applied sciences

MDPI

Article

Fast Pre-Diagnosis of Neoplastic Changes in Cytology Images Using Machine Learning

Jakub Caputa [1], Daria Łukasik [1], Maciej Wielgosz [1,2,*], Michał Karwatowski [1,2], Rafał Frączek [1,2], Paweł Russek [1,2] and Kazimierz Wiatr [1,2]

[1] Academic Computer Centre CYFRONET AGH, ul. Nawojki 11, 30-072 Kraków, Poland; jjakubcaputa@gmail.com (J.C.); daria.lukasik.vet@gmail.com (D.Ł.); mkarwat@agh.edu.pl (M.K.); rafalfr@agh.edu.pl (R.F.); russek@agh.edu.pl (P.R.); wiatr@agh.edu.pl (K.W.)

[2] Institute of Electronics, AGH University of Science and Technology, al. Adama Mickiewicza 30, 30-059 Kraków, Poland

* Correspondence: wielgosz@agh.edu.pl

Abstract: We present the experiment results to use the YOLOv3 neural network architecture to automatically detect tumor cells in cytological samples taken from the skin in canines. A rich dataset of 1219 smeared sample images with 28,149 objects was gathered and annotated by the vet doctor to perform the experiments. It covers three types of common round cell neoplasms: mastocytoma, histiocytoma, and lymphoma. The dataset has been thoroughly described in the paper and is publicly available. The YOLOv3 neural network architecture was trained using various schemes involving original dataset modification and the different model parameters. The experiments showed that the prototype model achieved 0.7416 mAP, which outperforms the state-of-the-art machine learning and human estimated results. We also provided a series of analyses that may facilitate ML-based solutions by casting more light on some aspects of its performance. We also presented the main discrepancies between ML-based and human-based diagnoses. This outline may help depict the scenarios and how the automated tools may support the diagnosis process.

Keywords: canines; neoplasms; detection; deep learning; YOLOv3

Citation: Caputa, J.; Łukasik, D.; Wielgosz, M.; Karwatowski, M.; Frączek, R.; Russek, P.; Wiatr, K. Fast Pre-Diagnosis of Neoplastic Changes in Cytology Images Using Machine Learning. *Appl. Sci.* **2021**, *11*, 7181. https://doi.org/10.3390/app11167181

Academic Editor: Keun Ho Ryu

Received: 10 July 2021
Accepted: 30 July 2021
Published: 4 August 2021

1. Introduction

Veterinary oncology is a medical science field in which a precise diagnosis of the examined physical condition before introducing treatment is crucial for its effects and allows a doctor for a reasonable decision to be taken regarding the treatment of an oncological patient.

Many diagnostic methods (including clinical examination, imaging tests, or endoscopic examination) allow examiners for excellent visualization of the lesions, as well as to recognize their structure, size, number, and features of clinical malignancy (rapid growth, large volume of lesion, binding to oral tissues), invasive, infiltrative nature of growth, and destruction of adjacent structures [1]. A microscopic examination of tissue samples allows us to determine actual tumor nature [1].

Unfortunately, in many cases, the process involves having samples shipped to a dedicated laboratory for the examination, requiring additional time that may contribute to the deterioration of the patient's state. Alternatively, the microscopic examination may be conducted by a site physician. However, this approach is prone to errors, and not all doctors have appropriate training. Automating the initial diagnosis process using artificial intelligence methods (AI) may help make the process smoother.

Consequently, this paper aims to present a performance of the proof of concept system designed for automatic detection of neoplastic cells in cytology samples from canine skin tumors. The system is meant to ease the work of veterinarians and significantly shorten the time of diagnosis. The goal was also to prepare the auxiliary software tools that would

allow the veterinarian to perform an initial examination of the smeared cytology sample and support the traditional analysis (Figure 1). Three types of round cell neoplasms: mastocytoma, histiocytoma, and lymphoma, are covered in this work.

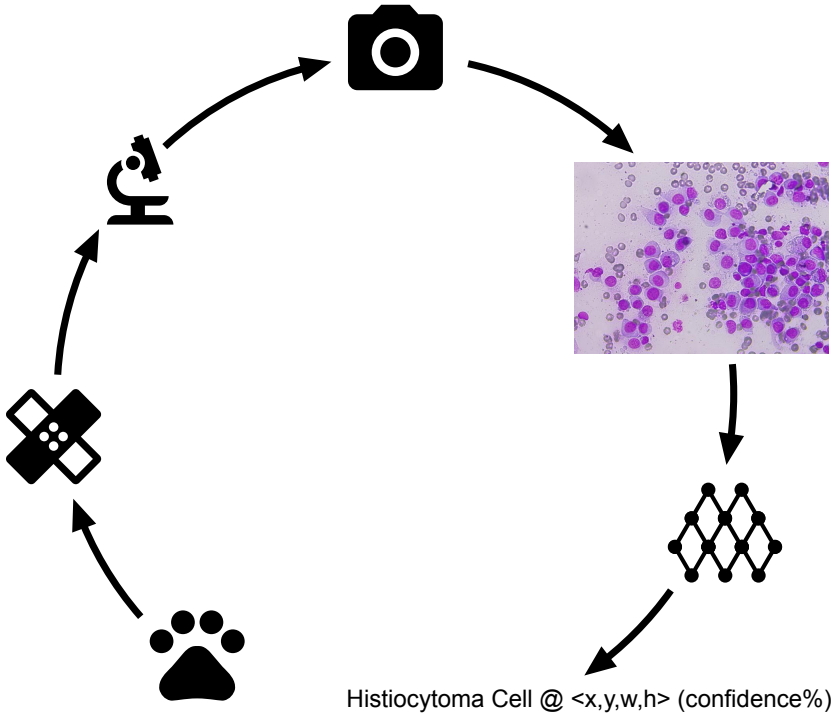


Figure 1. The basic stages of the experiment: obtaining the data, training the model, and getting the result. Graphic uses https://material.io/resources/icons (accessed on 10 March 2021).

The article is structured as follows: in Section 2, we introduce general issues related to the image classification and detection of objects; we present the architecture of a neural network used in our experiments and briefly describe the prepared custom dataset; we outline the metrics we used to assess the proposed model, and we give the essentials of the implementation of our detection system. The following Section 3 contains the experiments' description and obtained results. Finally, Section 4 compares various aspects of using machine learning algorithms with human performance to help physicians adopt this new technology.

2. Materials and Methods

2.1. State-of-the-Art Description

Up to date, several projects addressed the challenge of harnessing machine learning methods for cytological image analysis. They covered an array of tissue types, such as skin or lungs. Some methods employ algorithms created before the deep learning era [2]. The classical machine learning approaches involve the feature extraction step, which demands both parties, i.e., the medical practitioner and the ML researcher, to thoroughly comprehend features engineering requirements from both perspectives of applied algorithms and medicine.

Deep learning-based approaches do not need handcrafted features, only well-labeled data, and, therefore, the domain knowledge overlap does not need to be so complete as in traditional approaches. Another advantage of DL is adapting to different data types, such as training the model using a new, slightly different dataset or adding a new class when such a need occurs. As we focus on the deep learning domain, we provided few related sample research cases in Table 1.

Table 1. The summary of state-of-the-art in machine learning for cytological image analysis

	Title	Domain	Method	Metrics	Results
[3]	Classification of breast cancer cytological specimen using convolutional neural network	classification	CNN	accuracy	83%
[4]	Nasal cytology with deep learning techniques	classification	CNN	accuracy	94%
[5]	Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks	classification	CNN	accuracy	71%
[6]	Deep Learning-Based Quantification of Pulmonary Hemosiderophages in Cytology Slides	detection	RetinaNet	mAP	0.66
			human experts	hypothetical mAP	0.57–0.74
[7]	Automated Pap Smear Cervical Cancer Screening Using Deep Learning	detection	Mask R-CNN	mAP	0.57

The authors of [3] used two CNN architectures: GoogleNet and AlexNet. Their collection included samples from 50 patients with diagnosed cancer (25 malignant tumors and 25 benign tumors). As input images featured a very high resolution, they were divided into blocks of 256 × 256 pixels. The influence on the result of the concentration of the cells in the image was also checked. The primary metric for evaluating the model was the effectiveness, defined as the percentage of the correctly classified fragments of the photo taken from the entire area, which initial size was 200,000 × 100,000 pixels. The best results were achieved by the GoogleNet network, using images of the highest density. It was also found that the network was not efficient in diagnosing the malignant tumor.

The authors of [4] created their architecture based on the Agarap model. The process can be divided into two stages. The first was to extract cells from images and then to classify them. Contrary to the work mentioned above, it was not a binary classification problem, as there were seven types of cells in the dataset. Eventually, the pathologist selected 3423 objects with a size of 50 × 50 pixels.

The article [5] presents a classification for the three types of lung cancer. Approximately 300 photos of these tumors were collected. They were resized to the images of size 256 × 256 pixels. Additionally, an image-based augmentation was used. The proposed neural network architecture consisted of three convolutional and pooling layers and two fully connected layers. The authors emphasized that they were satisfied with the obtained results because the types of cancer they have selected were difficult for classification, even for pathologists.

The paper [6] presents research based on the 17 cytological slides of equine bronchoalveolar lavage fluid. The wholly annotated dataset contained 78,047 labeled macrophage cells. For the detection, Retina-Net was used, and the results were compared to Faster-RCNN with ResNet-50 backbone and SSD. The goal was to detect objects and determine the call's grade. The best results were achieved with the Retina-Net architecture.

The last studied article [7] applied a pre-trained Mask R-CNN for cervical cancer nuclei detection, segmentation, and classification into normal and abnormal ones. The authors used liquid-based histological slides.

Several ideas were used in these works, such as dividing a picture into smaller ones or augmenting a dataset. A detection algorithm has been chosen to avoid processing the photo before being fed into the network (single-cell extraction).

Contrary to the cited works that use the classification after detection approach, we propose a single-step method thanks to the you only look once version 3 (YOLOv3) neural network [8]. The YOLOv3 network finds the objects and then classifies them in a single pass. There is no need for extra image preparation. Additionally, this approach allows

us to count the actual number of cells in the given image. For additional description of YOLOv3 please see Appendix A.

2.2. Previous Work

As the first step of the research and applicability proof-of-concept, a simple classification model was proposed. It was based on residual learning concept [9], specifically Resnet-34 pre-trained on Imagenet.

The original dataset composed of a few hundred images per class was augmented using series of operations such as random flipping or blurring. The augmentation operation was guided by a genetic algorithm, which facilitated selecting the best images in terms of the quality assessment metrics (accuracy in this case). The guiding process was quite exhausting since, for each stage of optimization, a set of augmentation parameters were changed (e.g., crop size), and the training and evaluation were run.

However, the lack of information about the location of the examined tissue's pathological changes proved to be an obstacle in further adopting the classification scheme. Thus, we decided to apply a detection scheme.

2.3. The Detection Scheme

The system's primary function is to accept the cytology samples' images inputted by the veterinarians and return the detected and recognized cells' quantitative results. At this stage, we do not provide the diagnosis for the patient.

The beginning phase of the work involved the preparation of a dataset. This stage involved taking the hundreds of images from the newly acquired cytology samples and performing manual image labeling.

The AI model design step involved the selection of the neural network architecture and parameters tuning that allows the tool to obtain high-efficiency in the classification and detection of the targeted neoplastic cells.

The recurring activity was to perform many trial and error experiments for the neural network parameters tuning. The final accuracy of the model obtained in this stage is presented in this paper.

2.4. The Dataset

For dataset creation, the cell samples were obtained by the fine needle aspiration (FNA) procedure performed on dogs with skin masses. A needle with a small gauge (23–25 G) is usually adequate and reduces the possibility of vessel rupture [10]. After sampling, cellular material was spread over a microscopic slide. The slides were prepared with Diff-Quik staining methods. Due to short staining time, Diff-Quik is commonly used in veterinary clinics.Next, the samples were evaluated under the Delta Optical 300 microscope. The 40× lens has been chosen to obtain a total magnification power of 400×.

A suitable area of the sample should be selected to make the correct diagnosis. In this work, the diagnostic area identification was made by a human expert, and it is the subject of future research to make this critical step to be done automatically. In the future, we plan to use sample scanning to achieve this.

The analysis of several dozen images is needed to find a neoplasm, which will allow an expert to determine what type of cells is dominant in the collected sample and whether we are dealing with cancer. The sample belongs to a single patient.

The selected areas were digitally acquired using the 5-megapixel camera (Delta Optical DLT-Cam PRO 5 MP) attached to a microscope.

The prepared dataset aimed to detect three types of round cell tumors: lymphoma, histiocytoma, and mastocytoma (Figure 2). The dataset is available at https://git.plgrid.pl/projects/CYFROVET/repos/dataset-detection, accessed on 10 July 2021.

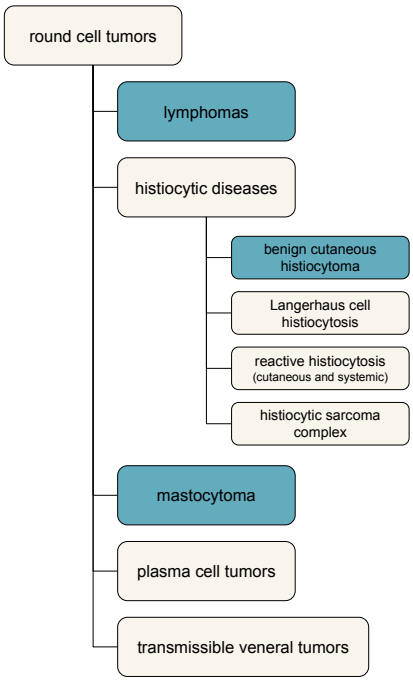
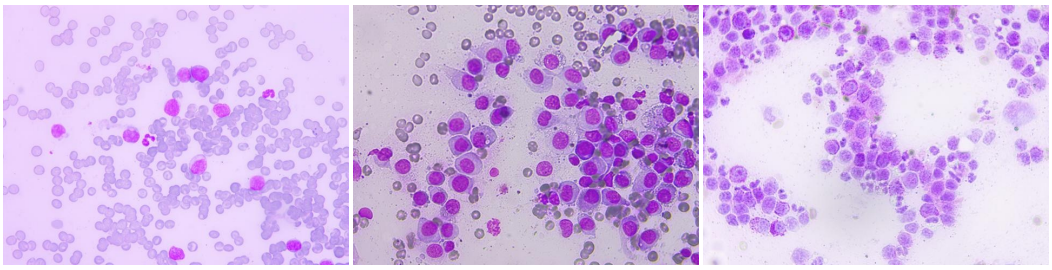


Figure 2. Round cell tumors classification based on [11]. Three classes represented in our dataset are highlighted.

Lymphoma (Figure 3a). Cells are round or oval. Some lymphoma cells may have a shape of a hand-held mirror. They do not have cytoplasm granularity. The nucleus, often slightly indented, fills almost the entire cell. The cytoplasm is usually present in small amounts, and it shows varying shades of blue: from deep basophilic to pale [10,12].

Histiocytoma (Figure 3b). The shape of histiocytoma cells resembles fried eggs"-cytoplasm is more abundant than lymphoma cells. Occasionally, tiny vacuoles can be observed inside the cytoplasm. The nucleus often resembles a characteristic kidney-shape [10,12].

Mastocytoma (Figure 3c). A mastocytoma may otherwise be called a mast cell tumor. Other cells (eosinophils) and collagen fibers may appear in mastocytoma preparations. The most distinctive feature is the presence of granules in the cells, which stain purple-pink. The cell nuclei (if obscured by the granules) are hardly visible. During the preparation of the samples, degranulation sometimes occurs. The release of granularity from the cells can make diagnosis difficult [10,12].



(**a**) (**b**) (**c**)

Figure 3. Example cancer cells. (**a**) Lymphoma. (**b**) Histiocytoma. (**c**) Mastocytoma.

The dataset is divided into two parts: detection and segmentation (Figure 4). The detection set is organized into four subsets: initial, divided, big, and balanced. The most important statistics of the dataset are presented in Figure 5.

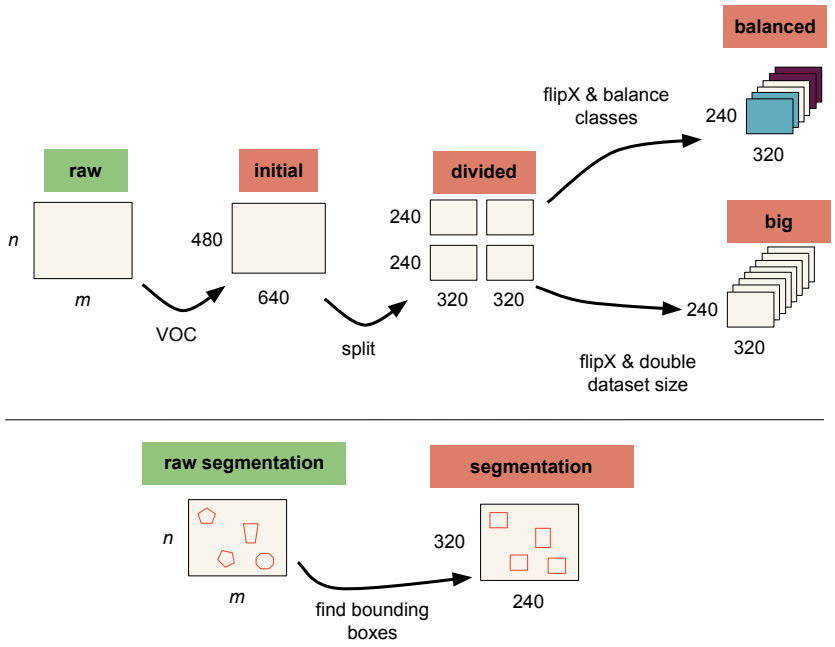
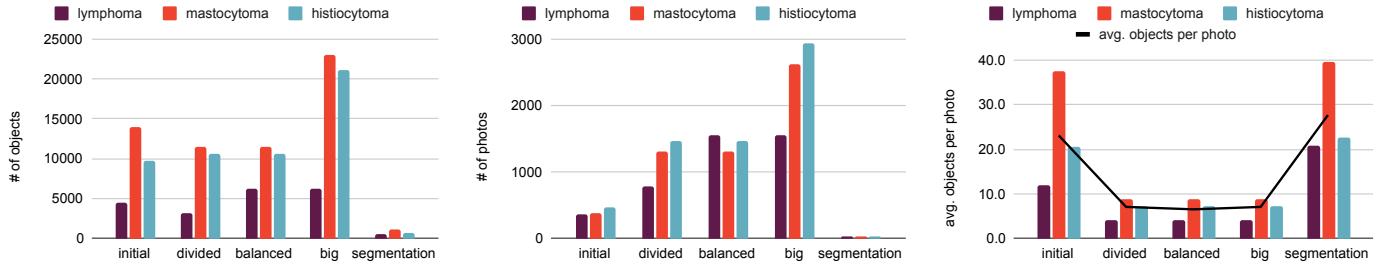


Figure 4. Overview of the datasets relationship.



(**a**) (**b**) (**c**)

Figure 5. Datasets comparison. (**a**) Total number of objects in each class. (**b**) Total number of photos in each class. (**c**) Objects per photo in each class.

2.4.1. The Detection Subsets

Initial Dataset

The original images were mapped to the standard resolution of 640 × 480 to create the 'initial' dataset. Pascal visual object classes (VOC) [13] format was used to standardize the set and make it digestible for future users. The primary image information is also included: the resolution, filename, and color depth.

When analyzing samples, the main problem is determining whether a given cell can be qualified as cancerous. During sample preparation, some cells may be damaged or distorted, making them non-diagnostic. Additionally, not all diagnostic cells were tagged in the 'initial' dataset.

The first models were trained with the 'initial' dataset and the default network parameters. However, the preliminary results were not satisfactory. The inferior performance was related to under-represented cells and the fact that the model detects objects in the context. Consequently, sparse cells and too few images in the selected category posed a challenge for the model. Therefore, so the 'divided' dataset concept was introduced.

Divided Dataset

Images with a resolution of 640 × 480 were divided into four parts, creating images with a resolution of 320 × 240 (Figure 6). This operation increased the number of samples in the dataset and removed areas where no cells or only one is selected. Unfortunately, the number of all objects decreased since the border cells were missed as they were split (see Figure 5).

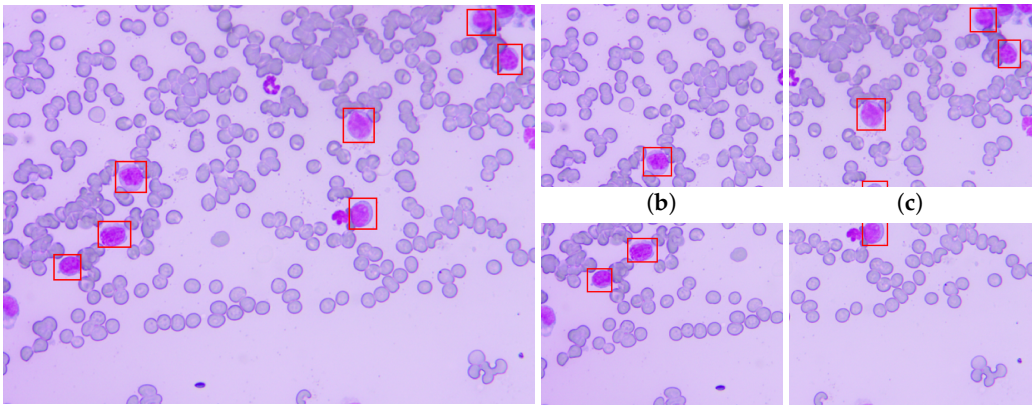


Figure 6. Image split for 'divided' dataset. (**a**) Original photo from the initial set with labels applied. (**b**) Discarded fragment. (**c**) Used fragment. (**d**) Used fragment. (**e**) Discarded fragment.

After splitting, the lymphoma image collection only doubled since no cells were selected in many areas of this class photos. It was different for the other two classes. They contained more tagged cells, and the cells were dispersed. There was a significant increase in the number of mastocytoma pictures as they form densely packed groups. The split of the images and discarding parts with no labeled data improved the model's performance, so we used this dataset as a basis for further experiments.

Balanced and Big Datasets

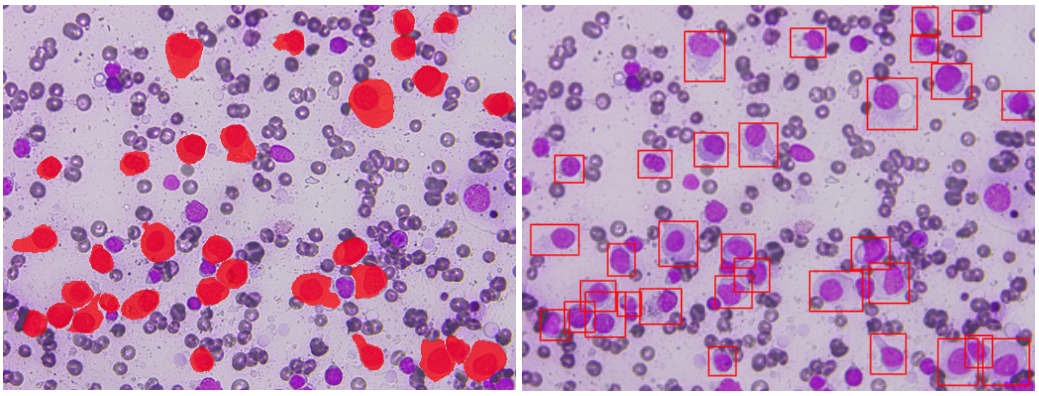
A standard solution to computer vision problems is the augmentation of the dataset that mitigates an unbalanced number of samples of different classes. The tool for image manipulation [14] was used that allowed us to edit not only the images but also the label information file. Using the option to mirror the photo along the X-axis, the two new datasets were prepared. The balanced set consists of a balanced number of photos for each class, and the big set is doubled using reflection along the axis.

2.4.2. Segmentation Dataset

Although the segmentation dataset was originally prepared for object segmentation, it was used in our later detection experiments. In segmentation, the objects are not labeled with rectangles but polygons. For our experiments, they were converted to rectangular frames according to the simple rule:

1. find the maximum and minimum values of X and Y-axis for each of the polygon points;
2. determine top, bottom, left, and right boundary;
3. add/subtract two pixels to slightly increase the frame and prevent overlapping with the object (Figure 7).

The converted segmentation dataset contains more detection samples than the original detection dataset. Images of the highest quality were selected. In the segmentation dataset, all tumor cells are marked in each image. It also contains information about damaged or cut cells. However, the set was still under development while presented experiments were conducted; therefore, this collection was used only for supplementary experiments. The basic information about the dataset is given in Figure 5.



(**a**) (**b**)

Figure 7. Conversion of the image from the segmentation dataset for the detection experiments. (**a**) Segmentation polygons. (**b**) Detection frames.

Summary of the created dataset and their designations can be found in Table 2.

Table 2. Summary of created datasets.

	Dataset Version				
	Initial	**Divided**	**Big**	**Balanced**	**Segmentation**
# of photos	1219	3560	7120	4337	87
# of objects	28,149	25,199	50,398	28,337	2408
Avg. objects per photo	23.1	7.1	7.1	6.5	27.7

3. Results

3.1. System Setup and Implementation

The neural network was implemented in Python, using Keras library with Tensorflow backend. In all experiments, the pre-trained YOLOv3 network was used [15] as a basis. For more information about YOLOv3 architecture please see Appendix A. Training and testing flow is depicted in Figure 8.

A series of experiments were conducted to assess the model and the approach proposed in this work. Using multiple metrics (please refer to Appendix B for details) and various tests enabled a comprehensive presentation of the model and the introduced methodology.

In the preliminary experiments and the model calibration, a range of hyper-parameters was chosen, making further exhaustive tests possible. They are specific to the dataset and the proposed approach. The most important ones are:

- anchor boxes values: {(18,22); (18,28); (22,31); (23,25); (24,38); (28,31); (30,43); (36,36); (40,48)}. They are consecutive pairs (the first one is an 18 × 22 frame), fitted using the K-Means algorithm. These are the predefined frames that best match those of the training set, and the network learns how to move and zoom them to match the labels;
- Objectness threshold—the threshold used to distinguish between object and non-object: 0.5;
- IoU The threshold used to decide when detection is positive or negative: 0.5.

These parameters are used across the experiments presented in this section unless other values are explicitly stated. The code used for the experiments is available at https://git.plgrid.pl/projects/CYFROVET/repos/vet-detection, accessed on 10 July 2021.

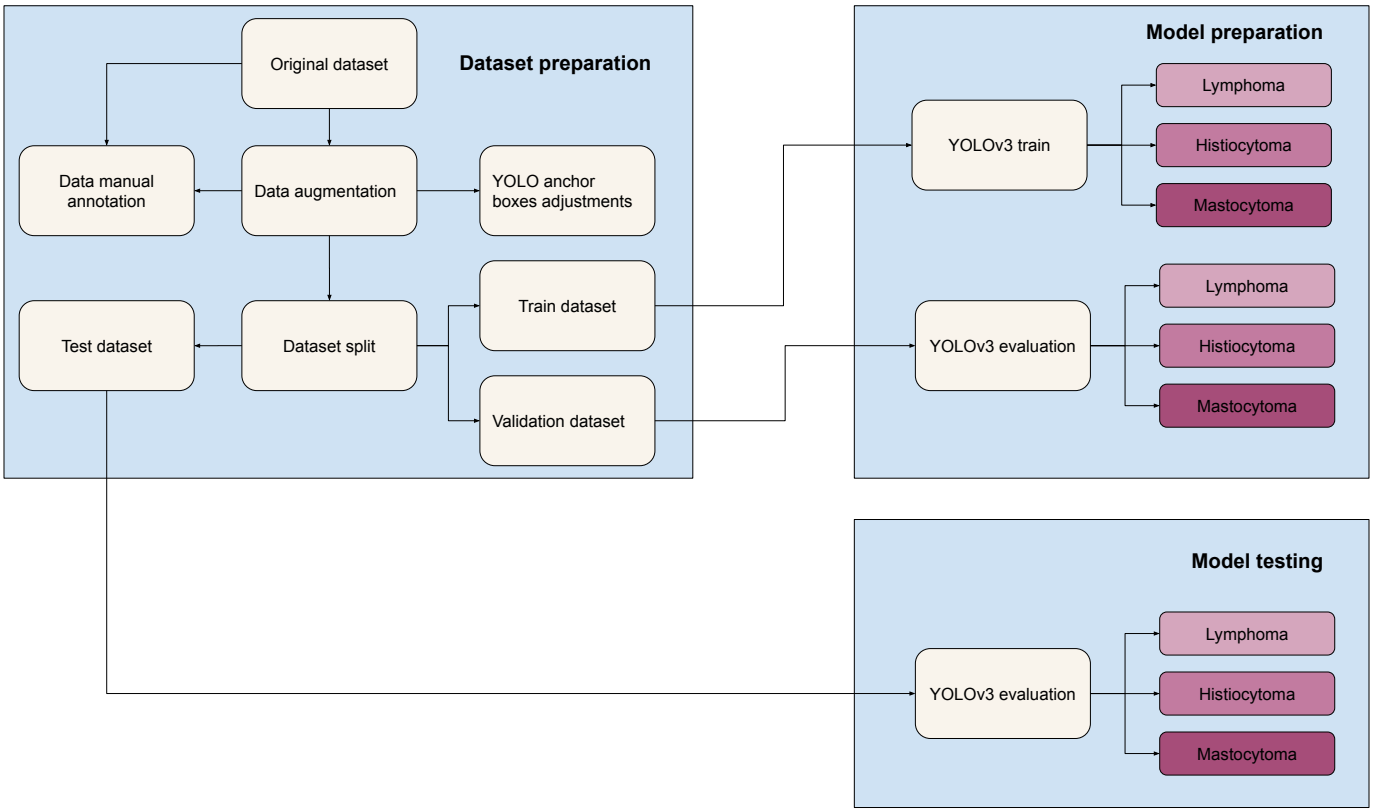


Figure 8. Training and testing flow.

All of the 'divided', 'big', and 'balanced' datasets were split into 70–20–10% training–validation–testing subsets. The presented proportion was selected to ensure that the protocol's validation step is stable. Usually, the validation set is smaller (e.g., 5%), but we wanted to guarantee the stability of the training process and appropriate coverage of all the classes, which was especially important for classes with a low sparsity of labeled cells.

For the 'divided' dataset we present the performance for all subsets (see Table 3). To compare different datasets we only use testing subset.

Table 3. Best model's Average Precision for each class using 'divided' dataset.

Class	Training	Validation	Testing
lymphoma	0.8749	0.6916	0.6628
mastocytoma	0.8303	0.7426	0.7230
histiocytoma	0.8597	0.7264	0.7021
mAP	0.8549	0.7202	0.6960

3.2. Extended Confusion Matrix

In order to compare the performance of the detector with the physician approach, the extended confusion matrix (ECM; please see Appendix B for details) was created (Table 4) for the 'divided' dataset. The table shows that the model falsely detected 309 objects as lymphoma. Those locations were populated by different objects outside of the dataset, i.e., unlabeled. By looking at the table's first row, we can notice that lymphoma was not confused with any other two classes in the dataset (neither mastocytoma nor histiocytoma). This mistake happens because our model analyzes images with objects that do not belong to any class and can be easily confused with one of the targets. The situation manifests itself in the last column of the table ('not present'), where objects that are not expected to be identified are detected as tumor cells. Additionally, we can identify the number of the diagnostic cells that were not detected by the model in the last row ('not detected') of ECM.

The advantage is that there are zeros outside of the classical matrix diagonal which means that the model is robust to misclassification within the dataset. That may also result from the fact that the objects of the three analyzed classes feature uniquely appearance.

Table 4. Extended confusion matrix for the best model for the 'divided' dataset.

		Ground Truth			
		Lymphoma	**Mastocytoma**	**Histiocytoma**	**Not Present**
Predictions	lymphoma	285	0	0	309
	mastocytoma	0	1010	0	1152
	histiocytoma	0	0	669	669
	not detected	44	116	71	

3.3. Selection of IoU Threshold

The model's stable performance within the wide range of IoU threshold is presented in Figure 9. Above 0.7 IoU threshold level, a slight decline of recall can be observed. However, above 0.8 IoU, a significant decline in the model performance can be noticed. This deterioration results from the fact that it is hard for the model to locate objects in the images with high precision.

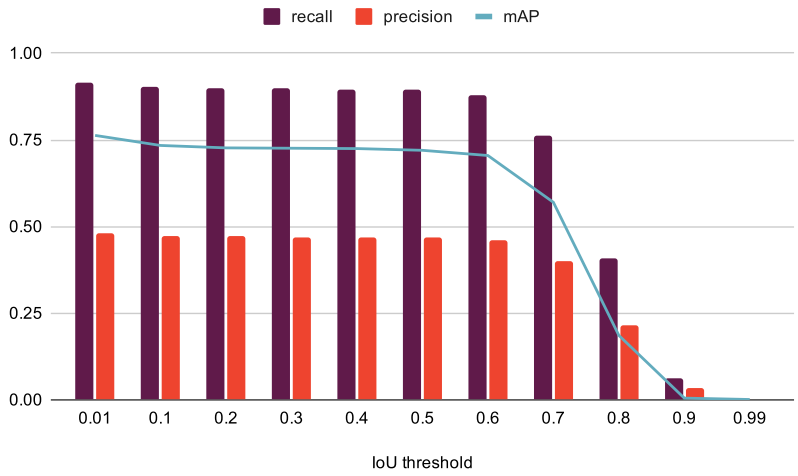


Figure 9. IoU threshold influence on model performance (YOLOv3 trained with the 'divided' dataset, results for validation subset).

3.4. Objectness Threshold Influence

Objectness may be considered as model confidence regarding the occurrence of an object in the image. As is expected, the objectness threshold has a significant impact on the precision of the model detection. It is worth noting that if the objectness threshold is too low, the false-positive number grows significantly. The optimal in terms of precision and recall trade-off value is approximately 0.9, which is presented in Figure 10. For the value of 0.9, precision, recall, and mAP have comparable values. However, beyond this value, there is a rapid decline in the performance. More detailed examination of Figure 10 reveals how deceptive mAP metrics may be. For the objectness threshold value of 0.1, mAP is almost 0.75, but precision is around 0.3, which is low. Therefore, it is important to consider several metrics to show the complete picture of the model performance. This ambiguity was also one of the reasons we decided to introduce the extended confusion matrix.

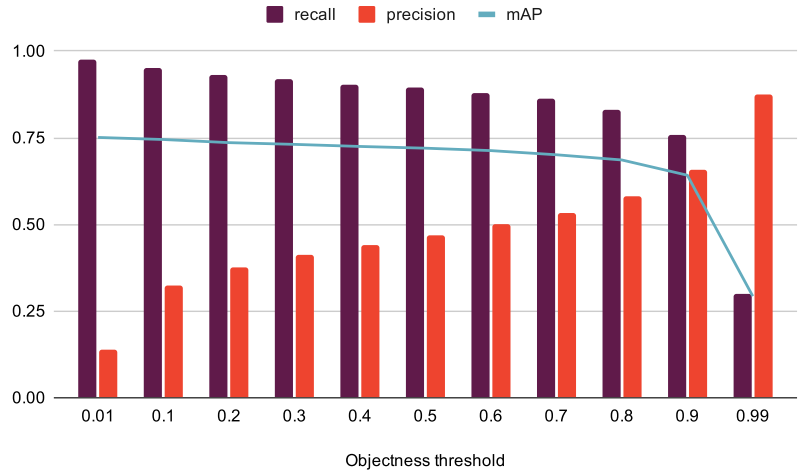


Figure 10. Objectness threshold influence on model performance (YOLOv3 trained with the 'divided' dataset, results for validation subset).

3.5. Dataset Balancing

As described in the previous section, the model was tested using different datasets. As presented in Table 5, the number of objects in the pictures is essential to achieve good performance of the model. The total number of objects is 'divided' dataset is the lowest. Consequently, the results are inferior as compared to the 'big' and 'balanced' datasets.

Table 5. Average precision for each class measured using testing subset of detection datasets.

	Dataset Version		
Class	**Divided**	**Balanced**	**Big**
lymphoma	0.6839	0.7347	0.7208
mastocytoma	0.6912	0.6834	0.7598
histiocytoma	0.7290	0.7350	0.7441
mAP	0.7014	0.7177	0.7416

3.6. The Model Validation with Segmentation Dataset

To better validate our model, we conducted a series of experiments with the segmentation dataset. Their results are presented in Table 6. It is worth noting that this dataset has a different color profile, making detection a much harder task for the model.

Table 6. Average Precision results for different experiments with segmentation dataset.

	Experiments with Segmentation Dataset		
Cell Class	**No Retraining**	**Grayscale Data**	**Color Data**
lymphoma	0.2831	0.0170	0.5204
mastocytoma	0.1207	0.1511	0.7821
histiocytoma	0.4315	0.3223	0.6089
mAP	0.2785	0.3240	0.6371

3.6.1. No Retraining Model Validation

In this validation experiment, the model is fed with the data without prior training with the segmentation dataset. As it can be seen in column two of Table 6, the results are poor (the mAP value is 0.2785).

3.6.2. Retrained Model with Grayscale Segmentation Data

Next, another experiment with data mapped into the grayscale representation. Training of the model was performed. The results are slightly better than in the previous experiment, but they are still relatively poor. The experiment shows that colors are essential for a model to perform well.

3.6.3. Retrained Model with COLOR Segmentation Data

The final experiment involved retraining the original model using a part of the segmentation dataset, leading to the best results of 0.6327 mAP. Based on the described experiments, we may expect that enriching the training dataset with the segmentation set members will lead to the further improvement of the model performance.

As the validation set came from the same dataset as a training set, the model achieved the best performance of 0.7416 mAP with this approach.

4. Discussion

In our work, we experimented with machine learning methods to support physicians in diagnosing tumors based on cytology samples. Often, when diagnosing, the doctors are inclined to treat ML-based model outputs as they treat their peers' suggestions. However, there are fundamental differences between the ML-based expert system and the human practitioners' approach. They need to be considered, especially when the ML-based diagnosis results are underestimated or overestimated sometimes. To understand it better, we need to scrutinize how humans and ML algorithms learn and examine the images.

4.1. Training

In general, the training process of a human-doctor is different from the one of an ML-based model. People are presented with a limited number of images, usually unique and distinctive, whereas the ML model is fed repeatedly with the same, vast set of unstructured images.

A few images were captured after the 5th, 13th, and 38th epoch as presented in Figure 11 to provide a better insight into the model's training process. It may be noticed that the training process is strongly non-linear, e.g., after 13 epochs, the model starts to distinguish between histiocytoma and lymphoma. We also observed that the model sometimes temporally loses skills to detect objects to regain them a few epochs later in the training process. Therefore, picking a specific number of training epochs is not a straightforward task.

It may be noticed that the model learns to localize objects before it actually can classify them correctly. The training is non-linear, which means that some valuable results may be obtained later on, after a few tens of epochs. Closer examination of Figure 11 allows us to notice that longer training tends to eliminate duplication of bounding boxes in histiocytoma and lymphoma cases. Additionally, a differentiation between the two is visible only in the later training stages.

However, the case of mastocytoma is quite the opposite. Its regular shape can be easily explained to an inexperienced doctor, resulting in physicians' high detection performance. On the other hand, the model learns to classify it during the first few epochs, but only in the latter part of the training process starts to increase the detected frames.

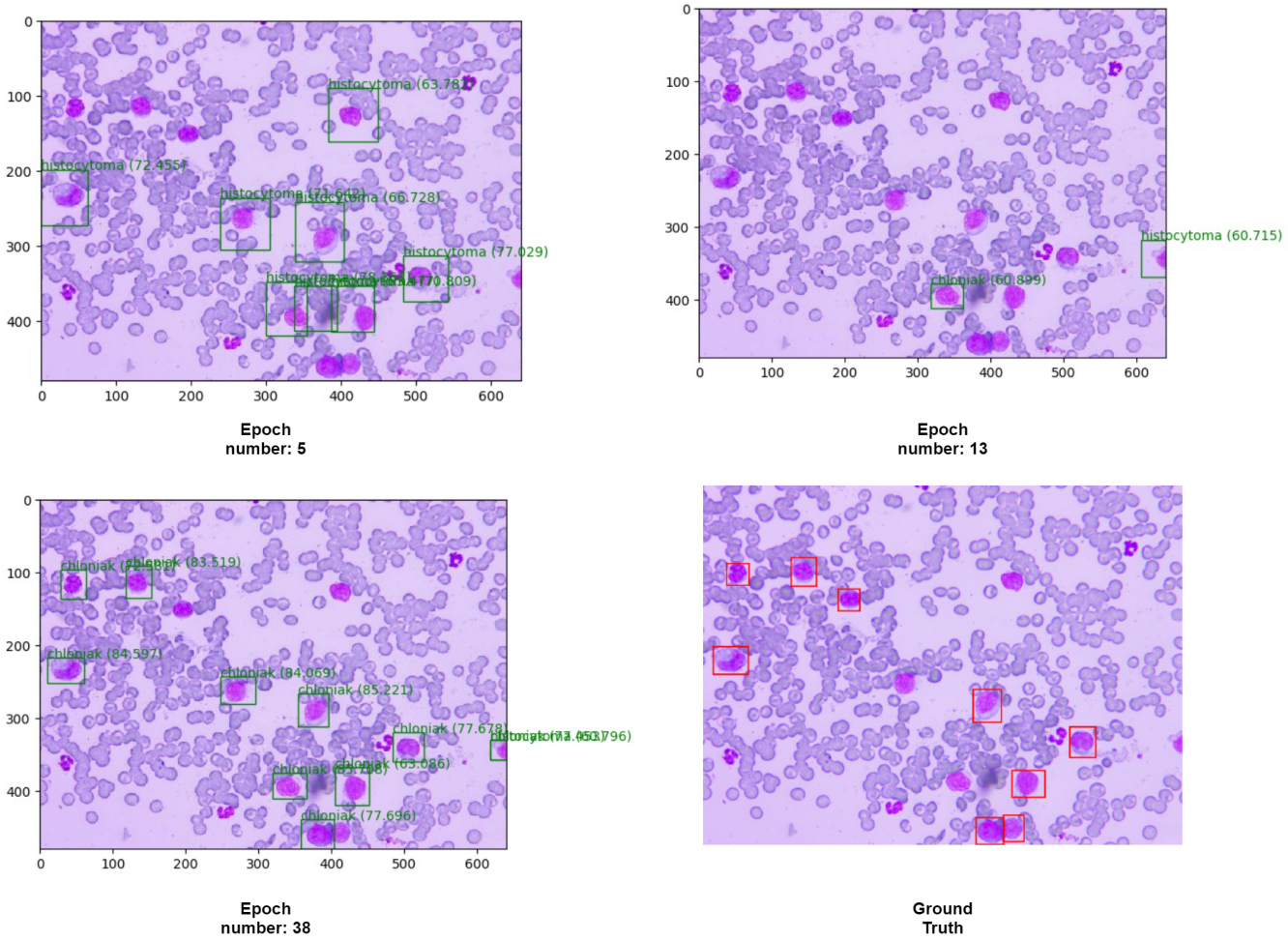


Figure 11. The visualization of the training process at various stages for the same testing image.

4.2. The Coloring of the Tissue Samples

We noticed that a change in the sample's coloring significantly affects the ML-model performance during the experiments. Even though the cells' shape did not change, only the color scheme was different, the model performance suffered. A panacea for this issue is a retraining process with data containing a new color scheme. The experiment in which we trained the model on grayscale images (Table 6) showed an inferior performance of the model, indicating the importance of a color scheme for the model.

The human-doctors diagnosis seems to be much less influenced by the color scheme of the images. They can handle detection across different images collected using different camera types.

One of the solutions to this challenge may be using a meta-learning approach [16].

4.3. Diagnosis

Physicians inspect several images from the same sample, whereas the basic model operates only on a single picture. We envision two approaches that could bridge this gap. The first one is a scenario in which a system is equipped with a scanning capability, enabling it to select a suitable tissue fragment to examine first. Another one is creating a voting system that would combine the outputs of the several image analysis. Implementation of both those approaches requires different datasets in which relations between images and their broader context are known.

4.4. Rare Cases

The ML-model cannot give direct feedback to the user (physician) when it does not know or cannot handle a given case. This information is indirectly provided as a probability score which informs how confident the model is that a given bounding box contains valuable information (object). Whenever the model encounters a challenging case, it is up to the user to analyze the score information and decide how reliable it is. It can be compared to when a doctor can ask his co-workers for their opinion and discuss the diagnosis. In regular or easy cases, the model results are trustworthy, but the doctor has to double-check when it comes to complex cases.

It is worth keeping in mind that cases easy for the model may not necessarily be simple for a human expert and vice versa. A series of tests with a dataset that would contain such known edge cases would have to be conducted to determine the relative difficulty in both cases.

4.5. Photo Background

In the analyzed cytological preparation, apart from neoplastic cells, there may also be inflammatory cells and erythrocytes in the background. Experienced vet doctors hardly ever confuse them with the cells of interest. On the other hand, the ML-based model sometimes confuses unimportant cells with ones subjected to pathological cancerous changes. A solution for this problem could be training a model using a dataset with all objects (including 'background' ones) annotated, either for detection or segmentation.

4.6. Future Considerations

4.6.1. Addressing Sparsity of the Data

In Section 2.4.1, we described the creation of 'divided' dataset. One of the reasons this step was necessary was the low amount of labeled objects in the images, with (relatively) big areas where nothing was annotated. This sparsity of the data can be addressed by applying a two-stage image processing procedure, which we plan to apply as future work. In the first step, a macro picture (from an extensive global scan of the sample) will be examined for spots of the high density of cells which are especially representative for potential pathological conditions. The detection algorithm (DL model) will be applied to these selected spots in the second step. Since we do not have access to a cytological scanner, and we intended our solution for regular clinics (which usually do not have an access to a costly scanner), we came up with a model which works well with images captured by physicians from the microscope. It is expected that a model trained on fragments selected by a scanner will have better performance than our current approach, in which the selection step for the training data was done manually by a physician.

4.6.2. Applying Different Models

The new versions of YOLO models (e.g., v4 and v5) can be adapted to our solution, and we may expect a slightly better performance based on the benchmarks. However, we are currently conducting intensive research on using models for semantic segmentation in this application since they address better-overlapping cells that are sometimes challenging to classify. In addition to this, we would also like to segment overlapping cell borders. We plan to apply a similar approach as the authors of [17].

5. Conclusions

This work presents a pilot study of the model for object detection in vet cytology. A new dataset composed of 1219 images with 28,149 objects was collected and annotated by a vet doctor for lymphoma, mastocytoma, and histiocytoma. The model using the proposed training protocol achieved a performance of 0.7412 mAP with a low cross-class variance of the result.

An in-depth analysis of the model performance was provided. It was mainly focused on the discrepancies between ML-based and vet doctor performance. We think it may help

bring ML solutions closer to real-life practice by exposing the weaknesses and strong sides of the ML-based solutions using the provided prototype and the dataset as an example.

Author Contributions: Conceptualization, D.Ł. and M.W.; Data curation, J.C., D.Ł. and R.F.; Formal analysis, M.W.; Funding acquisition, K.W.; Investigation, J.C. and D.Ł.; Methodology, M.W.; Project administration, P.R. and K.W.; Resources, K.W.; Software, J.C., M.K. and R.F.; Supervision, M.W.; Validation, M.W.; Visualization, J.C. and M.W.; Writing—original draft, J.C., D.Ł. and M.W.; Writing—review and editing, M.W. and P.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financed by Polish Government throughout Narodowe Centrum Badań i Rozwoju NCBiR (The National Centre for Research and Development) grant number Panda 2 September 2016 and by AGH University of Science and Technology subsidy funds No. 1616-230-434 from the Polish Ministry of Education and Science. The APC was funded by AGH University of Science and Technology subsidy funds No. 1616-230-434 from the Polish Ministry of Education and Science, with discount vouchers provided by M.W. The computations were carried out using high-performance computing resources of ACC Cyfronet AGH.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: https://git.plgrid.pl/projects/CYFROVET/repos/dataset-detection/browse, accessed on 10 July 2021.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AP	Average precision
ECM	Extended confusion matrix
IoU	Intersection over union
ML	Machine learning
VOC	Visual object classes
YOLO	You only look once

Appendix A. The YOLOv3 Architecture

Unlike earlier works on object detection that used classifiers or locators to detect objects, YOLOv3 [8] uses a window that moves through the image and locates objects in these fragments. In YOLOv3, a single neural network predicts the objects' locations and assesses the classes' probabilities. It is done directly from the picture with only one data pass. The YOLOv3 network divides the photo into regions and predicts the object frames and the category probabilities for each of them.

The network consists of 53 fully convolutional layers and 53 others, making a total of 106 layers. Each convolutional layer is followed by the batch normalization and ReLU normalization layers. The detection takes place in three layers: 82, 94, and 104. A picture subsampling is performed in convolution layers only. Such an approach prevents the loss of low-level features that are often excluded by pooling layers. This feature improves the detection of small objects and is particularly important in our project.

The resolution of the images used for network training and prediction is not specified for YOLOv3. They are automatically adjusted to the size of the network. The only requirement is that the image's width and height are divisible by 32.

Appendix B. Metrics

Appendix B.1. Extended Confusion Matrix

The confusion matrix is a primary method to describe a performance of a classification system. It allows us to quickly detect ground truth classes that are most problematic for the model, resulting in assigning objects to wrong prediction clusters.

One of the classic confusion matrix challenges is a lack of penalty measure when the system does not detect the objects existing in the ground truth data. Additionally, the method does not handle the cases when the system detects objects that are not in the ground truth data. Both cases were common in our experiments since only diagnostic tumor cells were marked in the data.

To better understand the results provided by the system, an extended confusion matrix was introduced. New columns were added to the classical confusion matrix: the 'not present' column and 'not detected' row, which allowed us to visualize both before-mentioned cases. The 'not present' column represents all objects detected by the model but not labeled by a human expert, and 'not detected' row are all object not found by the model but labeled by a human expert (Table 4).

Appendix B.2. Recall and Precision

In our experiments, we used the micro-averaged recall (also called sensitivity) and precision metrics. They are calculated as follows:

$$\text{recall} = \frac{tp}{tp + fn}, \tag{A1}$$

$$\text{precision} = \frac{tp}{tp + fp}, \tag{A2}$$

where (for a given IoU threshold):

- *tp*—true positives—objects correctly classified,
- *fp*—false positives—objects incorrectly classified,
- *fn*—false negatives—all labeled objects that were not found during detection.

Appendix B.3. Mean Average Precision

For a particular class, The average precision (AP) is the area under the precision-recall curve for a selected IoUs. For a given IoU threshold, the values of the recall and precision are calculated. Results of the detection are sorted starting from the ones with the highest IoU value. The function of precision vs. recall is obtained, which is declining towards rising recall values. The mean average precision (mAP) is the AP averaged for all classes, as shown in (A3).

$$\text{mAP} = \frac{1}{|\text{classes}|} \sum_i \text{AP}(\text{class}_i) \tag{A3}$$

References

1. Baba, A.I.; Câtoi, C. Cancer Diagnosis. In *Comparative Oncology*; The Publishing House of the Romanian Academy: Bucharest, Romania, 2007; Chapter 18.
2. Singh, Y.; Srivastava, D.; Chandranand, P.S.; Singh, D.S. Algorithms for Screening of Cervical Cancer: A Chronological Review. *arXiv* **2018**, arXiv:1811.00849.
3. Żejmo, M.; Kowal, M.; Korbicz, J.; Monczak, R. Classification of breast cancer cytological specimen using convolutional neural network. *J. Phys. Conf. Ser.* **2017**, *783*, 012060. [CrossRef]
4. Dimauro, G.; Ciprandi, G.; Deperte, F.; Girardi, F.; Ladisa, E.; Latrofa, S.; Gelardi, M. Nasal cytology with deep learning techniques. *Int. J. Med. Inform.* **2019**, *122*, 13–19. [CrossRef] [PubMed]
5. Teramoto, A.; Tsukamoto, T.; Kiriyama, Y.; Fujita, H. Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks. *BioMed Res. Int.* **2017**, *2017*, 4067832. [CrossRef] [PubMed]

6. Marzahl, C.; Aubreville, M.; Bertram, C.A.; Stayt, J.; Jasensky, A.K.; Bartenschlager, F.; Fragoso-Garcia, M.; Barton, A.K.; Elsemann, S.; Jabari, S.; et al. Deep Learning-Based Quantification of Pulmonary Hemosiderophages in Cytology Slides. *Sci. Rep.* **2020**, *10*, 9795. [CrossRef] [PubMed]
7. Sompawong, N.; Mopan, J.; Pooprasert, P.; Himakhun, W.; Suwannarurk, K.; Ngamvirojcharoen, J.; Vachiramon, T.; Tantibundhit, C. Automated Pap Smear Cervical Cancer Screening Using Deep Learning. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 7044–7048. [CrossRef]
8. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
10. Albanese, F. *Canine and Feline Skin Cytology: A Comprehensive and Illustrated Guide to the Interpretation of Skin Lesions via Cytological Examination*; Springer: Berlin/Heidelberg, Germany, 2017.
11. Moore, P.F. A Review of Histiocytic Diseases of Dogs and Cats. *Vet. Pathol.* **2014**, *51*, 167–184. [CrossRef] [PubMed]
12. Albanese, F. *Atlas of Dermatological Cytology of Dogs and Cats*; Merial Italia: Milan, Italy, 2010.
13. The PASCAL Visual Object Classes Homepage. Available online: http://host.robots.ox.ac.uk/pascal/VOC/ (accessed on 26 July 2021).
14. Kathuria, A. Data Augmentation for Object Detection. Available online: https://github.com/Paperspace/DataAugmentationForObjectDetection (accessed on 30 January 2021).
15. Anh, H.N. YOLO3 (Detection, Training, and Evaluation). Available online: https://github.com/experiencor/keras-yolo3 (accessed on 20 December 2020).
16. Lemke, C.; Budka, M.; Gabrys, B. Metalearning: A survey of trends and technologies. *Artif. Intell. Rev.* **2015**, *44*, 117–130. [CrossRef] [PubMed]
17. Jang, W.D.; Wei, D.; Zhang, X.; Leahy, B.; Yang, H.; Tompkin, J.; Ben-Yosef, D.; Needleman, D.; Pfister, H. Learning Vector Quantized Shape Code for Amodal Blastomere Instance Segmentation. *arXiv* **2020**, arXiv:2012.00985.