

Automated Deep Learning-Based Diagnosis and Molecular Characterization of Acute Myeloid Leukemia using Flow Cytometry

Joshua E. Lewis, MD PhD¹, Lee A.D. Cooper, PhD², David L. Jaye, MD³, Olga Pozdnyakova, MD PhD^{1*}

¹ Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA

² Department of Pathology, Northwestern University, Chicago, IL, USA

³ Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, USA

* Corresponding author:

Olga Pozdnyakova, MD PhD

Department of Pathology, Brigham and Women's Hospital

75 Francis Street, Boston, MA 02215, USA

opozdnyakova@bwh.harvard.edu; 617-732-5500

COMPETING INTERESTS STATEMENT

L.A.D.C. received honoraria from Roche Tissue Diagnostics and serves on an advisory committee for the NVIDIA MONAI project. O.P. is a consultant for Scpio and serves as a speaker for Sysmex.

ABSTRACT

Current flow cytometric analysis of blood and bone marrow samples for diagnosis of acute myeloid leukemia (AML) relies heavily on manual intervention in both the processing and analysis steps, introducing significant subjectivity into resulting diagnoses and necessitating highly trained personnel. Furthermore, concurrent molecular characterization via cytogenetics and targeted sequencing can take multiple days, delaying patient diagnosis and treatment. Attention-based multi-instance learning models (ABMILMs) are deep learning models which make accurate predictions and generate interpretable insights regarding the classification of a sample from individual events/cells; nonetheless, these models have yet to be applied to flow cytometry data. In this study, we developed a computational pipeline using ABMILMs for the automated diagnosis of AML cases based exclusively on flow cytometric data. Analysis of 1,820 flow cytometry samples shows that this pipeline provides accurate diagnoses of acute leukemia [AUROC 0.961] and accurately differentiates AML *versus* B- and T-lymphoblastic leukemia [AUROC 0.965]. Models for prediction of 9 cytogenetic aberrancies and 32 pathogenic variants in AML provide accurate predictions, particularly for $t(15;17)(PML::RARA)$ [AUROC 0.929], $t(8;21)(RUNX1::RUNX1T1)$ [AUROC 0.814], and *NPM1* variants [AUROC 0.807]. Finally, we demonstrate how these models generate interpretable insights into which individual flow cytometric events and markers deliver optimal diagnostic utility, providing hematopathologists with a data visualization tool for improved data interpretation, as well as novel biological associations between flow cytometric marker expression and cytogenetic/molecular variants in AML. Our

study is the first to illustrate the feasibility of using deep learning-based analysis of flow cytometric data for automated AML diagnosis and molecular characterization.

INTRODUCTION

The diagnosis of acute myeloid leukemia (AML) requires integration of myeloblast quantification and phenotypic characterization via flow cytometry and microscopic assessment, as well as genomic characterization via cytogenetics and targeted sequencing panels¹. Flow cytometry allows for rapid measurement of physical properties and cell surface marker expression of hundreds of thousands of hematopoietic cells from blood and bone marrow aspirate specimens within a few hours². This permits detection and characterization of aberrant myeloblast populations suggestive of AML on a time frame that provides clinicians the ability to expeditiously alter patient management. On the other hand, molecular characterization for detecting the presence of cytogenetic aberrancies and/or pathogenic variants in myeloblast populations often takes multiple days. This delay particularly complicates management of AML patients with certain genetic abnormalities requiring targeted therapies and/or changes in standard induction chemotherapy regimens^{1,3}.

Current flow cytometric analysis relies heavily on manual intervention in both data processing, including compensation and gating, and data interpretation; this limits analysis of flow cytometry data to highly trained laboratory technologists and hematopathologists. Having an automated approach for analysis of flow cytometry data which does not rely on manual intervention would expand its use to a wider group of laboratorians and diagnosticians, similar to what automated pre-classification of white

blood cell differentials has done for peripheral blood smears⁴. Additionally, manual data interpretation by hematopathologists introduces an element of subjectivity which could impact patient diagnosis and treatment; automated interpretation of data would eliminate such subjectivity. Finally, such automated software could be used as a diagnostic tool for improved data visualization and integration for hematopathologists, especially for complex cases.

Machine learning models have been extensively developed for applications in hematopathology, including for analysis of flow cytometry data⁴⁻⁶. Nonetheless, newer deep learning-based models, which have demonstrated improved predictive performance and interpretability compared to conventional machine learning techniques, have mainly been applied to morphologic assessment of blood and bone marrow aspirate samples rather than flow cytometric analysis⁵. A handful of studies have recently applied deep learning approaches to flow and mass cytometry data, including using convolutional neural networks (CNNs) on data histograms for detection of classic Hodgkin lymphoma⁷ and on CyTOF data for diagnosis of latent CMV infection⁸, combining self-organizing maps and CNNs for classification of mature B-cell neoplasms⁹, and using deep neural networks for minimal/measurable residual disease (MRD) analysis in acute lymphoblastic leukemia (ALL)¹⁰ and chronic lymphocytic leukemia¹¹. However, attention-based multi-instance learning models (ABMILMs), a deep learning approach used to identify individual sample components which provide the most diagnostic value for individual cases (i.e., which flow cytometry events and markers are most important for the diagnosis of AML for a particular patient), have yet to

be applied to flow cytometry data despite significant success in histopathology applications^{12,13}.

In this study, we utilize ABMILMs for automated diagnosis and molecular characterization of AML cases from flow cytometry performed on blood and bone marrow samples. We developed a computational pipeline consisting of a series of deep learning models for predicting the presence or absence of acute leukemia, differentiating between AML and ALL cases, and finally predicting the presence or absence of 9 cytogenetic aberrancies and 32 pathogenic variants in AML cases. Pipeline implementation on 1,820 flow cytometry samples shows that these models provide strong predictive performance overall, as well as for challenging cases. Additionally, analysis of model-outputted attention values for individual flow cytometry events and their correlations with cell surface marker expression identifies which markers provide optimal utility for diagnosis and molecular characterization of AML cases. Finally, we give a representative example of a model-generated flow cytometry report providing model predictions and analysis tools for hematopathologists to improve flow cytometry data interpretation. Together, our results demonstrate the feasibility of using deep learning to make automated diagnoses of AML and provide insight into clinically actionable molecular variants based solely on flow cytometry data.

MATERIALS AND METHODS

Flow cytometry samples

Flow cytometry was performed on blood and bone marrow specimens in the hematology laboratory at Brigham and Women's Hospital from 2019-2022 for routine patient care. Two-tube 10-color leukemia panels were run using the BD Biosciences (Franklin Lakes, NJ, USA) FACSCanto™ II Flow Cytometry System (Becton Dickinson) (**Supplementary Table 1**). 1,820 flow cytometry cases were analyzed, 732 with definitive absence of acute leukemia, and 736 with definitive presence of acute leukemia (**Supplementary Table 2**). Among acute leukemia cases, 568 were diagnosed as acute myeloid leukemia (AML), and 168 were diagnosed as B- or T-lymphoblastic leukemia (ALL). The dataset was composed of initial diagnoses as well as follow-up cases, including pre- and post-treatment. A 20% blast count threshold based on flow cytometric analysis was used for all acute leukemia cases regardless of the presence of AML-defining translocations as well as for follow-up cases; positive AML cases with defining translocations or follow-up leukemia cases with less than 20% blasts were excluded. Non-leukemic cases consisted of those with less than 20% blasts and no prior history of acute leukemia.

Cytogenetic and molecular genetic testing

Among 568 AML cases, 478 had concurrent cytogenetics available from karyotyping and/or fluorescence *in situ* hybridization analysis, and 476 had molecular genetic information available from Rapid Heme Panel testing at Brigham and Women's

Hospital¹⁴. **Supplementary Tables 3 and 4** provide the number of AML cases with different cytogenetic aberrancies and pathogenic variants, respectively. Cytogenetic aberrancies and pathogenic variants with a dataset prevalence of less than 2% were excluded from analysis due to lack of adequate sample size.

Flow cytometry data preprocessing

FCS files outputted by BD FACSCanto™ II flow cytometers were loaded into Python scripts using the FlowKit package¹⁵ and normalized to marker mean values of 0 and standard deviation values of 1. No further preprocessing steps, including compensation, gating, or doublet exclusion, were performed.

Encoder network development and pre-training

Two encoder neural networks were developed to analyze flow cytometry data from each of two tubes. These networks were designed as multilayer perceptrons (MLPs) with embedding modules based on periodic activation functions, which have been shown to perform similarly on tabular data to transformer models with significantly lower computational cost¹⁶. Specifically, MLP models with embedding function ReLU ◦ Linear ◦ Periodic (MLP-PLR) were used, with the Periodic function being:

$$f_i(x) = \text{Periodic}(x) = \text{concat}[\sin(v), \cos(v)], v = [2\pi c_1 x, \dots, 2\pi c_k x] \quad (1)$$

where c_i are trainable parameters initialized from $N(0, \sigma)$, and both σ and k are hyperparameters.

Encoder networks were pre-trained on all events from the respective tube of all 1,820 flow cytometry cases using Self-Supervised Contrastive Learning using Random Feature Corruption (SCARF)¹⁷. During pre-training, encoder hyperparameter optimization was performed using a random 1% subset of pre-training data for 100 iterations; **Supplementary Table 5** provides the hyperparameter search space and optimal values. Encoder networks with optimal hyperparameter values were then pre-trained on 100% of pre-training data.

Predictive model development

Models for the prediction of presence/absence of acute leukemia, AML vs ALL, and presence/absence of cytogenetic aberrancies and pathogenic variants were designed as attention-based multi-instance learning models (ABMILMs) with a gated attention mechanism, as described by Ilse et al¹². For each application, separate ABMILMs were developed for analysis of flow cytometry data from each tube; the outputs of each ABMILM were concatenated and inputted into a final MLP neural network to provide a class-level prediction. Hyperparameter optimization was performed for 50 iterations; **Supplementary Table 6** provides the hyperparameter search space. A negative log-likelihood loss function was used with class weights relative to the proportion of samples in each class. 5-fold cross validation of randomly-split training and validation datasets was performed for performance assessment of each model.

Due to memory constraints, a maximum threshold on the number of events per tube of 150,000 was implemented. For tubes with more than 150,000 events (impacting

151/1820 (8.3%) of tube 1 datasets and 152/1820 (8.4%) of tube 2 datasets), a random subset of 150,000 events was chosen.

Analysis of model predictions

Machine learning models output the percentage probability of a particular sample being within each respective class. Model performance on validation datasets was assessed by calculating the area under the receiver operating characteristic curve (AUROC), as well as the accuracy, sensitivity, and specificity using a probability threshold which maximizes the Youden's J statistic of the model.

Analysis of model attention values and predictive power scores

Attention values from the two tube ABMILMs were analyzed to identify the relative importance of each flow cytometry event towards a model's prediction for a particular sample. Attention values from each tube were normalized to range from (0,1) to improve visualization. Visualization of attention scores was achieved by coloring flow cytometry events based on normalized scores.

Predictive power scores (PPS), which are non-parametric scores detecting non-linear relationships between two variables, were calculated to assess the relationship between normalized attention weights and flow cytometry marker expression for individual cases¹⁸. PPS values rely on quantitative flow cytometric marker expression levels rather than arbitrary qualitative "positive/negative" gating of events. For a given model and flow cytometry marker, both the percentage of cases with non-zero PPS

values, as well as the mean PPS value among non-zero cases, were used to assess the relative diagnostic importance of the marker to that model.

Software and hardware

Python 3.8.10 and PyTorch 2.0.1+cu117 were used to develop and train neural network models¹⁹. All experiments were run on a Linux server with Intel Xeon Gold 6132 processor, 126 GB RAM, and 4 Tesla V100-PCIE-32GB GPUs.

RESULTS

Design of machine learning models for sample-level diagnosis and molecular characterization

We developed attention-based multi-instance learning models (ABMILMs) for each application of interest: (1) the prediction of presence *versus* absence of acute leukemia, (2) the differentiation of acute myeloid leukemia (AML) *versus* acute lymphoblastic leukemia (ALL), and (3) the prediction of presence *versus* absence of individual cytogenetic aberrancies and pathogenic variants (**Figure 1A**). These models are designed to input two-tube flow cytometry data from an individual sample, subsequently outputting the percentage probability of a sample being represented by each class (e.g., percentage probability that a case does *versus* does not represent acute leukemia). Data from each tube is first analyzed by an encoder, a neural network that identifies new features from the flow cytometry data for improved predictive capabilities. These new features are then analyzed by an attention module, which

decides which individual flow cytometry events are important *versus* not important for the diagnosis, and accordingly assigns each event a weight based on this perceived importance. Using these weights along with the encoder-outputted features, pooled features from each tube are obtained. The pooled features from both tubes are concatenated and finally analyzed by an integrating neural network to provide the final prediction for the entire sample.

A stepwise workflow for the diagnosis and molecular characterization of individual cases was developed, first identifying which cases are positive for acute leukemia, which of those represent AML cases, and finally which cytogenetic aberrancies and pathogenic variants each AML case has (**Figure 1B**). **Figure 1C** shows the specific cytogenetic aberrancies and pathogenic variants for which predictive models were developed. Only variants with dataset prevalence of at least 2% were chosen to ensure adequate sample size.

Machine learning model for the diagnosis of acute leukemia

A machine learning model was first developed to distinguish between non-leukemic cases and cases with acute leukemia. This model demonstrated strong predictive performance overall (AUROC 0.961 ± 0.011 , accuracy 0.903 ± 0.013), as well as individually on leukemic (sensitivity 0.881 ± 0.020 at Youden's J) and non-leukemic (specificity 0.925 ± 0.027 at Youden's J) cases (**Figure 2A**). **Figure 2A** also shows that by adjusting the predictive probability threshold along the ROC curve, model sensitivity can be significantly increased with a relative tradeoff in specificity; this could allow the model to be used in more of a triaging/screening role²⁰. Analysis of flow cytometry event

attention values for individual leukemia cases shows that, analogous to manual diagnosis by a hematopathologist, the model learns to pay most attention to blast cells for diagnosis, with mixed attention to monocytic cells, and less attention to mature lymphocytes and granulocytes (**Figure 2B**).

To assess the overall importance of each marker towards model predictions, predictive power scores (PPS) were calculated to quantify the relationship between attention values and flow cytometry marker expression for individual cases; by fitting a non-parametric model to the data which can capture non-linear relationships, this approach provides a numerical score analogous to a correlation coefficient which represents the strength of the relationship between attention values and marker expression (**Figure 2C**). PPS values across cases were averaged by assessing both the percentage of non-zero PPS values per flow cytometry marker, as well as the mean non-zero PPS value for each marker (**Figure 2D**).

PPS value analysis for this model shows that the machine learning model utilized forward scatter (FSC) and side scatter (SSC) information in a larger percentage of cases than any other flow cytometry markers. Among cell surface markers, CD15 (sialyl Lewis x), a marker for human myeloid cells, was associated with both the highest percentage of non-zero PPS values, as well as the highest magnitude PPS values²¹. CD19, critical for the diagnosis of B-ALL and AML with t(8;21)(*RUNX1*::*RUNX1T1*), also showed strong predictive performance²². Interestingly, flow cytometry markers often used for manual blast identification in AML, including CD45, CD34, and HLA-DR, were among the least important cell surface markers used by the model²³. Finally, while CD3 was among the least used markers by the model, it displayed the second highest mean

value among non-zero cases; this likely represents the use of CD3 in identification of the limited number of T-ALL cases in the dataset.

Machine learning model for classification of AML *versus* ALL cases

Performance of a machine learning model for classification of AML *versus* ALL (including B-ALL and T-ALL) showed equally strong performance (AUROC 0.965 ± 0.015 , accuracy 0.922 ± 0.025) as initial diagnosis of acute leukemia (**Figure 3A**). While events representing blast cells still had comparably high attention values for most cases, other cell types appeared to provide similar or larger importance to model predictions, including mature granulocytes with high SSC values, as well as a subset of mature lymphocytes (**Figure 3B**). This suggests that the immunophenotypic profile of cell types other than blast cells may provide an underexplored role in the diagnosis of acute leukemia cases.

Despite the model's strong accuracy, misclassifications were still made for cases with lower model confidence. An analysis was performed to determine at what model-outputted confidence level the machine learning model can make highly accurate predictions for classification of AML *versus* ALL (**Figure 3C**). For cases with model confidence of at least 95% (560/736, 76.1% of total cases), the model's cumulative accuracy across all validation sets increased from 91.4% to 97.9%. Furthermore, at a confidence threshold of 99.9% (254/736, 34.5% of total cases), only one misclassification was made, with a model accuracy of 99.6%. Thus, while employing such a high confidence threshold may limit model generalizability, it may be favored to make highly accurate predictions in particularly challenging cases.

Analysis of model PPS values shows that FSC values carried small but non-zero importance in most cases, representative of the fact that myeloblasts and lymphoblasts often exhibit subtle differences in cell size²⁴ (**Figure 3D**); interestingly, side scatter showed lower relative importance despite differences in granularity between myeloblasts and lymphoblasts. CD123 (IL3 receptor), which shows expression in most AML and B-ALL cases but not T-ALL cases, as well as CD10, which is commonly expressed in ALL but not AML, yielded relatively high importance values^{25,26}. Finally, CD15 as well as T-cell markers (CD3, CD5, and CD7), demonstrated high importance in a small subset of AML and T-ALL cases.

Prediction of cytogenetic aberrancies and pathogenic variants in AML cases

Individual machine learning models predicting the presence or absence of 9 cytogenetic aberrancies and 32 pathogenic variants among AML cases were developed (**Figure 4A**). The model for prediction of $t(15;17)(PML::RARA)$, which has a distinct immunophenotypic profile with negative CD34 in the hypergranular variant, negative HLA-DR, and positive CD117 expression, showed the strongest performance among all variants tested (AUROC 0.929 ± 0.032 , accuracy 0.885 ± 0.071)²⁷; the model performed well for both positive (sensitivity 0.925 ± 0.061) and negative (specificity 0.882 ± 0.083) cases (**Figure 4B**). Other models with strong performance are associated with variants with known immunophenotypic signatures, including $t(8;21)(RUNX1::RUNX1T1)$ (AUROC 0.814 ± 0.050) which is associated with CD19 and CD56 expression²⁸, and *NPM1* variants (AUROC 0.807 ± 0.020) which are associated with CD19 and CD4 expression as well as monocytic markers²⁹. 9/41 (22.0%) of variant models

demonstrated an AUROC > 0.7, and 32/41 (78.0%) of variant models demonstrated an AUROC > 0.6.

PPS score analysis across all predictive models shows that flow cytometry markers can be placed into three groups based on relative importance for model predictions (**Figure 4C**). A select few markers, including forward scatter (FSC-A/H), and CD33, demonstrated strong predictive utility in almost all models. While CD33 expression has been implicated in *NPM1*- and *FLT3*-mutated AML^{30,31}, these results appear to associate CD33 with many other cytogenetic and mutational variants, with implications for targeted therapy³². Many other markers, including B-cell (CD19 and CD20) and T-cell (CD3, CD5, and CD7) markers, exhibit very little predictive value for most models. Finally, some markers show strong importance for individual models but otherwise weak importance. For example, monocytic markers CD14 and CD64 had high importance scores for prediction of *inv(16)/t(16;16)(CBFB::MYH11)*, which often exhibits monocytic differentiation³³. Additionally, CD117 had a relatively high importance score for prediction of *IDH2* mutations; prior studies have associated *IDH2* R172 mutations with high CD117 expression³⁴, and *IDH2*-targeting therapies have been shown to decrease CD117 expression in AML³⁵.

Visualization and analysis of model outputs for individual patient cases

Figure 5 provides an example of a model-generated report for a representative case of acute promyelocytic leukemia with confirmed presence of *t(15;17)(PML::RARA)* by cytogenetics/FISH, but which had significant diagnostic uncertainty during manual interpretation. The model correctly identified the presence of leukemia (predicted

probability $p = 98.6\%$) and AML rather than ALL ($p = 100.0\%$). The model also correctly predicted the presence of the $t(15;17)(PML::RARA)$ translocation ($p = 75.4\%$), as well as a *FLT3* pathogenic variant (albeit with low predictive confidence; $p = 50.4\%$).

To illustrate the utility of this computational pipeline for identification of cytogenetic aberrancies/pathogenic variants in challenging cases, flow cytometry marker expression and corresponding event attention values were further analyzed. Significant model attention was placed on the population of CD45-dim immature cells that were characterized as SSC(increased)/CD45(dim)/CD34(variable)/HLA-DR-/CD117+/CD13(variable)/CD33+/CD14-/CD64+/CD3-/CD19-. Although most of this immunophenotype is consistent with hypergranular APML, the variable CD34 positivity is quite infrequent in hypergranular APML (only approximately 16% of cases)²³ and likely led to the diagnostic uncertainty about the presence of $t(15;17)(PML::RARA)$.

PPS value analysis showed that many flow cytometry markers not commonly utilized in assessment of APML, including CD123, CD10, and CD38, were associated with high feature importance for diagnosis of this case. **Figure 5** shows that the population of promyelocytes demonstrates high expression of these markers, and the model's attention values are highly correlated with marker expression. Additionally, conventionally used markers for APML including CD45, CD34, and HLA-DR surprisingly carried small or no importance in the case diagnosis. Thus, the machine learning model was able to correctly predict the presence of $t(15;17)(PML::RARA)$ by developing a previously-uncharacterized association between this translocation and increased expression of CD123, CD10, and CD38, rather than relying on markers more commonly used during manual analysis.

DISCUSSION

While statistical and machine learning approaches have been extensively applied to analysis of flow and mass cytometry data⁶, these prior approaches carry significant shortcomings. Dimensionality reduction methods for data visualization (including PCA, tSNE, and UMAP), as well as unsupervised methods for cell population clustering (including flowSOM³⁶, FLOCK³⁷, and PhenoGraph³⁸) are capable of identifying clinically-relevant cell populations, but are not primarily designed for classification applications such as predicting the presence or absence of a disease state or cytogenetic/molecular aberrancies. Past supervised learning methods including linear discriminant analysis³⁹ and support vector machines⁴⁰ often underperform current state-of-the-art deep learning methods in classification tasks. Most deep learning approaches to flow and mass cytometry data, including CellCNN⁴¹ and Deep CNN⁸, have utilized convolutional neural networks, a machine learning architecture that is well suited for image classification tasks, but is not designed for analysis of tabular data such as flow cytometric data⁴². Deep learning approaches specifically designed for tabular data have yet to be applied for the diagnosis and molecular characterization of hematolymphoid malignancies from flow cytometry data. Additionally, attention-based multi-instance learning, a deep learning approach which learns to identify (or pay “attention” to) particular data components (e.g., individual flow cytometric events within a sample) which carry the most diagnostic value in sample classification, has yet to be applied to analysis of flow cytometry data.

This study is, to our knowledge, the first to apply attention-based multi-instance learning to flow cytometry data for automated hematopathology diagnostics (**Figure 1**).

We demonstrated that this approach provides highly accurate predictions on the presence/absence of acute leukemia (**Figure 2**), as well as for discriminating between AML and ALL (**Figure 3**). We additionally show the feasibility of using these models trained solely on flow cytometry data to provide insight on the presence or absence of clinically relevant cytogenetic aberrancies and pathogenic variants in AML-positive samples (**Figure 4**). Finally, we provide a representative example of how model outputs can be visualized and incorporated into a hematopathologist's workflow for improved patient diagnosis, and for identification of novel associations between flow cytometry marker expression and particular molecular aberrancies (**Figure 5**).

This automated approach to flow cytometry-based diagnosis and molecular characterization of AML yields several potential use cases. For clinicians, the computational pipeline provides accurate and rapid predictions (within minutes after flow cytometry data collection) on disease status, including insight into the molecular subtype, of their patients, allowing for faster initiation of optimal treatment protocols. Additionally, model output visualizations can be tailored for clinicians to provide succinct yet insightful information into patients' immunophenotype with therapeutic implications. For hematopathologists, the pipeline can be used as a triaging tool to rule out cases with very low predicted probabilities of acute leukemia, as well as a diagnosis assistance tool providing unbiased predictions of disease state with associated data visualizations for data analysis. For hematology laboratories as well as low-resource settings, the pipeline provides a software- and personnel-agnostic approach to flow cytometry-based diagnosis without need for manual intervention. Finally, for biomedical researchers, this tool provides information about novel associations between myeloblast

cell surface marker expression and AML molecular subtypes which could be further explored as therapeutic targets.

Future efforts will continue to improve the generalizability of this approach toward particular technical considerations, use cases, and application areas. Currently, our computational pipeline does not discriminate between new *versus* recurrent cases of AML, and our model for predicting the presence or absence of acute leukemia requires a 20% blast count for positive cases. Modifications to model architecture including utilizing the patient's previously characterized immunophenotype could be made to improve performance on recurrent cases, as well as for minimal/measurable residual disease assessment of minute cell populations. Additionally, these models were trained on a set flow cytometry panel with no changes in markers, antibodies, or fluorochromes. Incorporation of additional model elements which learn associations between flow cytometry markers using cross-institutional data could be a potential approach to allow for variance in flow cytometry panel construction, the number of colors per tube, and the use of multiple markers per channel. Lastly, addition of CNNs into the model architecture would allow for integration of flow cytometry data with cell morphologic data from blood and bone marrow aspirate smears, which could provide improved diagnostic accuracy.

Overall, this study represents a significant advancement in digital hematopathology by demonstrating the feasibility of using deep learning toward automated diagnosis and molecular characterization of AML cases solely using flow cytometry data. This automated approach provides many advantages and potential use cases compared to manual flow cytometric assessment for clinicians,

hematopathologists, and the hematology laboratory. Continued improvements and integration with other data modalities including cell morphologic assessment will further enhance model performance and generalizability, ultimately moving these automated approaches closer to clinical use.

FIGURES

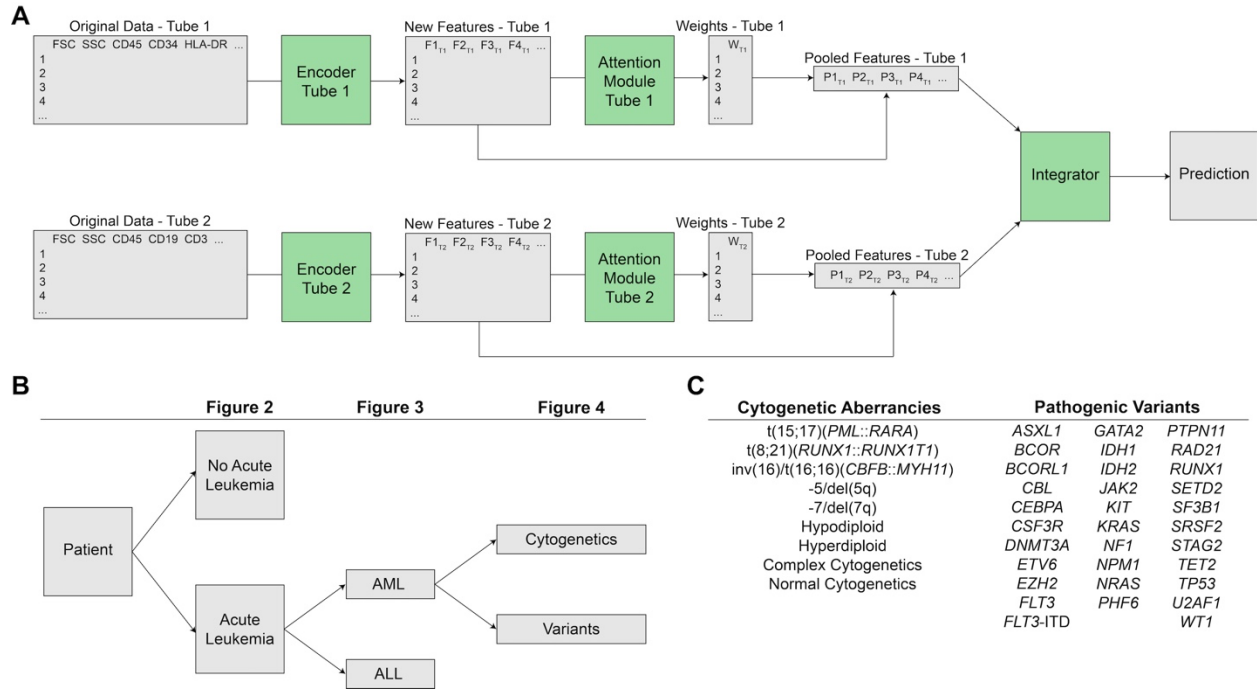


Figure 1. Workflow for automated machine learning-based diagnosis and molecular characterization of acute myeloid leukemia from flow cytometry data.

(A) Schematic of attention-based multi-instance learning models. Flow cytometry data extracted from FCS files from two-tube leukemia panels are provided as input to encoder neural networks which extract new features from the data in each respective tube. These new features are analyzed by attention modules which designate a weight for each flow cytometry event based on the perceived importance of the event in overall case diagnosis by the model. These weights, along with the new features, are used to extract pooled features from each tube. Pooled features from the two tubes are combined and analyzed by a final neural network which provides a sample-level prediction for the sample. **(B)** Workflow for sample diagnosis and molecular characterization. The first model predicts the presence or absence of acute leukemia.

Acute leukemia cases are then analyzed by the next model to differentiate between AML and ALL cases. Finally, AML cases are analyzed by a set of individual models that predict the presence or absence of cytogenetic aberrancies and pathogenic variants.

(C) List of cytogenetic aberrancies and pathogenic variants tested for in AML cases.

Only variants with dataset prevalence of at least 2% were chosen to ensure adequate sample size.

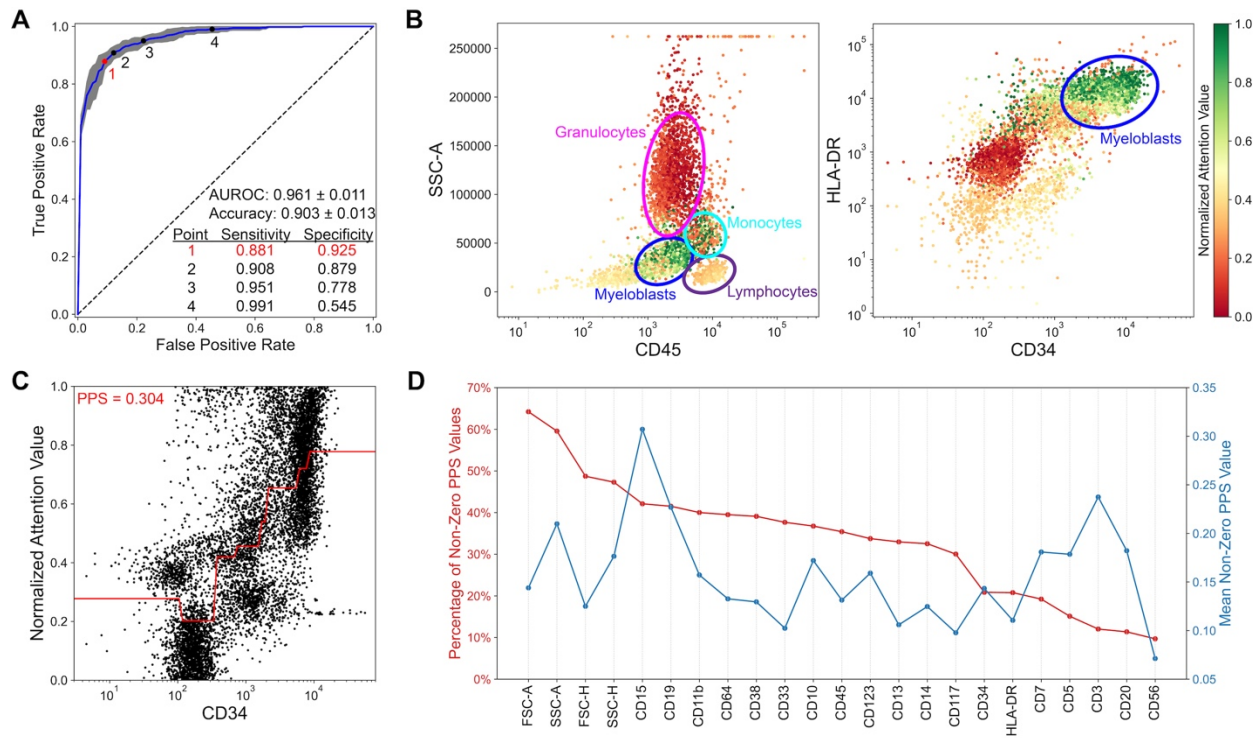


Figure 2. Machine learning model for the diagnosis of acute leukemia. (A)

Receiver operating characteristic (ROC) curve for the model predicting the presence or absence of acute leukemia from flow cytometry samples. Blue line represents the mean of five training/validation data splits, and gray outline represents the standard error. Model AUROC and accuracy are provided. Numbered points along the ROC curve indicate varying sensitivity and specificity values with different predictive probability threshold values. Point #1 (red) indicates the point along the ROC curve which maximizes Youden's J, whereas points #2-4 (black) indicate those with sensitivity values of 90%, 95%, and 99%, respectively. **(B)** Analysis of model attention values for a representative AML case, which was correctly identified by the model as being positive for acute leukemia. Plots of side scatter (SSC-A) *versus* CD45, and HLA-DR *versus* CD34, are provided. Individual events are colored based on the normalized attention value assigned by the model to each flow cytometry event. Cell populations of interest

are illustrated. **(C)** Diagram demonstrating an example calculation of predictive power scores (PPS) for quantifying the relationship between attention values and flow cytometry marker expression for an individual case. Red line represents the non-linear relationship identified between the two variables. The PPS score that quantifies this relationship on a scale from 0-1 is provided. **(D)** Diagram summarizing the PPS values for all flow cytometry markers in the prediction of the presence or absence of acute leukemia. Values show (red, left) the percentage of non-zero PPS values across all samples, and (blue, right) the mean non-zero PPS value across all samples. Markers are ordered based on descending value of the percentage of non-zero PPS values.

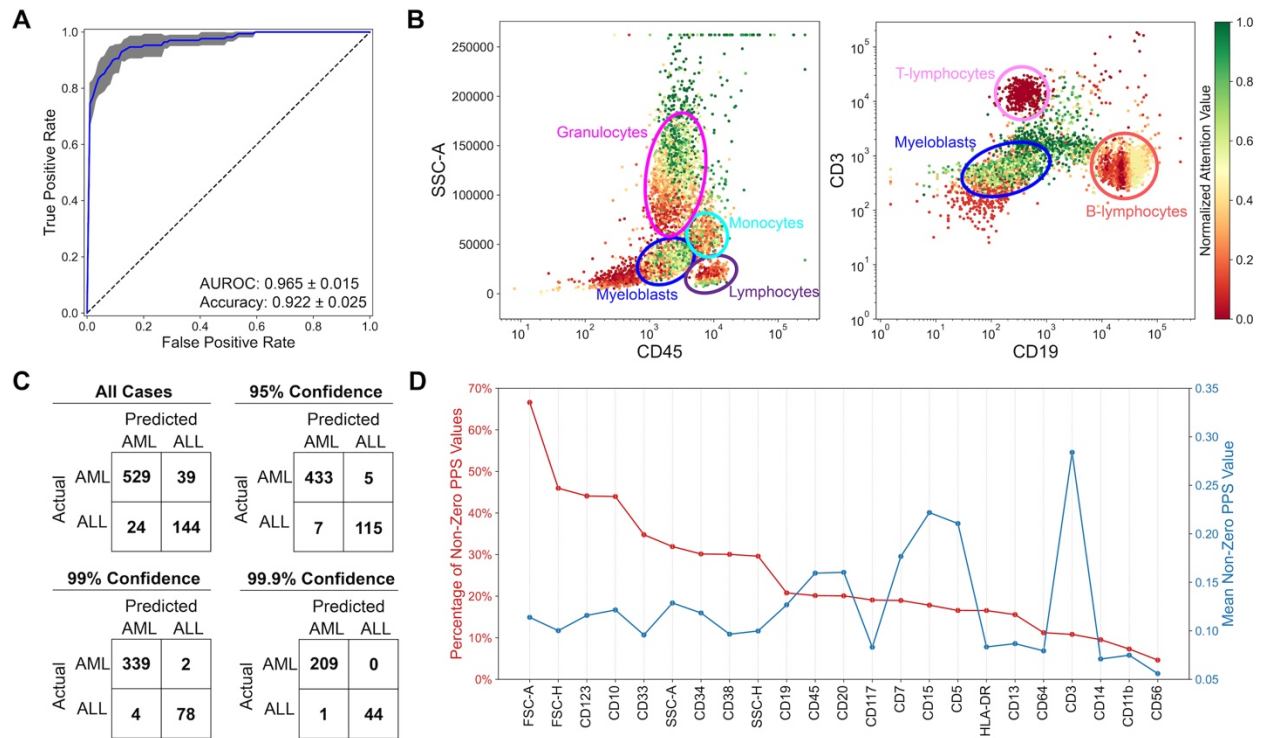


Figure 3. Machine learning model for the classification of AML versus ALL. (A)

Receiver operating characteristic (ROC) curve for the model predicting the presence of AML versus ALL from flow cytometry samples. Blue line represents the mean of five training/validation data splits, and gray outline represents the standard error. AUROC and accuracy values for the model are provided. **(B)** Analysis of model attention values for a representative AML case, which was correctly identified by the model as being positive for AML. Plots of side scatter (SSC-A) versus CD45, and CD3 versus CD19, are provided. Individual events are colored based on the normalized attention value assigned by the model to each flow cytometry event. Cell populations of interest are illustrated. **(C)** Confusion matrices showing the accuracy of model predictions at different model confidence thresholds. Cases where the predicted probability of a case being either an AML or ALL case is above the provided threshold are included. Rows represent the actual class and columns represent the predicted class. **(D)** Diagram

summarizing the PPS values for all flow cytometry markers in the prediction of the presence of AML *versus* ALL. Values show (red, left) the percentage of non-zero PPS values across all samples, and (blue, right) the mean non-zero PPS value across all samples. Markers are ordered based on descending value of the percentage of non-zero PPS values.

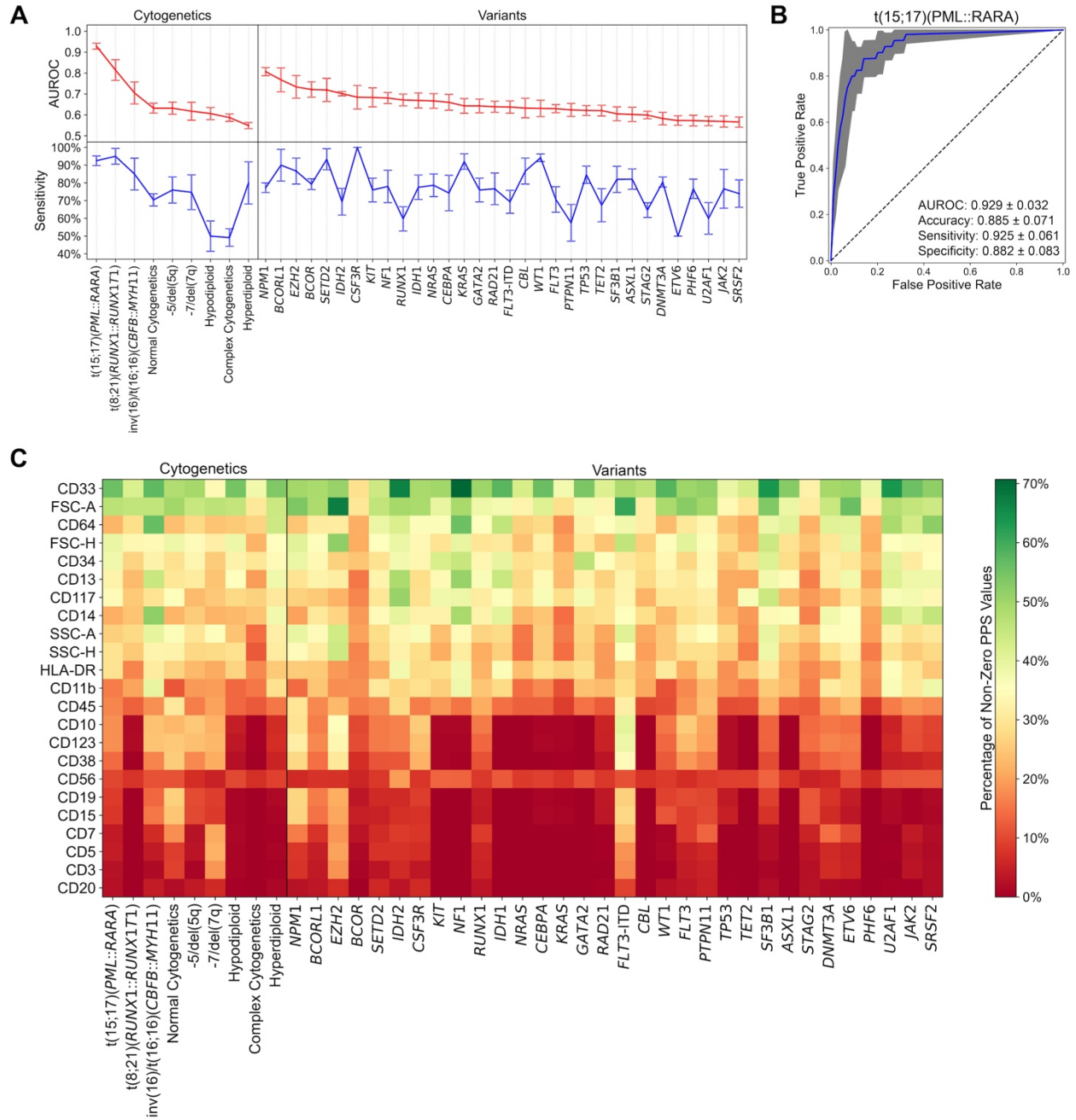


Figure 4. Machine learning models for predicting the presence or absence of cytogenetic aberrancies and pathogenic variants in AML cases. (A) Performance metrics (AUROC: red, top; sensitivity: blue, bottom) of individual machine learning models predicting the presence or absence of a particular cytogenetic aberrancy or pathogenic variant. Bar graph represents the mean \pm one standard error. Cytogenetic

aberrancies and pathogenic variants are ordered based on descending AUROC value.

(B) Receiver operating characteristic (ROC) curve for the model predicting the presence or absence of the $t(15;17)(PML::RARA)$ translocation from flow cytometry samples. Blue line represents the mean of five training/validation data splits, and gray outline represents the standard error. AUROC, accuracy, sensitivity, and specificity values for the model are provided. **(C)** Heatmap showing the percentage of non-zero PPS values representing the importance of each flow cytometry marker for all predictive models. Flow cytometry markers are ordered (top to bottom) based on decreasing mean value of the percentage of non-zero PPS values among all models.

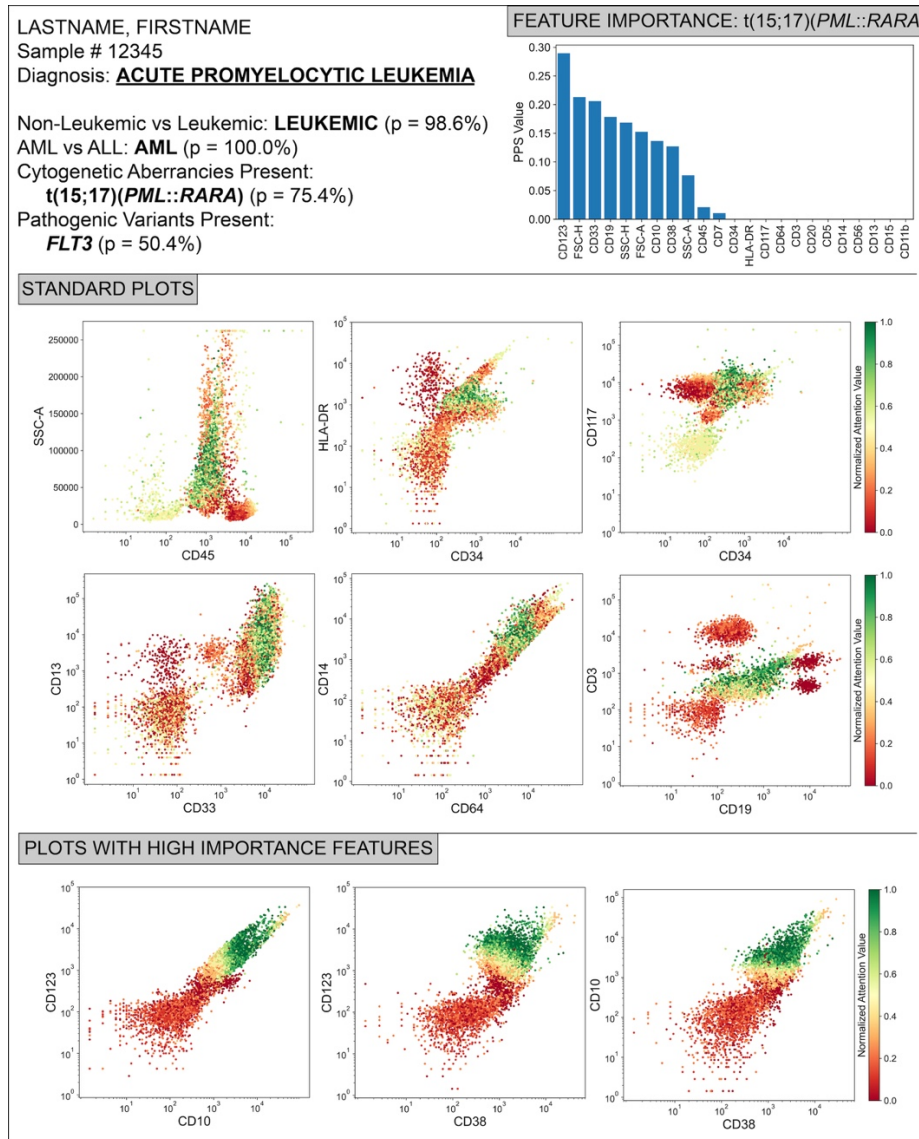


Figure 5. Example model-generated report of representative acute promyelocytic leukemia case. (Top-Left) The final diagnosis of acute promyelocytic leukemia, as well as model predictions and associated confidence values along the entire computational pipeline, are shown. **(Top-Right)** PPS values for flow cytometry markers in the prediction of t(15;17)(PML::RARA) for this case. **(Bottom)** Analysis of model attention values and expression of flow cytometry markers conventionally used in diagnosis of acute promyelocytic leukemia, as well as those identified as diagnostically important for this particular case.

Supplementary Table 1. Flow cytometry parameters and markers used in two-tube acute leukemia panel from this study.

Tube 1		Tube 2	
<i>Marker</i>	<i>Fluorophore</i>	<i>Marker</i>	<i>Fluorophore</i>
FSC-A	-	FSC-A	-
FSC-H	-	FSC-H	-
SSC-A	-	SSC-A	-
SSC-H	-	SSC-H	-
CD45	V500-C	CD45	V500-C
CD34	FITC	CD3	BV421
HLA-DR	PerCP-Cy5.5	CD5	PE-Cy7
CD11b	APC	CD7	BV605
CD13	PE-Cy7	CD10	APC-R700
CD14	APC-H7	CD15	FITC
CD33	PE	CD19	PE
CD56	BV421	CD20	APC-H7
CD64	APC-R700	CD38	PerCP-Cy5.5
CD117	BV605	CD123	APC

Supplementary Table 2. Number of patients in respective diagnostic categories from this study.

Category	# Cases	% Cases
All cases (used for encoder pre-training)	1820	
Cases used for model training	1468	
No acute leukemia	732	49.9%
Acute leukemia	736	50.1%
ALL	168	22.8%
AML	568	77.2%
Cytogenetic aberrancy data available	478	84.2%
Pathogenic variant data available	476	83.8%

Supplementary Table 3. Number of cases based on the presence of cytogenetic aberrancies.

Cytogenetic Aberrancy	# Cases	% Cases
Normal Karyotype	159	33.3%
Complex Karyotype	104	21.8%
-5/del(5q)	46	9.6%
-7/del(7q)	42	8.8%
t(15;17)(<i>PML::RARA</i>)	40	8.4%
Hypodiploid Karyotype	34	7.1%
t(8;21)(<i>RUNX1::RUNX1T1</i>)	22	4.6%
inv(16)/t(16;16)(<i>CBFB::MYH11</i>)	19	4.0%
Hyperdiploid Karyotype	15	3.1%

Supplementary Table 4. Number of cases based on presence of pathogenic variants.

Variant	# Cases	% Cases
<i>FLT3-ITD</i>	131	27.5
<i>NPM1</i>	123	25.8
<i>DNMT3A</i>	98	20.6
<i>TET2</i>	97	20.4
<i>RUNX1</i>	87	18.3
<i>FLT3</i>	75	15.8
<i>NRAS</i>	75	15.8
<i>IDH1</i>	72	15.1
<i>TP53</i>	65	13.7
<i>SRSF2</i>	61	12.8
<i>IDH2</i>	58	12.2
<i>ASXL1</i>	57	12.0
<i>WT1</i>	53	11.1
<i>PTPN11</i>	39	8.2
<i>BCOR</i>	39	8.2
<i>CEBPA</i>	31	6.5
<i>SF3B1</i>	28	5.9
<i>STAG2</i>	26	5.5
<i>KRAS</i>	26	5.5
<i>GATA2</i>	25	5.3
<i>KIT</i>	24	5.0
<i>NF1</i>	23	4.8
<i>U2AF1</i>	20	4.2
<i>JAK2</i>	19	4.0
<i>PHF6</i>	17	3.6
<i>EZH2</i>	15	3.2
<i>CBL</i>	13	2.7
<i>SETD2</i>	13	2.7
<i>RAD21</i>	12	2.5
<i>ETV6</i>	10	2.1
<i>BCORL1</i>	10	2.1
<i>CSF3R</i>	10	2.1

Supplementary Table 5. Hyperparameter values from encoder pre-training.

Hyperparameter	Search Range	Optimal Value – Tube 1	Optimal Value – Tube 2
k (periodic activation function, # features / 2)	Int(1,128)	50	102
σ (periodic activation function, $\max(c_i)$)	Float(0.01,100), log range	17.875532	0.344150
# neurons in linear layer of PLR embedding	$e^{\text{Int}(0,7)}$	e^4	e^6
# MLP blocks	Int(1,16)	1	7
# neurons in linear layer of MLP block	$e^{\text{Int}(0,10)}$	e^6	e^9
Dropout rate of MLP block	Float(0,0.5)	0.425532	0.038182
Learning Rate	Float($5e^{-5}$, $5e^{-3}$)	$1.824962e^{-4}$	$8.712371e^{-5}$
Weight decay	Float($1e^{-6}$, $1e^{-3}$)	$2.252650e^{-5}$	$4.066816e^{-5}$
Corruption rate	Float(0.2,0.8)	0.210043	0.201649

Supplementary Table 6. Hyperparameter values from multi-instance learning model training.

Hyperparameter	Search Range
# hidden layers, attention module, tube 1	Int(0,10)
# hidden layers, attention module, tube 2	Int(0,10)
# layers, integrator MLP	Int(1,16)
# neurons per layer, integrator MLP	$e^{\text{Int}(0,10)}$
Dropout rate of MLP block	Float(0,0.5)
Learning Rate	Float($5e^{-5}$, $5e^{-3}$)
Weight decay	Float($1e^{-6}$, $1e^{-3}$)

REFERENCES

- 1 Döhner, H., Wei, A. H., Appelbaum, F. R., Craddock, C., DiNardo, C. D., Dombret, H. *et al.* Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood, The Journal of the American Society of Hematology* **140**, 1345-1377 (2022).
- 2 Craig, F. E. & Foon, K. A. Flow cytometric immunophenotyping for hematologic neoplasms. *Blood* **111**, 3941-3967 (2008).
- 3 Sanz, M. A., Fenaux, P., Tallman, M. S., Estey, E. H., Löwenberg, B., Naoe, T. *et al.* Management of acute promyelocytic leukemia: updated recommendations from an expert panel of the European LeukemiaNet. *Blood, The Journal of the American Society of Hematology* **133**, 1630-1643 (2019).
- 4 Lewis, J. E. & Pozdnyakova, O. Digital assessment of peripheral blood and bone marrow aspirate smears. *International Journal of Laboratory Hematology* (2023).
- 5 Dehkharghanian, T., Mu, Y., Tizhoosh, H. R. & Campbell, C. J. Applied machine learning in hematopathology. *International Journal of Laboratory Hematology* (2023).
- 6 Hu, Z., Bhattacharya, S. & Butte, A. J. Application of machine learning for cytometry data. *Frontiers in immunology* **12**, 787574 (2022).
- 7 Simonson, P. D., Wu, Y., Wu, D., Fromm, J. R. & Lee, A. Y. De novo identification and visualization of important cell populations for classic Hodgkin lymphoma using flow cytometry and machine learning. *American Journal of Clinical Pathology* **156**, 1092-1102 (2021).

- 8 Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep learning model for cytometry data. *Proceedings of the National Academy of Sciences* **117**, 21373-21380 (2020).
- 9 Zhao, M., Mallesh, N., Höllein, A., Schabath, R., Haferlach, C., Haferlach, T. *et al.* Hematologist-level classification of mature B-cell neoplasm using deep learning on multiparameter flow cytometry data. *Cytometry Part A* **97**, 1073-1080 (2020).
- 10 Woedlinger, M., Reiter, M., Weijler, L., Maurer-Granofszky, M., Schumich, A., Sajaroff, E. O. *et al.* Automated identification of cell populations in flow cytometry data with transformers. *Computers in Biology and Medicine* **144**, 105314 (2022).
- 11 Salama, M. E., Otteson, G. E., Camp, J. J., Seheult, J. N., Jevremovic, D., Holmes III, D. R. *et al.* Artificial intelligence enhances diagnostic flow cytometry workflow in the detection of minimal residual disease of chronic lymphocytic leukemia. *Cancers* **14**, 2537 (2022).
- 12 Ilse, M., Tomczak, J. & Welling, M. in *International conference on machine learning*. 2127-2136 (PMLR).
- 13 Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M. & Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5**, 555-570 (2021).
- 14 Kluk, M. J., Lindsley, R. C., Aster, J. C., Lindeman, N. I., Szeto, D., Hall, D. *et al.* Validation and Implementation of a Custom Next-Generation Sequencing Clinical Assay for Hematologic Malignancies. *J Mol Diagn* **18**, 507-515 (2016).

- 15 White, S., Quinn, J., Enzor, J., Staats, J., Mosier, S. M., Almarode, J. *et al.* FlowKit: A Python Toolkit for Integrated Manual and Automated Cytometry Analysis Workflows. *Front Immunol* **12**, 768541 (2021).
- 16 Gorishniy, Y., Rubachev, I. & Babenko, A. On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems* **35**, 24991-25004 (2022).
- 17 Bahri, D., Jiang, H., Tay, Y. & Metzler, D. Scarf: Self-supervised contrastive learning using random feature corruption. *arXiv preprint arXiv:2106.15147* (2021).
- 18 Wetschoreck, F., Krabel, T. & Krishnamurthy, S. 8080labs/ppscore: zenodo release. *Zenodo: London, UK* (2020).
- 19 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).
- 20 Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine* **4**, 627 (2013).
- 21 Gadhoum, S. Z. & Sackstein, R. CD15 expression in human myeloid cell differentiation is regulated by sialidase activity. *Nature chemical biology* **4**, 751-757 (2008).
- 22 Francis, J., Dharmadhikari, A. V., Sait, S. N., Deeb, G., Wallace, P. K., Thompson, J. E. *et al.* CD19 expression in acute leukemia is not restricted to the

- cytogenetically aberrant populations. *Leukemia & lymphoma* **54**, 1517-1520 (2013).
- 23 Promsuwicha, O. & Auewarakul, C. U. Positive and negative predictive values of HLA-DR and CD34 in the diagnosis of acute promyelocytic leukemia and other types of acute myeloid leukemia with recurrent chromosomal translocations. *Asian Pacific Journal of Allergy and Immunology* **27**, 209 (2009).
- 24 Chiaretti, S., Zini, G. & Bassan, R. Diagnosis and subclassification of acute lymphoblastic leukemia. *Mediterranean journal of hematology and infectious diseases* **6** (2014).
- 25 Bras, A. E., de Haas, V., van Stigt, A., Jongen-Lavrencic, M., Beverloo, H. B., Te Marvelde, J. G. *et al.* CD123 expression levels in 846 acute leukemia patients based on standardized immunophenotyping. *Cytometry Part B: Clinical Cytometry* **96**, 134-142 (2019).
- 26 Bavikatty, N. R., Ross, C. W., Finn, W. G., Schnitzer, B. & Singleton, T. P. Anti-CD10 immunoperoxidase staining of paraffin-embedded acute leukemias: comparison with flow cytometric immunophenotyping. *Human pathology* **31**, 1051-1054 (2000).
- 27 Dong, H. Y., Kung, J. X., Bhardwaj, V. & McGill, J. Flow cytometry rapidly identifies all acute promyelocytic leukemias with high specificity independent of underlying cytogenetic abnormalities. *American journal of clinical pathology* **135**, 76-84 (2011).

- 28 Khoury, H., Dalal, B. I., Nevill, T. J., Horsman, D. E., Barnett, M. J., Shepherd, J. D. *et al.* Acute myelogenous leukemia with t (8; 21)—identification of a specific immunophenotype. *Leukemia & lymphoma* **44**, 1713-1718 (2003).
- 29 Dalal, B. I., Mansoor, S., Manna, M., Pi, S., Di Sauro, G. & Hogge, D. E. Detection of CD34, TdT, CD56, CD2, CD4, and CD14 by flow cytometry is associated with NPM1 and FLT3 mutation status in cytogenetically normal acute myeloid leukemia. *Clinical Lymphoma Myeloma and Leukemia* **12**, 274-279 (2012).
- 30 Liu, J., Tong, J. & Yang, H. Targeting CD33 for acute myeloid leukemia therapy. *BMC cancer* **22**, 1-7 (2022).
- 31 De Propriis, M. S., Raponi, S., Diverio, D., Milani, M. L., Meloni, G., Falini, B. *et al.* High CD33 expression levels in acute myeloid leukemia cells carrying the nucleophosmin (NPM1) mutation. *haematologica* **96**, 1548 (2011).
- 32 Maakaron, J. E., Rogosheske, J., Long, M., Bachanova, V. & Mims, A. S. CD33-Targeted Therapies: Beating the Disease or Beaten to Death? *The Journal of Clinical Pharmacology* **61**, 7-17 (2021).
- 33 Mangaru, Z., Shetty, S., Visconte, V., Liu, H. D. & Rogers, H. J. Acute myeloid leukemia with inv (16)(p13. 1q22), abnormal eosinophils, and absence of peripheral blood and bone marrow blasts. *American Journal of Hematology* **91**, E273-274 (2016).
- 34 Bill, M., Jentsch, M., Bischof, L., Kohlschmidt, J., Grimm, J., Schmalbrock, L. K. *et al.* Impact of IDH1 and IDH2 mutation detection at diagnosis and in remission

- in patients with AML receiving allogeneic transplantation. *Blood Advances* **7**, 436-444 (2023).
- 35 Golub, D., Iyengar, N., Dogra, S., Wong, T., Bready, D., Tang, K. *et al.* Mutant isocitrate dehydrogenase inhibitors as targeted cancer therapeutics. *Frontiers in oncology* **9**, 417 (2019).
- 36 Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* **87**, 636-645 (2015).
- 37 Dorfman, D. M., LaPlante, C. D. & Li, B. FLOCK cluster analysis of plasma cell flow cytometry data predicts bone marrow involvement by plasma cell neoplasia. *Leukemia Research* **48**, 40-45 (2016).
- 38 Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184-197 (2015).
- 39 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**, 769-781 (2019).
- 40 Toedling, J., Rhein, P., Ratei, R., Karawajew, L. & Spang, R. Automated in-silico detection of cell populations in flow cytometry readouts and its application to leukemia disease monitoring. *BMC bioinformatics* **7**, 1-11 (2006).
- 41 Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature communications* **8**, 14825 (2017).

- 42 Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* **34**, 18932-18943 (2021).

ACKNOWLEDGEMENTS

This study was supported by National Institutes of Health National Cancer Institute grants U01CA220401 and U24CA19436201.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Institutional Review Board. The study was performed in accordance with the Declaration of Helsinki.

AUTHORSHIP CONTRIBUTIONS

J.E.L. and O.P. conceived the study; J.E.L. collected data and developed software; L.A.D.C. provided computational resources; J.E.L., L.A.D.C., D.L.J., and O.P. analyzed and interpreted the results; J.E.L. and O.P. wrote the manuscript. All authors read and approved the final paper.

FUNDING

This study was supported by National Institutes of Health National Cancer Institute grants U01CA220401 and U24CA19436201. L.A.D.C. received honoraria from Roche Tissue Diagnostics and serves on an advisory committee for the NVIDIA MONAI project.

DATA SHARING STATEMENT

Data sharing requests should be sent to Olga Pozdnyakova (opozdnyakova@bwh.harvard.edu).