WGAN Latent Space Embeddings for Blast Identification in Childhood Acute Myeloid Leukaemia

Roxane Licandro*†‡, Thomas Schlegl*‡,
Michael Reiter†, Markus Diem†, Michael Dworzak§¶, Angela Schumich¶, Georg Langs‡, Martin Kampel†
* contributed equally to this publication
†TU Wien, Institute of Computer Aided Automation, Computer Vision Lab, Austria, Vienna
‡Medical University of Vienna, Department of Biomedical Imaging and Image-guided Therapy
Computational Imaging Research Lab, Austria, Vienna
§Children's Cancer Research Institute and Medical University of Vienna, St. Anna Children's Hospital, Austria, Vienna
¶Labdia Labordiagnostik GmbH, Austria, Vienna

Email: roxane.licandro@tuwien.ac.at, thomas.schlegl@gmail.com

***Abstract*— Acute Myeloid Leukaemia (AML) is a rare type of childhood acute leukaemia. During treatment, the assessment of the number of cancer cells is particularly important to determine treatment response and consequently adapt the treatment scheme if necessary. *Minimal Residual Disease (MRD)* is a diagnostic measure based on Flow CytoMetry (FCM) data that captures the amount of blasts in a blood sample and is a clinical tool for planning patients' individual therapy, which requires reliable blast identification. In this work we propose a novel semi-supervised learning approach, which is acquired whenever large amounts of unlabeled data and only a small amount of annotated data is available. The proposed semi-supervised learning approach is based on Wasserstein Generative Adversarial Network (WGAN) latent space embeddings learned in an unsupervised fashion and a simple Fully connected Neural Network (FNN) trained on labeled data leveraging the learned embedding. We apply our proposed learning approach for semi-supervised classification of blasts vs. non-blasts. We compare our approach with two baseline approaches, 1) semi-supervised learning based on Principal Component Analysis (PCA) embedding, and 2) a deep FNN that is trained only on the annotated data without leveraging an embedding. Results suggest that our proposed semi-supervised WGAN embedding outperforms semi-supervised learning based on PCA embeddings and if only small amounts of annotated data is available it even outperforms an FNN classifier.**

I. Introduction

Minimal Residual Disease (MRD) is a prognostic value used for assessing therapy response in leukaemia (blood cancer) and is defined as the proportion between cancer cells (blasts) among the amount of all cells observed. Thus, for its computation the identification of blast cells among non blast cells is essential. In recent years the MRD value has been identified as powerful and guiding diagnostic tool for planning therapeutic intensity of an individual patient [1]–[4].

A. Childhood Acute Leukaemia

Acute Leukaemia is a blood cancer, which is caused by genetic alterations of hematopoetic progenitor cells, which

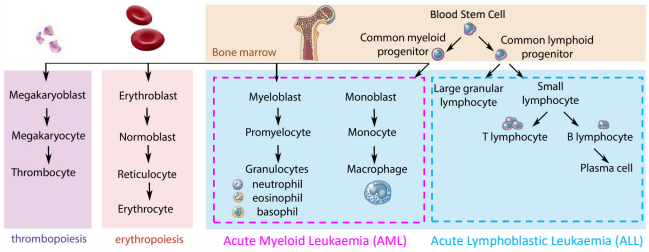

Fig. 1. Affected Blood Generation Pathways in Acute Leukaemia

are involved in the blood generation process (hematopoiesis). Consequently, this leads to the proliferation of leukaemic (undifferentiated) cells. Dependent on the affected progenitor cell type, we can differentiate between Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML).

1) Acute Lymphoblastic Leukaemia (ALL): Acute Lymphoblastic Leukemia is the most common leukaemia type in children and has a 25% fraction of cancer diseases in the pediatric population [5]. The involved progenitor cells can influence the differentiation process of T-cells or B-cells, why ALL is divided into two main sub categories namely T-ALL and B-ALL. The T-ALL's prevalance has its peak during childhood at age of 10 years and the prevalence's peak of B-ALL occurs in the age range between 2 to 5 years [6] [7]. Fig.1 illustrates the general blood cells generation pathway. Cell groups affected in ALL lie within the blue dotted rectangle on the right side. Since the 1970s the cure rates improved to a 85% survival rate for childhood ALL in developing countries [6].

2) Acute Myeloid Leukaemia (AML): Acute Myeloid Leukaemia is a rare leukaemia type in children and accounts for 20 percent of leukaemias in this population [8]. Its incidence increases with age [9] and the peaks of its prevalence

lie between 18.4 at age range 0 to 1 and 4.3 million at the age range 10 to 14 years in the United States [10]. Dependent on the AML subtype children have approximately a 70 percent five year survival rate, if they are younger than 15 years [8]. This blood cancer type affects the myeloid progenitor cells, which consequently leads to a reduction of mature blood cells and an increase of the number of malignant cells [10]. Cell groups affected in AML lie within the pink dotted rectangle in Fig. 1.

B. MRD Assessment over Leukaemia Therapy

The therapy for acute leukaemia is guided by treatment protocols, and evaluated by performing international clinical trials over several years, to ensure quality and safety [11]. In the case of ALL data used in this work the AIEOP-BFM 2009 protocol[1] is followed and for AML the AML BFM 2004 treatment protocol [8][2] is used. For the reliable assessment of MRD over leukaemia therapy, FlowCytometry (FCM) acquisitions of a patient's blood sample are obtained. This technique is more cost- and time effective compared to polymerase chain reaction [13]. FCM uses a laser-based technology to observe blood cell specific antigen patterns on a blood cell's surface (immunophenotypes) [14], [15]. Therefore, in a staining procedure specific antibodies with an attached fluorochrom are used to mark these antigens in a first step and subsequently, FCM is used to capture the emission spectra of fluorochroms triggered by the lasers' hit on the stained cells' surface. In Fig. 2 an example of a FCM image is visualised. Every dot represents a cell and its position is determined by the expression level of two antibodies, in this case CD 317 (y-axis) and CD 33 (x-axis). The scale of each axis is of logarithmic scale. Since FCM data is multi-dimensional (number of dimensions = number of antibodies in the panel) clinicians draw polygons (gates) around cell clusters of interest by observing different combinations of antibodies in two dimensional graphical representations (dot plots) and follow a specific gating hierarchy. The manual gating procedure strongly relies on the operator's skills and expertise, is highly subjective and time-consuming. The MRD assessment is especially challenging and important in the late therapy phases, where small populations of leukaemic cells (0.1% of all observed cells) have to be accurately detected, to estimate the risk of relapse and if necessary adapt the therapy. Additional challenges arise by the limited number of cells in a test tube available for FCM measurements and treatment, age or phenotypic variances, which influence the regeneration status of bone marrow precursors and the MRD value [13].

[1]AIEOP-BFM 2009 is a conducted randomized clinical trial for ALL between age 1-18 years in 10 countries in- and outside Europe, with approximately 1000 patients observed per year [12]) http:/www.bfm-international.org/ [accessed 2018-01-05]

[2]AML BFM 2004 is a conducted randomized clinical trial for children and adolescents with AML between age 0-18 years with 722 patients https://www.kinderkrebsinfo.de/ health_professionals/clinical_trials/ closed_trials/aml_bfm_2004/index_eng.html [accessed 2018-01-05]

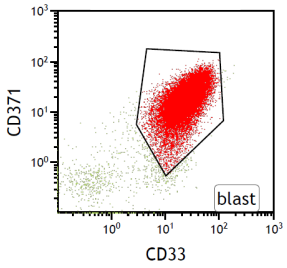

Fig. 2. Illustration of a sample obtained by a flow cytometer and the manual drawn blast gate (polygon) composed by the features CD 371 and CD 33. Leukaemic cells are illustrated in red, normal cells in green.

C. Related Work

Recent automated gating approaches proposed focus on modelling leukaemic and non-leukaemic cells for ALL [16], [17], [18], [19], [20], [21] since it is the more common leukaemia in children. The main goal lies in the observation of the multidimensional space at once to replace the multiple observation of 2D feature representations. However, visualisation and interpreting multi-dimensional data remains challenging. Recent approaches [22] use embedding techniques to represent multidimensional FCM data in reduced dimensionality. They use subsampled data of 10 control subjects to create t-SNE maps (cf. [23] for more details) for the subsequent visual MRD assessment in ALL by projecting a patient's data into the embedding space using the transformation learned. The limit lies in the restricted amount of cells that are observable per subject ($2 * 10^4$ out of 10^6) according to increasing computational burden with increasing cell counts. Van Unen et al. [24] address this issue by providing a hierarchical stochastic neighbourhood embedding of gastrointestinal disorders mass cytometry data. In [25] machine learning approaches were evaluated regarding their ability to automatically assign a biologically meaningful population label to every observed cell in childhood AML. Because of the rarity of the disease and consequently of available data containing cancer cell populations, they consider to focus on modelling the distribution of background cell populations (non cancer cells) using Gaussian Mixture Models. Cancer cells are detected subsequently, by estimating the anomaly of cells of a new sample estimated on bases of the learned control distribution. They showed that non-cancer populations of non AML cases and AML cases are combinable and can be used to increase the dataset in case of lack of available data to predict MRD in AML.

D. Contribution

The contribution of this work is three fold. First, we propose to utilize Wasserstein GAN (WGAN) for unsupervised learning and a simple Fully connected Neural Network (FNN) classifier trained on labeled data leveraging the learned embedding. Based on large amounts of unlabeled data (comprising both blood samples of control subjects and AML patients), we train a WGAN and leverage the learned latent space, semantically meaningful low-dimensional

embedding space, on which we base subsequent classifier training for classification of blasts vs. non-blasts. Second, we provide extensive evaluations to assess the performance of our proposed semi-supervised learning approach in comparison to two baseline approaches, 1) semi-supervised learning based on Principal Component Analysis (PCA) embedding, and 2) a deep FNN that is trained only on the annotated data without leveraging an embedding. The new formulation provides a dimensionality reduction scheme to compute z-embeddings of FCM data, considering the whole amount of cells in a patient's sample without applying subsampling techniques compared to recent proposed embedding approaches [22] [23]. Third, we apply these methods on FCM data that have not been used yet for blast identification and in further consequence to assess MRD in AML patients.

In Section II a brief overview of the methodology proposed is given and the embedding techniques as well as the used neural network classifiers are introduced. In Section III the datasets are presented and the evaluation setup is summarized. This work concludes in Section IV with a discussion of directions for future work.

II. Methodology

In this section a brief summary of the approaches designed for the automatic MRD assessment of AML are introduced. For this task the solving of a binary classification problem (blast, non-blast) for every measured blood cell is necessary. According to the definition of MRD (cf. Equation 1) the counts of blast and non-blast cells have to be estimated, to be able to compute this therapeutic value. In Equation 1 MRD is represented by the ratio between the number of blast cells $N^{(x)}_{blasts}$ and number of all cells in a sample $N^{(x)}_{allCells}$, where x refers to the x-th sample.

$$MRD^{(x)} = \frac{N^{(x)}_{blasts}}{N^{(x)}_{allCells}} \tag{1}$$

The proposed automatic gating strategies observe the multi-dimensional feature space in contrast to manual gating procedures. Nine features are measured per cell which consist of two physical FCM measures and seven distinct antibody types (cf. Section III-B). The data are annotated per medical experts and dependent on the condition of the patient, approximately $3{*}10^5$ - 10^6 cells are extracted per subject. In the following, we describe two approaches for leaning low-dimensional feature embeddings.

A. Generative Adversarial Networks (GAN)

We propose to utilize WGANs for unsupervised learning of a low-dimensional embedding. This unsupervised-training approach is inspired by the work of [26], where a deep convolutional generative adversarial network (DCGAN) [27] is utilized for unsupervised learning to perform anomaly detection. We first train a WGAN [28], following the improved WGAN training procedure [29]. In the original GAN training procedure [30], the Jensen-Shannon divergence is minimized. In contrast, during WGAN training, the Wasserstein distance – a smoother metric – is intended to stabilize (and thus to improve) the training procedure. A fully connected network architecture is used for both components of the WGAN, the generator and the discriminator. The generator is implemented as fully connected decoder network, which maps from low-dimensional space (3 dimensions) to original higher-dimensional data space (9 dimensions). The discriminator maps from original data space (9 dimensions) to low-dimensional space (3 dimensions). Both networks comprise 4 hidden layers with 128 neurons and an output layer. The generator utilizes rectified linear unit (ReLU) activation functions [31] and a *tanh*-function on top of the linear output layer, whereas the discriminator utilizes leaky ReLU activation functions and a linear output layer. During WGAN training, the generator is trained to generate realistic "looking" data samples from random inputs $z \in \mathcal{Z}$ that are sampled from the normal distribution. WGAN training yields a parametrized *generator* and *discriminator* and a semantically meaningful (low-dimensional) *latent space* $\mathcal{Z}$, from which realistic data samples can be generated. For further processing steps, we only utilize the *generator* and freeze the learned parameterization, i.e. the parameters are not updated in the subsequent steps. Following the *mapping procedure* from data space to the latent space, introduced in [26], for a given data sample x, we find the optimal $z \in \mathcal{Z}$ in an iterative process via n backpropagation steps, where only z coordinates are adapted. We consider that z optimizes that maps to the generated sample $\hat{x}$ with minimum residual $|x - \hat{x}|$. All data samples $x \in \mathcal{X}$ are mapped via this procedure to latent space $\mathcal{Z}$, which represents our low dimensional embedding space.

B. Principal Component Analysis (PCA)

As baseline approach we utilize Principal Component Analysis to compute low-dimensional FCM data representations in a feature space of reduced dimensionality (3 dimensions). This technique is used in an unsupervised way to learn a new representation of decorrelated components (principal components) in terms of a linear combination of the orginal variables. It has been used for the analysis for FCM data since 1984 [32], but for MRD assessment of AML in developing cohorts, to our knowledge, no application has been reported yet. PCA is used as a baseline for an embedding technique in the evaluation scheme proposed.

C. Neural Network Classifier (NN)

For the final discrimination between blast and non-blast, we train a neural network classifier on top of the embeddings computed with WGAN or with the PCA embedding method described above. The WGAN or the PCA embeddings are fed into a classifier, which is a simple *fully connected neural network* (NN) comprising two hidden layers with 128 neurons each and a classification layer. We use rectified linear units as activation function. The NN classifier is trained using cross entropy loss.

TABLE I
BLAST IDENTIFICATION PERFORMANCE OF OUR PROPOSED WGAN-NN CLASSIFIER AND TWO ALTERNATIVE APPROACHES, PCA-NN AND FNN, TRAINED ON SMALL OR LARGE ANNOTATED DATASETS.

	Scenario *small* dataset					Scenario *large* dataset				
Methodology	Precision	Recall	fscore	Specificity	AUC	Precision	Recall	fscore	Specificity	AUC
PCA-NN	0.5471	0.7690	0.5829	0.8151	0.7952	0.5986	0.9213	0.6893	0.7916	0.9008
FNN	**0.6037**	0.7543	0.6164	**0.8594**	0.8336	**0.6526**	**0.9250**	**0.7305**	0.7933	0.8892
WGAN-NN	0.5502	**0.9049**	**0.6483**	0.7452	**0.8592**	0.6320	0.8942	0.7147	**0.8364**	**0.9139**

D. Fully Connected Network Classifier (FNN)

In contrast to the simple neural network classifier that is trained on top of the low-dimensional embeddings, we implement a deep fully connected feed forward neural network, which is used as reference approach. Based on this model, we study the achievable performance of a state-of-the-art classifier trained on a large annotated dataset. This networks comprise 4 hidden layers with 128 neurons followed by 2 hidden layers with 32 neurons and a classification layer. The activations of the hidden units are computed applying ReLU activation functions. This network is soley trained on original data without any prior embedding or preprocessing, and thus the optimal discriminative feature representation has to be learned during classifier training.

III. RESULTS

In this section first the dataset and the preprocessing steps are introduced and second a description of the evaluation setup and results are presented.

A. Preprocessing

Every subject's sample is preprocessed and annotated at the national diagnostic reference center for paediatric AML, following the international standard operating procedure for MRD detection using FCM. According to the partial overlap of different fluorochrom spectra, *spillover compensation* is used to obtain statistical independence of the data, by using a correction matrix. The preprocessing concludes with a normalization step of the measured parameters to obtain a value range between 0 and 1.

B. Datasets

As introduced in Section I-B specific antibodypanels are used to stain the cell specific antigen patterns on the cells' surfaces. In this study nine different features are measured for every cell: two optical parameters (FSINT, SSINT) and seven fluorescence based parameters (CD38, CD34, CD117, CD33, CD123, CD45RA, CD45). One feature represents a dimension in the multidimensional data space. All participants' guardians (parents) and patients were informed about the aim of the study and gave their written, informed consent prior to inclusion. According to Licandro et al. [25] *healthy* cell populations of ALL and AML cases are combinable and can be used to increase the dataset in case of lack of available data to learn a good data representation of AML data during unsupervised learning. For training of the simple NN classifier and for training of the deep FNN classifier only annotated AML data is used. Thus, we considered to follow this strategy to achieve a higher number of samples and use a combination of AML and ALL dataset as well.
The AML dataset is acquired using FCM measurements of in total 15 AML diagnosed subjects containing blast and non-blasts. The ALL dataset consists of measurements of 24 patients that have been diagnosed with ALL in the remission phase, i.e. where no blasts are present and thus no annotations on cell-level are available.

C. Evaluation Strategies

To keep comparability of the embedding approaches evaluated the same training set consisting of ALL and AML samples are used to learn an embedding without subsampling of the data in an unsupervised way. As PCA implementation we used the Incremental Principal Component Analysis (IPCA) toolbox integrated in scikit learn Python framework[3] with the number of components set to 3. The WGAN is trained in an unsupervised way with an latent space of three dimensions. We evaluated the blast identification performance using two different scenarios of annotated training data sizes: In *Scenario small dataset* 32 cells per case (i.e. patient) were used and in *Scenario large dataset* 50000 cells per case, which reflect scenarios of low and high annotation burden, respectively. For both strategies, 3-fold cross-validation was performed with 5 cases per split and every patient occurs only in one split. In every iteration, 2 splits were used for training and 1 split was used for validation. In the testing phase no subsampling was performed.

D. Blast Identification

We use the following measures for blast identification performance evaluation: precision, recall, f-score, specificity and AUC [34]. In Table I the evaluation parameters for both scenarios and every approach are illustrated and computed via averaging the values of the patient specific clinical performance measures. In scenario *large dataset*, WGAN-NN and FNN outperform the PCA-NN approach. This suggests, whenever large amounts of annotated data for supervised classifier training is available, there is no additional performance gain, when classification training is performed on an embedding learned in an preceding unsupervised training step, since the FNN is already capable to learn the data representation. In contrast to this, in scenario *small dataset* (i.e., a case of a small number of available annotated data) WGAN outperforms the

[3]http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.IncrementalPCA.html [33]

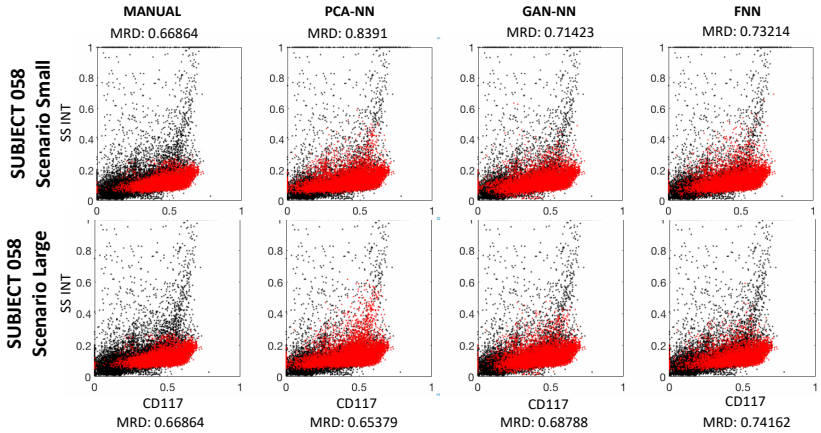

Fig. 3. Qualitative results for *scenario small* (first row) and *scenario large* second row) of identified blasts by PCA-NN, GAN-NN and FNN of a subject's testsample compared to manual blast annotations in AML data for the features SS INT and CD117.

FNN classifier. In Fig. 3 an example for qualitative results for a test sample of a subject on *scenario small* (top) and *scenario large* (bottom) are visualised with corresponding MRD values. Blasts are shown in red and non-blasts in black. The manual annotations are presented in the first column, while the identified labels are shown in the second column for PCA-NN, third column for GAN-NN and for FNN in the fourth column. In Fig. 5 the predicted MRD values are plotted versus the true MRD values. Every point represents a subject and should lie on the diagonal in the optimal case. It is observable that the samples of PCA-NN and FNN deviate more from the optimal line compared to the WGAN-NN approach, especially in the small scenario. Fig. 4 illustrates the Receiver Operating Characteristic (ROC) curve for the blast identification performance for all patients' cells observed of a simple neural network classifier that has been trained on WGAN embedding (red), on a PCA embedding (blue) and a fully connected network classifier (green), that has been trained on original input data. The results of the *small* annotated dataset scenario are visualised left and on the right the *large* annotated dataset scenario is shown.

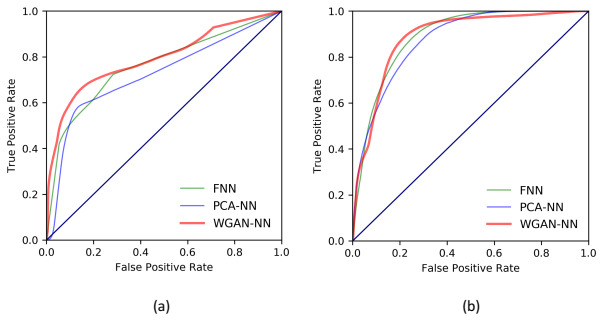

Fig. 4. Receiver Operating Characteristic (ROC) curves. Binary classification performances of a simple neural network classifier trained on a WGAN embedding (WGAN-NN) or trained on a PCA embedding (PCA-NN), and a deep fully connected neural network classifier (FNN) soley trained on original input data. (a) Scenario *small* annotated dataset. (b) Scenario *large* annotated dataset.

IV. Conclusion

We propose a novel semi-supervised learning approach based on a combination of WGAN latent space embedding and a simple fully connected neural network. We provided two evaluation schemes: In the first scenario large annotated datasets are used and results show, that WGAN-NN and FNN outperform the PCA-NN approach. According to the rareness of the disease, only a limited number of annotated data is available. We created a second evaluation scenario comprising a small dataset simulating this fact, where the results show that WGAN outperforms both, the PCA-NN and FNN approach. For future work we aim to use data from different countries, machines and background samples and focus on obtaining machine and country independent data representations.

Acknowledgment

This work was supported by ZIT Life Sciences 2014 (1207843), by the Marie Curie Alumni Association and has received funding from IBM (Fellowship and Faculty Grant) and FWF (I2714-B31).

References

[1] C. Eckert, G. Henze, K. Seeger, N. Hagedorn, G. Mann, R. Panzer-Grümayer, C. Peters, T. Klingebiel, A. Borkhardt, M. Schrappe *et al.*, "Use of allogeneic hematopoietic stem-cell transplantation based on minimal residual disease response improves outcomes for children with relapsed acute lymphoblastic leukemia in the intermediate-risk group," *Journal of Clinical Oncology*, vol. 31, no. 21, pp. 2736–2742, 2013.

[2] D. Campana, "Minimal residual disease in acute lymphoblastic leukemia," *ASH Education Program Book*, vol. 2010, no. 1, pp. 7–12, 2010.

[3] M. Borowitz, D. Pullen, J. Shuster, D. Viswanatha, K. Montgomery, C. Willman, and B. Camitta, "Minimal residual disease detection in childhood precursor–b-cell acute lymphoblastic leukemia: relation to other risk factors. a children's oncology group study," *Leukemia*, vol. 17, no. 8, pp. 1566–1572, 2003.

[4] P. Theunissen, E. Mejstrikova, L. Sedek, A. J. van der Sluijs-Gelling, G. Gaipa, M. Bartels, E. Sobral da Costa, M. Kotrová, M. Novakova, E. Sonneveld, C. Buracchi, P. Bonaccorso, E. Oliveira, J. G. te Marvelde, T. Szczepanski, L. Lhermitte, O. Hrusak, Q. Lecrevisse, G. E. Grigore, E. Froňková, J. Trka, M. Brüggemann, A. Orfao, J. J. M. van Dongen, and V. H. J. van der Velden, "Standardized flow cytometry for highly sensitive MRD measurements in B-cell acute lymphoblastic leukemia," *Blood*, vol. 129, no. 3, pp. 347–357, 2017.

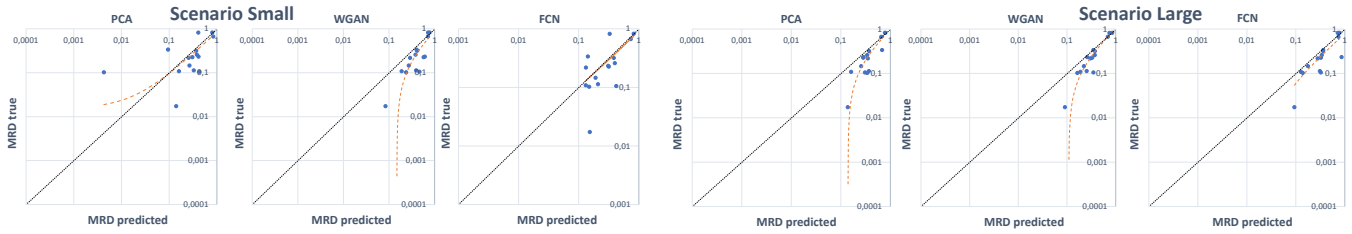

Fig. 5. Visualisation of MRD assessment in AML with *scenario small* (first column) and *scenario large* (second column).

[5] M. Karsa, L. Dalla Pozza, N. Venn, T. Law, R. Shi, J. Giles, A. Bahar, S. Cross, D. Catchpoole, M. Haber, G. Marshall, M. Norris, and R. Sutton, "Improving the identification of high risk precursor B acute lymphoblastic leukemia patients with earlier quantification of minimal residual disease." *PloS one*, vol. 8, no. 10, p. e76455, January 2013.

[6] C.-H. Pui, C. G. Mullighan, W. E. Evans, and M. V. Relling, "Pediatric acute lymphoblastic leukemia: where are we going and how do we get there?" *Blood*, no. 6, pp. 1165–1174, 2012.

[7] H. Inaba, M. Greaves, and C. Mullighan, "Acute lymphoblastic leukaemia." *Lancet*, vol. 381, no. 9881, pp. 1943–55, June 2013.

[8] U. Creutzig, M. Zimmermann, M. N. Dworzak, J. Ritter, G. Schellong, and D. Reinhardt, "Development of a curative treatment within the AML-BFM studies," *Klinische Padiatrie*, vol. 225, no. SUPPL1, pp. 79–86, 2013.

[9] G. Juliusson, P. Antunovic, A. Derolf, S. Lehmann, L. Möllgård, D. Stockelberg, U. Tidefelt, A. Wahlin, and M. Höglund, "Age and acute myeloid leukemia: real world data on decision to treat and outcomes from the Swedish Acute Leukemia Registry." *Blood*, vol. 113, no. 18, pp. 4179–87, apr 2009.

[10] S. E. Puumala, J. A. Ross, R. Aplenc, and L. G. Spector, "Epidemiology of childhood acute myeloid leukemia." *Pediatric blood & cancer*, vol. 60, no. 5, pp. 728–33, may 2013.

[11] U. Creutzig, M. Zimmermann, J.-p. Bourquin, M. N. Dworzak, G. Fleischhack, N. Graf, T. Klingebiel, B. Kremens, T. Lehrnbecher, C. V. Neuhoff, A. Sander, A. V. Stackelberg, J. Star, and D. Reinhardt, "Randomized trial comparing liposomal daunorubicin with idarubicin as induction for pediatric acute myeloid leukemia : results from study," *Blood*, vol. 122, no. 1, pp. 37–44, 2013.

[12] M. Dworzak, "Minimal residual disease in pediatric acute lymphoblastic leukemia: Bfm experience," *Hematología*, vol. 17, October 2013.

[13] G. Gaipa, G. Cazzaniga, M. Valsecchi, R. Panzer-Grümayer, B. Buldini, D. Silvestri, L. Karawajew, O. Maglia, R. Ratei, A. Benetello, S. Sala, A. Schumich, A. Schrauder, T. Villa, M. Veltroni, W.-D. Ludwig, V. Conter, M. Schrappe, A. Biondi, M. Dworzak, and G. Basso, "Time point-dependent concordance of flow cytometry and real-time quantitative polymerase chain reaction for minimal residual disease detection in childhood acute lymphoblastic leukemia." *Haematologica*, vol. 97, no. 10, pp. 1582–93, October 2012.

[14] G. Basso, M. Veltroni, M. Valsecchi, M. Dworzak, R. Ratei, D. Silvestri, A. Benetello, B. Buldini, O. Maglia, G. Masera *et al.*, "Risk of relapse of childhood acute lymphoblastic leukemia is predicted by flow cytometric measurement of residual disease on day 15 bone marrow," *Journal of Clinical Oncology*, vol. 27, no. 31, pp. 5168–5174, 2009.

[15] M. Dworzak, G. Fröschl, D. Printz, G. Mann, U. Pötschger, N. Mühlegger, G. Fritsch, and H. Gadner, "Prognostic significance and modalities of flow cytometric minimal residual disease detection in childhood acute lymphoblastic leukemia," *Blood*, vol. 99, no. 6, pp. 1952–1958, 2002.

[16] R. Licandro, P. Rota, M. Reiter, and M. Kampel, "Flow Cytometry based automatic MRD assessment in Acute Lymphoblastic Leukaemia: Longitudinal evaluation of time-specific cell population models," in *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, jun 2016, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7500274/

[17] I. Naim, S. Datta, J. Rebhahn, J. Cavenaugh, T. Mosmann, and G. Sharma, "Swift - scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: Algorithm design," *Cytometry Part A*, vol. 85, no. 5, pp. 408–421, 2014.

[18] H. Zare, P. Shooshtari, A. Gupta, and R. Brinkman, "Data reduction for spectral clustering to analyze high throughput flow cytometry data," *BMC bioinformatics*, vol. 11, no. 1, p. 403, 2010.

[19] N. Aghaeepour, G. Finak, H. Hoos, T. Mosmann, R. Brinkman, R. Gottardo, R. Scheuermann, F. Consortium, D. Consortium *et al.*, "Critical assessment of automated flow cytometry data analysis techniques," *Nature methods*, vol. 10, no. 3, pp. 228–238, 2013.

[20] A. Bashashati and R. Brinkman, "A survey of flow cytometry data analysis methods." *Advances in bioinformatics*, vol. 2009, pp. 584 603–584 603, 2009.

[21] M. Reiter, P. Rota, F. Kleber, M. Diem, S. Groeneveld-Krentz, and M. Dworzak, "Clustering of cell populations in flow cytometry data using a combination of Gaussian mixtures," *Pattern Recognition*, vol. 60, pp. 1029–1040, dec 2016.

[22] J. A. DiGiuseppe, M. D. Tadmor, and D. Pe'er, "Detection of minimal residual disease in B lymphoblastic leukemia using viSNE," *Cytometry Part B: Clinical Cytometry*, vol. 88, no. 5, pp. 294–304, sep 2015.

[23] L. van der Maaten, "Barnes-Hut-SNE," *arXiv*, jan 2013.

[24] V. van Unen, T. Höllt, N. Pezzotti, N. Li, M. J. T. Reinders, E. Eisemann, F. Koning, A. Vilanova, and B. P. F. Lelieveldt, "Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types," *Nature Communications*, vol. 8, no. 1, p. 1740, dec 2017.

[25] R. Licandro, M. Reiter, M. Diem, M. Dworzak, A. Schumich, and M. Kampel, "Application of Machine Learning for Automatic MRD Assessment in Paediatric Acute Myeloid Leukaemia," in *ICPRAM2018*. IEEE, jan 2018.

[26] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," in *Information Processing in Medical Imaging. IPMI 2017. Lecture Notes in Computer Science*. Springer, Cham, jun 2017, pp. 146–157.

[27] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[28] M. Arjovsky, S. Chintala, and B.L., "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[31] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[32] E. Lugli, M. Roederer, and A. Cossarizza, "Data analysis in flow cytometry: the future just started." *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, vol. 77, no. 7, pp. 705–13, jul 2010.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.