POSSIBLE JOURNALS: AJCP, MODERN PATHOLOGY

AUTHORS:

Ramraj Chandradevan
Ahmed A. Abdulrahman
Bradley R. Drumheller
Nilakshan Kunananthaseelan
Mohamed Amgad
David A. Gutman
Lee A.D. Cooper
David L. Jaye

TITLE: DEVELOPING A DIGITAL PATHOLOGY TOOL THROUGH MACHINE LEARNING TO AUTOMATE BONE MARROW ASPIRATE DIFFERENTIAL COUNTS: PROMISING INITIAL RESULTS WITH DETECTION AND CLASSIFICATION OF BENIGN BONE MARROW CELLS

INTRODUCTION

Examination of the bone marrow is an essential part of the hematological work-up for many blood and bone marrow diseases and a common laboratory procedure. As part of this examination, a nucleated differential cell count (DCC) is obtained by microscopy on Wright-stained bone marrow aspirate (BMA) smears. This procedure entails quantification of cells of different lineages to determine the proportions of each, the findings of which aid in the classification of numerous benign and malignant hematologic disorders. In fact, disease defining criteria are based on cutoff percentages of myeloblasts for myeloid malignancies, such as acute myeloid leukemia (AML) and myelodysplastic syndromes (MDS), and the percentage of plasma cells for plasma cell neoplasms, such as monoclonal gammopathy of undetermined significance (MGUS) and smoldering myeloma [1].

Several factors render manual DCC analysis suboptimal, as currently performed in clinical laboratories [2]. First, DCCs are labor intensive and time consuming. Second, interobserver and intraobserver variability in terms of cell identification and choice of cells for counting represent ongoing sources of error. Third, there is inherent statistical imprecision due to the relatively small number of cells generally counted. If successfully developed, automation of DCCs could obviate most of these concerns. Traditional automated hematology analyzers that do not employ digital images have been explored for performing DCCs. Major problems to this

approach included failure to count nucleated red blood cells, to differentiate stages of cell development, and interference by bone marrow lipid [3, 4]. These issues are perhaps unsurprising given the complex nature of bone marrow compared to blood for which these instruments were designed. However, a computerized method using digital pathology images could potentially perform DCCs on all pertinent bone marrow cells on a smear. Aside from increasing throughput and reducing labor costs, such an approach could potentially improve accuracy, reproducibility, and objectivity and provide much needed standardization for DCCs.

Cell detection and classification are perhaps the most widely studied problems in computational pathology, with most efforts focused on the analysis of hematoxylin and eosin stained solid tumor sections. While commercial blood analyzers (e.g. CellaVision™) have begun utilizing automated image analysis of Wright stained smears [5], their accuracy largely depends on precise control of preanalytical variables to minimize staining variations and cell crowding while maximizing preservation of cytologic details. Detection and classification in BMA smears is significantly more challenging due to the high density of touching and overlapping cells, as well as the greater diversity and complexity of cell morphologies. Cell and nuclei detection algorithms often rely on circular or axial symmetry and may fail to detect cells with irregular or multilobed nuclei or may incorrectly interpret these as multiple cells. Classification is difficult without accurate detection and segmentation, and further compounded by the subtlety of differences in cytologic characteristics used to distinguish many cell types found in bone marrow.

Machine-learning approaches have emerged as the dominant paradigm in analyzing histology images [6-11]. Whereas traditional image analysis methods are engineered using domain knowledge or mathematical models, machine-learning algorithms that utilize neural networks are adaptive and can learn from data in an unbiased manner [12]. While neural networks typically exhibit superior performance in tasks like detection and classification, realizing these benefits can require thousands of labeled examples for training algorithms to recognize variations in staining and morphology and to reach diagnostically-meaningful accuracy. This demand for labeled data places significant emphasis on the process of image annotation, with efficient protocols and software interfaces being key additional ingredients for developing highly accurate, deep learning algorithms. Current literature on image analysis of BMA smears has not adequately addressed the detection of cells, a particularly challenging problem in BMA smears, and has demonstrated success with only a few cytological classes, limiting potential clinical use [13, 14].

In this paper, we describe our initial steps towards the development of a machine-learning digital pathology system to perform DCCs and describe promising initial results in detecting and classifying all bone marrow cellular constituents of the DCC. Our software prototype achieves a high degree of accuracy in cell detection and classification tasks, using a two-stage system, based on convolutional neural networks. This system is, moreover, able to reliably localize closely packed cells and classify diverse cytomorphologies. A large-scale annotation effort to produce data for training and validation was critical in achieving these results. This study outlines a promising prototype system for automating bone marrow DCCs and provides a basis for further development and validation studies.

METHODS

Bone marrow aspirate smears. Wright-stained BMA smears, collected as part of routine patient care from 27 patients, were de-identified and scanned at 40X objective magnification using an Aperio AT2 scanner™ to generate whole-slide images. All smears were prepared in the bone marrow clinical laboratory at Emory University Hospital using the same procedure and reagent vendors. Inclusion criteria included having cellular particles with at least 500 bone marrow hematopoietic cells from which manual analysis showed generally reference range DCCs, minimal cellular degeneration, and a paucity of smearing artifacts, and lastly pathologic analysis failed to disclose morphologic, immunophenotypic or genetic abnormalities. This study was approved by the Institutional Review Board.

Cell annotation. Whole-slide images were uploaded to a Digital Slide Archive (DSA) server for visualization and annotation. The DSA enables web-based viewing, allowing users to pan and zoom through large whole-slide images, and features a collection of annotation tools for marking and labeling regions and structures [15]. The annotation interface is shown in **Figure 1A.** Regions-of-interest (ROIs) for annotation were first selected using the rectangle or polygon tool and included areas of the smears with a high concentration of intact and more separated cells, located in non-hemodilute regions, adjacent to bone marrow spicules that best represent the spectrum of hematopoiesis. In addition, other regions were selected using the polygon tool to exclude erythrocytes and non-counted cells (macrophages, stromal cells, mast cells, etc.), that typically would not be included in DCCs (**not shown**). Individual cells for the DCC were annotated using the point annotation tool by placing a single point at the cell center-of-mass and placing them in one of the 13 classes shown in **Table 1**. The cells within each region were exhaustively annotated to enable accurate assessment of cell detection algorithms. Cells of uncertain class, such as those with smudged and/or naked nuclei, were assigned to an "unknown" class. Megakaryocytes were also annotated, but not included in the cell detection or classification analyses since they are relatively few and not typically included in DCCs.

Following point annotations, rectangular bounding boxes were drawn to demarcate the extent of each cell (RC, NK). These bounding boxes are required to train and validate the detection algorithm. Additional point annotations were generated outside ROIs to augment the number of examples of cell types, such as basophils, that inherently occur less frequently in bone marrow (**Table 1**). These latter annotations were utilized only during classifier training and neither for classifier validation nor for training and validation of the cell detection algorithm. Cell annotations were based on well-established cytomorphological criteria used for the microscopic identification of each cell type [16].

Subsequently, all annotated cells were examined for cytologic quality and appropriateness of classification through a consensus review by 3 pathologists (AAA, BRD, DLJ). To accomplish this review, the DSA application-programming interface was used to extract a 96x96 pixel thumbnail image of each annotated cell. These thumbnail images were next organized into folders by assigned cell type. The few initially misclassified cells were identified, and corrections were made to the annotation database. Representative examples of the cytologic classes used in our analysis are displayed in **Figure 1B**.

Cell detection algorithm. Our cell detection algorithm is based on the Faster Region-Based Convolutional Network (Faster-RCNN) [17]. This network combines bounding box regression for predicting bounding box locations, region pooling, and a residual convolutional network for extracting feature maps from the input images. Detection was approached by treating all cells as a single 'object' class, without regard to actual cytomorphologic class. The residual network was trained using two equally weighted loss functions: 1. A cross entropy loss for object classification and 2. An L1 loss on the bounding box coordinates and sizes. Proposed regions were then pooled for computational efficiency, since many proposals are generated for each object. A pre-trained model was used to initialize the residual convolutional net [18], where the remaining network components were random normal initialized (zero mean, variance 1e-4). The entire network was trained for 500 epochs, where an epoch represents one training pass through all training instances. Training employed a momentum-based gradient optimization with momentum 0.9, learning rate 3e-4, weight decay 5e-4, and dropout fraction 0.2. Non-max suppression with a threshold of 0.5 was applied to reduce duplicate proposals.

Cell classification algorithm. Cell classes were predicted using the VGG16 convolutional network [19]. Cell images, sized at 96x96 pixels, were cropped from the center of each manually-generated bounding box. These bounding boxes were mapped to the point annotations using the Hungarian algorithm applied to the pairwise distances between each box and each point. These images were unit normalized to the range [0, 1]. A cross entropy loss for the 12 classes (including "unknown" and excluding megakaryocytes) was used for network optimization. A pre-trained network was used for initialization and then trained using the gradient descent optimizer with 100 cell batches and a learning rate of 1e-4 for 500 epochs. Dropout fraction 0.3 was applied to the fully connected layers.

Data augmentation. Augmentation techniques were utilized in cell detection and classification to improve prediction accuracy. For cell detection, we generated randomly cropped 600 x 600 pixel regions from the ROIs and randomly flipped these horizontally and vertically. For cell classification, we applied standard augmentation techniques to manipulate the orientation, brightness, and contrast of the cropped cell images. Each cell image was randomly mirror-flipped along the horizontal and vertical axes, rotated by an increment of 90 degrees, brightness adjusted (random_brightness, delta=0.25), and contrast adjusted (random_contrast, range [0.9, 1.4]). To simulate errors in the detection algorithm, we performed a random translation of up to 5 pixels horizontally and vertically. Testing time augmentation was also performed to improve classification performance. During inference, 16 augmented instances of each cell were generated. The softmax values for these augmented versions were then aggregated to generate a single prediction for each cell.

Detection and classification validation. We performed six-fold cross validation to measure the prediction accuracy of our cell detection and classification methods. Using 12 annotated slides that had ROIs selected, training sets were composed of 10 slides and validation test sets the remaining 2 slides. Each training set was used to develop a cell detection and a separate

cell classification model. These models were evaluated on the 2 validation test slides, yielding six total measurements of detection accuracy and of classification accuracy.

Cells and ROIs from each training slide set were used to train the detection and classification models using the manual point and bounding box annotations. These models were applied to the test slides as follows: 1. The detection model was applied to the test slides to generate prediction bounding boxes and their probabilities and the detection accuracy was measured (see details next paragraph) 2. For detections regarded as true positives, cell images were cropped and centered at the predicted bounding box locations. These cells were then used to evaluate the accuracy of the cell classification model. A 6-fold cross validation was performed to evaluate cell detection and classification accuracy.

Detection accuracy was measured using precision-recall and intersection-over-union (IoU) analysis. IoU is defined for any pair of predicted and manually-annotated bounding boxes, the latter representing the ground truth (gold standard) bounding box, as the area of box intersection over the area of box union. This reaches 1 for perfect overlap and 0 for non-overlapping boxes. The following definitions were used for precision-recall analysis: 1. True positive (TP) where a manually-annotated box has a corresponding predicted box meeting the IoU threshold 2. False negative (FN) where a manually-annotated box has no predicted box meeting the IoU threshold 3. False positive (FP) where a predicted box does not have a corresponding manually-annotated box meeting the IoU threshold. The Hungarian algorithm was used to generate a correspondence between manually-annotated and predicted boxes that maximizes the sum of IoUs to avoid double counting of manually-annotated boxes in accuracy calculations. Each predicted bounding box has an associated confidence score and so a precision-recall curve is generated using TP, FN, FP for the range of detection confidence thresholds from 0 to 1. The area under this precision recall curve measures detection accuracy over a broad range of detection sensitivities [20]. In addition, we measured error in the positioning of predicted bounding boxes as the difference in location between the predicted box centers and the matched manually-annotated box centers. Using the **TP** correspondence from above, we calculated the Euclidean distance between box centers and normalized by the manually-annotated bounding box size (using half the length of the annotated box diagonal).

Classification accuracy was measured using receiver-operating characteristic (ROC) analysis. For each classification model, we measured the sensitivity and specificity of a binary classifier for each cell type (this cell type versus all others) to generate an ROC curve. The area under the ROC curve (AUC) was measured for each cell type, along with the macro average (average performance over all classes, not weighted by class prevalence) to measure class specific and overall accuracy [21].

Software.

We employed the following software tools. Tensorflow 1.8 served as the basic framework for the entire system. Luminoth v0.2.0 is Tensorflow-based and provided the detection framework. The Hungarian algorithm was implemented in Scipy v1.1.0. The OpenCV 3.1.0 library was used to handle png/jpeg images.

RESULTS

Study overview. An overview of our approach is presented in **Figure 2**. The software consists of two convolutional networks, one for cell detection and another for classification (**Figures 2A, 2B**). The cell detection network identifies the locations of cells, and image patches are then extracted at these locations to predict cytological class of each detected cell. The accuracy of these algorithms was evaluated through a 6-fold cross validation that strictly separates cases into either training or validation sets (**Figure 2C**). Data for training and validating algorithms were generated using a web-based DSA annotation system (see **Figure 1**, **Table 1**). Our cell detection and classification analyses included 11 cytological classes that constitute those of standard DCCs and an unknown class (**Figure 1B**).

Large-scale annotation. Convolutional networks can deliver outstanding performance given large training datasets of thousands of examples that represent the morphological and staining variations observed in clinical practice [6]. To generate annotations at sufficient scale, we developed a protocol using the DSA [15]. A screenshot of the DSA annotation interface is presented in **Figure 1A** with example ROIs, cell annotations, and bounding boxes. We used the DSA and the annotation protocol to annotate 9269 cells that are specified in **Table 1,** and included those within ROIs and a smaller subset outside ROIs. The latter subset increased the representation of less common cell classes. Annotation efficiency was improved using a tiered approach that took into account the expertise and availability of annotators, and the effort involved. Labeling cell types requires expertise in the cytomorphology of bone marrow cells. A simple and efficient point and click method utilizing a mouse was developed for cell type labeling by pathologists. Since point annotations alone are not adequate for training detection algorithms, students performed the more laborious task of placing rectangular bounding boxes around the pre-identified cells. Although the bounding boxes alone could be utilized for both localization and cell type labeling, this tiered approach proved more efficient and allowed us to generate a much larger number of annotations.

Cell detection with region-proposal networks. Detection results for one representative ROI are presented in **Figure 3A**. Cells that were missed often had a corresponding detection bounding box that was close, but did not have adequate overlap, based on IoU analysis, to be called a match (**Figure 3A**, **subpanels 1-4**). A number of false positives correspond to cells that were mistakenly not annotated by our human observers (**Figure 3A**, **subpanels 5 and 6**). A precision-recall analysis was performed to evaluate detection performance from the most sensitive to the most selective detection threshold. The detectors generated in cross validation simultaneously achieved high precision and recall with only minor variation in performance from fold to fold, as displayed in **Figure 3B**. The median area under these precision-recall curves was 0.959 +/- 0.008 (see **Table S1**). In addition to these discrete detection errors, we also measured the positional errors in the placement of predicted bounding boxes. Correct bounding box placement is important for the classification stage, since the center of the predicted boxes is used to extract cell images for classification. Since the tolerance for bounding box placement is higher when detecting larger cells, we developed a relative error measure that considers both cell size and predicted bounding box position (see **Figure S1**). The median relative placement

error was 6% (**Figure 3C**), indicating good coincidence between the centers of predicted and actual cell centers.

Cell classification and augmentation strategies. Classification results for one representative ROI are presented in **Figure 4A**. In this example, two unknowns were misclassified as erythroids, as shown in **subpanels 1 and 2**. Classifier accuracy was evaluated by examining a one-versus-all classifier for each cytologic class. The classification threshold was varied from the most sensitive to the most specific, generating an ROC profile for each class. These classifiers achieve a high AUC for each class in each fold as shown in **Figure 4B**. The median total AUC (all classes weighted equally) was 0.982 +/- 0.03. Median AUC for each class ranges from 0.960 (monocyte) to 1.00 (basophil) (see **Table S2A** for per-fold AUC measurements). A confusion matrix describing the misclassifications is presented in **Figure 4C**. As shown, the most common source of classification errors were defined cell types, particularly monocytes (14%) and lymphocytes (21%), being predicted to be unknown cell types. Other common errors included adjacent cell classes in the myeloid series: blasts being misclassified as promyelocytes (10%), myelocytes being predicted to be promyelocytes (8%), and promyelocytes being predicted to be blasts (7%). Lastly, lymphocytes were predicted to be erythroid cells (6%) and monocytes were misclassified as metamyelocytes (7%) or myelocytes (6%).

To better understand the potential impact of these errors on DCC-based diagnoses, we calculated the expected ranges of cell counts for plasma cells and blasts for several common diagnostic scenarios using the errors presented in Figure 4C (see **Table S3**). Given 8.6% plasma cells in a sample, the expected range of our computational DCC would be 7.7-8.8% based on misclassification of plasma and other cell types. For 57.2% plasma cells, the expected range would be 50.9-57.3%. The variance for blasts is higher. A sample containing 6.4% blasts would have an expected range 5.4-7.4%, a sample with 10% blasts corresponds to an expected range 8.2-12.1%, and a sample with 20% blasts would have an expected range 22.1-28.0%.

Handling detection and classification with two separate networks provided more flexibility in network design, and enabled us to employ advanced strategies for data augmentation that had a significant impact on classification accuracy. During class inference, each detected cell was augmented to generate 16 versions with different orientations and intensity transformations as displayed in **Figure 5A**. The classification network was applied to these augmented versions, and the predicted class probabilities were aggregated. This procedure improved the total AUC an average of 5.0% over all folds as displayed in **Figure 5B** (see also **Table S2B**). This increase in classification accuracy was statistically significant (**p=3.12e-2**, Wilcoxon signed rank).

DISCUSSION

BMA DCC is routinely performed to assess hematopoietic activity, to compare the proportions of the different cell lineages with reference ranges, and to quantify abnormal cells when present. It is generally performed by pathologists and/or the laboratory technical staff

depending on workflow and the laboratory case volume. While publications vary in the total number of cells recommended for performance of DCCs, they generally fall between 300-500 cells, but can vary based on specific clinical circumstances [1, 22, 23]. At the high end, counts of more than 500 cells have been recommended based on theoretical work that considered the odds of unacceptable error in classification when initial counts fall near diagnostic cutoffs for critical cells classes [24, 25]. Yet, manual DCCs suffer from being labor intensive with inherent inter- and intra-observer variability in cell classification and choice of cells counted. Automation of the DCC could not only obviate these issues, including the ability to readily analyze the many hundreds to thousands of pertinent cells on a smear, but also could also afford standardization.

One promising method of automation that we explored in this work is digital image analysis with machine learning. Images of BMA smears present significant technical challenges for image analysis algorithms. BMA smears contain a diversity of cell cytomorphologies with some displaying only subtle differences. A large number of cells in any given smear may be ambiguous and the boundaries between cells indistinguishable, particularly in areas with clumping. Traditional image analysis techniques that rely on models of cell appearance and morphology are difficult to apply in these scenarios, and may fail to accurately detect and distinguish closely packed cells (i.e. segmentation). Reliable segmentation is absolutely necessary to extract shape, texture, and color features that are used for classification, and often difficulties in segmentation will be reflected in poor classification performance. A data-driven approach based on machine learning with convolutional networks can avoid the problem of explicit segmentation, and does not rely on a-priori definitions of cell features for classification, but requires extensive amounts of annotated data for training and validation.

To generate sufficient data for such convolutional network approaches, we developed an efficient tiered annotation protocol using the DSA. This web-based platform facilitated de-centralized annotation and review, and helped to scale our labeled dataset. The tiered protocol utilized experts to classify cells using a simple point annotation tool, and students to do the more laborious work of placing bounding boxes, enabling us to annotate over 9,000 cells. This large dataset allowed us to engineer an analysis pipeline based on convolutional networks for cell detection and classification. This pipeline achieved high accuracy in detection (0.959 +/- 0.008 precision-recall AUC) and classification (0.982 +/- 0.03 ROC AUC) in a 6-fold cross validation. Decoupling the detection and classification steps provided significant benefits in our system. As we demonstrated, the ability to perform augmentation of detected cells significantly improved classification accuracy, increasing the accuracy from 0.917 +/- 0.027 to 0.982 +/- 0.03 ($p = 3.12e-2$). Using separate networks for detection and classification also improved flexibility in design. Detection and classification tasks have very different design requirements and creating a single convolutional network to perform both tasks is difficult and will likely result in suboptimal overall performance. Nonetheless, these two networks would appear seamless to users of the software whether pathologist/hematologist or medical technologist.

Limited studies have evaluated image analysis in automating DCCs. Choi *et al* [13] published promising preliminary results using convolutional networks for cell classification in DCCs. Their dataset comprised 2174 cells of non-neoplastic erythroid and myeloid precursors, and did not include other cells types important in DCCs including eosinophils, basophils, monocytes, lymphocytes, and plasma cells. This study focused on classification and did not address detection, utilizing only manually cropped images of cells to develop and validate the

classifier. Moreover, noise due to the detection process was not accounted for in their classification. Their reported classification performance was 0.971 precision at 0.971 recall. This is comparable to the overall classification performance for our system that included analysis of all relevant cell types in the DCC. Reta *et al* [14] developed a cell detection and classification framework for classification of acute leukemia subtypes. Their dataset comprised 633 cells from acute lymphoblastic leukemias and acute myeloid leukemias. They developed an elaborate software pipeline to detect and segment leukocytes in digital images of Wright stained BMA smears captured at 100X. Detected cells were characterized using a set of features that describe the shape, color, and texture of each cell. These cells were classified individually using basic machine learning algorithms, and the cell classifications were aggregated to provide a single diagnosis for the sample. While their application is narrow and focuses only on a few cell types, their reported segmentation accuracy has a high precision (95.75%) and their subtype classification accuracy ranged from 0.921 to 0.784 ROC AUC. Our findings expand the work in these earlier publications and point to the promise of machine learning approaches towards automation of DCCs.

In this study, we present preliminary results in developing a computational BMA DCC system. Our approach combines state-of-the-art detection and classification algorithms based on convolutional networks, and achieved excellent performance in detection and classification tasks. This success was enabled primarily by extensive annotation and curation of training and validation data using the DSA. While our results are promising, this study currently has some important limitations. First, our experiments were performed on non-neoplastic bone marrow samples; performance on abnormal/disease samples was not established. Neoplastic or abnormal cells often exhibit cytomorphologic differences that will require additional annotation and the development and validation of new detection and classification models in future studies. Second, the performance on large ROIs encompassing marrow particles that are enriched in bone marrow cells that would typically be analyzed by pathologists has not been assessed. These areas contain more highly dense overlapping marrow cells and stromal cells that will need to be addressed by detection and classification models. Fourth, we have not established performance criteria for clinical validation of this novel method that is still early in development, but this will certainly be required before deploying for clinical use, as it has for automated analysis of blood smears [5, 26]. Any automated approach will ultimately have to be shown to be at least as reliable and accurate clinically as manual microscopic review of slides and faster than the manual DCC performance, even after reclassification of any cells wrongly classified categorized by pathologists/technologists. Future annotation efforts will include an inter-observer variability study to better understand the ranges for classification and detection performance of human observers. The final software application will also require a convenient graphical interface that allows users to identify errors and to manually override the algorithm. Lastly, while our algorithms performed well on samples processed in our lab, variations in pre-analytic factors like smearing and staining quality will impact generalization to other sites, and additional data collection would be required to deploy the system in other labs. Nonetheless, our annotation system and protocol offers an exciting template for others to generate similar training and validation data and results.

AUTHOR CONTRIBUTIONS

AAA and BRD generated and reviewed annotations. RC and NK developed algorithms and performed experiments. MA and DAG provided technical support for the annotation platform and database. LADC directed development and implementation of the annotation protocol, all computational approaches, and designed experiments. DLJ reviewed annotations, provided slides, conceived of the problem, and directed the project. RC, LADC, AAA and DLJ wrote and edited the manuscript.

ACKNOWLEDGEMENTS

This research was supported by the National Cancer Institute Informatics Technology for Cancer Research grants U01CA220401 and U24CA19436201. We thank ___ for helpful suggestions and comments on the manuscript.

FIGURES AND TABLES

Table 1. Counts of annotated cells by cytological class. Annotations outside ROIs were performed to increase counts for rare classes.

Cytological class	Total annotated	Inside ROI	Outside ROI
Erythroid	1526	1396	130
Blast	571	288	283
Promyelocyte	295	112	183
Myelocyte	613	414	199
Metamyelocyte	547	443	104
Band/neutrophil	1036	1005	31
Eosinophil	412	156	256
Basophil	62	21	41
Monocyte	178	109	69
Lymphocyte	544	363	181
Plasma	283	86	197
Megakaryocyte*	39	14	25
Unknown	3163	3162	1
Total	9269	7569	1700

ROI: Region-of-interest. * Annotated but not used in the cell detection or classification analyses.

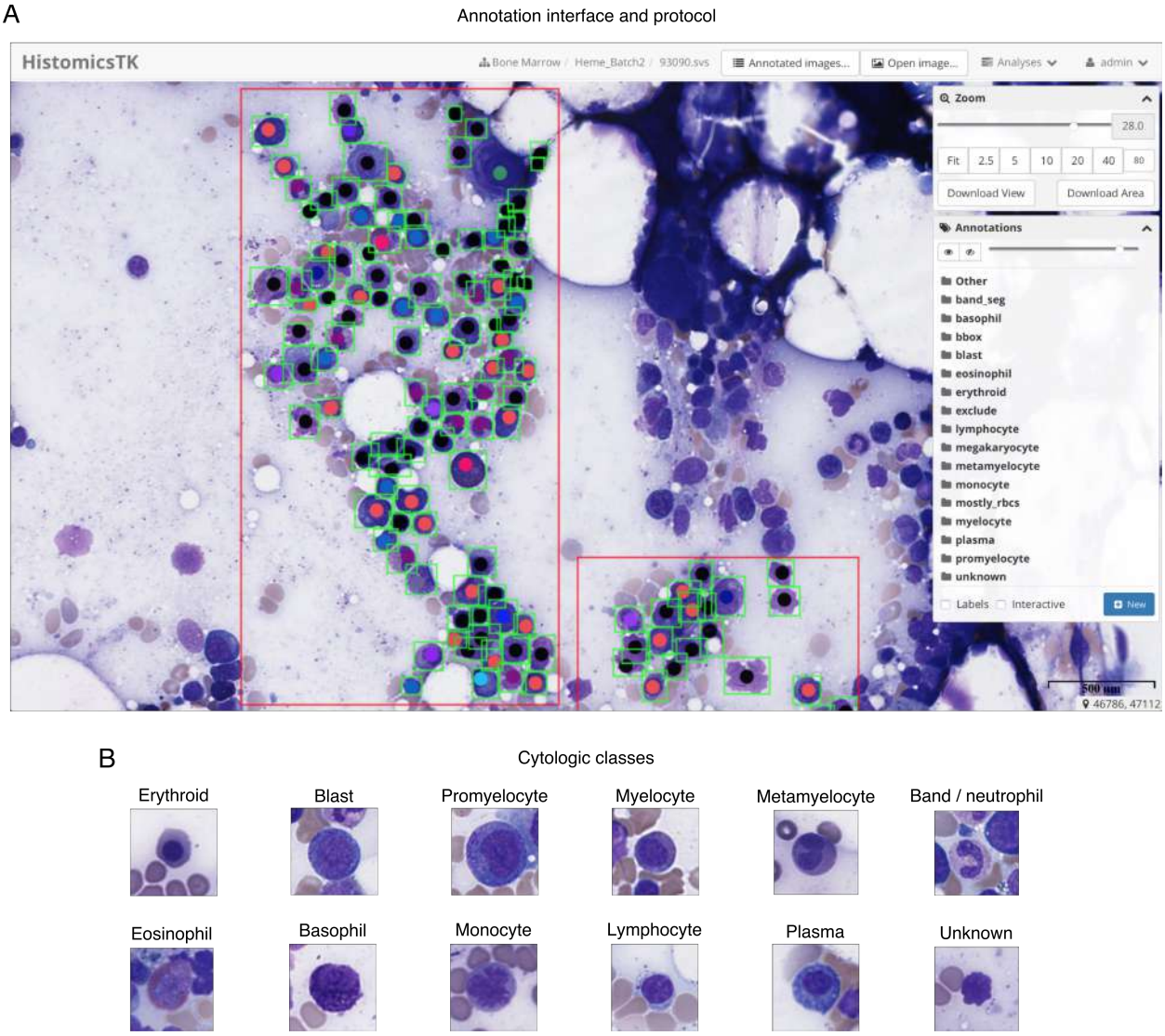

Figure 1. The Digital Slide Archive (DSA) annotation interface. (A) ROIs were defined using the rectangle draw tool, shown in red. Cells within these regions were then annotated exhaustively using the dot tool to encode cytologic class. Finally, bounding boxes, shown in green, were drawn around each annotated cell to delineate the cell boundary for detection algorithm training. The annotations are organized in layers in the *Annotation* menu, at right, where colors, transparency, and visibility of the annotation markers can be controlled. In addition to the layers for cytological classes, other layers are provided for the region-of-interest (“Other”), regions to exclude due to artifacts (“exclude”), and regions containing mostly red blood cells (“mostly_rbcs”). (B) Thumbnail images, representing examples of the 11 cell classes encountered in generating DCCs plus an unknown class, are shown.

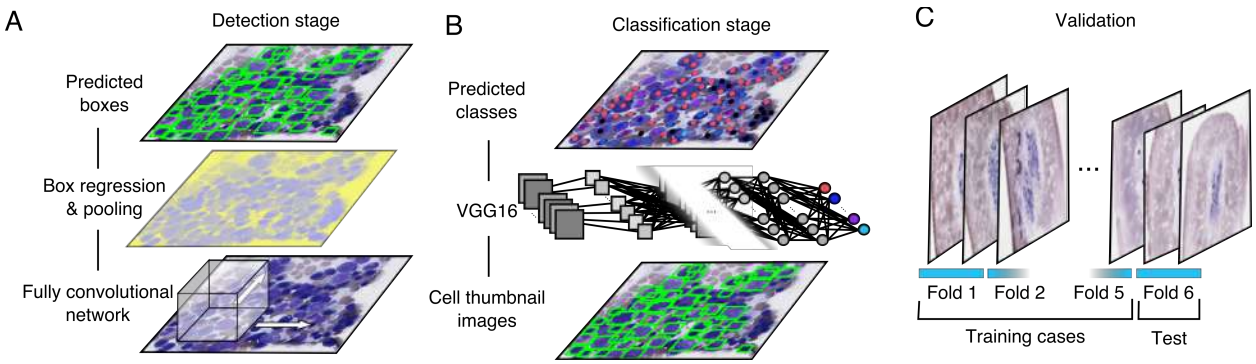

Figure 2. Computational detection and classification of cells in bone marrow aspirate smears. (A) Cell detection was performed using a Faster R-CNN network built on the resnet101 fully convolutional network. (B) Following cell detection, a separate convolutional network was used to classify the detected cells into 12 cytological classes. (C) The performance of this framework was evaluated through 6-fold cross validation to measure detection and classification accuracy using human annotations of cytological class and bounding box location. Validation was performed by at the case level, so that annotated cells from each case were assigned entirely to either the training or testing set.

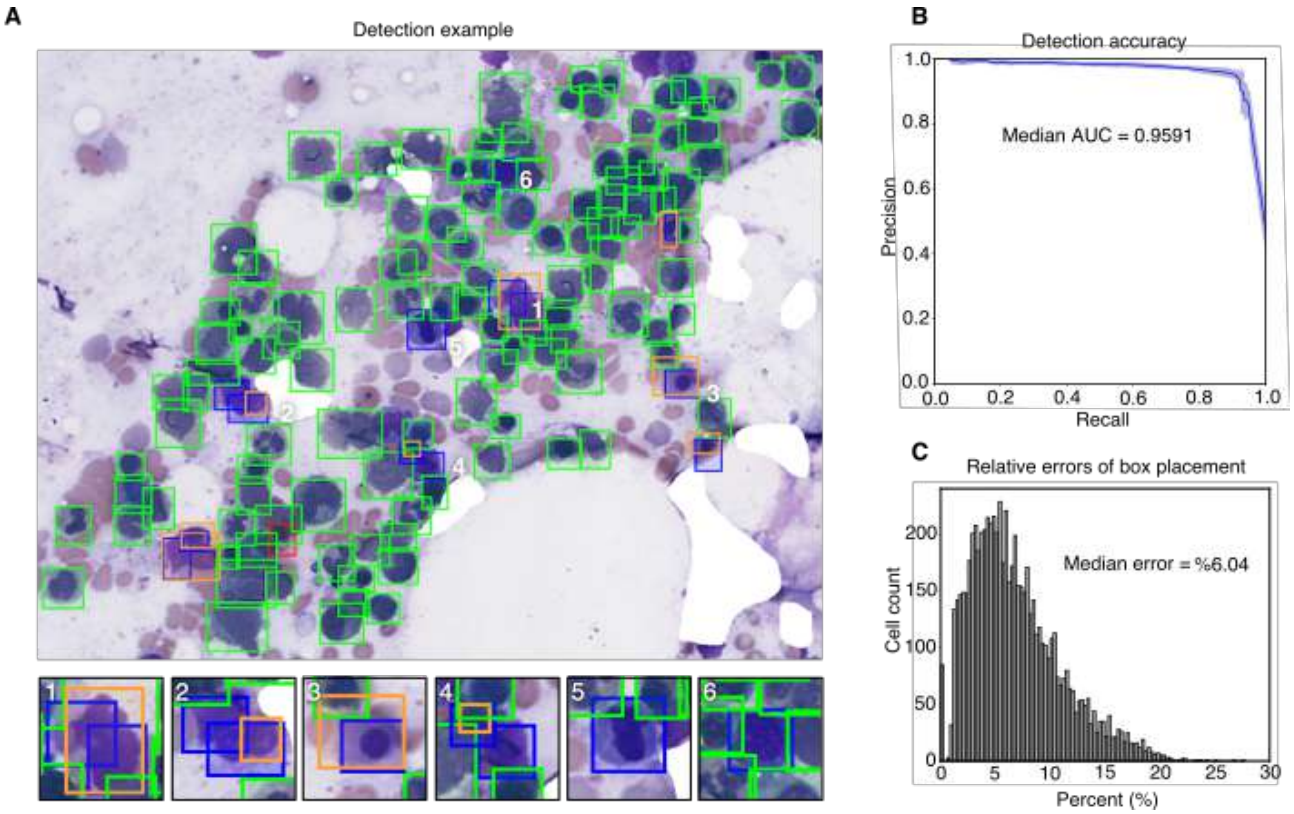

Figure 3. Cell detection results. (A) Sample detection result on test ROI. Here, green boxes indicate true-positive detections, red boxes false negatives missed by the detector, blue boxes false positives where a detection does not match ground truth, and orange boxes the ground-truth annotations that best correspond to false positives. In many cases (examples in panels 1-4), false positives result due to insufficient overlap with a ground truth annotation (intersection-over-union at least 0.5). Some false positives correspond to cells correctly detected by the algorithm but that were missed during the annotation process (examples in panels 5 and 6). (B) Precision-recall of detection algorithm on cross validation test sets. Shaded region indicates standard deviation of precision-recall over the six sets. (C) Histogram of bounding box placement relative error. For true positive detections this error measures the distance between predicted and actual bounding box centers relative to actual bounding box size.

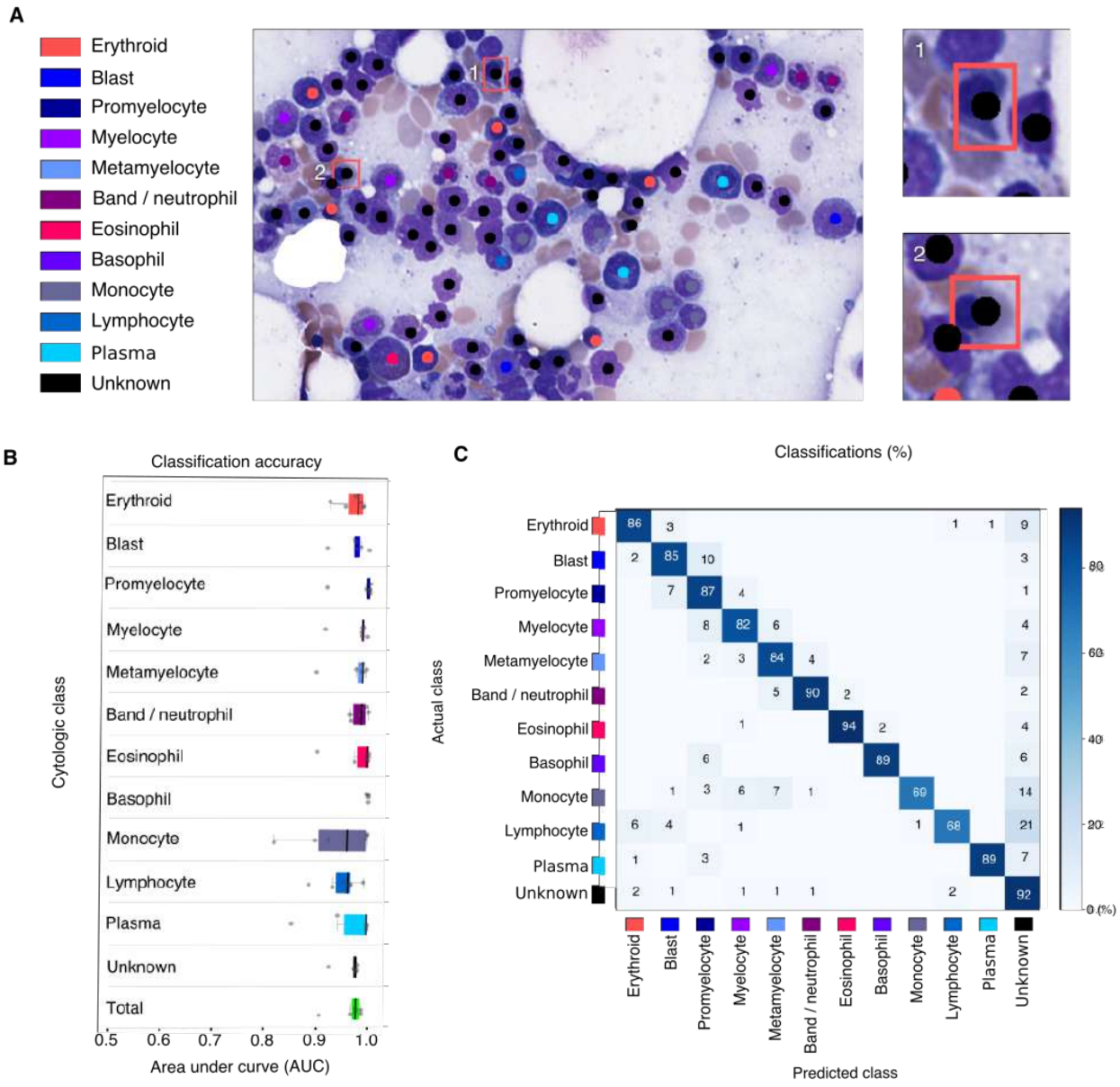

Figure 4. Cell classification results. (A) Sample classification result on test ROI. Predicted class and location of detected cells is indicated with color-code dots. Classification errors are indicated with a bounding box in sub panels 1 and 2, colored to indicate the annotated cell class. (B) Classification area-under-curve on cross validation test sets. Each point represents the AUC of one class in one testing fold. Total AUC was calculated as the average AUC over all classes (unweighted by class proportions). (C) Classification confusion matrix indicating the prediction errors that were made for each class. Results presented are aggregated over all folds.

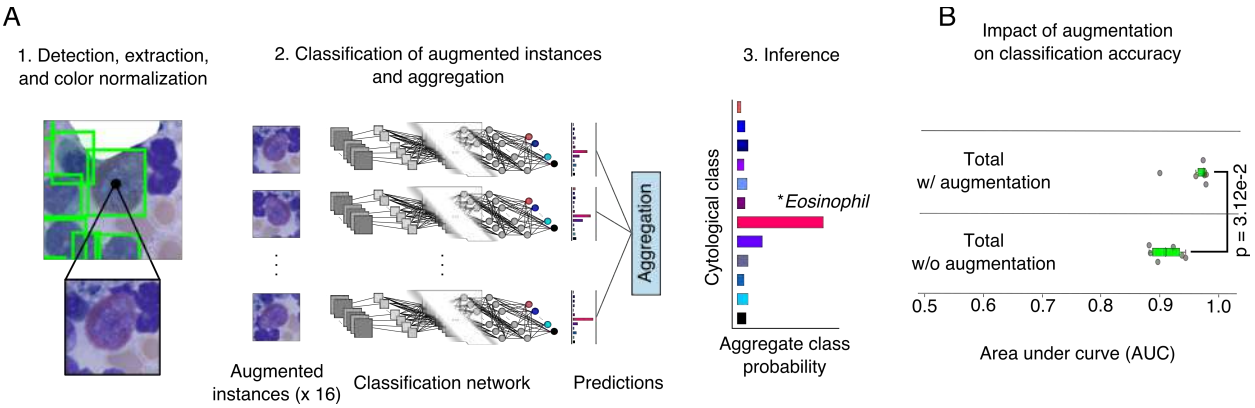

Figure 5. Impact of data augmentation on classification performance. (A) Data augmentation procedure for inference. (A.1) At inference time, for each detected cell we extract an image centered at the predicted bounding box location. (A.2) This image is transformed using rotations, translations, and pixel intensity transforms to generate an "augmented" set of 16 images for inference. These images are passed through the classification network to generate 16 total predictions of cytological class. Each prediction describes the probabilities that the image belongs to each of the 12 cytological classes. These predictions are aggregated to smooth out noise and to improve robustness. (A.3) The cell in question is assigned to the highest-probability cytological class using the aggregated predictions. (B) This augmentation procedure significantly improves classification accuracy (Wilcoxon signed rank test). Each dot represents the accuracy from one fold in the cross validation.

SUPPLEMENT

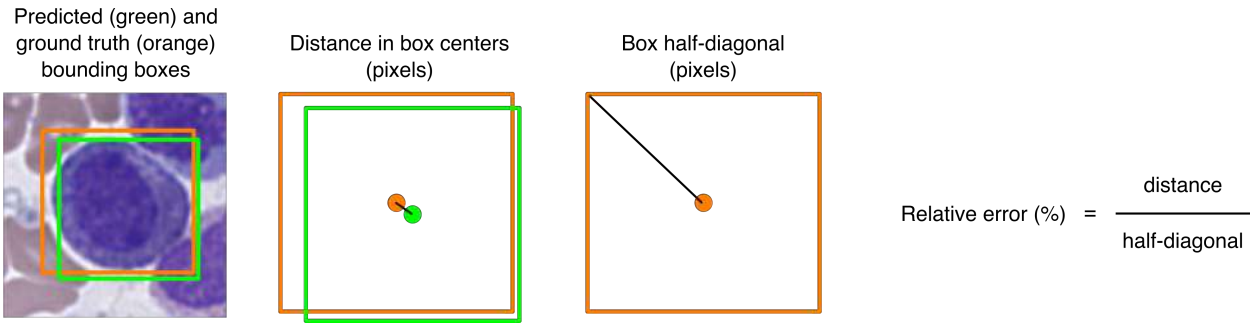

Figure S1. Definition of relative error for predicted bounding box position. Since larger cells can tolerate more error in the positioning of the bounding box, we developed a relative error that takes into account both the placement of the predicted bounding box and the size of the cell. The distance in pixels between the centers of the predicted and ground truth bounding boxes are first calculated. Then the diagonal of the ground truth bounding box is measured to represent the cell size. The relative error is then calculated as the ratio of box position error and box size.

Table S1. Precision recall values for data presented in Figure 3B.

Table S2. Classification performance data. (A) Classification accuracy for each class and cross-validation fold without augmentation at inference time. (B) Classification accuracy with augmentation at inference time.

Table S3. Error analysis for sample DCC cell counts. Expected ranges for over and under estimates of blast and plasma cell counts are presented for several scenarios. These ranges were calculated using the error estimates presented in the confusion matrix in Figure 4C. Each scenario was selected to represent the diagnostic threshold for a disease diagnosed by DCC.

REFERENCES

1. Swerdlow, S.H., et al., eds. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. 4th ed. Vol. 2. 2017, IARC publications: Lyon, France. 585.
2. Abdulrahman, A.A., et al., *Is a 500-Cell Count Necessary for Bone Marrow Differentials?: A Proposed Analytical Method for Validating a Lower Cutoff.* Am J Clin Pathol, 2018. **150**(1): p. 84-91.
3. d'Onofrio, G. and G. Zini, *Analysis of bone marrow aspiration fluid using automated blood cell counters.* Clin Lab Med, 2015. **35**(1): p. 25-42.
4. Mori, Y., et al., *Automation of bone marrow aspirate examination using the XE-2100 automated hematology analyzer.* Cytometry B Clin Cytom, 2004. **58**(1): p. 25-31.
5. Kratz, A., et al., *Performance evaluation of the CellaVision DM96 system: WBC differentials by automated digital image analysis supported by an artificial neural network.* Am J Clin Pathol, 2005. **124**(5): p. 770-81.
6. Janowczyk, A. and A. Madabhushi, *Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases.* J Pathol Inform, 2016. **7**: p. 29.
7. Sirinukunwattana, K., et al., *Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images.* IEEE Trans Med Imaging, 2016. **35**(5): p. 1196-1206.
8. Mobadersany, P., et al., *Predicting cancer outcomes from histology and genomics using convolutional networks.* Proc Natl Acad Sci U S A, 2018. **115**(13): p. E2970-E2979.
9. Saltz, J., et al., *Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images.* Cell Rep, 2018. **23**(1): p. 181-193 e7.
10. Bychkov, D., et al., *Deep learning based tissue analysis predicts outcome in colorectal cancer.* Sci Rep, 2018. **8**(1): p. 3395.
11. Senaras, C., et al., *DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning.* PLoS One, 2018. **13**(10): p. e0205387.
12. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
13. Choi, J.W., et al., *White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks.* Plos One, 2017. **12**(12).
14. Reta, C., et al., *Segmentation and Classification of Bone Marrow Cells Images Using Contextual Information for Medical Diagnosis of Acute Leukemias.* PLoS One, 2015. **10**(6): p. e0130805.
15. Gutman, D.A., et al., *The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research.* Cancer Res, 2017. **77**(21): p. e75-e78.
16. Glassy, E.F., *Color Atlas of Hematology; an illustrated field guide based on proficiency testing* 1998, Illinois, USA: College of American Pathologists.
17. Ren, S., et al. *Faster r-cnn: Towards real-time object detection with region proposal networks*. in *Advances in neural information processing systems*. 2015.
18. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
19. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
20. Davis, J. and M. Goadrich. *The relationship between Precision-Recall and ROC curves*. in *Proceedings of the 23rd international conference on Machine learning*. 2006. ACM.
21. Sokolova, M. and G. Lapalme, *A systematic analysis of performance measures for classification tasks.* Information Processing & Management, 2009. **45**(4): p. 427-437.

22. Bain, B.J., et al., *Dacie and Lewis Practical Hematology*. 11 ed. 2011, London: Churchill Livingstone. 668.
23. Ryan, D.H., *Examination of the Marrow*, in *Williams Hematology*, K. Kaushansky, et al., Editors. 2015, McGraw-Hill Education: New York, NY. p. 27-40.
24. Vollmer, R.T., *Blast counts in bone marrow aspirate smears: analysis using the poisson probability function, bayes theorem, and information theory.* Am J Clin Pathol, 2009. **131**(2): p. 183-8.
25. Lee, S.H., et al., *ICSH guidelines for the standardization of bone marrow specimens and reports.* Int J Lab Hematol, 2008. **30**(5): p. 349-64.
26. Briggs, C., et al., *Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system.* Int J Lab Hematol, 2009. **31**(1): p. 48-60.