Towards Self-Explainable Transformers for Cell Classification in Flow Cytometry Data

Florian Kowarsch[1], Lisa Weijler[1], Matthias Wödlinger[1,2], Michael Reiter[1,2], Margarita Maurer-Granofszky[2,3], Angela Schumich[2], Elisa O. Sajaroff[4], Stefanie Groeneveld-Krentz[5], Jorge G.Rossi[4], Leonid Karawajew[5], Richard Ratei[6], and Michael N. Dworzak[2,3]

[1] Computer Vision Lab, Faculty of Informatics, TU Wien, Vienna, Austria
[2] Immunological Diagnostics, St. Anna Children's Cancer Research Institute (CCRI), Vienna, Austria
[3] Labdia Labordiagnostik GmbH, Vienna, Austria
[4] Cellular Immunology Laboratory, Hospital de Pediatria "Dr. Juan P. Garrahan", Buenos Aires, Argentina
[5] Department of Pediatric Oncology/Hematology, Charité Universitätsmedizin Berlin, Berlin, Germany
[6] Department of Hematology, Oncology and Tumor Immunology, HELIOS Klinikum Berlin-Buch, Berlin, Germany

Abstract. Decisions of automated systems in healthcare can have far-reaching consequences such as delayed or incorrect treatment and thus must be explainable and comprehensible for medical experts. This also applies to the field of automated Flow Cytometry (FCM) data analysis. In leukemic cancer therapy, FCM samples are obtained from the patient's bone marrow to determine the number of remaining leukemic cells. In a manual process, called gating, medical experts draw several polygons among different cell populations on 2D plots in order to hierarchically sub-select and track down cancer cell populations in an FCM sample. Several approaches exist that aim at automating this task. However, predictions of state-of-the-art models for automatic cell-wise classification act as black-boxes and lack the explainability of human-created gating hierarchies. We propose a novel transformer-based approach that classifies cells in FCM data by mimicking the decision process of medical experts. Our network considers all events of a sample at once and predicts the corresponding polygons of the gating hierarchy, thus, producing a verifiable visualization in the same way a human operator does. The proposed model has been evaluated on three publicly available datasets for acute lymphoblastic leukemia (ALL). In experimental comparison it reaches state-of-the-art performance for automated blast cell identification while providing transparent results and explainable visualizations for human experts.

Keywords: self-explainable deep learning models · transformer · flow cytometry gating · acute lymphoblastic leukemia.

1 Introduction

Deep Learning models are applicable to a variety of problems arising in healthcare. However, since wrong predictions can have severe consequences, the interpretability of models in this domain is crucial. The output produced by a model needs to be transparent, even for clinicians without any knowledge about the interior of the model. This is also true for the field of automated cell detection in Flow Cytometry (FCM) data. FCM measures the antigen expression levels of blood or bone marrow cells. It is used in research as well as in daily clinical routines for tasks such as immunophenotyping or for monitoring residual numbers of cancer cells (minimal residual disease, MRD) during chemotherapy. A typical sample contains 50-500k cells (also called events) per patient with up to 15 different features (markers) measured. Each feature corresponds to either physical properties of a cell (cell size, granularity) or to the expression level of a specific antigen marker on the cell's surface [18]. While methods for automated MRD assessment already reach human expert level performance [30], they lack interpretability of their predictions. Regardless of a model's performance, clinicians have to manually verify the prediction in a time-consuming process. Using explainable methods could overcome this issue.

Molnar [19] divides existing explainable AI methods into two categories: **Intrinsically interpretable models** are interpretable due to their internal structures. Linear models, decision trees or naive Bayes are common examples of this category. **Post-hoc interpretation methods** analyze a model after training in order to gather explainable insights. Common examples of this category are methods that visualize inner structures of neural networks such as saliency maps [20] and CNN feature visualization techniques or methods, that analyze data input and output pairs of a model to build an explaining description such as LIME [23], shapely values [24,25] and partial dependence plots [9]. In [8] a third category **self-explaining AI** is described, according to which a self-explaining model yields two outputs: a decision and an explanation of that decision.

One way to obtain a self-explaining AI system is to reformulate a prediction task such that the model outputs the same kind of data a domain expert would create to solve or explain a particular problem instance. Instead of directly predicting the solution to a given problem instance, the model is asked to predict a *solution path*. For instance, to solve a linear equation, one can either directly state the solution or provide a series of coherent deductive steps that build an interpretable path to the solution. The latter approach strengthens the trust in the correctness of the solution. While not every data domain admits the modeling of such a solution path, in the field of FCM the **gating hierarchy** can be chosen as an explainable solution path for the problem of cell identification. The conventional procedure to analyze FCM data in the clinical routine is to look at 2D projections of the FCM data and label sub-populations of events by drawing polygons around them [18]. This procedure is called **gating** and the polygons are called **gates**. As illustrated in Figure 1, gates act as filters by defining the events that are subject to further analysis in other 2D projections (events inside a gate) and the events that will be discarded (events outside the

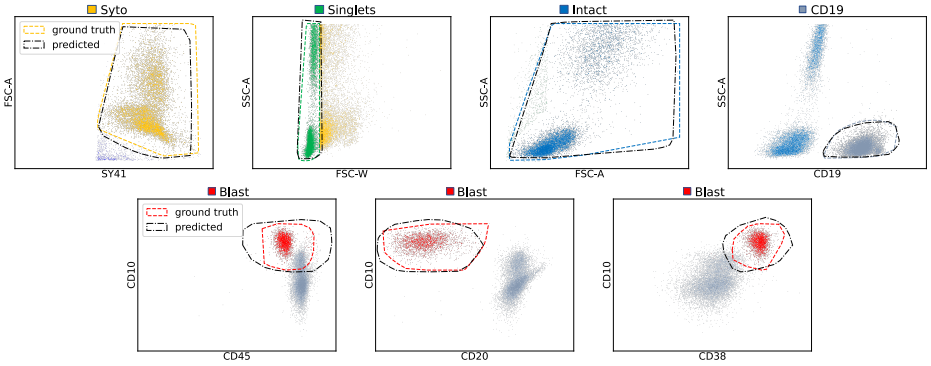

Fig. 1. All seven gates of the used gating hierarchy are depicted for an arbitrary FCM sample. Each plot shows a projection of the multidimensional data on two different features. The automated predicted polygons are drawn black and the human operator-created ground truth polygons are drawn in a different color per gate.

gate). The target population can then be identified by a boolean combination of gates. Gates drawn in specific projections are often applied in sequence, such that one plot only depicts the events selected by the previous plot's polygon. Sequentially applying these gates allows to identify cancer cell populations in the FCM sample. The 2D plots of the data space allow to explicitly depict antigen expressions of the cells in the sample, which are known to be relevant in particular diseases. For example, among other characteristics, CD19 is known to be higher expressed for B-cells [18]. Gating allows analyzing complex patterns of cell populations by a sequence of simpler intermediate steps, which are interpretable by clinicians. Thus, gating is not only a way for finding biologically meaningful sub-populations but has also become the standard for the communication and documentation of FCM sample assessment. Thus it is crucial that the output of machine learning model is compatible with this standard.

In this work we propose a novel method, based on the transformer network, that predicts the polygons of a gating hierarchy to identify cancer cells for MRD assessment in FCM samples of acute lymphoblastic leukemia (ALL) patients.

Contribution This work's contribution is two-fold:

1. A model for ALL blast cell identification is proposed that yields human interpretable visualizations by predicting the polygons of the gating hierarchy while reaching state-of-the-art performance.
2. The proposed model demonstrates how a self-explaining AI systems can be obtained in the medical domain by reformulating the objective function to mimic established human solution procedures.

The remainder of this work is structured as follows. Section 2 gives an overview of methods for automated MRD assessment in FCM data as well as of related architectures. In Section 3 the proposed model is described in detail.

Section 4 states the conducted experiments, compares the proposed approach to other methods and discusses the results.

2 Related Work

Numerous approaches have been established to automate the detection of cell populations in FCM data. The reader is referred to [7] for a more comprehensive review of current trends in automated FCM data analysis. We divide methods for the targeted analysis of FCM data into discriminative and holistic approaches. Approaches that process FCM data event-wise only learn fixed decision regions and are referred to as discriminative approaches. In contrast, holistic approaches process a whole FCM sample and, therefore can account for inter-sample variations, which has been identified as crucial for the correct classification of cell populations with high variability such as leukemic cells [28].

Discriminative Approaches In [1] linear discriminant analysis is proposed for the classification of cell populations as it allows for interpretable performance and reproducibility. Authors in [12] and [10] use a table of marker expression patterns in different cell types as a reference dictionary. Methods based on neural networks include [15,14].

Holistic Approaches FlowDensity [17] and FlowLearn [16] use an operator's 2D gating strategy as a guideline for detecting cell populations. Recently, a one-class classification approach based on Uniform Manifold Approximation was introduced [29]. Further, Gaussian mixture models (GMM) have proven to be well suited to model cell populations in FCM data [6,22]. Reiter et al. [22] fit a linear combination of GMMs with labeled components to an unseen sample by Expectation Maximization (EM). [31,4,30] are approaches based on neural networks that can process a whole sample at once. Authors in [31] use self-organized maps to obtain a 2D image that a CNN further processes. CellCNN [4] automatically learns a concise cell population representation with a 1D-convolution layer followed by a pooling layer to aggregate information. More recently, Wödlinger et al. [30] presented a method based on the transformer architecture [27] that performs classification on single-cell level, while processing a entire sample in a single neural network forward pass. The attention mechanism of the original transformer architecture [27] entails a quadratic complexity in the input length $\mathcal{O}(n^2)$ of both memory and time, which is unfavorable in the context of FCM data as one sample can contain up to millions of events. Wödlinger et al. thus use the concept of the Induced Set Attention Block (ISAB) as introduced in the set-transformer [13] that reduces the complexity to $\mathcal{O}(n)$.

Explainable Approaches With respect to explainability of results, [10,17,16] can be listed as their results rely on predicted thresholds and hence are interpretable. Algorithmic Population Descriptions (ALPODS), as proposed in [26], is designed to provide explainability by fuzzy reasoning rules in a Bayes decision network

expressed in visualizations similar to those generated by domain experts. Another approach related to explainable AI and the method presented in this work is GateFinder [2]. Its goal is to find the shortest yet most discriminative series of 2D polygon gates that lead to a previously specified target population. Although the goal of GateFinder is not targeted analysis, the underlying idea of mimicking the gating strategy of domain experts is similar to the approach presented.

3 Methods

The proposed method consists of a trained neural network that is based on the transformer architecture. The model expects a single FCM sample as input, i.e. a set of events $E \in \mathbb{R}^{N \times m}$. N defines the number of events $(50 - 500 \times 10^5)$ and m denotes the number of markers (typically $10 - 20$). The network's output are 7 polygons defined by $P = 20$ 2D points each. The polygons describe the gating hierarchy for MRD assessment in ALL data, which implies the cell's class membership. Table 1 displays the predicted gates and the used markers.

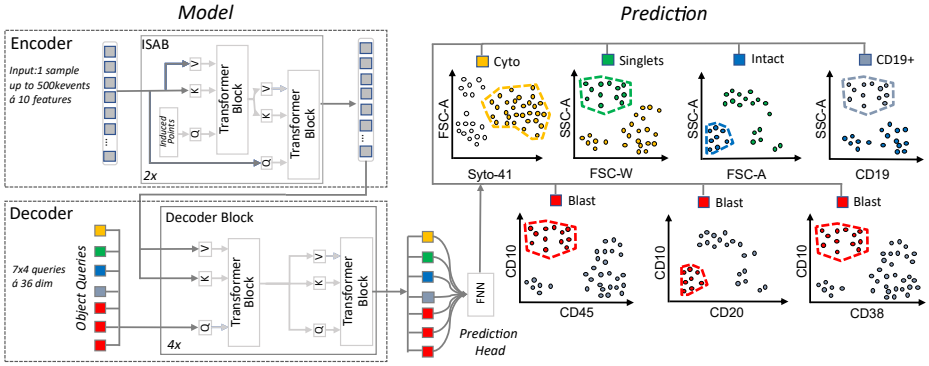

Fig. 2. The network architecture consists of the encoder, decoder, prediction head and the resulting polygons that form the gating hierarchy for a given input FCM sample.

Table 1. The gates and their used features of the predicted gating hierarchy

Name	Syto	Singlets	Intact	CD19	Blast-A	Blast-B	Blast-C
Marker y-Axis	FSC-A	SSC-A	SSC-A	SSC-A	CD10	CD10	CD10
Marker x-Axis	Syto41	FSC-W	FSC-A	CD19	CD45	CD20	CD38

3.1 Architecture

As depicted in Figure 2, the model's architecture follows an encoder-decoder schema as in [5]. A set-transformer similar to [30] is used for the encoder, consist-

ing of two ISAB blocks. The decoder design is inspired by [5]: for each predicted polygon, four static object queries are learned. The object queries are applied to the encoder's output via cross attention, which is followed by a self-attention layer. Each element of the 7-element long decoder output set is passed through a two-layer fully connected neural network called the prediction head. The resulting 20 2D points per element are used as gate polygon for each of the 7 gates in the ALL gating hierarchy. We empirically evaluated that 20 points are most suitable for the given task. More than 20 points only slightly increase the performance (max 1% median F1-Score) while drastically increasing the network size (see Table 4).

3.2 Preprocessing

The operator-annotated polygons comprise two issues regarding their usage as ground truth for training: polygons are typically only roughly estimated, with borders often far away from the nearest events inside the polygon. While this does not affect the effectiveness of the procedure during clinical routine, it introduces a source of ambiguity in the gating process by perturbing the relationship between polygon position and data points. Secondly, for different FCM samples different feature combinations for some of the plots in the gating hierarchy were used by the operator since different operators may use slightly different strategies to track down blast events. However, the model predicts the polygons for a statically predefined set of 2D plot feature combinations. The selected set reflects the most common feature combinations for each gate in the given datasets. We address both issues by computing the convex hull of all events inside the polygon during preprocessing for each gate. The resulting hull serves as adapted training ground truth, which can be created for any required combination of 2D plot features while tightly enclosing the events inside.

3.3 Loss Function

$$\mathcal{L}_{poly}(\hat{p}, p) = \sum_{i}^{P} \|\hat{p}_{\hat{\sigma}(i)}, p_i\|_1 \text{ with } \hat{\sigma} = \underset{\sigma \in \mathcal{S}_P}{\operatorname{argmin}} \sum_{i}^{P} \|p_i, \hat{p}_i\|_1 \tag{1}$$

The model is trained in a supervised manner. Since the number of polygon vertices differs from sample to sample in the ground truth but is fixed to $P = 20$ for the model prediction, we artificially insert or remove points in the ground truth polygons to obtain P points. Equation 1 states the loss for a predicted polygon $\hat{p}$ where $\hat{\sigma} \in \mathcal{S}_P$ defines a permutation of the polygon points such that every predicted point is matched to one corresponding ground truth point using the Hungarian method [11]. The distance between two points is calculated via L1 norm. Similar to [5,3] we experienced, an auxiliary loss benefits the model convergence. The auxiliary loss performs the same computation as the main loss but after each intermediate layer the following intermediate layers are skipped.

3.4 Data Augmentation

To address the low number of training samples (e.g.: ≤ 60 for the BUE dataset), to overcome inter-laboratory differences and to facilitate learning the relationship between polygon and cell cluster position, four different data augmentation steps are applied to the FCM samples during training: For all events and polygons random linear translations of randomly selected features are applied. For randomly selected gates linear scaling (stretching and squeezing in relation to the center), linear translation and shearing of polygons and their corresponding events are used. Further information is given in the Supplementary material 7.

4 Experiments

The same experiments as in [30] have been conducted. In all experiments the proposed model's ability to generalize to new unseen FCM samples (in most cases from different institutes) is tested. The model is implemented in Pytorch 1.10 [21] and trained using the Adam optimizer with a batch size of 12 and a learning rate of 1×10^{-3}. It consists of 32892 parameters and has been trained on a NVIDIA Gefore RTX 2080 Ti. One model forward pass takes $\approx 400ms$ on the used GPU and $\approx 3000ms$ on an Intel i7-10750H CPU. Details about the training setup can be found in the provided code on GitHub[7].

4.1 Data

The proposed model is evaluated on four different datasets collected across three distinct institutions, measured on three different FCM devices, consisting of over 600 samples in total. From all four datasets, the three datasets VIE14, BLN, BUE are publicly available[8]. All samples have been obtained from the bone marrow of pediatric B-ALL patients on day 15 after induction therapy. The following markers are used in the experiments as they are shared upon all samples: CD10, CD19, CD20, CD34, CD38, CD45 and Syto41 as well as FSC-A, FSC-W and SSC-A. For a detailed dataset description, the reader is referred to [22] for VIE14, BLN and BUE, and to [30] for VIE20. The experiments have been evaluated by training one network for each dataset.

4.2 Results

Table 2 displays the results compared to [22] and [30]. For each experiment the cell classification performance (blast cell vs. non-blast cell) of each sample is summarized with the mean and median F1-Score of all samples in the corresponding test set. The results show that the proposed model is able to reach state-of-the-art performance for blast identification tested on data across different institutes. However, the model under-performs on small training datasets

[7] Github Repository
[8] flowrepository.org

such as BLN and BUE with 70 and 60 training samples. In these cases, the model overfitted during training and was not able to generalize well onto new samples from different sources: Qualitatively inspections revealed that while the cluster positions were mostly correctly predicted, the model failed to predict the correct form of unseen polygon shapes.

Table 2. Experiment results of the proposed method compared to GMM [22] and set-transformer [30]. The table reports mean F1-Score / median F1-Score.

Train	**Test**	**GMM** [22]	**Transformer** [30]	**Proposed**
VIE14	**BLN**	0.72/0.81	0.77/**0.90**	0.79/0.88
	BUE	0.75/0.90	0.82/**0.95**	0.78/0.89
	VIE20	0.77/0.90	0.80/**0.91**	0.78/0.87
VIE20	**BLN**	0.53/0.58	0.68/0.83	0.73/**0.85**
	BUE	0.74/0.88	0.75/0.88	0.82/**0.92**
	VIE14	0.80/0.91	0.84/**0.93**	0.73/0.88
BLN	**BUE**	0.65/0.76	0.66/**0.87**	0.69/0.84
	VIE14	0.48/0.48	0.82/**0.92**	0.58/0.73
	VIE20	0.53/0.60	0.82/**0.91**	0.50/0.55
BUE	**BLN**	0.62/0.73	0.64/**0.78**	0.57/0.69
	VIE14	0.66/0.73	0.83/**0.92**	0.62/0.69
	VIE20	0.67/0.78	0.79/**0.90**	0.65/0.75

The explainable and hierarchical processing of FCM samples in the proposed model elicits two main benefits: first, during model development, unwanted model behaviors such as learned biases can be spotted and addressed. For instance, all applied data augmentation steps were motivated during inspection of the prediction results in the early development stages. Secondly, during inference, the model's prediction can be interpreted. For example, a medical expert can spot and correct a fault in the blast cell classification due to a miss-positioning of a specific polygon in the predicted hierarchy. Take, for example, the CD10CD45-Blast-Gate in Figure 1: a clinician could adjust the predicted polygon such that no events of the seconded cluster are included in the gate.

5 Conclusion

This work proposes a novel transformer-based approach for blast cell detection in FCM samples of ALL patients. The model visually reveals which cells it identifies as blast cells by predicting the polygons of the gating hierarchy for a given FCM sample. This imitates the construction of a gating hierarchy by a human expert in clinical practice and therefore explains why certain events are detected as blast cells. While the proposed model fails to generalize well when trained on small datasets (≤ 70 samples), its performance is comparable to non-explainable state-of-the-art approaches on more populated datasets (≥ 180 samples). Future work could address this issue by pretraining the model on artificially generated data.

Since the model mimics the decision process of domain experts, it is suitable to be included in the clinical gating routine in the future. The proposed model is designed for pediatric ALL, but the underlying concept could be applied to any disease for which standardized FCM gating hierarchies exist.

6 Acknowledgement

We thank Dieter Printz (FACS Core Unit, CCRI) for flow-cytometer maintenance and quality control, as well as Daniela Scharner and Susanne Suhendra-Chen (CCRI), Jana Hofmann (Charité), Marianne Dunken (HELIOS Klinikum), Marianela Sanz, Andrea Bernasconi, and Raquel Mitchell (Hospital Garrahan) for excellent technical assistance. We are indebted to Melanie Gau, Roxane Licandro, Florian Kleber, Paolo Rota and Guohui Qiao (all from TU Vienna) for valuable contributions to the AutoFLOW project. We thank Markus Kaymer and Michael Kapinsky (both from Beckman Coulter Inc.) for kindly assisting in the provision of customized DuraClone^{TM} tubes for this study as designed by the authors. Notably, Beckman Coulter Inc. did not have any influence on study design, data acquisition and interpretation, or manuscript writing. The study has received funding from the European Union's H2020 Research and Innovation Program through Grant number 825749 "CLOSER: Childhood Leukemia: Overcoming Distance between South America and Europe Regions", the Vienna Business Agency under grant agreement No 2841342 (Project MyeFlow) and by the Marie Curie Industry Academia Partnership & Pathways (FP7-MarieCurie-PEOPLE-2013-IAPP) under grant no. 610872 to project "AutoFLOW" to MND.

References

1. Abdelaal, T., van Unen, V., Höllt, T., Koning, F., Reinders, M.J., Mahfouz, A.: Predicting cell populations in single cell mass cytometry data. Cytometry Part A **95**(7), 769–781 (2019)
2. Aghaeepour, N., Simonds, E.F., Knapp, D.J.H.F., Bruggner, R.V., Sachs, K., Culos, A., Gherardini, P.F., Samusik, N., Fragiadakis, G.K., Bendall, S.C., Gaudilliere, B., Angst, M.S., Eaves, C.J., Weiss, W.A., Fantl, W.J., Nolan, G.P.: GateFinder: projection-based gating strategy optimization for flow and mass cytometry. Bioinformatics **34**(23), 4131–4133 (05 2018)
3. Al-Rfou, R., Choe, D., Constant, N., Guo, M., Jones, L.: Character-level language modeling with deeper self-attention. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3159–3166 (2019)
4. Arvaniti, E., Claassen, M.: Sensitive detection of rare disease-associated cell subsets via representation learning. Nature Communications **8**(14825), 2041–1723 (2017)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Chen, X., Hasan, M., Libri, V., Urrutia, A., Beitz, B., Rouilly, V., Duffy, D., Patin, É., Chalmond, B., Rogge, L.: Automated flow cytometric analysis across large numbers of samples and cell types. Clinical Immunology **157**(2), 249–260 (2015)

7. Cheung, M., Campbell, J.J., Whitby, L., Thomas, R.J., Braybrook, J., Petzing, J.: Current trends in flow cytometry automated data analysis software. Cytometry Part A pp. 1–15 (2021)
8. Elton, D.C.: Self-explaining ai as an alternative to interpretable ai. In: International conference on artificial general intelligence. pp. 95–106. Springer (2020)
9. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. arXiv preprint arXiv:1805.04755 (2018)
10. Ji, D., Nalisnick, E., Qian, Y., Scheuermann, R.H., Smyth, P.: Bayesian trees for automated cytometry data analysis. bioRxiv (2018)
11. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
12. Lee, H.C., Kosoy, R., Becker, C.E., Dudley, J.T., Kidd, B.A.: Automated cell type discovery and classification through knowledge transfer. Bioinformatics **33**(11), 1689–1695 (jan 2017)
13. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: International Conference on Machine Learning. pp. 3744–3753. PMLR (2019)
14. Li, H., Shaham, U., Stanton, K.P., Yao, Y., Montgomery, R.R., Kluger, Y.: Gating mass cytometry data by deep learning. Bioinformatics **33**(21), 3423–3430 (2017)
15. Licandro, R., Schlegl, T., Reiter, M., Diem, M., Dworzak, M., Schumich, A., Langs, G., Kampel, M.: Wgan latent space embeddings for blast identification in childhood acute myeloid leukaemia. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3868–3873. IEEE (2018)
16. Lux, M., Brinkman, R.R., Chauve, C., Laing, A., Lorenc, A., Abeler-Dörner, L., Hammer, B.: flowlearn: fast and precise identification and quality checking of cell populations in flow cytometry. Bioinformatics **34**(13), 2245–2253 (feb 2018)
17. Malek, M., Taghiyar, M.J., Chong, L., Finak, G., Gottardo, R., Brinkman, R.R.: flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics **31**(4), 606–607 (oct 2014)
18. McKinnon, K.: Flow cytometry: An overview. Current protocols in immunology **120**(1), 5–1 (2018)
19. Molnar, C.: Interpretable machine learning. Lulu. com (2020)
20. Nie, W., Zhang, Y., Patel, A.: A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: International Conference on Machine Learning. pp. 3809–3818. PMLR (2018)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32**, 8026–8037 (2019)
22. Reiter, M., Diem, M., Schumich, A., Maurer-Granofszky, M., Karawajew, L., Rossi, J.G., Ratei, R., Groeneveld-Krentz, S., Sajaroff, E.O., Suhendra, S., et al.: Automated flow cytometric mrd assessment in childhood acute b-lymphoblastic leukemia using supervised machine learning. Cytometry Part A **95**(9), 966–975 (2019)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
24. Shapley, L.S.: A value for n-person games, contributions to the theory of games, 2, 307–317 (1953)

25. Sundararajan, M., Najmi, A.: The many shapley values for model explanation. In: International conference on machine learning. pp. 9269–9278. PMLR (2020)
26. Ultsch, A., Hoffmann, J., Röhnert, M., Von Bonin, M., Oelschlägel, U., Brendel, C., Thrun, M.C.: An Explainable AI System for the Diagnosis of High Dimensional Biomedical Data. arXiv e-prints arXiv:2107.01820 (Jul 2021)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
28. Weijler, L., Diem, M., Reiter, M., Maurer-Granofszky, M.: Detecting rare cell populations in flow cytometry data using umap. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4903–4909 (2021)
29. Weijler, L., Kowarsch, F., Wödlinger, M., Reiter, M., Maurer-Granofszky, M., Schumich, A., Dworzak, M.N.: Umap based anomaly detection for minimal residual disease quantification within acute myeloid leukemia. Cancers **14**(4) (2022)
30. Wodlinger, M., Reiter, M., Weijler, L., Maurer-Granofszky, M., Schumich, A., Groeneveld-Krentz, S., Ratei, R., Karawajew, L., Sajaroff, E., Rossi, J., Dworzak, M.N.: Automated identification of cell populations in flow cytometry data with transformers. Computers in Biology and Medicine p. 105314 (2022)
31. Zhao, M., Mallesh, N., Höllein, A., Schabath, R., Haferlach, C., Haferlach, T., Elsner, F., Lüling, H., Krawitz, P., Kern, W.: Hematologist-level classification of mature b-cell neoplasm using deep learning on multiparameter flow cytometry data. Cytometry Part A **97**(10), 1073–1080 (2020)

7 Supplementary material

Equation of the event and polygon scaling data augmentation:

$$\hat{x} = (x - c) \cdot (1 \pm s) + c \tag{2}$$

with $c = \min(x) + \frac{\Delta x}{2}$, where $\Delta x = \max(x) - \min(x)$ and $s \sim \mathcal{U}(0, 0.3)$

Table 3. Due to missing intermediate or blast gates, not all samples as in [30] could be used in this work. This table compares the number of used samples per dataset to [30]. In Table 2 the same samples were used to evaluate all 3 methods.

Dataset	**# Transformer [30]**	**# Proposed**
VIE14	200	186
VIE20	319	291
BLN	72	70
BUE	65	60

Table 4. Median F1-Score of the artificially generated convex hull polygons compared to the operator ground-truth for different polygon lengths.

Dataset	5	10	20	30	40	60
VIE14	72.02	92.65	94.81	94.98	95.07	95.38
VIE20	67.90	92.57	92.96	93.35	93.51	94.07
BLN	61.38	88.97	90.35	90.41	90.79	91.18
BUE	72.27	96.75	97.54	97.46	97.71	97.97

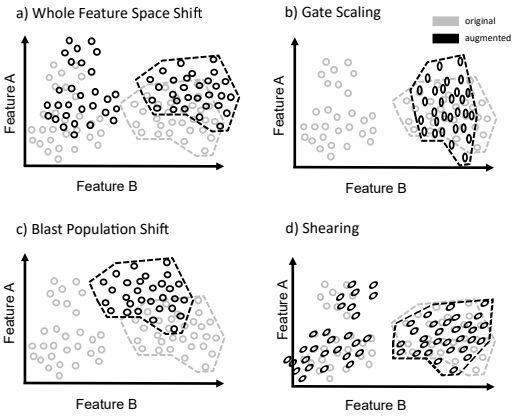

Fig. 3. The different augmentation steps applied to an FCM sample: a) Random linear shifts of the whole feature space. b) Scaling of the blast population's shape. c) Random linear shifts of the blast events. d) Shearing of gates and events along single features.

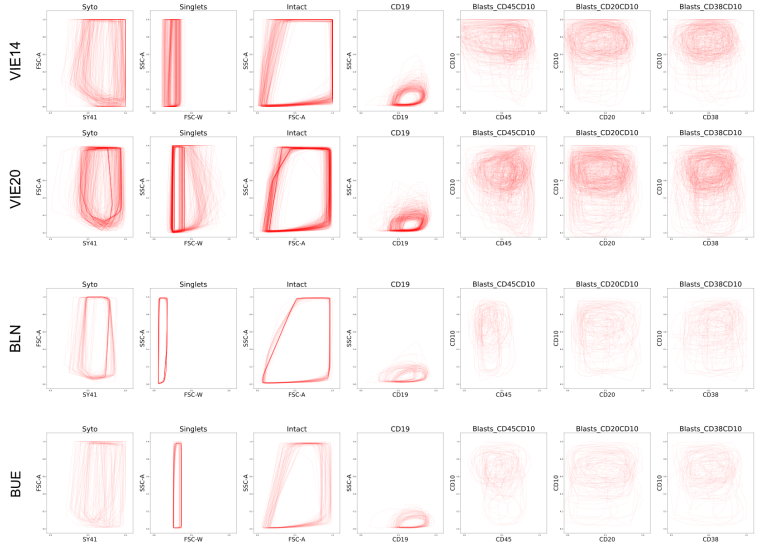

Fig. 4. The ground truth polygons constructed from convex data hulls. Each row shows the gates of one dataset. Each column shows one of the 7 gates of the used ALL gating hierarchy. This plot highlights the cell clusters shifts among the different laboratories.