Cell Selection-based Data Reduction Pipeline for Whole Slide Image Analysis of Acute Myeloid Leukemia

Jacqueline Kockwelp[1,2] Sebastian Thiele[1] Pascal Kockwelp[1] Jannis Bartsch[2]
Christoph Schliemann[2] Linus Angenendt[2] Benjamin Risse[1]
[1]University of Münster [2]University Medical Centre of Münster
[1]{firstname.lastname}@uni-muenster.de [2]{firstname.lastname}@ukmuenster.de

Abstract

Computer-aided analyses of cells in Whole Slide Images (WSIs) have become an important topic in digital pathology. Despite the recent success of deep learning in biomedical research, these methods are still difficult to apply to multi-gigabyte WSIs. To overcome this difficulty, a variety of patch-based solutions have been introduced, which however all suffer from certain limitations compared to manual examinations and often fail to meet the specificities of cytological inspections. Here we introduce an alternative scheme which incorporates clinical expertise in the selection process to automatically identify the clinically relevant areas. By using a bone marrow smear dataset containing 22*-gigapixel images of* 153 *patients, we introduce a novel pipeline combining unsupervised and supervised methodologies to gradually select the most appropriate single-cell regions, which are subsequently used in multiple medically crucial Acute Myeloid Leukemia (AML) predictions. Our approach is capable of dealing with a variety of common WSI challenges, massively limits the manual annotation effort, reduces the data by a factor of up to* 99.9% *and achieves super-human performance on the final cytological prediction tasks.*

1. Introduction

Digital Whole Slide Images (WSIs) have been widely adopted in pathology resulting in ever growing datasets of multi-gigapixel images. As a consequence, the demand for fully automatic analysis strategies has dramatically risen over the last decade [22]. Especially deep learning algorithms have proven to be powerful systems in biomedical image classification tasks and have also been used in histopathology [15]. For example deep learning has been used for tumor classification [21], cancer prediction [10], metastases detection [1], survival analysis [28], or the prediction of associated mutated genes [2].

Unfortunately, the analysis of WSI images using deep machine learning strategies is still challenging for mainly two reasons. Firstly, the extremely high resolution of WSIs (often more than $100,000 \times 100,000$ pixel) renders a direct application of state-of-the-art machine learning approaches impossible [5]. Secondly, the heterogeneity and variance within WSI data often include huge areas of non-informative, redundant, or even erroneous image proportions [30]. For example, non-optimal fixations, staining variabilities, inappropriate illumination, challenging image tiling or focus and bleed-through artefacts can lead to huge image areas with insufficient information [5].

1.1. Related Work

To overcome the above mentioned limitations a variety of patch-based selection strategies have been proposed in the past. Depending on the level of annotation three different selection strategies can be identified. On the lowest level multiple WSIs are annotated on the level of patients. These methods usually focus on non-local information by sampling candidate patches from all related WSIs which are further aggregated to extract representative features [31]. If the usage of multiple WSIs does not provide enough patches, data augmentation has been used to further increase the training data [29].

On the next level individual WSIs are annotated and subsequent patch selection aggregates the information within the image. For example Hou *et al.* used multi instance learning to aggregate the information across glioma WSI for histophathological image analysis [11]. In this two-stage approach a convolutional neural network (CNN) was used for patch selection followed by a support vector machine (SVM) to classify images based on these patches. This approach has been extended by an attention-based aggregation method [12] and further optimised to enable training with limited GPU memory capacities [27].

The finest labelling granularity is achieved by patch level annotations. For example, Korbar *et al.* used CNNs for individual patch prediction followed by a majority voting-

based colorectal polyp classifications [18]. A similar CNN-based approach was also used to detect lymphocytes on breast cancer WSIs [20] and other machine learning techniques have been compared for multi-class breast cancer histophathology based on patch level annotations [25].

All above mentioned strategies are patch exhaustive methods meaning that they incorporate the patches in a straightforward way to classify the image. Since these methods have a tendency to include non-informative and redundant regions and are also often computationally inefficient, patch selective methods have been introduced. These methods usually use low resolution representations of the WSIs to identify the relevant patches. For example Maksoud *et al.* utilises down-scaled WSIs to determine patch labels for the majority of autoimmune liver disease investigation cases [24]. In a different approach Katharopoulos *et al.* extended the multi instance learning algorithm from Hou *et al.* by selecting patches based on attention maps for low resolution WSI representations [16]. Hashimoto *et al.* use multi instance, domain adversarial, and multi-scale learning to evaluate random patches on multiple magnification levels [8]. To handle spatial and magnification based patch selection in a unified framework Zhang *et al.* suggest a novel attention mechanism on down-sampled WSIs followed by patch selection mechanisms that also consider the most informative magnification [30].

All above mentioned approaches either suffer from inaccurate labels (i.e. patient or WSI level annotations) or require tedious patch level annotations. Especially in the latter annotation strategy medical experts have to inspect a huge amount of often obviously non-informative image regions which further aggravates the labelling situation. Moreover, many attention-based patch selection strategies have to operate on down-sampled images which is unacceptable in situations where comparatively small structures such as individual cells carry the most value. Especially in cytology, where cell entities rather than broader image regions (i.e. patches) are inspected by clinicians, none of the above mentioned approaches incorporates cell selection strategies from medical experts directly into the data reduction framework.

1.2. Contribution

In this paper we present a novel cell selection-based WSI pipeline to classify Acute Myeloid Leukemia (AML) in multi-gigabyte images for end-to-end cytological analysis. Our data reduction approach incrementally combines unsupervised and supervised processing steps, allowing to incorporate unbiased medical expert knowledge at the very end of our cell selection process by means of a cell image quality grading network, which ultimately limits the annotation effort. We evaluated our approach using a novel bone marrow smear dataset containing huge WSIs ($218,944 \times 103,704$ pixel resolution, approx. 42GB per image) of 153 patients (total dataset size 6.43TB) which comprises a variety of well-known challenges (e.g. staining variability, out-of-focus regions, image tiling artefacts etc.). The algorithmically selected cells were analysed in a clinical evaluation to predict two AML associated mutations (*NPM1* [3], *FLT3* [4]) and the overall genetic risk according to the European LeukemiaNet (*ELN2017* [6]) classification. *NPM1* mutations can often be identified by blasts with so-called cup-shaped nuclei, whereas *FLT3* mutations and *ELN2017* genetic risk categories cannot be assessed by the clinicians from the images directly. Moreover all predictions are extremely relevant indicators for oncological treatment decisions, rendering these characteristics ideal candidates for our study. Our results indicate that our cell-selection pipeline is capable of reducing the data by more than 99.9% while effectively addressing the above mentioned challenges and enabling modern deep learning image analysis procedures for fully-automatic cytological AML prognostication with super-human performance.

2. Method

Our approach can be separated into a data reduction pipeline (subsection 2.1) and clinical prediction architectures (subsection 2.2), which are described in more detail below.

2.1. Data Reduction

To enable deep learning-based WSI analysis massive data reduction is necessary, which is usually achieved by patch selection procedures [5]. Here we introduce an alternative cell selection pipeline for the inference of clinicopathological features from AML cytomorphology which is motivated by cytological WSI examinations. Medical experts mainly inspect individual and well-separated blasts and residual haematopoietic cells to infer morphological features. Avoiding dense cell clusters might help to mitigate the risk of possible data biases since mutated phenotypes should be scattered sparsely accross the scan while cell clusters likely belong to the same clone, thus not necessarily adding any additional information in single-cell inspections.

This expert knowledge is reflected by our pipeline which reduces the amount of data in three incremental levels as illustrated in Figure 1:

Level One: 512×512 pixel multi-cell selections, in which empty, blurry or too dense areas are filtered out.

Level Two: 80×80 pixel single-cell selections, additionally filtered by compactness.

Level Three: 80×80 pixel single-cell selections, additionally filtered by a small neural network (called grad-

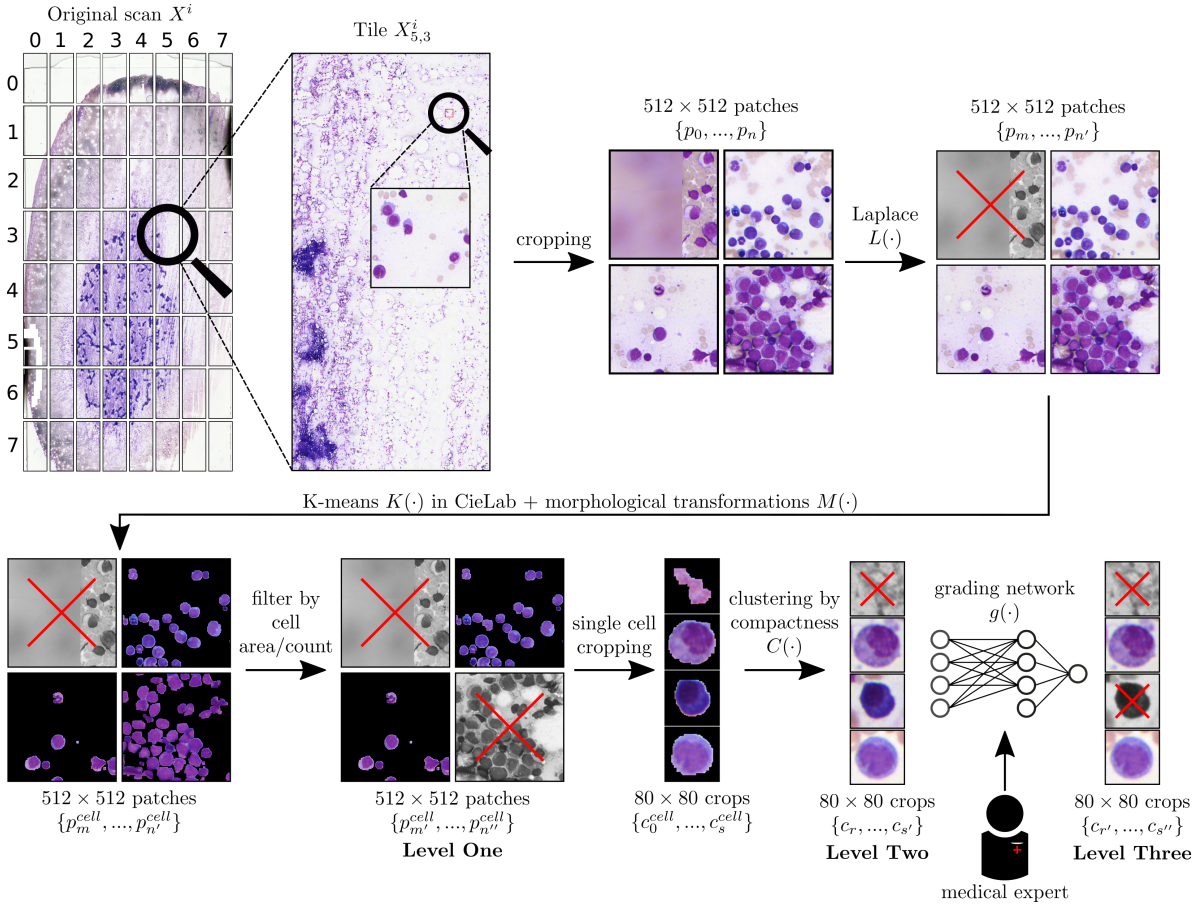

Figure 1. Starting with a WSI X^i the inner most 24 tiles $\{X^i_{x,y} \mid x \in \{2, ..., 5\}, y \in \{1, ..., 6\}\}$ are kept for further processing. Each tile is split up into 512×512 pixel patches. These patches are successively filtered by blurriness with a Laplace filter $L(\cdot)$, segmented by K-means $K(\cdot)$ in the CieLab colour space and refined with morphological transformations $M(\cdot)$ and filtered by cell area and counts, resulting in the Level One dataset $\{p_{m'}, ..., p_{n''}\}$. Afterwards, 80×80 pixel crops are extracted from the patches and filtered by compactness using K-means clustering $C(\cdot)$. The remaining crops $\{c_r, ..., c_{s'}\}$ constitute the Level Two dataset. Finally, the Level Two data is filtered by a grading network $g(\cdot)$, creating the final Level Three dataset $\{c_{r'}, ..., c_{s''}\}$. These steps are executed for all WSIs and their respective tiles.

ing network) based on task agnostic expert annotations of few single-cell images. These annotations classify a cell's overall condition, e.g. whether it's a cell at all or if it's heavily damaged.

For each level different image representations are extracted, namely a raw image crop and a crop in which the non-cell background pixels are set to zero. Since the different data reduction levels operate in a hierarchical fashion we will describe our algorithm in an incremental fashion.

2.1.1 Level One

In a first step, the $218,944 \times 103,704$ original images $\{X^0, ..., X^N\}$ are divided into 8×8 grids, resulting in $12,963 \times 27,368$ patches $X^i_{x,y}$ with $i \in \{0, .., N\}$ and $x, y \in \{0, ..., 7\}$. From each grid we dismiss the outer four columns and the outer two rows, which are mostly blank or consist of out of focus regions, so that the innermost 24 large patches remain. The images are further subdivided into 512×512 pixel patches $\{p_0, ..., p_n\}$ and the remaining out-of-focus regions are removed by converting the patches into single channel images which are convolved with a Laplace filter $L(\cdot)$. Blurry regions are excluded by removing patches with low variance values resulting in a reduced set of patches $\{p_m, ..., p_{n'}\}$ with $n' \leq n$ and $m \in \{0, ..., n'\}$.

Next, a segmentation strategy is used to separate the bone marrow cells from the background leading to image patches p^{cell}_i. Our bone marrow smear images imposed several challenges regarding the cell segmentation, namely the non-optimal separability in RGB colour space, staining variability and significant intensity differences due to different illuminations, aging corners of the slides and different camera settings. We therefore converted the image crops

into the CieLab space [14] and reduced the colour information to two dimensions a and b by omitting the L component (i.e. the luminescence). The subsequent segmentation step is motivated by the approach of Kumar and Udwadia [19]. First, the patches p_i are clustered using the K-Means algorithm $K(\cdot)$ [23] on the a and b channels. In contrast to Kumar *et al.*, in which the number of clusters was set to $k = 4$ to identify the cytoplasm, nuclei, background and other cells such as erythrocytes [19], we had to segment entire blasts (cell nucleus including the cytoplasm). Therefore we initialised K-Means with $k = 3$ and selected blast cluster centroids based on the highest instead of the lowest red colour value intensity. The cell masks p_i^{cell} were created by setting the non-cell clusters (i.e. neither cell nuclei nor cytoplasm) to zero followed by morphological operations $M(\cdot)$ [7] to close small gaps within cells and remove artefacts in the cytoplasm.

As mentioned above, only areas with well spread out blasts are of interest so that crops containing cell clutter, medullary nodule or other unwanted components have to be excluded. This is done by extracting the individual contours within p_i^{cell} and assessing their absolute number and respective area coverage. Experimentally determined thresholds for the minimum and maximum number of clusters as well as the maximum area of a cell were used to extract the final Level One 512×512 pixel image crops $\{p_{m'}, ..., p_{n''}\}$ and $\{p_{m'}^{\text{cell}}, ..., p_{n''}^{\text{cell}}\}$ with $n'' \leq n'$ and $m' \in \{m, ..., n''\}$.

2.1.2 Level Two

The already reduced 512×512 pixel image crops p_i^{cell} of Level One can be further processed into single-cell images. For this purpose, 80×80 pixel crops centered around each cell in the images are extracted resulting in the set $\{c_0^{\text{cell}}, ..., c_s^{\text{cell}}\}$. For each single-cell crop its compactness, which is defined as $\frac{\text{perimeter}^2}{\text{area}}$, is additionally calculated. This value indicates the *roundness* of the cell.

Based on the calculated values, a K-Means clustering $C(\cdot)$ with $k = 2$ is performed. The resulting model can then be used to filter images that do not contain cells based on an experimentally derived threshold to compute the Level Two dataset $\{c_r^{\text{cell}}, ..., c_{s'}^{\text{cell}}\}$ with $s' \leq s$ and $r \in \{0, ..., s'\}$. To also include cell crops with background intensities the same regions were extracted from p_i to derive $\{c_r, ..., c_{s'}\}$.

2.1.3 Level Three

Up to now our cell selection scheme only utilised colour information and simple geometric priors while substantially reducing the amount of data to single-cell 80×80 image crops. Even though this reduction was based on straightforward medical intuition, single-cell assessments as performed by a clinical expert have not been included so far resulting in crops containing damaged cells or staining artefacts. Since the incorporation of this type of knowledge is far from trivial for conventional computer vision methods we applied a supervised deep learning paradigm for the Level Three data reduction. Moreover and due to the unsupervised selection scheme described above crop annotations can be done efficiently on the selected areas of Level Two. To further accelerate the labelling process medical experts had to grade the selections into six discrete classes ranging from *no cell* (6) over *damaged cell* (3) to *valid cell* (1).

Gradings	1	2	3	4	5	6
Quantity	36	103	78	148	296	845

Table 1. Gradings of Level Two single-cell images by a medical expert.

This way $1,506$ crops were annotated in 2.5 hours and the resulting gradings are given in Table 1. Subsequently, we trained a ResNet18 [9] on this data, which was used as a grading network $g(\cdot)$ to filter out single-cell images of the grades 5 and 6. This results in the Level Three datasets $\{c_{r'}^{\text{cell}}, ..., c_{s''}^{\text{cell}}\}$ and $\{c_{r'}, ..., c_{s''}\}$ with $s'' \leq s'$ and $r' \in \{r, ..., s''\}$.

2.2. Architectures

To demonstrate the advantages of our incremental processing pipeline we evaluated our data reduction strategy by analysing the results of all crop levels in three highly relevant AML classification tasks, namely the official *ELN2017* [6] genetic risk classification and two related mutations called *NPM1* [3] and *FLT3* [4]. We tested four different network architectures as backbone feature extractors: InceptionV3 [26], ResNet50 [9], ResNet18 [9] and a modified version of ResNet18. In the modified ResNet18 the first 7×7 convolution was replaced by a 3×3 filter and an additional residual block with a projection shortcut replaced the first max pooling to enable the learning of fine grained textual features of the cell.

Image crops resulting from Level Two and Three were used in a multi instance learning approach (i.e. several images are used as an input at the same time) since not all cell selections might show the respective phenotype. Since Level One crops already incorporate multiple cells these images were used in a single instance learning paradigm, which also reduced the computational requirements for these higher resolution crops.

For our multi instance learning approach we followed a similar principle to Ilse *et al.* [13]. Each input crop is processed by a backbone feature extractor with no classification head, which outputs a $1 \times 1 \times d$ representation of the image, where d stands for the dimensionality of the representation depending on the backbone that was used.

	NPM1		FLT3		ELN2017	
	Reduction	#Images	Reduction	#Images	Reduction	#Images
Level One	-	-	-	-	84.5%	157,142
Level Two	-	-	-	-	98.2%	710,944
Level Three	99.9%	200,276	99.8%	156,342	99.9%	142,270

Table 2. Overview of datasets used in the experiments. "Reduction" indicates by what percentage the number of pixels has been reduced compared to the original WSI selection.

In line with recent work on patch-based WSI analysis [30] we evaluated two alternatives of multi instance learning, namely with and without an attention mechanism. For the attention-based approach, each representation is passed through a Multilayer Perceptron (MLP). The MLP consists of two dense layers with 2048 and 1 neuron respectively. The first layer uses ReLU and the second layer uses a Sigmoid activation function. As a consequence, the output of the MLP is a scalar that is multiplied by the corresponding image representation of the used backbone, effectively weighing its importance and therefore providing attention to the different samples. In the non attention-based approach no additional weighting is used. Finally, all (weighted or non-weighted) representations are passed through a global average pooling layer and a last dense layer to provide the respective classification result.

3. Experiments

For all our experiments we used the TensorFlow framework. The data reduction was performed in parallel on an *AMD Ryzen 9 5950X* and a *RTX 2080* was used to perform training and inference of the deep learning architectures. For optimisation, we used the Adam [17] optimiser with a learning rate of $lr = 0.001$. The models for the Level One dataset were trained for 10 epochs, while the multi-input models for Level Two and Level Three were trained for 30 epochs. For all multi-input trainings the number of input images has been set to 32. In addition, the training data of a WSI is shuffled every epoch, meaning that the multi instance learning input stacks are random combinations of cell selections to avoid identical stacks during training. This also decreases the tendency of having neighboring cells in the same stack and also reduces potential overfitting, since the network cannot memorize features of particular input combinations.

3.1. Datasets

For our clinical study we evaluated three different and highly relevant AML classification tasks:

ELN2017 stands for the genetic risk classification [6], which was introduced by the European LeukemiaNet in 2017. It is used to divide the molecular and cytogenetic alterations into the following three risk classes with distinct outcome: *favorable*, *adverse* and *intermediate*.

NPM1 is the abbreviation for the *nucleophosmin* gene. It can be either *mutated* or *wildtype*. This genotype has a major influence on the prognosis. In the past, it has been shown that *NPM1* mutations show subtle morphological anomalies (i.e. blasts with cup shaped nuclei) which can be recognised by a trained oncologist. [3]

FLT3 is the abbreviation for the *fms like tyrosine kinase 3* gene. It can be either *mutated* or *wildtype*. To date, this mutation cannot be recognised in the WSIs by a clinician. [4]

Since not all labels were available for each patient, the number of patients and thus the number of WSI slides used differs between the experiments. An overview of resultant dataset sizes is given in Table 2.

For training, we used a 5-fold cross-validation with patient-wise splits. Each fold is therefore separated into an 80% training and a 20% validation subset. Since the number of extracted single-cell crops can vary greatly between scans, WSIs for which less than 100 cell crops were found are not considered for the training. Moreover, for all cross-validation sets for the binary classifications the training and validation sets were chosen to be close to uniform. For the *ELN2017* categorisation, the *intermediate* class was strongly underrepresented, which prevented balanced splits. Instead, we tried to balance this dataset with the main focus on the classes *favorable* and *adverse*.

3.2. Results

In order to demonstrate the incremental benefit of our multi-level processing pipeline we evaluated each level independently.

Level One Both of the Level One datasets (with and without background) were used in combination with the *ELN2017* label to train a ResNet18. The larger ResNet50 architecture was trained, but only for the fifth cross-validation fold due to the high computation time and since we only wanted to examine, if a much deeper network would pose a significant performance gain. The results of these first trainings are listed in Table 3. The best overall result of 0.567 validation accuracy was achieved by a ResNet18 (for this 3-class classification problem) on the dataset without background. Tendentiously the images without background yielded a slightly higher validation accuracy than the ones with the full background information. For all folds it is noticeable that the values of the training accuracy and loss with background show a much better performance on the training set than those without background.

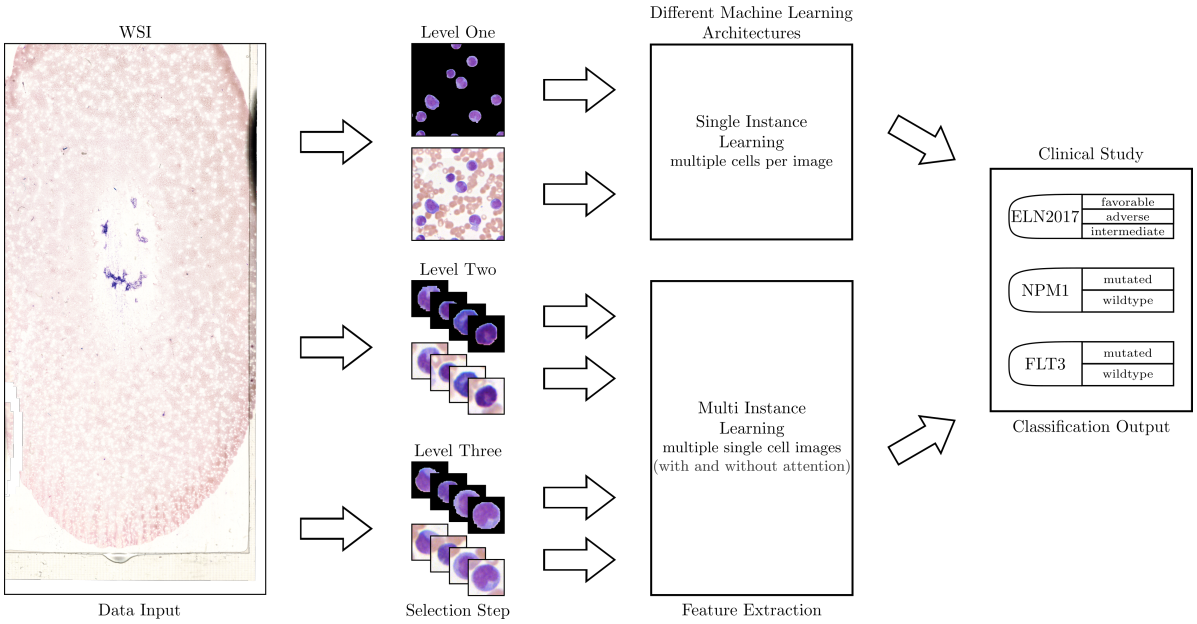

Figure 2. Overview of experimental design. Each WSI is reduced by three different levels of data reduction. The resulting datasets (with and without background) are evaluated using different machine learning strategies to derive three clinically relevant labels from the images.

	ResNet18 (without background)				ResNet18 (with background)			
	Training		Validation		Training		Validation	
	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
0	0.810	0.623	2.075	**0.567**	0.404	0.869	4.713	0.475
1	0.794	0.657	1.109	0.493	0.448	0.850	3.004	0.504
2	0.781	0.656	1.073	0.558	0.472	0.838	1.395	**0.553**
3	0.795	0.644	0.984	0.535	0.363	0.891	2.449	0.505
4	0.735	0.688	1.432	0.481	0.354	0.900	2.775	0.475

Table 3. Results of the 5-fold cross-validation of the Level One datasets without and with background for the *ELN2017* label.

Additionally, the validation loss values are also a lot higher in the experiments with background than in the trials without background. These first results confirm the medical expert's intuition, namely that only the cell parts of the images contain relevant information. Moreover these results also indicate background information can lead to more overfitting, which shows itself in the better training but worse validation performance of the models. To test whether deeper architectures could yield better accuracies on the relatively complex Level One crops we also evaluated a ResNet50 on this data [9]. However, the ResNet50's results were in line with the other Level One experiments. Even though a better validation accuracy was achieved compared to the ResNet18, its effect was very low. Considering the much higher training time and memory overhead (especially for the subsequently used multi instance learning) we discarded the ResNet50 for the Level Two and Three experiments.

Level Two The results of the Level Two evaluation with and without background for *ELN2017* are listed in Table 4 and Table 5 respectively. A ResNet18 and the modified version of ResNet18 were used as the backbone for non attention-based multi instance learning with and without background inclusion. The modified model provided overall better results compared to the conventional ResNet18. This is why all further tests of Level Three do not include the original ResNet18 version from [9].

The best result with a 0.763 validation accuracy was achieved by an InceptionV3 network on the Level Two dataset with background. Overall, however, it is not possible to tell which network performed best, as the best model varies from fold to fold. Compared to the non attention-based multi instance learning approaches, the attention-based modified ResNet18 version seems to perform worse or at least does not provide a significant performance improvement.

In evaluations on other labels, such as *NPM1*, the attention-based approach also sometimes did not manage to converge at all, thus only providing random predictions. Since the attention mechanism has more computational overhead without any measurable benefit, we did not evaluate it further on the Level Three data. With regard to the question whether a dataset with or without background provides better results or is more prone to overfitting, no clear answer can be given. The test results are similar and there is no indication of the Level Two dataset with background to be more prone to overfitting than the one without. We hypothesise, that this is because large parts of the back-

	ResNet18				modified ResNet18				modified ResNet18 Attention based				InceptionV3			
	Training		Validation		Training		Validation		Training		Validation		Training		Validation	
	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
0	0.342	0.924	2.961	0.568	0.339	0.926	3.461	0.539	0.350	0.921	3.027	0.530	0.693	0.693	1.034	0.608
1	0.499	0.865	5.510	0.397	0.603	0.813	2.700	0.492	0.584	0.838	6.415	0.433	0.451	0.809	4.323	0.444
2	0.277	0.938	3.129	0.498	1.050	0.700	2.555	0.500	1.028	0.679	1.619	0.523	0.193	0.922	4.105	0.481
3	0.272	0.945	2.866	**0.632**	0.253	0.940	1.893	0.653	0.620	0.778	1.467	0.573	0.654	0.733	1.113	0.652
4	0.475	0.858	3.459	0.614	0.256	0.942	3.564	**0.759**	0.295	0.926	2.539	**0.688**	0.410	0.832	1.650	**0.705**

Table 4. Results of the 5-fold cross-validation of the Level Two dataset without background for the *ELN2017* label.

	ResNet18				modified ResNet18				modified ResNet18 Attention based				InceptionV3			
	Training		Validation		Training		Validation		Training		Validation		Training		Validation	
	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
0	0.201	0.975	4.068	0.535	0.249	0.958	5.827	0.577	0.430	0.886	4.197	0.545	0.365	0.852	1.418	0.632
1	0.246	0.964	7.401	0.307	0.255	0.958	6.291	0.353	0.350	0.918	4.653	0.416	0.485	0.803	2.184	0.484
2	1.007	0.746	4.034	0.453	0.281	0.937	2.881	0.506	0.924	0.658	1.356	0.514	0.184	0.930	6.668	0.524
3	0.233	0.956	2.082	0.599	0.295	0.934	2.444	0.691	0.499	0.858	1.777	**0.701**	0.578	0.777	1.340	0.558
4	0.218	0.973	2.296	**0.710**	0.215	0.970	2.485	**0.707**	0.586	0.813	3.306	0.632	0.249	0.904	1.923	**0.763**

Table 5. Results of the 5-fold cross-validation of the Level Two dataset with background for the *ELN2017* label.

ground are already filtered out at this point, and only small portions of the background around the single-cells are still visible, largely removing a model's tendency to overfit on the training set.

Grading Network The unsupervised cell selection mechanism used to extract the Level Two crops is based on colour and contrast heuristics and does not incorporate any expert intuition on the descriptive potential of these cells. For example, damaged cells or staining artefacts can also be included so that these images need to be removed, so that only samples that would have been chosen by a medical expert remain. To integrate expert knowledge on the appearance of informative cells we incorporated a dedicated grading network as mentioned in subsubsection 2.1.3. Importantly, the clinicians were only asked to grade cell selections, which have been automatically selected by the Level Two pipeline based on qualitative criteria such as no cell at all (grading 6) or valid cell (grading 1). No disease-related assessments were considered so that gradings 1 to 4 refer to good to medium overall cell crops and gradings 5 and 6 indicate generally inappropriate cell representations.

The results of the expert gradings in absolute numbers are shown in Table 1. These results indicate that there are still many uninformative single-cell images in the dataset, which can potentially impede the classification performance. For example, about 75% of all crops were considered inappropriate for cytological assessments and could therefore impede the machine learning based classification task.

The grading network was trained by aggregating gradings 1 to 4 (useful condition) and gradings 5 and 6 (unsuitable condition). Given the imbalanced nature of these two classes a loss weighting of 3 to 1 was used for training and a validation accuracy of 0.881 was achieved. This performance again indicates that medical experience can be objectified and serves as a powerful selection criterion for cytological WSI analyses. Once trained the grading network was used to predict cell selections to complete our Level Three data reduction.

Level Three For the final training iteration, the Level Three dataset without background was evaluated on the *ELN2017*, *NPM1* and *FLT3* labels since Level Two studies indicated superior results using this type of data. For each label the modified ResNet18 and the InceptionV3 network have been trained. The results are listed in Table 6, Table 7 and Table 8. The results on the Level Three *ELN2017* dataset show an improvement compared to Level Two. Especially the InceptionV3 architecture benefited from the additional filtering and also appeared to be less volatile across the different folds. In fact, the average validation loss over all folds improved from 2.45 to 1.23 and the average validation accuracy increased from 0.578 to 0.659. The modified ResNet18 backbone improved with a reduced average validation loss (2.83 to 2.41) and a slightly increased average validation accuracy could be measured (0.588 to 0.591). The performance difference between the InceptionV3 and ResNet18 backbone might be caused by the advantages of the InceptionV3 model with respect to overfitting [26].

An overview of the *ELN2017* classification accuracy for each level (with and without background) is given in Fig-

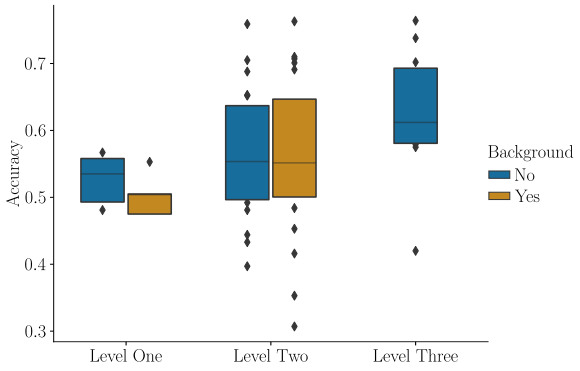

Figure 3. Boxenplot of validation accuracy for *ELN2017* by data reduction level including all tested backbones and background modalities.

ure 3. Evidently Level One cell images achieve mediocre performances across all experiments, which can be increased by using individual cell images (Level Two). Due to erroneous cell selections the overall performance is however more erratic resulting in a higher variance for Level Two, especially when the background is included. The grading network-based filtering can overcome this limitation by stabilising the results while further improving the performance so that the Level Three selections achieve the best accuracy measured in our experiments.

In a similar fashion the classification of *NPM1* mutations benefited from the more selective Level Three data reduction. Especially the InceptionV3 model achieved better performances compared to Level Two cell selections and also showed more stable trainings and less volatility with respect to the different training folds. These findings support the hypothesis that expert knowledge in the selection of cells is advantageous for cytological examinations.

The results of the *FLT3* classification are of particular interest from a clinical point of view, since this type of mutation cannot be inferred from manual WSI inspections directly. Surprisingly, classification performances similar to *NPM1* could be achieved with validation accuracies of more than 0.7 in all folds with the only exception of one fold in the InceptionV3 architecture (the model always predicted the same label for this fold). This suggests that subtle spectral features such as fine grained textual pattern are indicative for *FLT3* and are available in the WSI data which however escape current clinical examinations.

4. Discussion & Conclusion

In this paper we introduced a novel cell selection scheme to analyse histological multi-gigabyte WSI dataset. Instead of using patch-based selection schemes, we used a combination of supervised and unsupervised machine learning techniques to extract the most informative image regions making it particularly suitable for cytological investigations. We evaluated our pipeline on a novel bone marrow smear dataset and classified three important characteristics for acute myeloid leukemia diagnostics. To the best of our knowledge our algorithm is the first approach to achieve very good performances on those tasks while automatically processing entire WSI data. Furthermore it is able to reduce the amount of data by more than 99.9%.

In the future we will evaluate our cell selection scheme on different datasets and will extend the classification outputs based on the clinical records of the patients. We will also investigate alternative processing steps along our pipeline and will use our model in an extended clinical study to demonstrate its translational applicability.

	modified ResNet18				InceptionV3			
	Training		Validation		Training		Validation	
	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
0	0.402	0.908	2.641	0.625	0.541	0.774	2.015	0.666
1	0.629	0.808	4.756	0.599	0.948	0.548	0.754	0.702
2	0.796	0.662	1.543	0.575	0.610	0.727	1.357	0.579
3	0.738	0.728	2.124	0.420	0.997	0.452	1.301	0.586
4	0.526	0.849	1.004	**0.738**	0.840	0.596	0.732	**0.764**

Table 6. Results of the 5-fold cross-validation of the Level Three dataset without background for the *ELN2017* label.

	modified ResNet18				InceptionV3			
	Training		Validation		Training		Validation	
	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
0	0.184	0.970	1.464	0.762	0.240	0.903	0.700	0.728
1	0.330	0.881	0.594	0.781	0.161	0.935	0.625	0.772
2	0.265	0.904	0.640	0.782	0.161	0.933	0.598	**0.834**
3	0.250	0.917	0.591	**0.791**	0.476	0.765	0.409	0.831
4	0.324	0.911	0.671	0.772	0.439	0.801	0.555	0.815

Table 7. Results of the 5-fold cross-validation of the Level Three dataset without background for the *NPM1* label.

	modified ResNet18				InceptionV3			
	Training		Validation		Training		Validation	
	Loss	Acc.	Loss	Acc.	Loss	Acc.	Loss	Acc.
0	0.587	0.872	1.097	0.787	0.383	0.818	0.756	**0.832**
1	0.256	0.913	0.457	**0.836**	0.438	0.804	0.466	0.755
2	0.481	0.834	0.842	0.748	-	-	-	-
3	0.403	0.854	0.658	0.734	0.345	0.847	1.507	0.786
4	0.789	0.816	0.722	0.812	0.371	0.849	2.159	0.723

Table 8. Results of the 5-fold cross-validation of the Level Three dataset without background for the *FLT3* label.

Acknowledgements

JK and BR would like to thank the Deutsche Forschungsgemeinschaft (DFG) CRU326. BR would also like to thank the Deutsche Forschungsgemeinschaft (DFG) – CRC 1450 – 431460824.

References

[1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 1

[2] Mingyu Chen, Bin Zhang, Win Topatana, Jiasheng Cao, Hepan Zhu, Sarun Juengpanich, Qijiang Mao, Hong Yu, and Xiujun Cai. Classification and mutation prediction based on histopathology h&e images in liver cancer using deep learning. *NPJ precision oncology*, 4(1):1–7, 2020. 1

[3] Weina Chen, Sergej Konoplev, L. Jeffrey Medeiros, Hartmut Koeppen, Vasiliki Leventaki, Saroj Vadhan-Raj, Dan Jones, Hagop M. Kantarjian, Brunangelo Falini, and Carlos E. Bueso-Ramos. Cuplike nuclei (prominent nuclear invaginations) in acute myeloid leukemia are highly associated with flt3 internal tandem duplication and npm1 mutation. *Cancer*, 115(23):5481–5489, 2009. 2, 4, 5

[4] Naval Daver, Richard F Schlenk, Nigel H Russell, and Mark J Levis. Targeting flt3 mutations in aml: review of current knowledge and evidence. *Leukemia*, 33(2):299–312, 2019. 2, 4, 5

[5] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, page 264, 2019. 1, 2

[6] Hartmut Döhner, Elihu Estey, David Grimwade, Sergio Amadori, Frederick R. Appelbaum, Thomas Büchner, Hervé Dombret, Benjamin L. Ebert, Pierre Fenaux, Richard A. Larson, Ross L. Levine, Francesco Lo-Coco, Tomoki Naoe, Dietger Niederwieser, Gert J. Ossenkoppele, Miguel Sanz, Jorge Sierra, Martin S. Tallman, Hwei-Fang Tien, Andrew H. Wei, Bob Löwenberg, and Clara D. Bloomfield. Diagnosis and management of aml in adults: 2017 eln recommendations from an international expert panel. *Blood*, 129(4):424–447, Jan 2017. 2, 4, 5

[7] R. M. Haralick, S. R. Sternberg, and X. Zhuang. Image analysis using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(4):532–550, 1987. 4

[8] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3852–3861, 2020. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 4, 6

[10] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta, and Joel H Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2019. 1

[11] Le Hou, Kunal Singh, Dimitris Samaras, Tahsin M Kurc, Yi Gao, Roberta J Seidman, and Joel H Saltz. Automatic histopathology image analysis with cnns. In *2016 New York Scientific Data Summit (NYSDS)*, pages 1–6. IEEE, 2016. 1

[12] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1

[13] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018. 4

[14] COLORIMETRY - PART 4: CIE 1976 L*A*B* COLOUR SPACE. Cie standard, International Comission on Illumination, 2007. 4

[15] Monika Jyotiyana and Nishtha Kesswani. Deep learning and the future of biomedical image analysis. In *Deep Learning Techniques for Biomedical and Health Informatics*, pages 329–345. Springer, 2020. 1

[16] Angelos Katharopoulos and François Fleuret. Processing megapixel images with deep attention-sampling models. In *International Conference on Machine Learning*, pages 3282–3291. PMLR, 2019. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Bruno Korbar, Andrea M Olofson, Allen P Miraflor, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Deep learning for classification of colorectal polyps on whole-slide images. *Journal of pathology informatics*, 8, 2017. 2

[19] Preetham Kumar and Shazad Maneck Udwadia. Automatic detection of acute myeloid leukemia from microscopic blood smear image. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1803–1807, 2017. 4

[20] Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Luke Torre-Healy, Richard A Moffitt, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, et al. Utilizing automated breast cancer detection to identify spatial distributions of tumor-infiltrating lymphocytes in invasive breast cancer. *The American journal of pathology*, 190(7):1491–1504, 2020. 2

[21] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 470–479. Springer, 2020. 1

[22] Chen Li, Xintong Li, Md Rahaman, Xiaoyan Li, Hongzan Sun, Hong Zhang, Yong Zhang, Xiaoqi Li, Jian Wu, Yudong Yao, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification, and detection approaches. *arXiv preprint arXiv:2102.10553*, 2021. 1

[23] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982. 4

[24] Sam Maksoud, Kun Zhao, Peter Hobson, Anthony Jennings, and Brian C Lovell. Sos: Selective objective switch for rapid immunofluorescence whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3862–3871, 2020. 2

[25] Shallu Sharma and Rajesh Mehra. Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. *Journal of digital imaging*, 33(3):632–654, 2020. 2

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. 4, 7

[27] Shusuke Takahama, Yusuke Kurose, Yusuke Mukuta, Hiroyuki Abe, Masashi Fukayama, Akihiko Yoshizawa, Masanobu Kitagawa, and Tatsuya Harada. Multi-stage pathological image classification using semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10702–10711, 2019. 1

[28] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PloS one*, 15(6):e0233678, 2020. 1

[29] Xingzhi Yue, Neofytos Dimitriou, and Ognjen Arandjelovic. Colorectal cancer outcome prediction from h&e whole slide images using machine learning and automatically inferred phenotype profiles. *arXiv preprint arXiv:1902.03582*, 2019. 1

[30] Jingwei Zhang, Ke Ma, John Van Arnam, Rajarsi Gupta, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3776–3784, 2021. 1, 2, 5

[31] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017. 1