FlowCat: Automated classification of B-cell lymphoma subtypes

Zhao M.[1], Mallesh N.[1], Hertzberg J.[1], Höllein A.[2], Schabath R.[2], Haferlach T.[2], Haferlach C.[2], Krawitz P.[1], Kern W.[2]
1 – Institute for Genomic Statistics and Bioinformatics, Bonn, Germany;
2 – MLL Munich Leukemia Laboratory, Munich, Germany

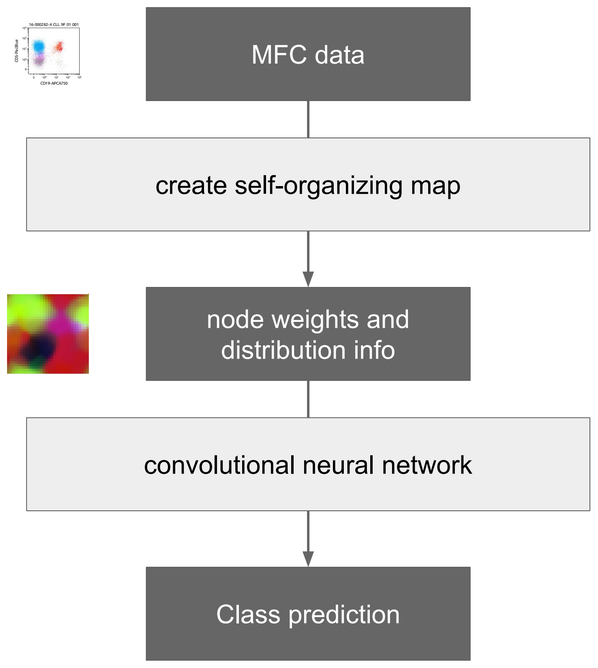
Introduction

Leukemia and lymphoma are haematological disorders of the bone marrow and lymphatic system involving malignant neoplasms of leukocytes. There exists a large variety of subtypes, which can differ significantly in their therapeutical consequences and their prognosis. Differentiations between subtypes can only be done on the basis of histological findings in the affected tissues, karyotyping, genetic sequencing or immunological analysis via flow cytometry. Multi-channel flow cytometry (MFC) is an integral diagnostic method in diagnosis and monitoring of haematological disorders. With the availability of ever more powerful measuring devices, a large number of computational approaches have been developed to automate the manual process of gating and interpreting cell populations in flow cytometry data.

Materials and Methods

Flowcat is an automated approach to classify flow cytometry data into diagnosis labels without further human processing.
Data (FCS files) was obtained from routine diagnostic MFC of 16,543 patients and included 9 different B-cell lymphoma subtypes. Two 11-colour MFC tubes were used to measure up to 50,000 cells per case.
The Convolutional neural network (CNN) processes individually generated self-organizing map (SOM) for each input tube, which are average-merged after two convolutional/max pooling layers and processed in two dense layers producing soft maxed class predictions for the given 8 classes.
In order to visualize the decision process, we utilized a saliency analysis of the trained model, which is used to label the original input data to define populations of special interest to the resulting classification decision, shown in c).
We demonstrate the feasibility of automated classification of flow cytometry data into subtype labels in dataset both larger in the number of samples and classes than previously described with high accuracy.

a) **Classification workflow**



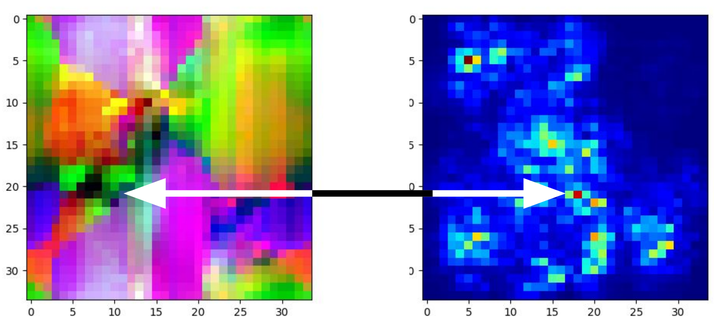
b) **8-class confusion matrix**

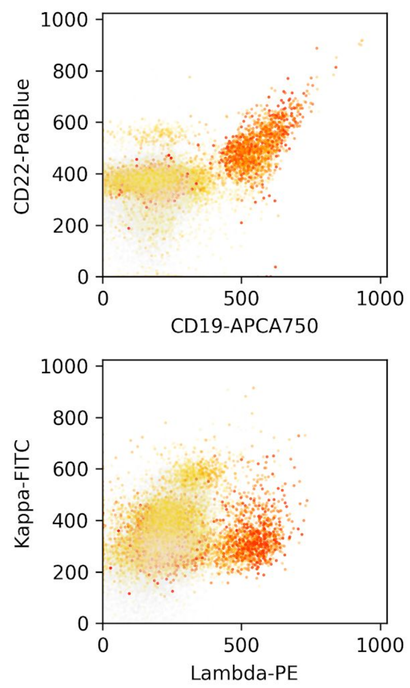
	CLL	MBL	MCL	PL	LPL	MZL	FL	HCL	normal
CLL	0.94		0.00	0.03	0.00	0.00	0.00	0.00	0.02
MBL									
MCL	0.01		0.69	0.26	0.01	0.03	0.00	0.00	0.00
PL	0.13		0.10	0.66	0.04	0.07	0.00	0.00	0.00
LPL	0.04		0.00	0.04	0.53	0.28	0.00	0.00	0.10
MZL	0.01		0.02	0.04	0.10	0.71	0.03	0.00	0.08
FL	0.00		0.00	0.00	0.05	0.02	0.89	0.00	0.05
HCL	0.00		0.00	0.00	0.00	0.00	0.00	0.98	0.02
normal	0.00		0.00	0.00	0.01	0.01	0.00	0.00	0.98

Cohort structure of B-cell lymphoma

CM		MP		LM				
CLL	MBL	MCL	PL	LPL	MZL	FL	HZL	norm
3356	1458	415	597	623	968	218	187	8442

c) **Saliency analysis**





Results

In an initial validation of our approach on a 10% stratified test set of our data, the classifier performed with a class-size weighted f1-score of 0.92 and an unweighted f1-score of 0.79. Per class results are shown in a confusion matrix in b). We analyzed model classification behavior by calculating saliency using gradCAM on a per case basis, we further used the saliency maps in analysis of misclassified samples.

Conclusion

Computational analysis can provide considerable advantages in the current hematological diagnostic process by reducing the need for manual gating and interpretation in most common cases. We demonstrate the possibility for automated analysis in a proof-of-concept for the classification of B-cell lymphoma. Further improvements in generalizability and visualization strategies can improve the system towards usability for routine clinical diagnostics.

Acknowledgments
Patient data and advice on medical interpretation from Munich Leukemia Laboratory (MLL)
Financial support by Berlin Institute of Health (BIH) MD Stipend

Correspondence: nanditha.mallesh@uni-bonn.de