnostic process involving clinical, laboratory, genetic, and histological data. In particular, the assessment of cellularity in bone marrow biopsies (BMBs) is a key diagnostic feature in MPNs [11]. However, the manual nature of the assessments, typically conducted through an optical microscope, results in significant inter-observer and intra-observer discrepancies [12–15]. Human-based cellularity assessments are further confounded by the presence of inflammation, necrosis, fibrosis, and specimen artifacts [16]. Consequently, concordance for MPN diagnoses, even between expert pathologists, ranges from 53%, based only on morphology, to 88% when paired with clinical and mutational data [17], and is poorly reproducible and dependent on pathologist experience [18,19]. Automated assessments of cellularity, like other computational pathology tools [20], would be advantageous not only due to providing objective assessments, but also due to improving laboratory workflow, reducing pathologist assessment time, and enhancing the training and skill development of young pathologists [21,22].

Prior work has explored automated assessment of BMBs in heterogeneous patient groups. Nielsen et al. [23] developed a segmentation system for BMB components, classifying them as yellow bone marrow (YBM) and red bone marrow (RBM), and then calculating cellularity in specific areas of interest in BMBs from eight subjects. Van Eekelen et al. [24] developed a segmentation system for different cell lines present in the bone marrow (erythropoiesis, myelopoiesis, megakaryocytes, lipocytes, trabecular bone, and erythrocytes), estimating the variation in cellularity in BMBs with age. Both studies achieved an optimal degree of agreement with a human-based evaluation of cellularity, proving that hematopoietic tissue can be quickly and easily distinguished from adipose tissue and digital image processing can provide an accurate and objective measurement of the examined region. However, they did not address the utility of such tools in a clinical scenario across multiple users with different levels of expertise or across different scanners for digitized image generation. The aim of this work was to develop and train an AI-based tool for the automated quantification of cellularity in patients with Ph- MPNs, validate it against multiple expert histopathologist assessments, and use it to evaluate the performance of pathology trainees. In addition, model usability was appraised by users and model performance was compared between digitized slides from two different scanners.

2. Materials and Methods

2.1. Cases Selection and Digital Images Scanning

We obtained bone marrow biopsy (BMB) histological slides diagnosed as Ph-MPN between January 2018 and June 2023 from the archives of the Pathology Unit of University “Luigi Vanvitelli” in Naples (Italy). The inclusion criteria were the following: age of the patient ≥ 18 years; biopsy frustule length ≥ 1 cm; availability of clinical data (age and histological and clinical diagnosis), and histological diagnosis of Ph-MPN. We chose BMBs with diagnoses of Ph-MPN for their significant variations in bone marrow cellularity. Conditions such as PV and pre-fibrotic PM typically present with increased cellularity, whereas other forms, such as fibrotic PM, can exhibit reduced cellularity [11].

A total of 86 BMBs were initially retrieved from archives and reviewed by an experienced hematopathologist (AR). Of these, 55 cases met the inclusion criteria and were included in the study. A histological slide stained with hematoxylin and eosin was scanned for each case with a Ventana DP200 Slide Scanner (Roche Diagnostics S.p.A., Monza, Italy) and exported as a whole-slide image (WSI). Subsequently, scanned slides were randomly divided into a training set, including 15 BMBs, and a validation set, including 40 BMBs.

2.2. Production of the Training Set

An expert pathologist (AR), assisted by a trained resident (GDA), annotated hematopoietic and adipose tissues on the 15 BMBs selected for the training set using QuPath v 4.3 software [25]. Then, annotations were exported as .tiff images (resolution 256×256 , $20\times$ magnification) and .png labels (Figure 1) with three main colors (black for hematopoietic tissue; yellow for adipose tissue; and white for regions that must be ignored during the training, such as bone tissue, red blood cells, artifacts, etc.). A specific script for Qupath [26] was used to export the annotation, providing 2617 images and 2617 labels that constituted our training set.

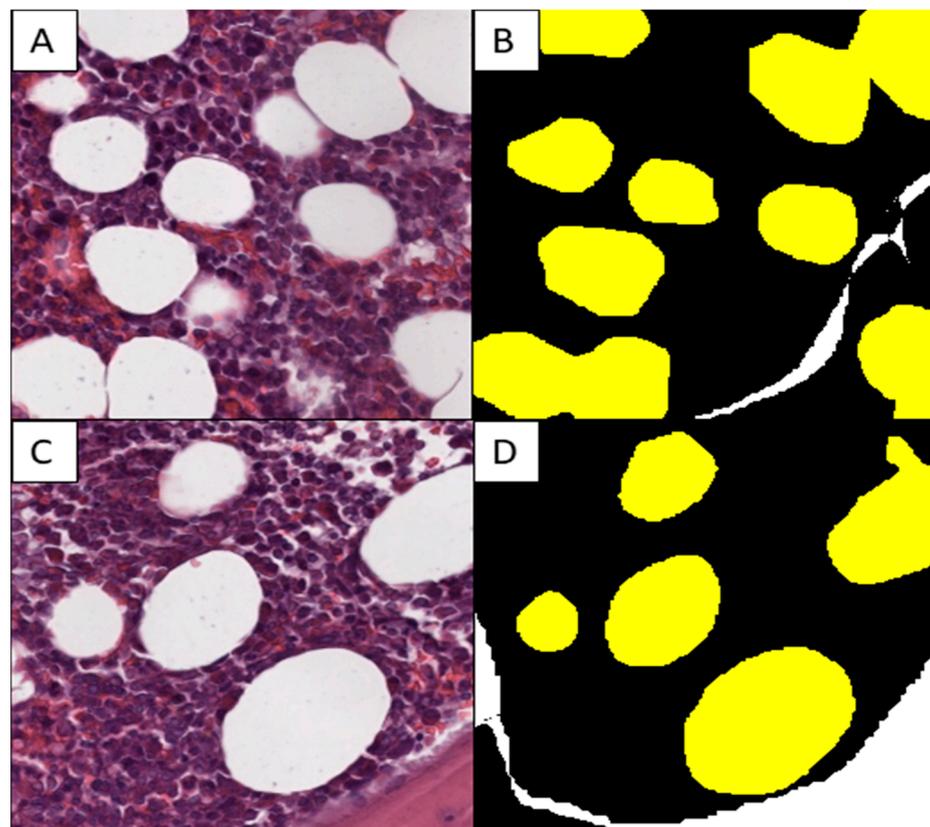


Figure 1. Example of a training set image with tiles ((A–C); H&E, $20\times$) and corresponding labels (B–D).

A semantic segmentation model for hematopoietic tissue and adipose tissue was developed using a fully convolutional network (FCN) based on InceptionV3 architecture, utilized up to the “mixed7” layer. This architecture was enhanced with a series of convolutional and deconvolutional layers, concluding with a 1×1 convolution for pixel-by-pixel classification using softmax activation. The model featured a total of nine main custom layers post-inception and was trained over 25 epochs using the Adam optimizer with categorical cross-entropy as the loss function.

2.3. Model Testing

Firstly, the cellularity of the 40 BMBs selected for the validation set was evaluated by 5 expert pathologists (RF, ADA, ADC, LP, LS) and 3 young pathologists (Resident1, Resident2, Resident3) using an optical microscope, as is routinely performed in clinical practice. Subsequently, two scripts for model application and tissue segmentation were developed. The first script was developed using the FAST library for Python [27] and segmented the WSI image into 256×256 patches at $20\times$ magnification. It applied the model to each patch, identifying adipose and hematopoietic tissues; extracted the corresponding labels; and reconstructed the slide with the segmentation results overlaid on the tissue (see Figure 2). The second script calculated the cellularity as the ratio of the area segmented as hematopoietic tissue to the total area of both hematopoietic and adipose tissues.

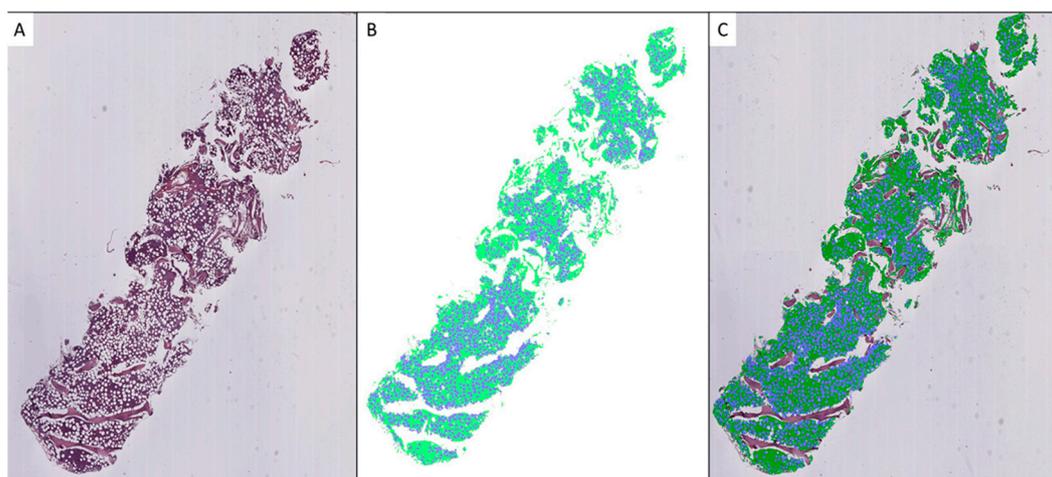


Figure 2. Workflow of segmentation process: WSI (A), segmentation result (B), WSI with overlaid segmentation image (C). Green indicates hematopoietic tissue and blue indicates adipose tissue.

$$\text{Cellularity} = \frac{\text{Segmented area of hematopoietic tissue}}{\text{Segmented area of hematopoietic tissue} + \text{Segmented area of adipose tissue}} \times 100\%$$

The agreement between the hematopathologists and the model, as well as between the residents and the model, was evaluated using Lin’s concordance correlation coefficient. Furthermore, agreement between the hematopathologist and the model was also evaluated using the Bland–Altman plot.

Six months later, the expert pathologist who annotated the training images (A.R.-user1), one of the expert pathologists who was compared with the model (R.F.-user2), and a resident (Resident2-user3) were trained to use the proposed model on a set of 20 cases from the BMB series. Subsequently, we asked these three users to rate the model on a scale from 1 to 5 (1: insufficient–5: optimal) in terms of usability (learning difficulties and autonomy in using the model), rapidity, and accuracy of evaluations. Finally, in order to test the performance of our model with different scanners, we randomly selected 5 BMBs from the validation set and rescanned them using a different scanner (EpreDia P1000) (EpreDia,

Portsmouth, NH, USA). We then compared the performance of our model with the WSIs obtained from both the Ventana DP200 and the EpreDia P1000 in terms of segmentation (assessed visually) and cellularity.

2.4. Statistical Analysis

To verify the effectiveness of our tool, we compared the performance of our model with the cellularity assessment performed by 5 pathologists (R.F.; A. D'antonio; A. De Chiara, L.S., and L.P.) with at least ten years of experience in the field of hematopathology. We asked them to express a single percentage value of cellularity in each BMB. Firstly, we measured the agreement between hematopathologists in pairs in order to test for interobserver variability in the assessment of cellularity; then, we measured the agreement between our model and each pathologist.

Furthermore, in order to show the poor reproducibility of evaluations of cellularity in BMBs, we asked medical residents (Resident 1, Resident 2, and Resident 3) to express their cellularity assessments on the same cases submitted to the expert pathologists, and we compared their evaluation with the performance of our model.

To measure agreement between the hematopathologists, between the hematopathologists and the model, and between the residents and the model, we chose to use Lin's concordance correlation coefficient (CCC) [28], which considers both the correlation of the assessments (Pearson correlation) and the average and variance of the measurements:

$$CCC : \frac{2p\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x + \mu_y)^2}$$

p : Pearson correlation coefficient

$\sigma_x^2; \sigma_y^2$: variance of measurement

$\mu_x; \mu_y$: average of measurement

In this way, the agreement between pathologists' model measurements was evaluated in terms of the evaluated percentages, penalizing both systematic and random differences.

According to Altman's definition [29], we considered optimal agreement to be characterized by a CCC greater than 0.8 and substantial agreement by a CCC between 0.6 and 0.8. Subsequently, in order to further evaluate the agreement between our model and the pathologists, we performed another statistical test using Bland–Altman plots.

Bland–Altman plots compare the averages of measurements, provide graphical representations of the agreement between the system of measurement, and identify any systematic trends [30]. In detail, agreement ranges were determined by calculating the averages and standard deviations of differences between measurements provided by the AI system and the pathologist. Then, we calculated the upper limit of agreement (ULA) and the lower agreement limit (LAL) according to the following formulas:

$$ULA = \bar{d} + 1.96 \times S_d$$

$$LAL = \bar{d} - 1.96 \times S_d$$

\bar{d} : average between measurement difference;

S_d : standard deviation of measurement difference.

The calculation was based on the assumption of a normal distribution of the differences between measurements, covering 95% of the expected differences between the measurements of the model and the measurements of the pathologists.

2.5. Hardware and Software

The annotations and classifications of the training set were carried out using QuPath software, version 4.3; the images and masks were exported using a specific script in Groovy language. All segmentation models were developed using Python 3.10.12 on the Google Colaboratory platform with a pro profile for the use of additional RAM and GPUs provided by the

platform (V100). The scripts for model verification, cellularity calculation, and statistical analysis were developed in Python 3.10.12 using the Anaconda® distribution, with Jupyter Notebook as the programming environment. All computer processing procedures were performed on an Alienware Aurora® R11 hardware (Alienware, Dell Inc., Round Rock, TX, USA) with an Intel(R) Core (TM) i5-10600KF CPU at 4.10 GHz, 64 GB of RAM, and an AMD Radeon RX 5700 XT graphics card with 8 GB of dedicated memory.

3. Results

3.1. Model Training

During the training process, our segmentation model showed a positive performance. It began with an initial loss of function of 0.6258, but quickly converged, showing a significant reduction in the loss of function with each subsequent epoch. In the 25th epoch, the loss had decreased to an impressive value of 0.0348. Simultaneously, the model's accuracy consistently increased, reaching 98.61% in the 25th epoch. This progress was also confirmed through visual verification with the pathologists' annotations, which demonstrated almost perfect correspondence (Figure 3). These results indicate not only the model's capability to effectively learn from the provided dataset, but also its stability, as evidenced by the absence of any overt signs of overfitting (Figure 4).

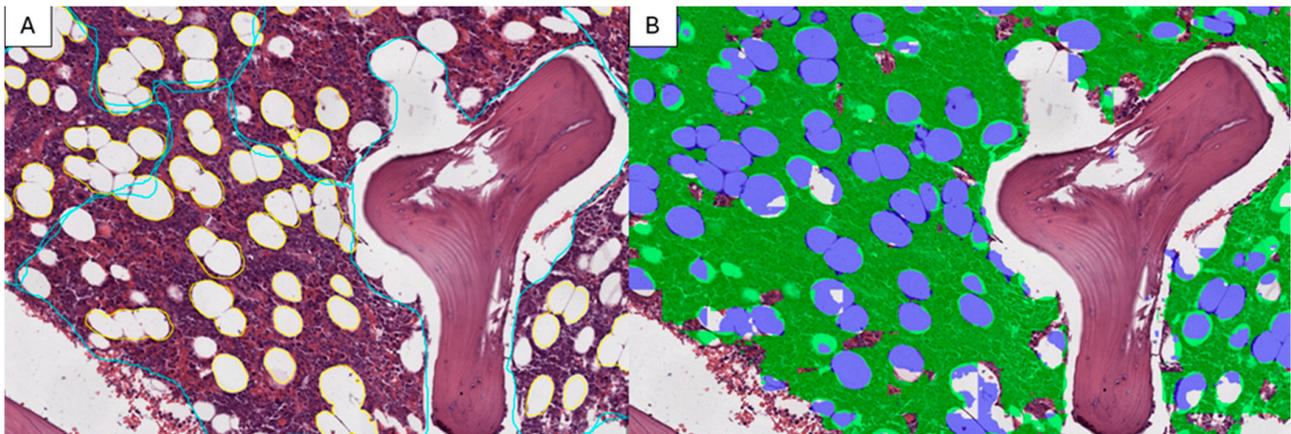


Figure 3. Comparison between pathologists' annotations ((A) cyan: hematopoietic tissue; yellow: adipose tissue) and the model's result ((B) green: hematopoietic tissue; blue: adipose tissue).

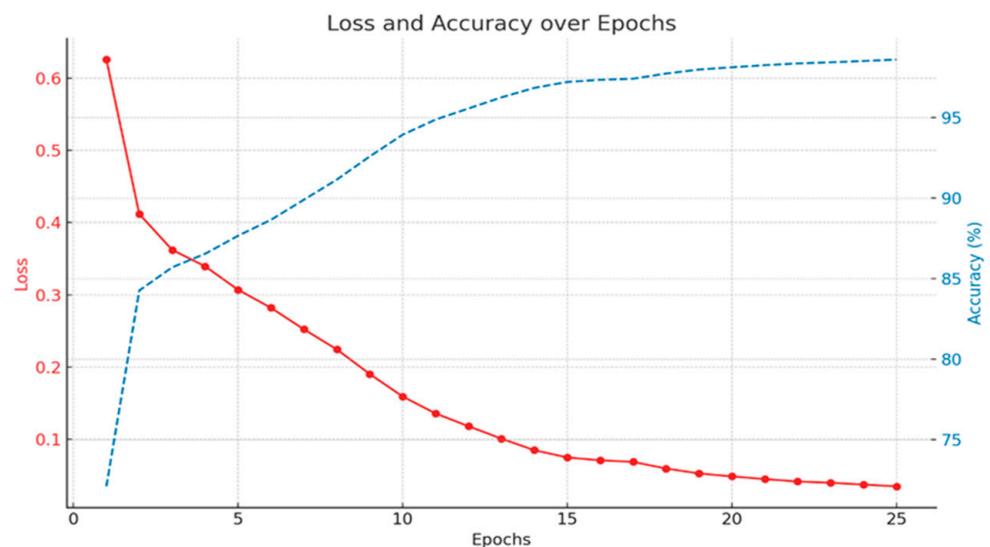


Figure 4. Graph of the training process: loss of function (red) and accuracy (blue).

3.2. Model Validation

The interobserver agreement between pathologists through Lin's concordance correlation coefficient was optimal in eight out of ten cases and substantial in the remaining two cases, as detailed in Table 1.

Table 1. Agreement between pathologists (B-F).

| Couple of Pathologist | Lin's Concordance Correlation Coefficient | Confidence Interval |
|-----------------------|---|---------------------|
| B-C | 0.887 | 0.789, 0.938 |
| B-D | 0.967 | 0.941, 0.982 |
| B-E | 0.886 | 0.788, 0.944 |
| B-F | 0.899 | 0.813, 0.944 |
| C-D | 0.877 | 0.784, 0.925 |
| C-E | 0.951 | 0.906, 0.973 |
| C-F | 0.773 | 0.663, 0.860 |
| D-E | 0.878 | 0.794, 0.929 |
| D-F | 0.902 | 0.715, 0.952 |
| E-F | 0.784 | 0.648, 0.873 |

According to Altman [29], the agreement between the model and pathologist B was optimal (CCC: 0.9400, IC: 0.8736, 0.9728). Similar results were obtained for the comparison between the model and pathologists D and F, whose CCCs were 0.9291 (IC: 0.8560, 0.9680) and 0.9170 (IC: 0.8813, 0.9672), respectively. Furthermore, the agreement between the model and pathologist C was substantial (CCC: 0.7918; IC: 0.6529, 0.8774), as well as agreement between the model and pathologist E (CCC: 0.7785; IC: 0.628, 0.8759). Finally, the agreement between the model and the mean of the pathologists was optimal, with CCC 0.9170 (IC: 0.8399, 0.9622). The results are summarized in Table 2.

Table 2. Agreement between model (A), each pathologist (B-F), and the mean of pathologists (G).

| | Lin's Concordance Correlation Coefficient (CCC) | Confidence Interval | |
|-----|---|---------------------|--------|
| A-B | 0.9400 | 0.8736 | 0.9728 |
| A-C | 0.7918 | 0.6529 | 0.8774 |
| A-D | 0.9291 | 0.8560 | 0.9680 |
| A-E | 0.7785 | 0.6288 | 0.8759 |
| A-F | 0.9386 | 0.8813 | 0.9672 |
| A-G | 0.9170 | 0.8399 | 0.9622 |

Similarly, the agreement between our model and the five expert pathologists, verified using Bland–Altman's plots, was also satisfactory overall. Indeed, in all comparisons, all bias lines were near 0, while all observation were within the agreement ranges (Figures 5–9).

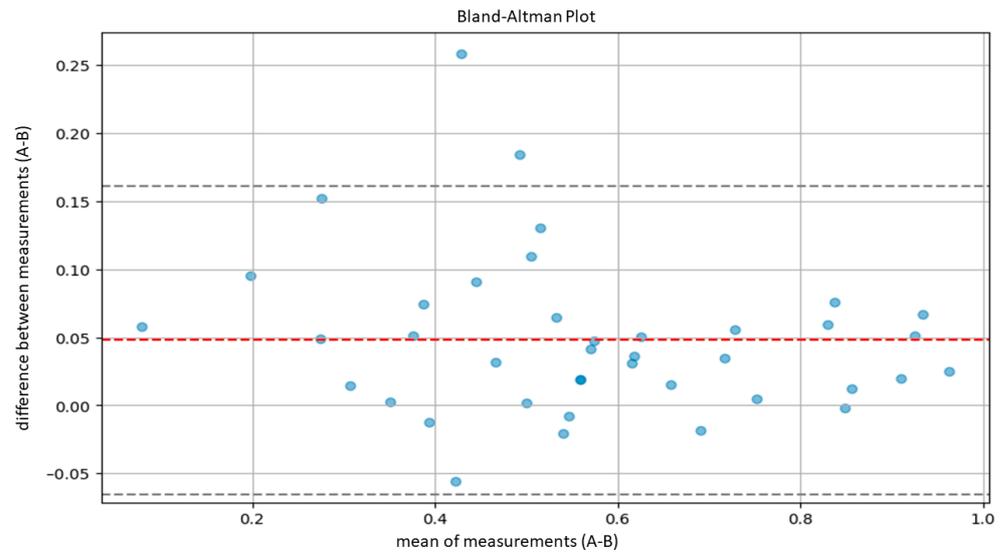


Figure 5. Bland–Altman plot to measure agreement between the model (A) and pathologist B.

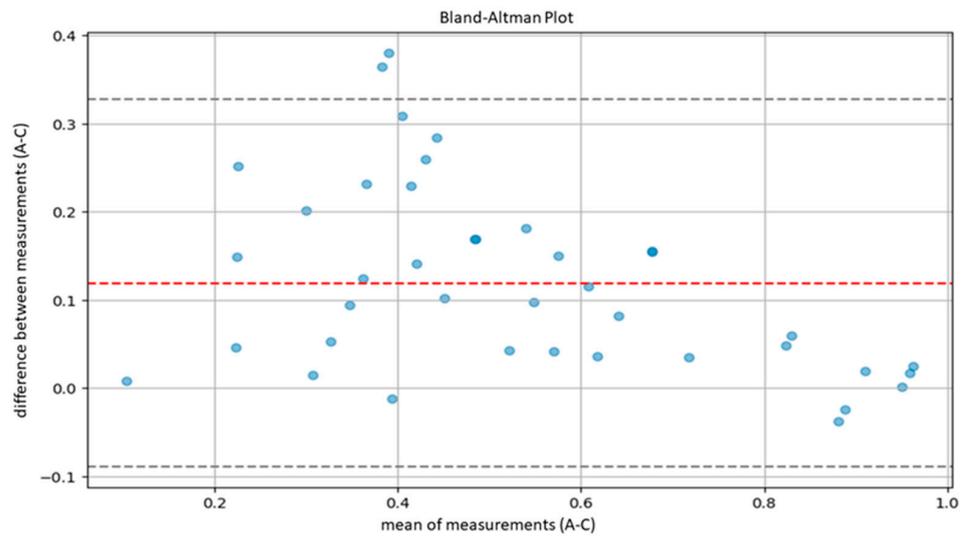


Figure 6. Bland–Altman plot to measure agreement between the model (A) and pathologist C.

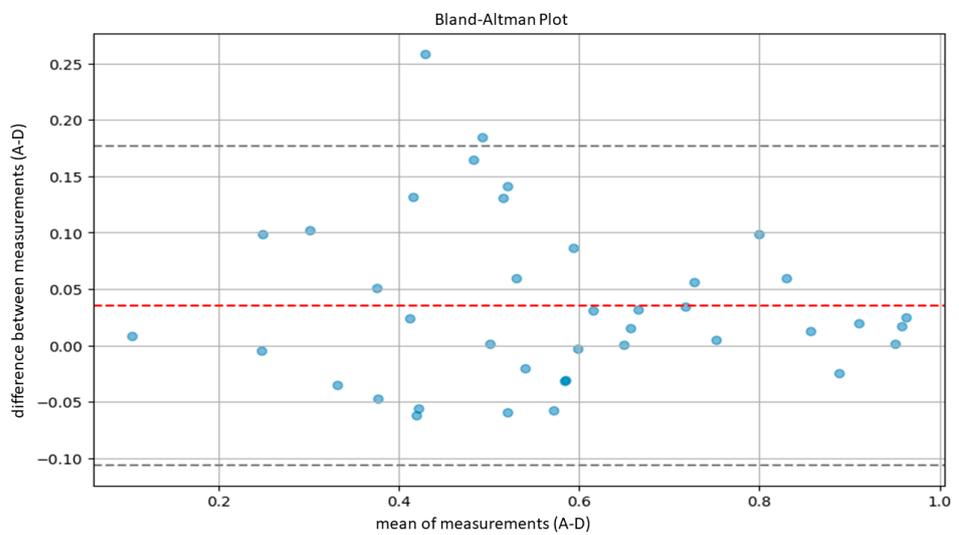


Figure 7. Bland–Altman plot to measure agreement between the model (A) and pathologist D.

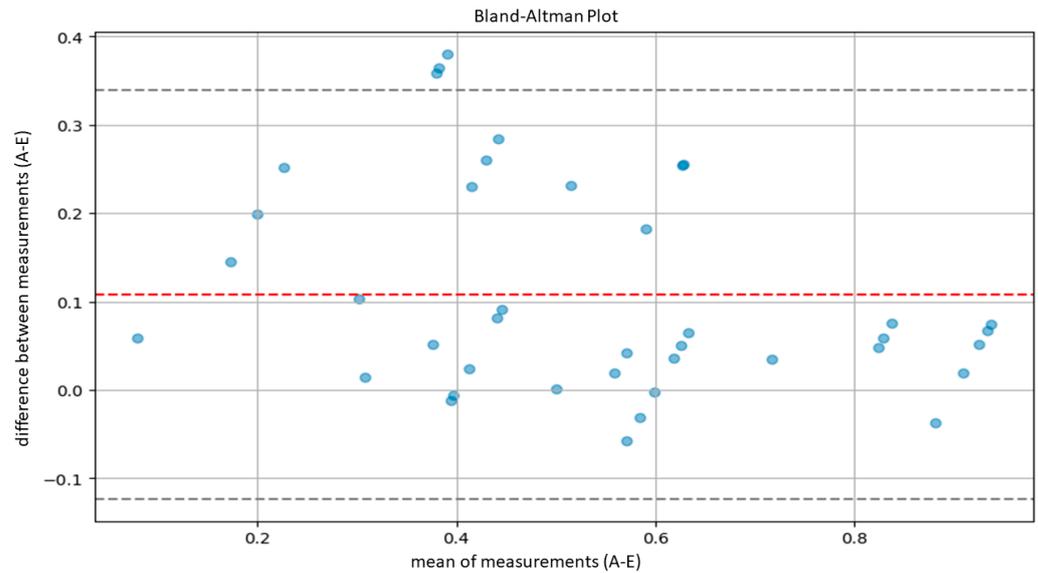


Figure 8. Bland–Altman plot to measure agreement between the model (A) and pathologist E.

Agreement between our model and the residents was not as satisfactory. In fact, in all cases, the agreement was only substantial (Table 3): the agreement between the model and Resident 1 was 0.659; the agreement between the model and Resident 2 was 0.721; and the agreement between the model and Resident 3 was 0.615, while the agreement between the model and the mean of the residents was 0.776.

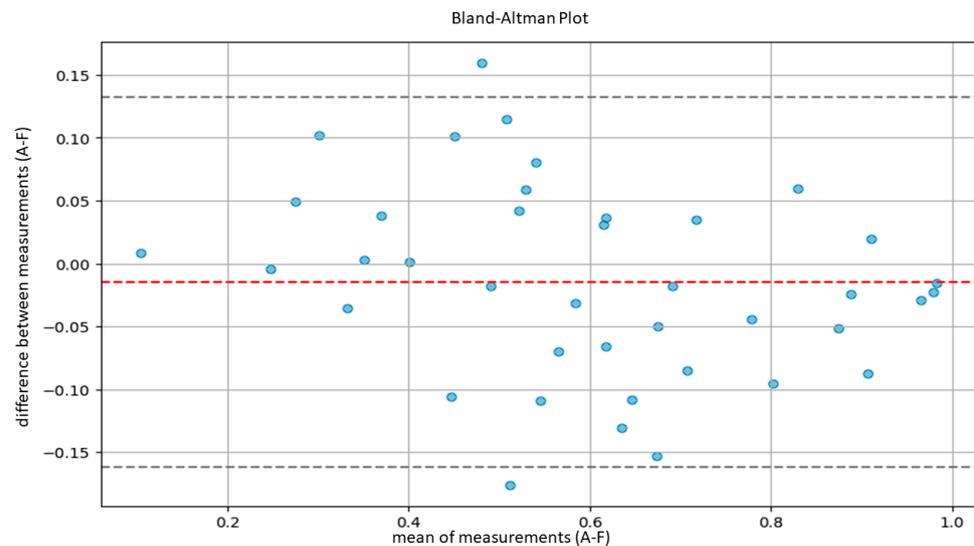


Figure 9. Bland–Altman plot to measure agreement between the model (A) and pathologist F.

Table 3. Agreement between model, each resident, and mean of residents.

| | Lin’s Concordance Correlation Coefficient (CCC) | Confidence Interval | |
|-------------------------|---|---------------------|-------|
| Model–Resident 1 | 0.659 | 0.508 | 0.771 |
| Model–Resident 2 | 0.721 | 0.546 | 0.845 |
| Model–Resident 3 | 0.615 | 0.432 | 0.755 |
| Model–mean of residents | 0.664 | 0.513 | 0.776 |

After approximately six months, two expert pathologists and a resident learned to use the proposed algorithm. Upon applying the algorithm to a set of 20 cases from the BOMs series, all three users agreed with the model's automated assessments and decided to adopt these evaluations as the definitive. This unanimous agreement on the model's evaluation effectively eliminated inter-observer variability.

Regarding the opinions of the users concerning usability, the rapidity of evaluation, and the correctness of evaluation of our model, the results are summarized in Table 4.

Table 4. Opinion of the users about the model (1: completely insufficient; 2: acceptable; 3: sufficient; 4: good; 5: optimal).

| | Usability | Rapidity | Correctness |
|-------|-----------|----------|-------------|
| User1 | 2 | 5 | 5 |
| User2 | 2 | 5 | 4 |
| User3 | 3 | 4 | 5 |

Finally, evaluation of the performance of our model across different scanners provided superimposable results both in terms of segmentation (Figure 10) and in terms of the cellularity assessment performed by our model (Table 5).

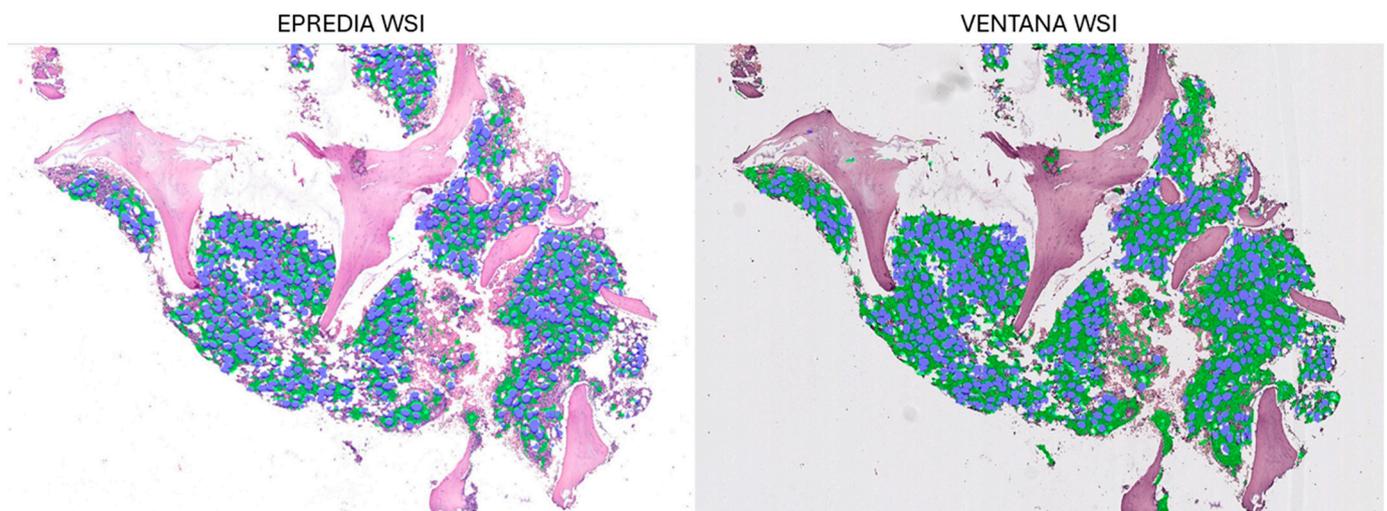


Figure 10. Details of a selected WSI for rescanning with the EpreDia P1000. A comparison between the segmentation results of the WSI processed by EpreDia and those processed by Ventana.

Table 5. Cellularity assessment performed by the model on EpreDia WSI and on Ventana WSI.

| | EpreDia WSI | Ventana WSI |
|---------|-------------|-------------|
| Slide 1 | 50.85% | 50.15% |
| Slide 2 | 64.98% | 65.05% |
| Slide 3 | 99.35% | 97.47% |
| Slide 4 | 58.62% | 59.14% |
| Slide 5 | 84.77% | 86.26% |

4. Discussion

Our study shows that the tool we developed quantifies the cellularity of BMBs equivalently to experienced pathologists, and allows for the assessment of the performance of trainee pathologists with good user acceptability. Regarding the technical aspects of our

tool, we chose to develop a fully convolutional network (FCN) that classifies every single pixel, thus allowing for precise differentiation between various types of tissues. Additionally, the use of the InceptionV3 architecture up to the “mixed7” layer provides a robust foundation for feature extraction, which has proven effective in processing histological images, particularly in the hematological field [31]. Furthermore, to ensure maximal reproducibility of our study, we trained our model using WSIs obtained from a slide scanner system (Ventana DP 200) approved for in vitro diagnostic use [4].

Unfortunately, despite high concordance among experienced hematopathologists, our results showed that inter-observer variability was evident, with a wide range in Lin’s coefficient. This may in part be explained by the fact that hematopathologists typically express assessments of bone marrow cellularity in increments of 5% (e.g., 10%, 15%, 20%, etc.), while the AI system provides a precise numerical value to the second decimal place. Consequently, the concordance between hematopathologists, as well as between hematopathologists and the model, could be influenced by this specific procedure of evaluation performed by the hematopathologist, which represents a more generalized and “rounded” assessment value.

Regarding the agreement between the AI-based tool and the human eye, our results were optimal. The models demonstrated high concordance with the assessments of five expert hematopathologists. Additionally, the concordance between the AI model and the average assessments of the five hematopathologists was also optimal. Furthermore, evaluations using Bland–Altman analysis were substantially positive across all comparisons, with a bias line near zero and most measurements falling within the limits of concordance. In addition, the lower agreement observed between the model and the residents indicates that, for unexperienced pathologists, cellularity assessment can be less reproducible.

Despite the involvement of a higher number of observers in our study (five expert pathologists and three residents), our results are consistent with the existing literature. In their respective studies, Nielsen et al. [23] and van Eekelen et al. [24] developed automated tools for assessing cellularity in BMBs, achieving optimal concordance with the human observations. However, their tests were limited to single observer, whereas our model’s performance was compared with that of five expert hematopathologists. Sarkis et al. [32] also compared the performance of a model for BMB cellularity assessment with that of four pathologists, obtaining a high level of concordance with all four. However, the model developed by Sarkis et al. required semi-automated interaction, necessitating user involvement for performing a white color balance, making three annotations (for the tissue to be evaluated), and selecting the background and artifacts before analyzing the entire BMB.

A strength of our work is that, following training and a lag time of 6 months, when we asked three users to apply the model to 20 randomly selected BMBs, all concurred with the assessments provided by the model and were willing to adopt the assessments as definitive. This unanimous agreement to adopt the model removes the inter-observer variability of human assessments.

Evaluation of the model’s usability (usability (learning difficulties and autonomy in using the model) indicated that, while users found the rapidity and accuracy of the model to be favorable, they were not satisfied with its general usability. These findings align with those from previous studies evaluating pathologist-reviewed models. In Dy et al.’s work [33], which assessed a tool for KI 67 evaluation in breast cancer, and Steiner et al.’s study [34], which evaluated a tool for prostate biopsy assessments, users consistently praised the models’ accuracy and supported their integration into clinical practice. However, neither study addressed the models’ usability or the learning curves associated with their use. It is important to note that both studies developed user interfaces that allowed pathologists to easily select cases for examination.

Since previous studies have demonstrated how the optical and computed properties of WSIs are affected by technical differences in slide scanners, potentially influencing the performance of AI-based tools [9], we found it necessary to evaluate our model’s performance using WSIs obtained from different scanners. Duenweg et al. [10] demonstrated

that the optical and computed properties of WSI from different scanners can affect AI-based tools, particularly when applied to low-resolution images (i.e., low magnification). In our study, we worked (for both training and validation of the model) on patches obtained at an intermediate magnification (20×), which ensured that the performance of our model was not adversely affected by the differing properties of the two scanners. Therefore, unlike these previous data, where model performance was influenced by the scanner type, our data confirm that the model's performance was remarkably consistent, in terms of both the method of segmentation and the cellularity assessment.

A major limitation of our model was usability by clinicians, particularly those with little computer experience, and particularly when under time pressure. It would benefit from a user-friendly interface that requires the pathologist to simply select the case to be analyzed. Moreover, our model's validation was conducted in a single center, which may limit the generalizability of our findings. For real use in clinical practice, our model should be tested in a multicenter study in order to test its reproducibility and its efficacy on a larger number of cases under different conditions of use (hardware, network, clinical routine, etc.). The sole assessment of cellularity is also limiting. The model's utility could be enhanced if other important features of Ph-MPNs could be added, such as individuation of the proliferation of one or more cellular lines, evaluation of increasing numbers, and morphological details of the megakaryocytes and of a shift in maturation and BM fibrosis. However, automated assessment of these features is more complex than cellularity evaluation; therefore, future studies would require the development of more complex neural networks capable of carrying out this type of evaluation.

5. Conclusions

In conclusion, our automated system for assessing cellularity in BMBs compared favorably with the evaluations by five experienced pathologists using an optical microscope, demonstrating optimal results and the ability to eliminate interobserver variability in the evaluation of cellularity from BMBs of Ph-MPNs. The low variability, accuracy, and speed of our model potentially make it a useful adjunct to expert histopathologist assessments, reducing the time needed for quantitative assessments and acting as a learning resource for pathology trainees.

Author Contributions: G.D., A.R. and R.F. conceptualized the study; A.D. (Anna Diluvio), A.S. and G.S. collected the clinical data; A.R., G.D. and A.D. (Anna Diluvio) performed preliminary analysis of the WSIs and annotated training images; G.D. wrote the code and performed statistical analysis; R.F., A.D. (Antonio D'Antonio), A.D.C., L.P. and L.S. analyzed WSIs for cellularity assessment; G.D. and A.R. wrote the manuscript; R.F., A.S. and G.S. reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and with the approval of the Internal Review Board of the I.N.T. Pascale (Approval Code: CEI 556/10. Approval Date: 12 March 2010).

Informed Consent Statement: Images for the training and test sets are anonymous; however, informed consent for the use of biological material for scientific purposes was preliminarily obtained from the patients.

Data Availability Statement: All codes and data are available upon reasonable request by email to the corresponding author.

Acknowledgments: Thanks to Stefano Lucà, Emma Carrautro, and Alessandro Cioce for their essential and effective contributions.

Conflicts of Interest: All the authors declare that they have no conflicts of interest.

References

1. Angrisani, L.; Annuzzi, G.; Arpaia, P.; Bozzetto, L.; Cataldo, A.; Corrado, A.; De Benedetto, E.; Di Capua, V.; Prevete, R.; Vallefucio, E. Neural Network-Based Prediction and Monitoring of Blood Glucose Response to Nutritional Factors in Type-1 Diabetes. In Proceedings of the 2022 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Ottawa, ON, Canada, 16–19 May 2022; pp. 1–6.
2. Nartowt, B.J.; Hart, G.R.; Muhammad, W.; Liang, Y.; Stark, G.F.; Deng, J. Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification. *Front. Big Data* **2020**, *3*, 6. [CrossRef] [PubMed]
3. Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices—Guidance for Industry and Food and Drug Administration Staff. Available online: <https://www.fda.gov/media/90791/download?attachment> (accessed on 17 March 2024).
4. García-Rojo, M.; De Mena, D.; Muriel-Cueto, P.; Atienza-Cuevas, L.; Domínguez-Gómez, M.; Bueno, G. New European Union Regulations Related to Whole Slide Image Scanners and Image Analysis Software. *J. Pathol. Inform.* **2019**, *10*, 2. [CrossRef]
5. Pantanowitz, L.; Hartman, D.; Qi, Y.; Cho, E.Y.; Suh, B.; Paeng, K.; Dhir, R.; Michelow, P.; Hazelhurst, S.; Song, S.Y.; et al. Accuracy and Efficiency of an Artificial Intelligence Tool When Counting Breast Mitoses. *Diagn. Pathol.* **2020**, *15*, 80. [CrossRef] [PubMed]
6. Whitney, J.; Corredor, G.; Janowczyk, A.; Ganesan, S.; Doyle, S.; Tomaszewski, J.; Feldman, M.; Gilmore, H.; Madabhushi, A. Quantitative Nuclear Histomorphometry Predicts Oncotype DX Risk Categories for Early Stage ER+ Breast Cancer. *BMC Cancer* **2018**, *18*, 610. [CrossRef]
7. Najdawi, F.; Sucipto, K.; Mistry, P.; Hennek, S.; Jayson, C.K.B.; Lin, M.; Fahy, D.; Kinsey, S.; Wapinski, I.; Beck, A.H.; et al. Artificial Intelligence Enables Quantitative Assessment of Ulcerative Colitis Histology. *Mod. Pathol.* **2023**, *36*, 100124. [CrossRef]
8. Food and Drug Administration Approval for PAIG Prostate. Available online: <https://www.fda.gov/news-events/press-announcements/fda-authorizes-software-can-help-identify-prostate-cancer> (accessed on 1 March 2024).
9. Shrestha, P.; Kneepkens, R.; Vrijnsen, J.; Vossen, D.; Abels, E.; Hulsken, B. A Quantitative Approach to Evaluate Image Quality of Whole Slide Imaging Scanners. *J. Pathol. Inform.* **2016**, *7*, 56. [CrossRef] [PubMed]
10. Duenweg, S.R.; Bobholz, S.A.; Lowman, A.K.; Stebbins, M.A.; Winiarz, A.; Nath, B.; Kyereme, F.; Iczkowski, K.A.; LaViolette, P.S. Whole Slide Imaging (WSI) Scanner Differences Influence Optical and Computed Properties of Digitized Prostate Cancer Histology. *J. Pathol. Inform.* **2023**, *14*, 100321. [CrossRef]
11. WHO Classification of Tumours Editorial Board. *Haematolymphoid Tumors*, 5th ed.; WHO Classification of Tumours Series; [Internet; beta version ahead of print]; International Agency for Research on Cancer: Lyon, France, 2022; Volume 11. Available online: <https://tumourclassification.iarc.who.int/chapters/63> (accessed on 22 July 2023).
12. Wilkins, B.S.; Erber, W.N.; Bareford, D.; Buck, G.; Wheatley, K.; East, C.L.; Paul, B.; Harrison, C.N.; Green, A.R.; Campbell, P.J. Bone Marrow Pathology in Essential Thrombocythemia: Interobserver Reliability and Utility for Identifying Disease Subtypes. *Blood* **2008**, *111*, 60–70. [CrossRef]
13. Pozdnyakova, O.; Rodig, S.; Bhandarkar, S.; Wu, K.; Thiele, J.; Hasserjian, R. The Importance of Central Pathology Review in International Trials: A Comparison of Local versus Central Bone Marrow Reticulin Grading. *Leukemia* **2015**, *29*, 241–244. [CrossRef]
14. Kvasnicka, H.M.; Thiele, J.; Bueso-Ramos, C.E.; Sun, W.; Cortes, J.; Kantarjian, H.M.; Verstovsek, S. Long-Term Effects of Ruxolitinib versus Best Available Therapy on Bone Marrow Fibrosis in Patients with Myelofibrosis. *J. Hematol. Oncol.* **2018**, *11*, 42. [CrossRef]
15. Ng, W.Y.; Erber, W.N.; Grigg, A.; Dunne, K.; Perkins, A.; Forsyth, C.; Ross, D.M. Variability of Bone Marrow Biopsy Reporting Affects Accuracy of Diagnosis of Myeloproliferative Neoplasms: Data from the ALLG MPN01 Registry. *Pathology* **2023**, *56*, 75–80. [CrossRef] [PubMed]
16. Hebeda, K. Necrosis, Stromal Changes and Artefacts. In *Diagnostic Bone Marrow Haematopathology*; Cambridge University Press: Cambridge, UK, 2020; pp. 26–41.
17. Madelung, A.B.; Bondo, H.; Stamp, I.; Loevgreen, P.; Nielsen, S.L.; Falensteen, A.; Knudsen, H.; Ehinger, M.; Dahl-Sørensen, R.; Mortensen, N.B.; et al. World Health Organization-defined Classification of Myeloproliferative Neoplasms: Morphological Reproducibility and Clinical Correlations—The Danish Experience. *Am. J. Hematol.* **2013**, *88*, 1012–1016. [CrossRef] [PubMed]
18. Alvarez-Larrán, A.; Ancochea, A.; García, M.; Climent, F.; García-Pallarols, F.; Angona, A.; Senín, A.; Barranco, C.; Martínez-Avilés, L.; Serrano, S.; et al. WHO-histological Criteria for Myeloproliferative Neoplasms: Reproducibility, Diagnostic Accuracy and Correlation with Gene Mutations and Clinical Outcomes. *Br. J. Haematol.* **2014**, *166*, 911–919. [CrossRef] [PubMed]
19. Buhr, T.; Hebeda, K.; Kaloutsi, V.; Porwit, A.; Van der Walt, J.; Kreipe, H. European Bone Marrow Working Group Trial on Reproducibility of World Health Organization Criteria to Discriminate Essential Thrombocythemia from Prefibrotic Primary Myelofibrosis. *Haematologica* **2012**, *97*, 360–365. [CrossRef] [PubMed]
20. Kim, I.; Kang, K.; Song, Y.; Kim, T.-J. Application of Artificial Intelligence in Pathology: Trends and Challenges. *Diagnostics* **2022**, *12*, 2794. [CrossRef] [PubMed]
21. Sarwar, S.; Dent, A.; Faust, K.; Richer, M.; Djuric, U.; Van Ommeren, R.; Diamandis, P. Physician Perspectives on Integration of Artificial Intelligence into Diagnostic Pathology. *NPJ Digit. Med.* **2019**, *2*, 28. [CrossRef] [PubMed]
22. Hassell, L.A.; Absar, S.F.; Chauhan, C.; Dintzis, S.; Farver, C.F.; Fathima, S.; Glassy, E.F.; Goldstein, J.A.; Gullapalli, R.; Ho, J.; et al. Pathology Education Powered by Virtual and Digital Transformation: Now and the Future. *Arch. Pathol. Lab. Med.* **2023**, *147*, 474–491. [CrossRef]

23. Nielsen, F.S.; Pedersen, M.J.; Olsen, M.V.; Larsen, M.S.; Røge, R.; Jørgensen, A.S. Automatic Bone Marrow Cellularity Estimation in H&E Stained Whole Slide Images. *Cytom. Part A* **2019**, *95*, 1066–1074. [[CrossRef](#)] [[PubMed](#)]
24. van Eekelen, L.; Pinckaers, H.; van den Brand, M.; Hebeda, K.M.; Litjens, G. Using Deep Learning for Quantification of Cellularity and Cell Lineages in Bone Marrow Biopsies and Comparison to Normal Age-Related Variation. *Pathology* **2022**, *54*, 318–327. [[CrossRef](#)]
25. 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. *Sci. Rep.* **2017**, *7*, 16878. [[CrossRef](#)]
26. Pettersen, H.S.; Belevich, I.; Røyset, E.S.; Smistad, E.; Simpson, M.R.; Jokitalo, E.; Reinertsen, I.; Bakke, I.; Pedersen, A. Code-Free Development and Deployment of Deep Segmentation Models for Digital Pathology. *Front. Med.* **2022**, *8*, 816281. [[CrossRef](#)] [[PubMed](#)]
27. Smistad, E.; Bozorgi, M.; Lindseth, F. FAST: Framework for Heterogeneous Medical Image Computing and Visualization. *Int. J. Comput. Assist. Radiol. Surg.* **2015**, *10*, 1811–1822. [[CrossRef](#)] [[PubMed](#)]
28. Lin, L.I. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255–268. [[CrossRef](#)] [[PubMed](#)]
29. Altman, D.G. *Practical Statistics for Medical Research*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1990; ISBN 9780429258589.
30. Bland, J.M.; Altman, D.G. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* **1986**, *1*, 307–310. [[CrossRef](#)]
31. Guan, Q.; Wan, X.; Lu, H.; Ping, B.; Li, D.; Wang, L.; Zhu, Y.; Wang, Y.; Xiang, J. Deep Convolutional Neural Network Inception-v3 Model for Differential Diagnosing of Lymph Node in Cytological Images: A Pilot Study. *Ann. Transl. Med.* **2019**, *7*, 307. [[CrossRef](#)] [[PubMed](#)]
32. Sarkis, R.; Burri, O.; Royer-Chardon, C.; Schyrr, F.; Blum, S.; Costanza, M.; Cherix, S.; Piazzon, N.; Barcena, C.; Bisig, B.; et al. MarrowQuant 2.0: A Digital Pathology Workflow Assisting Bone Marrow Evaluation in Experimental and Clinical Hematology. *Mod. Pathol.* **2023**, *36*, 100088. [[CrossRef](#)] [[PubMed](#)]
33. Dy, A.; Nguyen, N.-N.J.; Meyer, J.; Dawe, M.; Shi, W.; Androustos, D.; Fyles, A.; Liu, F.-F.; Done, S.; Khademi, A. AI Improves Accuracy, Agreement and Efficiency of Pathologists for Ki67 Assessments in Breast Cancer. *Sci. Rep.* **2024**, *14*, 1283. [[CrossRef](#)]
34. Steiner, D.F.; Nagpal, K.; Sayres, R.; Foote, D.J.; Wedin, B.D.; Pearce, A.; Cai, C.J.; Winter, S.R.; Symonds, M.; Yatziv, L.; et al. Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies. *JAMA Netw. Open* **2020**, *3*, e2023267. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.