Automatic Minimal Residual Disease Assessment in Acute Myeloid Leukemia on Flow Cytometry Data: A Comparison

Ivan Lukšić
Computer Vision Lab
TU Wien
Vienna, Austria
e12235571@student.tuwien.ac.at

***Abstract*—Automatic Minimal Residual Disease (MRD) Assessment in Acute Myeloid Leukemia (AML) on Flow Cytometry (FCM) samples entails a machine-learning classification model that classifies the cell samples as non-blast (non-cancerous) or blast (cancerous). This paper explains and compares two automatic AML-MRD assessment approaches on FCM samples. The first approach includes multiple state-of-the-art algorithms and a novel Gaussian Mixture Model (GMM) where said approaches rely on learning non-cancer cell populations. The second approach uses a semi-supervised UMAP algorithm that trains on non-cancer cells and classification based on detecting anomalies (blast cells) in cell populations. The performance and results of these models are compared regarding precision, recall, and F-score metrics with a variation of the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm achieving the best results. In conclusion, the advantages and limitations of said approaches and the problem that impedes both approaches are discussed.**

***Index Terms*—Automatic MRD assessment, Machine learning, Cancer cell detection, Anomaly detection, Gaussian Mixture Model, Semi-supervised learning**

I. Introduction

Acute Myeloid Leukemia (AML) is a cancer that starts in the bone marrow and moves to the blood and is the second most common leukemia (with the first being Acute Lymphoblastic Leukemia (ALL)) found in children/adolescents and accounts for 20% of all childhood leukemia [2].

Minimal (Measurable [5]) Residual Disease (MRD) denotes malignant cells that remain in the blood during or after the chemotherapy/radiotherapy [6]. In addition to being used to determine the intensity of the subsequent therapy steps, MRD has a strong correlation with the outcome of the treatment and possible disease relapse [5], [6].

Morphological techniques (eg. microscopy [11]) don't provide adequate information about the treatment as they can only detect malignancy when the malignant cells account for 1%-5% of the sample cells [13]. Figure 1 illustrates this as can be seen that only leukemia is above the microscopic-morphologic threshold and MRD assessment treshold sits under it, implying other methods need to be used to detect MRD.

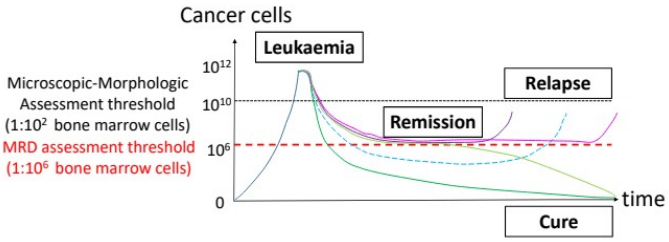

Fig. 1. Schematic illustration of the method proposed for automated MRD detection in AML FCM data (image taken from [7]).

MRD is most commonly detected by immunophenotyping using multiparameter Flow CytoMetry (FCM) (also abbreviated as MFC [5]) [5]. FCM produces multiparameter tabular data where each parameter corresponds to fluorescence marker data value [4]. FCM sample is considered MRD positive if it has $\geq 0.1\%$ of CD45-expressing cells with CD45 being a core marker [5]. Relapse is defined as conversion from MRD negative to MRD positive sample results [5].

FCM data is manually interpreted by clinical experts in a process known as "gating" [12]. Gating is a "selection of successive subpopulations of cells" [12] on two-dimensional plains (dot plots [7]). Figure 2 shows a single FCM AML sample with each plot using a different combination of parameter values on its axis.

The underlying problem of gating is that it is based on expert knowledge and among experts there are disagreements about how to apply gates [12]. That is why the manual gating approach is being regarded as subjective gating [4]. Regarding that, it is hard to achieve consistency in gating outside of each sample being reviewed by the same individual or a few coordinated experts [8]. Additionally, the high heterogeneity of AML cell populations makes manual gating resource heavy and complex [14].

Consequently, there is a need for automatic cell classification for AML FCM samples. In contrast to manual gating, which can utilize just 3 features at once, automatic cell classification can exploit the whole feature space [14].

This paper compares two related works on the topic of

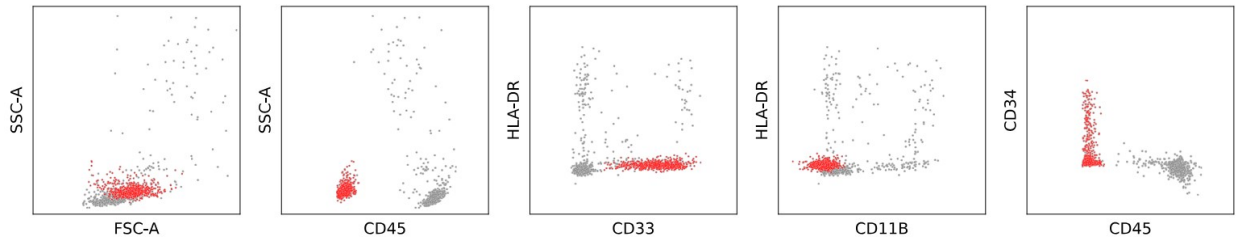

Fig. 2. Dot plots of an FCM AML sample with gray cells representing healthy cells and red being malignant cells (image taken from [14]).

automatic MRD AML assessment on FCM data. Automatic MRD assessment in AML on FCM data in both compared papers entails a machine-learning classification model that classifies the cell samples as background (non-cancerous) or blast (cancerous).

In [7] Licandro et al. propose solving the underlying problem of automatic AML-MRD detection by using a Random Forest classifier (RF), Support Vector Machine (SVM), Gaussian Mixture Model (GMM) and background formulations of these models where said approaches rely on learning non-cancer cell populations.

Weijler et al. [14] use Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), a semi-supervised approach in which the model is trained on non-cancer cells, and classification is based on detecting anomalies (blast cells) in cell populations.

The approaches in the two papers are compared based on precision, recall, and F-score. The best approach is determined for each comparison metric and overall best approach is discerned by the highest result across metrics with the emphasis being put on recall as it minimizes the chance of MRD not being detected. The results indicate that a variation of the UMAP approach performs best in AML-MRD detection on FCM data.

In the two subsequent sections, the two papers are examined. In the end, the results and shortcomings of these approaches are discussed and the conclusion is given.

II. Random Forest Classifier, Support Vector Machine, Gaussian Mixture Model and their background formulations

Licandro et al. in [7] explore three different machine-learning algorithms training them on simple (AML FCM data) backgrounds and introduce a complex background formulation where backgrounds of ALL and AML samples are combined to measure if background outlier distributions improve performance in comparison to simple backgrounds.

Figure 3 showcases two density distributions, first of the background and blast distributions of an AML dataset, and second of AML and ALL background distributions from separate datasets with the red line indicating where AML blast cells are. In Figure 3 it is shown that while blast cells do overlay background cells, they are mostly in regions of sparse density, especially in regard to ALL background distribution. Heed that this is only one of the possible parameter combinations and others may not evince the same similarity.

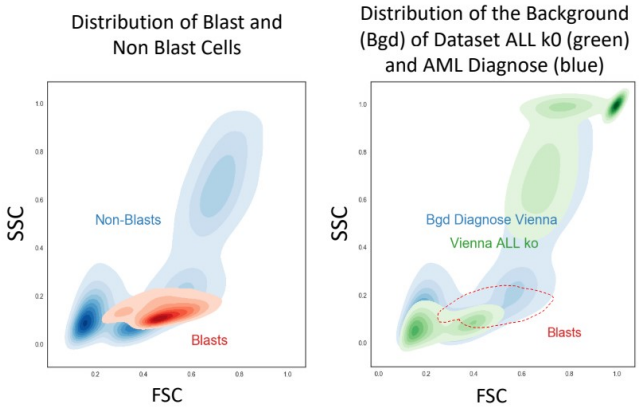

Fig. 3. Visualization of relations of AML background (blue) and AML blast (red) cells on the left and AML background and ALL background (green) with the red line enclosing a space where AML blast would lie on the right (image taken from [7]).

The paper aims to test different machine-learning algorithms on the task of AML MRD detection and to test if combining the backgrounds from different leukemia types, thus creating more training data, helps to overcome the problem of AML MRD FCM data scarcity and improve the model performance. In the results six models from this paper are compared, each of the algorithms and their combined background counterpart, denoted by adding Bgd suffix to the model name.

A. Random Forest Classifier

RF uses decision trees where each tree is comprised of a random subset of FCM training data. For each subset of randomly selected antibody features, a new node in the decision tree is constructed bearing in mind the upper tree levels. This can be illustrated by Figure 4 in which we see a simple example of how classification is done using RF algorithm. RF is trained in a supervised manner where each cell is annotated manually. The model produces binary output for each cell indicating non-blast or blast cell and the result metrics are derived from the accuracy of the output.

B. Support Vector Machine

The SVM approach in the work is used as a baseline and its performance is compared to the performance of RFC and

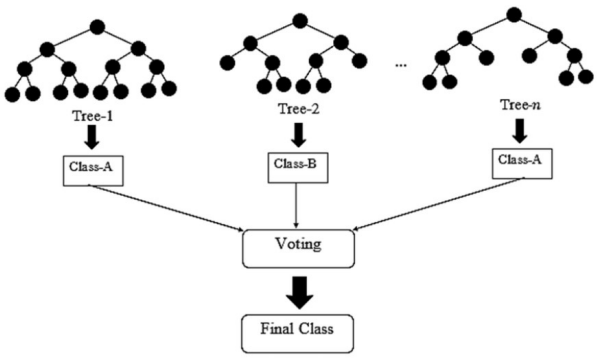

Fig. 4. Illustration of RF algorithm with two classes (image taken from [1]).

GMM. In Figure 5 the principle behind the SVM algorithm is seen. Hyperplane which aims to separate data into a discrete number of classes is constructed based on labeled training data [3]. The paper uses a more complex SVM formulation than in the illustration in Figure 5.

The SVM model in the paper is based on the Radial Basis Function (RBF) kernel and classification is based on events without providing information on neighboring events or different populations. The SVM is trained in a supervised manner and the test phase provides one output (non-blast or blast) for each input cell.

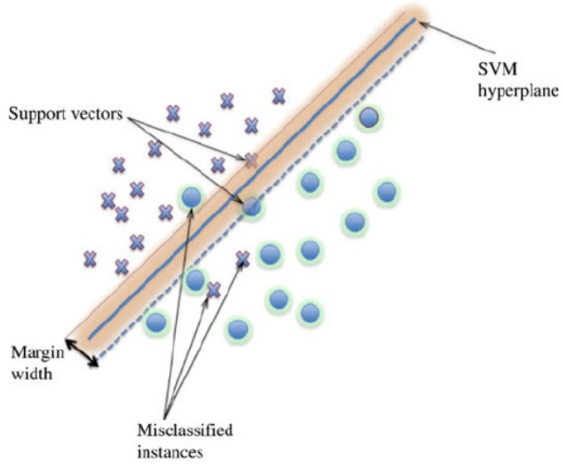

Fig. 5. Illustration of simple SVM example (image taken from [3]).

C. Gaussian Mixture Model

GMM clusters and classifies cells as blast or non-blast and as a model are widely used as GMMs are said to be flexible during the analysis of FCM data.

GMM "is a parametric probability density function represented as a weighted sum of Gaussian component densities" [10]. It is used in a setting where there is a continuous distribution of values [10]. GMM derives parameters from the training dataset by iterative Expected-Maximization (EM) algorithm [10]. In Figure 6 it is seen how simple GMM (green line) is made up out of two Gaussian distributions (blue and red lines) which are denoted as components.

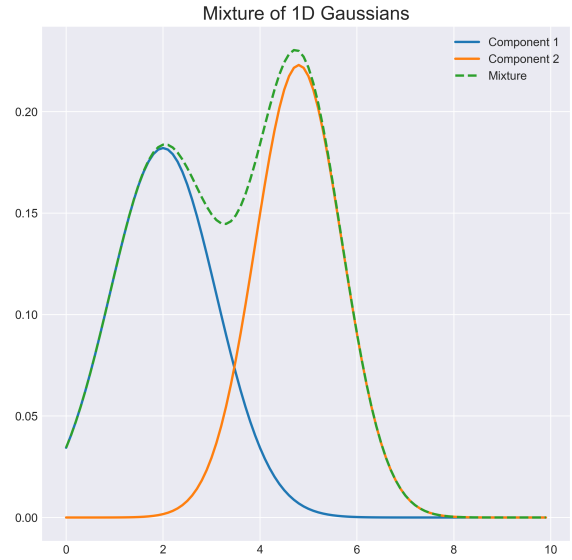

Fig. 6. Illustration of GMM using two Gaussian distributions (image taken from Deep A[2]).

The model in the paper is made using an adapted EM algorithm and 2 Gaussian distributions. Model is a two-step with the first step classifying a cell as non-blast or outlier, in the second step, outliers from the first step are classified as non-blast or blast. In the paper, only background populations are modeled as there is more data on background populations.

III. Uniform Manifold Approximation and Projection for Dimension Reduction based approach

The UMAP-based approach in [14] is a semi-supervised method for automatic AML-MRD detection using only background populations. It is used in a way that combines events of an FCM sample with randomly selected control sample (sample with no-blast cells) events and applies the UMAP algorithm to form two clusters where clusters with little to no control samples are declared as blast clusters. An advantage of this approach is that no labeled FCM data that contains leukemic cells is needed, only blast-free labeled FCM data is used.

The underlying idea is that by using healthy cell populations it is possible to detect anomalies that are presented in the form of blast cells. The approach also uses Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) for building the model. The model has the same

[2]DeepAI. Gaussian mixture models. https://deepai.org/machine-learning-glossary-and-terms/gaussian-mixture-models. Accessed May 22, 2023.

output as models in [7], binary classification for each input cell, where the cell can be non-blast or blast.

In Figure 7 the four-step method for AML MRD detection is seen. The first step consists of mixing a random subset of control data from multiple samples with input FCM data as denoted by 1 in Figure 7. In the second step, the UMAP embedding happens. Then HDBSCAN is used to identify clusters. Finally, the clusters with control sample presence $\leq$ 5% are labeled as blast as control samples are non-blast and lack of their presence indicates that the cells in the cluster are blast cells.

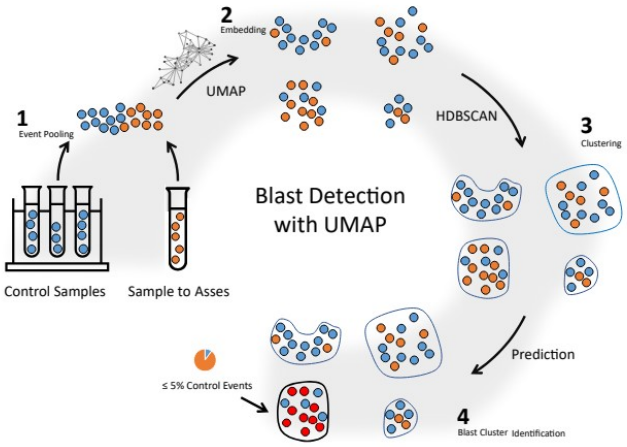

Fig. 7. Schematic illustration of the method proposed for automated MRD detection in AML FCM data (image taken from [14]).

UMAP is essentially a dimension reduction algorithm [9] that reduces the dimensionality of multiparameter spaces. In Figure 8 it is seen how dimensionality reduction looks on AML FCM data. In the first column, the sample composition after step two of the classification pipeline is seen. The second column showcases the same data after HDBSCAN. In the third column, the ground truth of where blast cells lie on the graph is given, and the fourth column contains all predicted blast cells.

Each row in Figure 8 corresponds to a different sample. Row A is an example of correctly detected blast cells. Row B sees a false negative classification due to too many control events in the cluster. Row C in contrast contains a false positive cluster. While row D is an example of poor separation of clusters.

In the paper, three different models are created, each using UMAP-HDBSCAN architecture, but training on using data from two different FCM tubes. Both tubes have 8 markers and 5 of them are used by both. One model trains on data acquired by the "Leukemia Associated ImmunoPhenotype (LAIP)-tube" while the other trains on the "Colony Forming Unit (CFU)-tube". The third model combines LAIP and CFU data.

IV. Results

Methodologies mentioned in both papers are used to automatically classify AML MRD on FCM data.

Regarding [7], it is seen that RF and GMM perform better than their combined background counterparts across all metrics, SVM and SVMBgd follow suit with the exception being precision, which is higher in SVMBgd.

Figure 9 visualizes automatic MRD assessment, the left column containing RF, SVM, and GMM, and the right column containing their background equivalents. Each point on the graph corresponds to a sample. Samples colored red are outside the accuracy threshold, which is determined by clinicians, the blue-colored samples are considered to be accurately predicted.

In [14] it is seen that UMAP-CFU outperforms the other two variations across metrics, and UMAP-CFU and LAIP achieves superior results to UMAP-LAIP. The performance difference between UMAP-CFU and UMAP-LAIP can be explained due to UMAP-CFU using fixed drop-in markers which UMAP-LAIP uses only in 25% of samples.

The comparison of all of the aforementioned methodologies with regards to precision, recall and F1 scored is reported in Table I. In terms of precision, the Random Forrest approach showcased in [7] is superior to others, for recall and F1-score, the UMAP-CFU approach tested in [14] outperforms the rest.

TABLE I
Results of MRD assessment performance

Methods	Precision	Recall	F-score
RF [7]	**0.762**	0.462	0.576
SVM [7]	0.620	0.580	0.600
GMM [7]	0.448	0.264	0.332
RFBgd [7]	0.742	0.396	0.516
SVMBgd [7]	0.680	0.531	0.597
GMMBgd [7]	0.439	0.261	0.327
UMAP-LAIP [14]	0.563	0.462	0.443
UMAP-CFU [14]	0.572	**0.812**	**0.607**
UMAP-CFU and LAIP [14]	0.567	0.612	0.514

When comparing the results it is important to note that these results aren't achieved on the exact same data regarding the markers as in [7] backgrounds of ALL samples are used and in [14] the control FCM data is used. Also, the UMAP approaches use more data than those in [7]. This paper compares the results of models in respective papers as they appear in the paper, and not modified versions of models for them to be trained on the same data as respective approaches are inherently different.

V. Conclusion

In this abstract, two papers and 9 different approaches to automatic AML-MRD assessment on FCM data are showcased.

The results show that in terms of recall and f-score the UMAP-CFU approach that is elaborated in [14] performs the best and can be regarded as the state-of-the-art in this task.

With 4 years between the papers, and the newer paper performing better, improvement in solving the task is seen. Albeit, even the best-performing model is not yet fit to be used in a clinical setting.

Also with the development of algorithms such as UMAP, which have found widespread usage in bioinformatics and

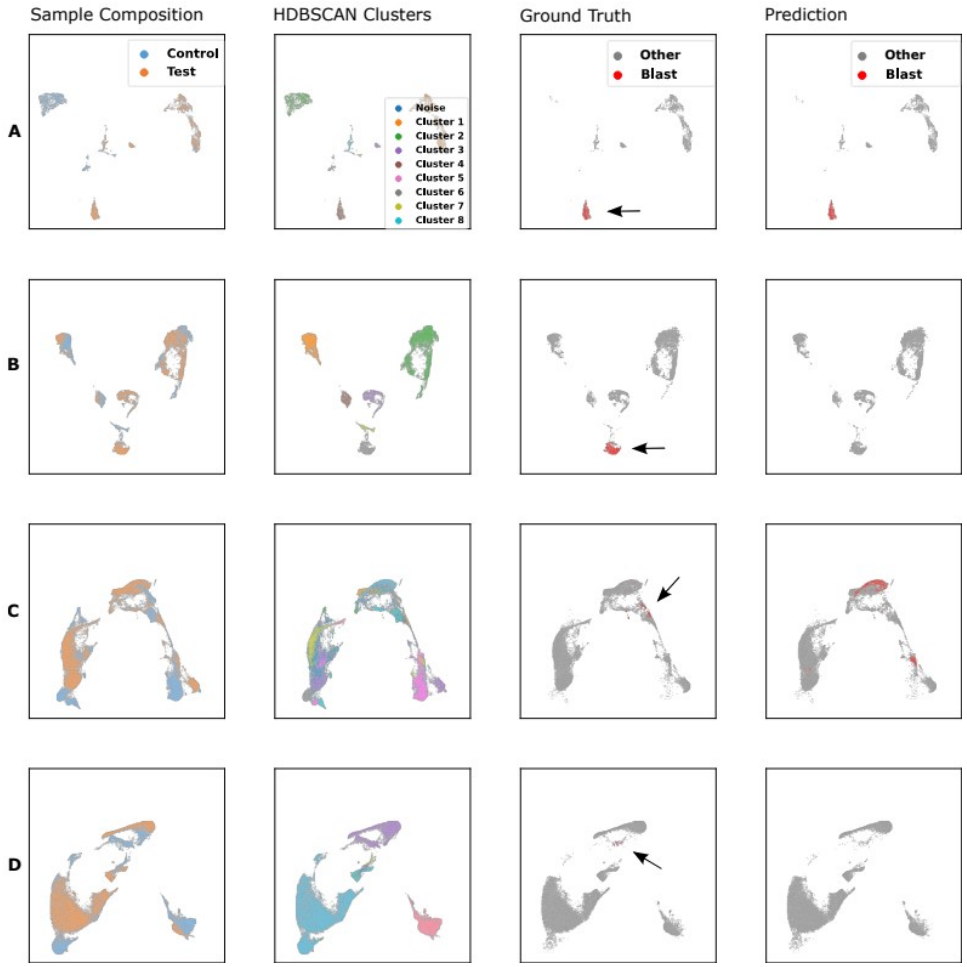

Fig. 8. Visualization of clusters in classification pipeline after applying UMAP algorithm with the ground truth in the third column. Each row represents different sample (image taken from [14]).

machine-learning in healthcare [9] rather than being more general machine-learning algorithms like RF or SVM, improvement in solving such tasks is exhibited.

Applications of machine learning in healthcare still have challenges to overcome, with the proposed Artificial Intelligence Act of the EU pushing for the explainability of AI for use in high-risk environments such as healthcare.

It is observed that the methodologies of both papers rely on learning the modality of healthy (non-cancerous) cells and that is indicative of a problem that this specific task faces and that is a lack of data for models to be trained on.

Paper [7] attempts to overcome the data scarcity by combining ALL backgrounds with existing AML backgrounds but to no avail. In [14], a similar principle is recognized by mixing control and FCM samples.

Novel algorithms that are based on outlier detection, with outliers, in this case, being blast cells, may improve the overall results, but due to the heterogeneity of AML data across samples, it is dubious whether the models will be clinically viable.

However, in the two papers, there is a difference in the number of samples, which is indicative of improvements in the data collection of AML-MRD FCM samples. With the further collection of AML-MRD FCM data, it is reasonable to expect models trained on the data to achieve better results.

The comparison suggests that further development of machine-learning algorithms combined with more standardized AML-MRD FCM data might be the stepstone towards the use of the models in clinical AML-MRD assessment.

References

[1] Ghea Apriliana, Titin Siswantining, Devvi Sarwinda, and Alhadi Bustamam. Analysis of data mining for classification of obstructive sleep apnea in chronic obstructive pulmonary disease patients. volume 2242, page 030023, 06 2020.

[2] D Dalbokova, M Krzyzanowski, and S Lloyd. *Children's health and the environment in Europe: a baseline assessment*. World Health Organization. Regional Office for Europe, 2007.

[3] Eric Guérin, Orhun Aydin, and Ali Mahdavi-Amiri. *Artificial Intelligence*, pages 357–385. 11 2019.

[4] Leonore Herzenberg, James Tung, Wayne Moore, Leonard Herzenberg, and David Parks. Interpreting flow cytometry data: a guide for the perplexed. *Nature immunology*, 7:681–5, 08 2006.

[5] Michael Heuser, Sylvie Freeman, Gert Ossenkoppele, Francesco Buccisano, Christopher Hourigan, Lok Ngai, Jesse Tettero, Costa Bachas, Constance Baer, Marie-Christine Béné, Veit Bücklein, Anna Czyż,

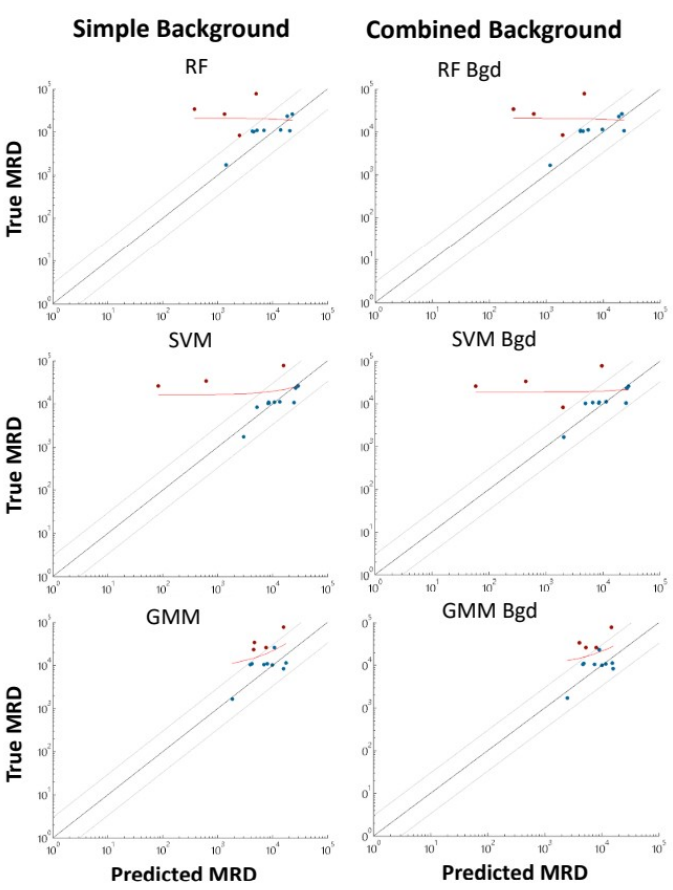

Fig. 9. Automatic MRD assessment of all models in [7] visualized (image taken from [7]).

Barbara Denys, Richard Dillon, Michaela Feuring, Monica Guzman, Torsten Haferlach, Lina Han, Julia Herzig, and Jacqueline Cloos. 2021 update measurable residual disease in acute myeloid leukemia: European leukemianet working party consensus document. *Blood*, 138, 11 2021.

[6] Aaron Kruse, Nour Abdel-Azim, Hye Na Kim, Yongsheng Ruan, Valerie Phan, Heather Ogana, William Wang, Rachel Lee, Eun Ji Gang, Sajad Khazal, and Yong-Mi Kim. Minimal residual disease detection in acute lymphoblastic leukemia. *International Journal of Molecular Sciences*, 21(3), 2020.

[7] Roxane Licandro., Michael Reiter., Markus Diem., Michael Dworzak., Angela Schumich., and Martin Kampel. Application of machine learning for automatic mrd assessment in paediatric acute myeloid leukaemia. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods - ICPRAM,*, pages 401–408. INSTICC, SciTePress, 2018.

[8] Holden Maecker, J. Mccoy, and Robert Nussenblatt. Standardizing immunophenotyping for the human immunology project. *Nature reviews. Immunology*, 12:191–200, 05 2012.

[9] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[10] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[11] Charles Schiffer and Richard Stone. Morphologic classification and clinical and laboratory correlates. In *Holland-Frei Cancer Medicine. 6th edition*, Cancer Medicine 6. BC Decker, 2003.

[12] Janet Staats, Anagha Divekar, J. McCoy, and Holden Maecker. *Guidelines for Gating Flow Cytometry Data for Immunological Assays*, volume 2032, pages 81–104. 09 2019.

[13] Tomasz Szczepanski, Alberto Orfao, Vincent Velden, Miguel Jimenez, and J.J.M. van Dongen. Minimal residual disease in leukaemia patients. *The lancet oncology*, 2:409–17, 08 2001.

[14] Lisa Weijler, Florian Kowarsch, Matthias Wödlinger, Michael Reiter, Margarita Maurer-Granofszky, Angela Schumich, and Michael N. Dworzak. Umap based anomaly detection for minimal residual disease quantification within acute myeloid leukemia. *Cancers*, 14(4,), Article 898., 2022.