# Outlier Detection using Self-Organizing Maps for Automated Blood Cell Analysis

**Stefan Röhrl** [*1] **Alice Hein** [*1] **Lucie Huang** [*1] **Dominik Heim** [2] **Christian Klenk** [2] **Manuel Lengl** [1] **Martin Knopp** [1 2] **Nawal Hafez** [1] **Oliver Hayden** [2] **Klaus Diepold** [1]

## Abstract

The quality of datasets plays a crucial role in the successful training and deployment of deep learning models. Especially in the medical field, where system performance may impact the health of patients, clean datasets are a safety requirement for reliable predictions. Therefore, outlier detection is an essential process when building autonomous clinical decision systems. In this work, we assess the suitability of Self-Organizing Maps for outlier detection specifically on a medical dataset containing quantitative phase images of white blood cells. We detect and evaluate outliers based on quantization errors and distance maps. Our findings confirm the suitability of Self-Organizing Maps for unsupervised Out-Of-Distribution detection on the dataset at hand. Self-Organizing Maps perform on par with a manually specified filter based on expert domain knowledge. Additionally, they show promise as a tool in the exploration and cleaning of medical datasets. As a direction for future research, we suggest a combination of Self-Organizing Maps and feature extraction based on deep learning.

## 1. Introduction

Nowadays, many diseases like leukemia are diagnosed by analyzing blood samples and detecting unhealthy distributions of different types of blood cells (Mittal et al., 2022). Therefore, analysis of cellular structures make up a large part of medical laboratory tests. However, currently used gold standards of hematological analysis either have the disadvantage that they cannot classify certain cell types or are associated with a high manual effort (Meintker et al., 2013; Filby, 2016). Computer vision and machine learning (ML) in combination with contrast-rich digital holographic microscopy has the potential to perform such hematological analyses in a more cost effective, flexible and faster way (Jo et al., 2018).

Unfortunately, during the process of data collection, outliers such as defocused cells, duplets and debris may occur due to activation, apoptosis, and aggregation of cells or insufficient flow focusing. In the training stage, including these outliers in one's dataset may deteriorate model performance, since there is also no industrial grade calibrator for this holographic flow cytometry assay. In a production environment, outliers may even pose a safety issue if the model cannot reliably recognize them as such, potentially leading to a wrong classification of, say, debris as an interesting event. In this work, we examine the suitability of Self-Organizing Maps (SOMs) as a tool for the detection of outliers in a dataset of holographic microscopic images of white blood cells (WBCs). We first provide some background on SOMs in Section 2 and describe our dataset and experimental setup in Section 3. Section 4 presents our results. We end with a brief discussion of related work in Section 5 and ways our approach could be expanded upon in Section 6.

## 2. Background

The SOM is an unsupervised artificial neural network first proposed by Teuvo Kohonen (1990) in early 1981. This dimensionality reduction technique groups data points into clusters on a 2D lattice according to their mutual similarity. The lattice space of a SOM consists of a predefined number of neurons. Each neuron has its own weight vector, which is initialized through some initialization function (e.g. principal component analysis). The weight vector of a neuron $j$ can be described as

$$w_j = [w_{j1}, w_{j2}, ..., w_{jd}]^T, \; j = 1, 2, ..., J,$$

where $J$ is the number of neurons and $d$ the number of input features.

The SOM is then trained for a set number of iterations by

[*]Equal contribution [1]Chair of Data Processing, Technical University of Munich, Germany [2]Heinz-Nixdorf Chair of Biomedical Electronics, Technical University of Munich, Germany. Correspondence to: Stefan Röhrl <stefan.roehrl@tum.de>.

*Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML)*. 

choosing an input data point $x \in \mathbb{R}^d$ from the training dataset and computing its activation distance to all other neurons. The index $c = c(x)$ of the neuron with the closest Euclidean distance to $x$, also called the Best Matching Unit (BMU), is determined using

$$c(x) = \underset{j}{argmin} \|x - w_j\|, \; j = 1, 2, ..., J.$$

Based on a predefined spread (e.g. standard deviation $\sigma$), a neighborhood kernel $h_{j,c(x)}(n)$ controls the update influence on the surrounding of the BMU. The weight vectors of the BMU and its neighbors $w_j$ are then updated according to a time-variant learning rate $\alpha(n)$ using

$$w_j(n+1) = w_j(n) + \alpha(n) \cdot h_{j,c(x)}(n) \cdot (x(n) - w_j(n)),$$

where $n$ stands for the current iteration. Algorithm 1 summarizes this process.

After successful training, the weight vectors of the SOM have adjusted to reflect the distribution of the input data in a topology-preserving manner: data points which are similar to each other in the input space are matched onto neurons close to each other in the lattice space (Kaski, 1997). This is a useful property for the detection of outliers within a large dataset, as inliers are expected to form large and dense clusters of neurons in the lattice space. Outliers on the other hand are expected to be scattered across the lattice space with a large distance from the dense clusters.

**Algorithm 1** SOM training algorithm

initialize weight vectors $w$ of all neurons
$N \leftarrow$ *number of iterations*
$J \leftarrow$ *number of neurons*
**for** $n \leftarrow 1$ to $N$ **do**
  $x \leftarrow$ random input data point from the input dataset
  **for** $j \leftarrow 1$ to $J$ **do**
    calculate distance $d_j(n) = \|x - w_j(n)\|$
  **end for**
  calculate index for BMU $c(x) = \underset{j}{argmin}\, d_j(n)$
  determine neighborhood function $h_{j,c(x)}(n)$ based on $\sigma$ and $c(x)$
  **for** $j \leftarrow 1$ to $J$ **do**
    update weights with $w_j(n+1)$
    $= w_j(n) + \alpha(n) \cdot h_{j,c(x)}(n) \cdot (x(n) - w_j(n))$
  **end for**
**end for**

# 3. Methods

## 3.1. Data

The dataset used in this work[1] consists of quantitative phase images of four types of WBCs (eosinophils, lymphocytes, monocytes, and neutrophils), taken by a digital holographic microscope. These images of size 512×384 pixels contain multiple cells per image and represent their optical density. Using threshold segmentation, the raw phase images are segmented to yield single cell image patches of size 50×50 pixels. Examples can be seen in Figure 2. For this work, we used three segmented datasets:

- **Unfiltered dataset**, 447,541 images
  This dataset contains images of 41,881 eosinophils, 77,672 lymphocytes, 58,760 monocytes and 269,228 neutrophils. Since it has not been manually cleaned, there are an unknown number of inliers and outliers.
- **Inlier dataset**, 82,056 images
  This dataset was created by filtering images based on predefined thresholds for the four morphological features *optical height max*, *circularity*, *area* and *equivalent diameter* of each cell. The four classes of WBCs are balanced to 20,514 images per class.
- **Outlier dataset**, 10,136 images
  The dataset contains 352 images captured with focus set 7.5 $\mu m$ over the ideal focus and 803 images with focus set 15 $\mu m$ over the ideal focus. 7,749 images contain high background noise and 1,232 images were captured at the border of the microfluidic channel, which leads to high interferences due to light scattering.

All segmented images were normalized to the range of the inlier dataset, and six ($d = 6$) morphological features were extracted, namely *area*, *circularity*, *equivalent diameter*, *optical height max*, *optical height variance*, and *energy* (Ugele et al., 2018; Röhrl et al., 2019).

## 3.2. Experiments

After preprocessing, we trained a SOM on the inlier dataset, then tested the model on the outlier and unfiltered dataset and evaluated the detected outliers and inliers. For evaluation, we used the *average quantization error*, which is the normed average of the quantization errors of all input samples, calculated using

$$E_{AQ} = \frac{1}{M} \sum_{i=1}^{M} \|x_i - w_c\| \quad \text{with } c = \underset{j}{argmin} \|x_i - w_j\|.$$

[1] All human samples were collected with informed consent and procedures approved by application 620/21 S-KK of the ethic committee of the Technical University of Munich.

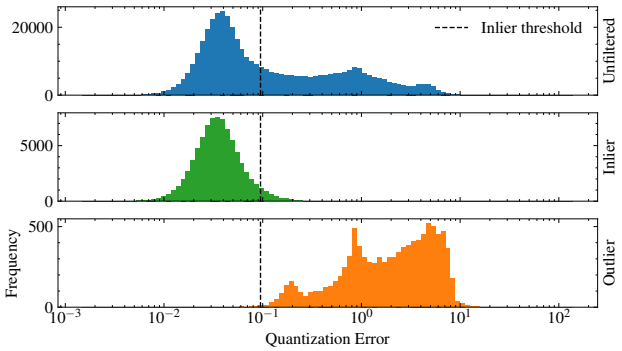



*Figure 1.* Quantization error distributions for the three datasets

Here, $M$ is the size of the input dataset and $j$ the index of the respective neuron. The smaller the average quantization error, the better a fixed-sized SOM reflects the input dataset. We defined samples with a quantization error greater than the $2\sigma$ deviation of all quantization errors as outliers.

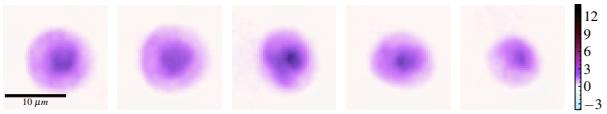
As per Kohonen's recommendation (1990), our SOM consisted of $5\sqrt{K}$ neurons, where $K$ is the cardinality of our inlier dataset. Its shape was chosen such that its ratio of height to width equaled the ratio of the two largest eigenvalues of its autocorrelation matrix (Ponmalai & Kamath, 2019), resulting in a $65\times22$ lattice. The SOM was trained with a sigma of 1, learning rate of 1, hexagonal topology, gaussian neighborhood function and euclidean activation distance, as this was found to be a suitable hyperparameter configuration in preliminary tests with a 5-fold cross-validation, leading to the lowest quantization error. All experiments were implemented in Python and made use of the Scikit-learn[2], OpenCV[3], TensorFlow[4], Keras[5], and MiniSOM[6] libraries.

## 4. Results

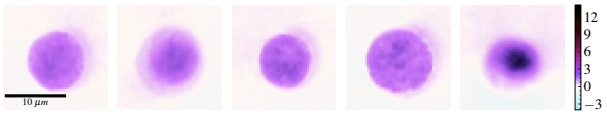
The middle graph of Figure 1 shows the distribution of all inlier quantization errors. As can be seen, most of the errors fall within a small range around 0.04, which indicates that the SOM was trained to fit the inlier dataset well. According examples for inliers are displayed in Figure 2(a). Next, we evaluate the quantization errors of the outlier dataset. If the SOM worked perfectly, all errors should be greater than the inlier threshold.

This is confirmed by Figure 1 (bottom), where 99.6% of all data are correctly detected as outliers. Finally, we tested the SOM on the unfiltered dataset consisting of an unknown number of unlabeled inliers and outliers. As expected, most of the quantization errors for the unfiltered dataset lay within the threshold of 0.095, while the rest stretches out to large quantization error ranges, yielding an outlier percentage of 43.26%. That is approximately the same amount as detected with the currently used filtering method, which relies on manually specified feature thresholds based on domain expertise.

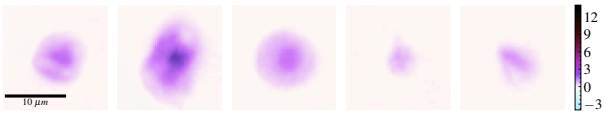
Taking a look at the inliers and outliers detected in the unfiltered dataset, we observe that in error range [0.5, 0.6], the detected outliers start to take on irregular shapes, such as too small or unclean circles. Cells in error range [1.0, 2.0] often have blurred and irregular contours. Range [3.0, 4.0] covers the case of double cells, which where mistaken for single cells in the segmentation process. Larger error ranges contain completely irregular cells or edge cases like the border of the microfluidic channel or air bubbles.



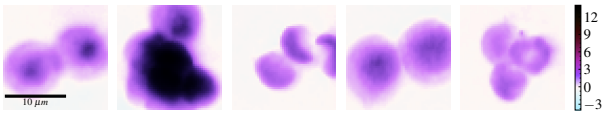
(a) Quantization error range 0.0 – 0.1



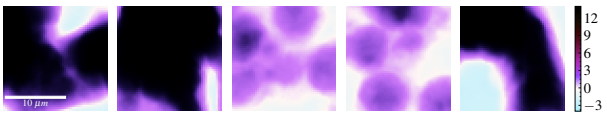
(b) Quantization error range 0.5 – 0.6



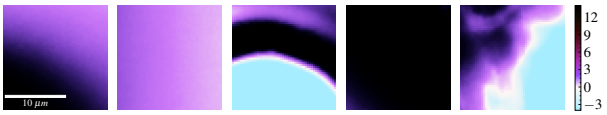
(c) Quantization error range 1.0 – 2.0



(d) Quantization error range 3.0 – 4.0



(e) Quantization error range 10.0 – 20.0



(f) Quantization error range 30.0 – 100.0

*Figure 2.* Examples of inliers and outliers detected by the SOM in the unfiltered dataset

[2] https://scikit-learn.org
[3] https://opencv.org
[4] https://tensorflow.org
[5] https://keras.io
[6] https://github.com/JustGlowing/minisom

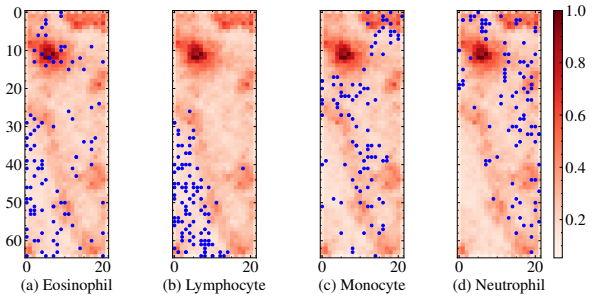



*Figure 3.* Positions of four inlier classes (a) eosinophil, (b) lymphocyte, (c) monocyte and (d) neutrophil on SOM distance map

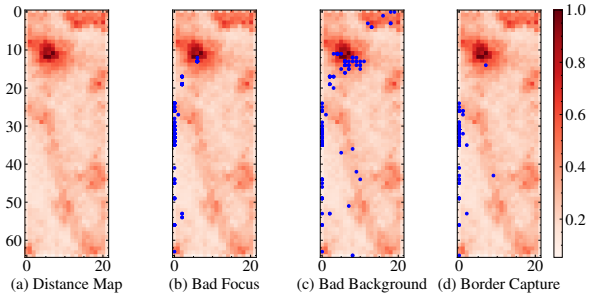



*Figure 4.* SOM distance map (a) and positions of outlier types (b) bad focus, (c) bad background and (d) border capture

A further evaluation technique we used was to inspect where on the SOM distance map the inliers and outliers were positioned. A distance map shows the distance of each neuron to its closest neighbors. The lighter the neuron, the smaller the distance to its neighbor neurons. Figure 3 displays the distance map as the aforementioned $65 \times 22$ lattice as a background pattern. Each sub-figure shows that almost all inliers were plotted in light regions of the distance map, confirming the assumption that clusters with many neurons close to each other represent dense inlier classes. Additionally, the winning neurons of input data points from the same white blood cell classes formed clusters, suggesting that the SOM had not only learned to distinguish inliers and outliers, but also to some extent the four different classes of inliers.

This pattern is also confirmed by Figure 4, which plots the positions of different types of outliers on the distance map. In contrast to the inlier data points, outliers tend to be positioned in darker, that is, less dense regions, or at the edge of the SOM.

## 5. Related Work

Out-Of-Distribution (OOD) detection methods provide important safety mechanisms to prevent real-world systems from failing when confronted with anomalous data and have thus been the focus of much research. The three main categories of OOD detection approaches are classification-based, nearest neighbor-based, and clustering-based techniques (Chandola et al., 2009), where SOMs can be said to belong to the latter category. Previous applications of SOMs for cluster-based OOD identification include intrusion detection (Labib & Vemuri, 2002), fault detection (Emamian et al., 2000) and fraud detection (Brockett et al., 1998). While this work uses a low-dimensional feature representation of the input objects, it is also possible to apply SOMs directly on pixel values as shown by Penn (2002) on hyperspectral imagery data. Xiao et al. (2018) extend this idea and combine the SOM with a deep neural network to obtain a *change graph* in synthetic aperture radar images used for environmental monitoring. In the domain of tissue cell analysis *in silico*, Yuan et al. (2021) presented a SOM for segmentation and classification. Rahmat et al. (2018) successfully demonstrate the morphological analysis of red blood cells which encourages the adaption of SOMs to WBCs in this work.

## 6. Discussion and Conclusion

In this work, we confirmed the suitability of a SOM-based OOD detection approach on a dataset of holographic blood cell images. The SOM reached an accuracy of 99.69% on a test set of outliers created through physical manipulations during the imaging or the sample preparation. When applied on a dataset with an unknown number of inliers and outliers, it performed similarly to a filter based on manually specified feature thresholds. Therefore, it spares the medical experts time consuming and expensive manual labor. The SOM-based method also enabled the observation of different types of outliers in different ranges of quantization errors, such as duplets and edge cases. This was not possible using the current filtering method. Hence, we achieved a more generalizable and robust approach to clean the vast holographic flow cytometry datasets. In addition, the optimized SOM could be used to distinguish between different classes of inliers, visible as separate clusters on the distance map.

However, the SOM still relies on extensive pre-processing to extract selected features. A next step would therefore be to take advantage of recent advances in deep learning by combining convolutional neural networks for feature extraction with SOMs for dimensionality reduction and OOD detection. Given the SOM's clustering abilities, we also envision further applications such as dataset exploration and efficient data annotation by labelling entire parts of the SOM rather than individual examples.

**Acknowledgement** The authors would like to especially honor the contributions of D. Heim for the sample preparation and recording of the measurements and L. Huang for the software implementation and experiments. This research was funded by the German Federal Ministry for Education and Research (BMBF) with the funding ID ZN 01 | S17049.

## References


Brockett, P. L., Xia, X., and Derrig, R. A. Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud. *Journal of Risk and Insurance*, 65(2):245–274, 1998.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3): 1–58, 2009.

Emamian, V., Kaveh, M., and Tewfik, A. H. Robust Clustering of Acoustic Emission Signals Using the Kohonen Network. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pp. 3891–3894. IEEE Computer Society, 2000.

Filby, A. Sample preparation for flow cytometry benefits from some lateral thinking. *Cytometry Part A: the journal of the International Society for Analytical Cytology*, 89 (12):1054–1056, 2016.

Jo, Y., Cho, H., Yun Lee, S., Choi, G., Kim, G., Min, H.-s., and Park, Y. Quantitative Phase Imaging and Artificial Intelligence: A Review. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1):1–14, 2018.

Kaski, S. *Data Exploration Using Self-organizing Maps*. Acta polytechnica Scandinavica. Finnish Academy of Technology, 1997.

Kohonen, T. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

Labib, K. and Vemuri, R. NSOM: A real-time network-based intrusion detection system using self-organizing maps. *Networks and Security*, 21(1), 2002.

Meintker, L., Ringwald, J., Rauh, M., and Krause, S. W. Comparison of automated differential blood cell counts from Abbott Sapphire, Siemens Advia 120, Beckman Coulter DxH 800, and Sysmex XE-2100 in normal and pathologic samples. *American journal of clinical pathology*, 139(5):641–650, 2013.

Mittal, A., Dhalla, S., Gupta, S., and Gupta, A. Automated analysis of blood smear images for leukemia detection: a comprehensive review. *ACM Computing Surveys (CSUR)*, 2022.

Penn, B. Using self-organizing maps for anomaly detection in hyperspectral imagery. In *Proceedings, IEEE Aerospace Conference*, volume 3, pp. 1531–1535, 2002.

Ponmalai, R. and Kamath, C. Self-Organizing Maps and Their Applications to Data Analysis. Technical report, Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States), 2019.

Rahmat, R. F., Wulandari, F. S., Faza, S., Muchtar, M. A., and Siregar, I. The morphological classification of normal and abnormal red blood cell using self organizing map. *IOP Conference Series: Materials Science and Engineering*, 308:012015, 2018.

Röhrl, S., Ugele, M., Klenk, C., Heim, D., Hayden, O., and Diepold, K. Autoencoder Features for Differentiation of Leukocytes based on Digital Holographic Microscopy (DHM). In *Computer Aided Systems Theory - EUROCAST*, pp. 281–288, 2019.

Ugele, M., Weniger, M., Stanzel, M., Bassler, M., Krause, S. W., Friedrich, O., Hayden, O., and Richter, L. Label-free high-throughput leukemia detection by holographic microscopy. *Advanced Science*, 5(12):1800761, 2018.

Xiao, R., Cui, R., Lin, M., Chen, L., Ni, Y., and Lin, X. SOMDNCD: Image Change Detection Based on Self-Organizing Maps and Deep Neural Networks. *IEEE Access*, 6:35915–35925, 2018.

Yuan, E., Matusiak, M., Sirinukunwattana, K., Varma, S., Kidziński, L., and West, R. Self-organizing maps for cellular in silico staining and cell substate classification. *Frontiers in Immunology*, 12:765923, 2021.