A Machine Learning Approach to the Classification of Acute Leukemias and Distinction From Nonneoplastic Cytopenias Using Flow Cytometry Data

Sara A. Monaghan, MD,[1,2,] Jeng-Lin Li,[3,*] Yen-Chun Liu, MD, PhD,[1,4] Ming-Ya Ko, MS,[3] Michael Boyiadzis, MD,[5,6] Ting-Yu Chang, MS,[7] Yu-Fen Wang, MS,[7] Chi-Chun Lee, PhD,[3] Steven H. Swerdlow, MD,[1,2] and Bor-Sheng Ko, MD, PhD[8,9]*

From the [1]Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [2]UPMC Presbyterian, Pittsburgh, PA, USA; [3]Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan; [4]Department of Pathology, St Jude Children's Research Hospital, Memphis, TN, USA; [5]Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [6]UPMC Hillman Cancer Center, Pittsburgh, PA, USA; [7]AHEAD Medicine, Taipei, Taiwan; [8]Department of Hematological Oncology, National Taiwan University Cancer Center, Taipei, Taiwan; and [9]Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan.

ABSTRACT

Objectives: Flow cytometry (FC) is critical for the diagnosis and monitoring of hematologic malignancies. Machine learning (ML) methods rapidly classify multidimensional data and should dramatically improve the efficiency of FC data analysis. We aimed to build a model to classify acute leukemias, including acute promyelocytic leukemia (APL), and distinguish them from nonneoplastic cytopenias. We also sought to illustrate a method to identify key FC parameters that contribute to the model's performance.

Methods: Using data from 531 patients who underwent evaluation for cytopenias and/or acute leukemia, we developed an ML model to rapidly distinguish among APL, acute myeloid leukemia/not APL, acute lymphoblastic leukemia, and nonneoplastic cytopenias. Unsupervised learning using gaussian mixture model and Fisher kernel methods were applied to FC listmode data, followed by supervised support vector machine classification.

Results: High accuracy (ACC, 94.2%; area under the curve [AUC], 99.5%) was achieved based on the 37-parameter FC panel. Using only 3 parameters, however, yielded similar performance (ACC, 91.7%; AUC, 98.3%) and highlighted the significant contribution of light scatter properties.

Conclusions: Our findings underscore the potential for ML to automatically identify and prioritize FC specimens that have critical results, including APL and other acute leukemias.

INTRODUCTION

Flow cytometry (FC) immunophenotypic analysis is a critical component of testing to establish precise diagnoses for hematolymphoid neoplasms and monitor therapeutic response.[1-3] Computational methods to evaluate cytometry data have been evolving for exploratory and discovery research,[4,5] but with the exception of tools that the EuroFlow Consortium has

KEY POINTS

- Machine learning (ML) approaches for clinical flow cytometry (FC) data can automatically and accurately distinguish acute leukemias from nonneoplastic cytopenias.
- ML approaches can accurately classify FC data using substantially fewer markers than currently employed and may help streamline antibody panels.
- Our ML approach differs from others recently proposed in that it preserves the full spectrum of FC data without employing dimensionality reduction.

KEY WORDS

Machine learning; Flow cytometry; Acute promyelocytic leukemia; Acute myeloid leukemia; B-cell lymphoblastic leukemia/lymphoma

Am J Clin Pathol XXXX 2021;XX:1–0
HTTPS://DOI.ORG/10.1093/AJCP/AQAB148

Received: April 29, 2021
Accepted: August 1, 2021
Advance publication: October 13, 2021

Corresponding authors: Sara A. Monaghan, MD; monaghans2@upmc.edu; Bor-Sheng Ko, MD, PhD; bskomd@ntu.edu.tw.

*First authors.

Funding: This work was supported by the Ministry of Science and Technology, Taiwan, Republic of China (MOST 108-2823-8-002-003) to Jeng-Lin Li, Ming-Ya Ko, Chi-Chun Lee and Bor-Sheng Ko.

Disclosure: Bor-Sheng Ko and Chi-Chun Lee hold shares for AHEAD Medicine. Ting-Yu Chang and Yu-Feng Wang are affiliated with AHEAD Medicine.

developed,[6,7] clinical software has primarily provided a user interface an analyst can use to manually inspect and manipulate data displayed on 2-dimensional plots through a complex, sequential gating process.[8] Because this approach is labor intensive, heavily dependent on specialized expertise, and difficult to standardize, data analysis has become a rate-limiting factor for providing the FC interpretations needed for patient care. A solution for this bottleneck would increase laboratory efficiency and permit more rapid diagnoses of acute leukemias and other hematologic malignancies.

Artificial intelligence (AI), including machine learning (ML), has the potential to substantially assist physicians caring for patients with hematolymphoid diseases with interpreting and using complex data for diagnosis, risk stratification, and response prediction.[9,10] ML models have demonstrated human-level performance using FC data to classify B-cell neoplasms[11,12] and detect residual leukemia.[13,14] Our ML approach to rapidly classifying FC data (~7 seconds) predicted residual acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) with promising accuracy (84.6%-92.4%) and was associated with survival.[14] Whether a similar approach could be used to distinguish leukemic from nonneoplastic bone marrow samples and to rapidly distinguish acute promyelocytic leukemia (APL) from AML and acute lymphoblastic leukemia (ALL) was uncertain.

To further investigate the application of ML approaches using clinical FC data, we aimed to build a model to classify acute leukemias, including APL, and distinguish them from nonneoplastic cytopenias. We also sought to illustrate a method to identify key FC parameters that contribute to the model's performance. Our findings highlight the potential for AI to support clinical FC laboratories to efficiently detect and classify hematolymphoid neoplasms.

MATERIALS AND METHODS

Case Selection and Ground-Truth Diagnostic Categories

This retrospective study was approved by the Institutional Review Board of the University of Pittsburgh and the Research Ethics Committee of National Taiwan University Hospital. Cases included were bone marrow specimens analyzed by the clinical FC laboratory with new diagnoses of APL, AML/not APL, and ALL and from patients with no history of hematolymphoid neoplasia who were evaluated for recent pancytopenia and whose bone marrow was negative for neoplasia (ie, nonneoplastic cytopenias). Cases were included if the 5-tube panel of markers for new acute leukemia had been performed (Supplemental Table S1 [all supplemental materials can be found at *American Journal of Clinical Pathology* online]) and if the bone marrow morphologic evaluation had been performed at UPMC Presbyterian. Patients with APL were diagnosed between January 2013 and December 2018, while the others had been evaluated between January 2015 and May 2018. The ground-truth diagnoses were determined by review of the bone marrow pathology reports, including morphologic evaluation, manual FC data interpretation, chromosome analysis (98% of cases), any other cytogenetic or molecular studies (eg, myeloid panel next-generation sequencing, other mutational studies), other relevant pathology reports, and electronic health records. Acute leukemia cases were excluded if patients received therapy beyond supportive care for a preceding myeloid neoplasm (eg, MDS) or if the diagnosis was mixed-phenotype acute leukemia. For nonneoplastic cytopenias, patients with current or prior overt hematolymphoid neoplasms according to the World Health Organization classification[15] were excluded, but those with monoclonal B-cell lymphocytosis were not.

FC Immunophenotypic Studies

FC data had been acquired predominantly on 1 of 2 FACSCanto II instruments (BD Biosciences); rare cases (1.5%) were acquired on a third FACSCanto II instrument. Initial instrument setup, according to standard procedures, used BD CompBeads (BD Bioscience) for fluorescent parameters and a normal peripheral blood specimen for light scatter parameters. Settings for light scatter parameters were adjusted to achieve optimal separation for lymphocytes, monocytes, and granulocytes and further adjusted to ensure appropriateness for all specimen types. Agreement across instruments was addressed by establishing targets for all fluorescent channels using Cytometer Setup and Tracking (CS&T) beads (BD Biosciences) and transferring them from the predicate instrument to the others (ie, instrument cloning). Static light scatter gates for lymphocytes, monocytes, and granulocytes from a normal peripheral blood specimen were established and also applied to all instruments. At 6-month intervals and after instrument service, voltages were adjusted as needed to achieve the laboratory-established targets for fluorescent channels using CS&T beads and to keep the initially set static light scatter gates for normal peripheral blood. Daily quality control (QC) included monitoring all channels, including for light scatter, using CS&T beads and using BD FACS 7-color setup beads (BD Biosciences). If a 20 V or more change was predicted for any channel, an instrument was serviced. Levy-Jennings plots were also monitored for all channels. Daily QC for light scatter parameters also included visual inspection to ensure appropriate scaling and separation for lymphocytes, monocytes, and granulocytes. Compensation, lot-to-lot reagent comparisons, and specimen preparation and staining have been previously described.[16] We acquired 30,000 events for each tube whenever possible (ie, 97.6% of cases) the same day as staining.

Machine Learning

Model Development

FC listmode data (Flow Cytometry Standard [FCS] version 3.1) was used from the 5-tube panel for new acute leukemia. We regarded each light scatter property and fluorescent marker as a unique FC parameter. Data for any parameter evaluated more than once in the same channel (eg, forward scatter area [FSC-A], side scatter area [SSC-A], CD45 V500) were aggregated and resampled to ensure that the same amount of data for all was used for model development. Consequently, the combined data for all 37 parameters served as the input to the ML framework; FC

data according to specific tubes of markers (ie, combinations of parameters) were not used as input to the model. Preprocessing of the data included compensation and z score normalization. The framework consisted of 2 stages inclusive of an unsupervised phenotype representation learning and a supervised discriminative classifier. To obtain the phenotype representation, we trained a gaussian mixture model (GMM) to capture the complex cellular distribution. Then, a Fisher gradient vectorization approach was applied to embed phenotype characteristics in terms of the learned probability distribution in the derived specimen level high-dimensional phenotype representation. Each set of the preprocessed FCS data $X \in \mathbb{R}^{T\times D}$ was used for multivariate GMM training, where T was the total cell number and D was the number of FC parameters. The multivariate GMM was trained through an expectation-maximization algorithm in an unsupervised manner to obtain a set of parameters λ,

$$\lambda = \omega_k; \mu_k; \sigma_k; \mathrm{k} = 1 \ldots \mathrm{K} \tag{1}$$

where ω_k, μ_k, σ_k denoted the weight, mean vector, and covariance vector of k-th gaussian cluster and K was a specified total number of mixtures. With a sufficient number of GMM clusters, the complexity of the cellular composition could be completely modeled. We then used the Fisher kernel method to estimate the sample-wise posterior on the GMM parameters as a gradient scoring function,

$$\nabla_\lambda log P(X \mid \lambda) \tag{2}$$

where $\mathrm{P}(x_t \mid \lambda) = \sum_{i=1}^{K} \omega_i P_i(x_t \mid \lambda)$ was the likelihood of the given GMM. The expansion form of Fisher scoring function in terms of the first and second derivatives could be written as follows:

$$g^X_{\mu_k} = \frac{1}{T\sqrt{\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{x_t - \mu_k}{\sigma_k} \right) \tag{3}$$

$$g^X_{\sigma_k} = \frac{1}{T\sqrt{2\omega_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\left(\frac{x_t - \mu_k}{\sigma_k} \right)^2 - 1 \right) \tag{4}$$

where $\mathrm{P}(i \mid x_t, \lambda) = \frac{\omega_i \mathrm{P}_i(x_t \mid \lambda)}{\sum_{j=1}^{N} \omega_j P_j(x_t \mid \lambda)}$ indicated the posterior probability for $x_t \in X$. The concatenated vector of $g^X_{\mu_k}$ and $g^X_{\sigma_k}$ was further normalized by power normalization and L2 normalization for better computational efficacy. This final representation, the phenotype representation, had 2 × K × D dimensions. We fed the specimen-level representation into a support vector machine (SVM), with a linear kernel to conduct the 4-category classification (APL, AML/not APL, ALL, and nonneoplastic cytopenias). The whole framework **FIGURE 1** was implemented in Python; the GMM and SVM were based on the open-source scikit-learn package. Hyperparameters, such as K for GMM and C for SVM, were selected by grid search.

Evaluation of Model Performance

We used a 5-fold cross-validation scheme, 80% of the data for training and tuning and the remaining 20% (ie, the testing set) used to evaluate predictions for the categories. This process was conducted

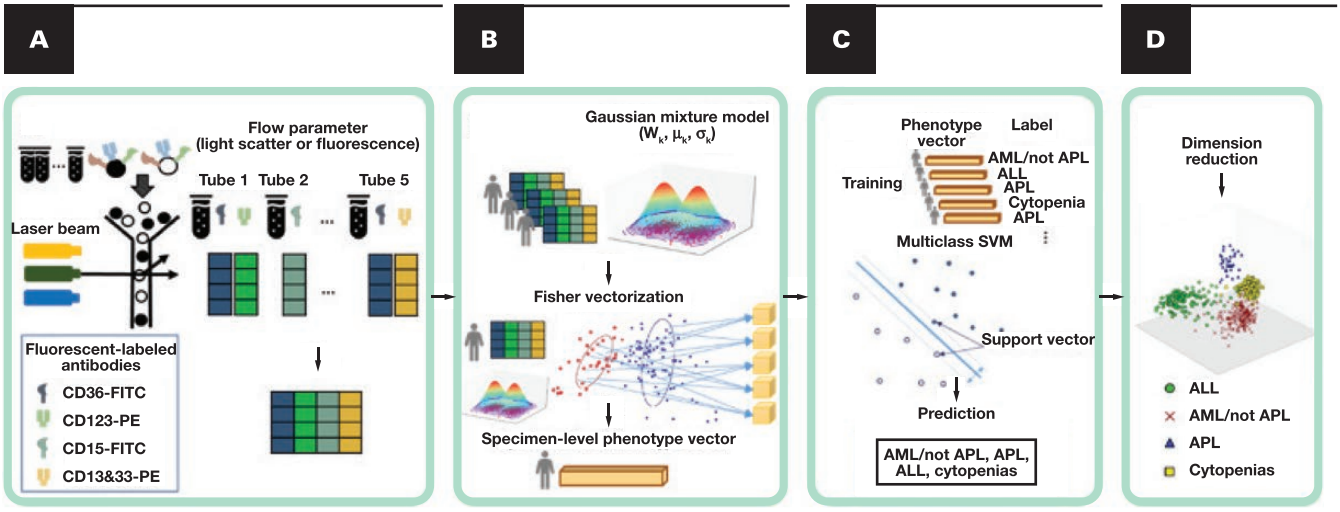

FIGURE 1 An overall schematic diagram of the machine learning (ML) framework. **A**, Flow cytometry (FC) listmode data were used as input for ML, including 37 FC parameters (light scatter properties and fluorochrome-labeled antibody binding) that had been individually evaluated for thousands of cells from each patient specimen; fluorescent parameters had been obtained from 5 different combinations of antibodies (ie, tubes), with some redundancy. Redundant parameters evaluated by more than 1 tube were aggregated and resampled so that the same amount of data for each parameter was used for ML model development. **B**, FC data were used to train an unsupervised gaussian mixture model (GMM) and encode it into a phenotype representation for each specimen, with a Fisher vectorization approach. In this encoding process, a specimen's FC data were transformed by computing the gradient distance with all the learned GMM cluster centers and aggregated as a specimen-level high dimensional representation (ie, vector). **C**, The specimen-level phenotype representations and their corresponding ground-truth labels were the input to train the supervised support vector machine (SVM) to classify the cases as acute promyelocytic leukemia (APL), acute myeloid leukemia (AML)/not APL, acute lymphoblastic leukemia (ALL), and nonneoplastic cytopenias (cytopenias). With the support vectors and the learned hyperplane, the multiclass prediction was performed on testing sets. **D**, Dimensionality reduction (ie, principal component analysis) was implemented on the specimen-level phenotype vectors and the decision score vectors of the SVM to illustrate the data distribution on a 3-dimensional plot. Each specimen was denoted as a dot, with different icons to indicate the ground-truth diagnoses.

5 times using different data randomly assigned to the testing set. Accuracy (ACC) and area under the receiver operating characteristic curve (AUC) were the evaluation metrics of performance.

Feature Selection Analysis

Cross-channel interaction was taken into account for the feature-selection experiments. The ACC was determined for the model by using each FC parameter alone to classify specimens into the 4 categories. The parameter that permitted the highest ACC was then paired with the individual remaining parameters to determine which pair provided the highest ACC. For each subsequent step, we added the remaining parameters individually to determine the combination with the next-highest ACC. Student *t* test analysis was used to determine the significance of performance difference for each step.

Evaluation of Cellular Composition and Potential Quality Indicators

For comparison of misclassified specimens with those specimens correctly classified using the ML model, proportions of populations out of total events were obtained from the most appropriate tube (Supplemental Table S2) from manual gating of FC data: lymphocytes, T lymphocytes, and natural killer cells (tube 4); B lymphocytes and hematogones (tube 3); granulocytes (tube 2); monocytic cells and erythroid cells (tube 1); and blasts (hematogones excluded) for nonneoplastic cytopenia cases (tube 1) and for APL (tube 2). Unless better isolated in another tube, blasts were from tube 3 for B-cell ALL (B-ALL) and tube 4 for T-cell ALL (T-ALL). The blasts for AML/not APL were from tube 1 unless better isolated by tube 2; monocytic cells were included with blasts for AML/not APL that had monocytic differentiation. AML/not APL "with monocytic differentiation" was recognized when supported by review of the bone marrow pathology report and other components that contributed to the final clinicopathologic diagnosis, which were reviewed in particular whenever monocytes were found to be 10% or more of the total events by manual FC data analysis. Percentages were obtained from dot plots created with BD FACSDiva v7.0, v8.0 (7/2013 - 8/2014), v8.0.1 (9/2014 - 12/2018) software (BD Biosciences) except when better isolated using Infinicyt software, version 2.0 (Cytognos): blasts and granulocytes for AML/not APL and APL, monocytic cells for AML/not APL, and erythroid cells for all specimens. Singlet cells were an average from the 5 tubes using FSC-A vs FSC height dot plots. Viable cells were those not staining for 7-aminoactinomycin D.

Three potential specimen quality indicators were also compiled. "Hypocellular BM biopsy" was defined as a bone marrow biopsy reported as adequate for interpretation with 20% or less cellularity. A designation of "less than optimal aspirate smears" was applied when the pathology report indicated that the smears were "inadequate," "limited," or "suboptimal." A "gross % blast underestimate by FC" was recorded when the aspirate smear manual differential blast percentage was 20% or more in the pathology report and the % blasts by manual FC data analysis was lower by a relative difference of 40%; no assessment was made when aspirate smears were reported as inadequate.

Continuous variables across groups were analyzed using the Kruskal-Wallis test, followed, when appropriate, by the Dunn multiple comparisons test. Categorical variables were analyzed using the χ^2 test and, when appropriate, the Fisher exact test. Statistical analyses were performed using GraphPad Prism software, version 8.0.1.

RESULTS

Diagnostic Categories

FC data were originally obtained as part of the clinical evaluation of bone marrow specimens from 531 patients with a new diagnosis of APL (n = 32 [6.0%]), AML/not APL (n = 200 [37.7%]), ALL (n = 131 [24.7%]; B-ALL, n = 118; T-ALL, n = 13) and patients evaluated for potential acute leukemia because of recent pancytopenia but whose comprehensive bone marrow evaluation was negative for hematolymphoid neoplasm (ie, nonneoplastic cytopenias, n = 168 [31.6%]). CBC data, manual differential blast percentages, % blasts by manual FC data analysis, and potential factors associated with specimen quality were summarized for the 4 categories **TABLE 1**.

ML Model Performance

The ML model classified FC list mode data into 4 categories, corresponding to the ground-truth diagnoses **FIGURE 1**. Performance was assessed in 5 rounds, with different data randomly held out to serve as testing sets (Supplemental Table S3). while classification of the whole data set using all 37 FC parameters, including light scatter properties and fluorescence, demonstrated 94.2% ACC and 99.5% AUC. For each category, the individual ACC ranged from 87.5% to 97.6%, sensitivity ranged from 87.5% to 97.6%, and specificity ranged from 95.6% to 100.0% **FIGURE 2**.

Feature Selection Analysis

The ML model ACC ranged from 47.5% to 77.0% when the 37 FC parameters were evaluated individually (Supplemental Table S4). The parameter yielding the highest ACC was then paired with each remaining parameter to determine which pair provided the highest ACC. Subsequent parameters were added individually according to the ACC of the combinations (Supplemental Figure S1). Model performance improved with each step ($P < .001$) up to 3 parameters (FSC-A, SSC height [SSC-H], CD117), but no significant improvement was gained by adding more markers, and there was no significant difference in the model's performance when using the full marker set and any other number of markers beyond the top 3. The top 3 parameters produced a performance (ACC, 91.7%; AUC, 98.3%) similar to that achieved for all 37 samples. The findings underscored the significant contribution of light scatter properties to model performance.

Specimens Misclassified by the ML Model

A total of 31 of 531 (5.8%) specimens were misclassified. Because numbers misclassified from one specific category into another were low **TABLE 2**, comparisons with correctly classified specimens were precluded. However, a somewhat larger group of misclassified specimens (n = 16) was assembled for comparisons by combining all

TABLE 1 Attributes of Cases According to Ground-Truth Diagnosis: CBCs, % Blasts by Manual Differentials, % Blasts by Manual Flow Cytometry Data Analysis, and Potential Specimen Quality Indicators

	AML/Not APL	APL	ALL	Nonneoplastic Cytopenias
Cases, No. (%)	200 (37.7)	32 (6.0)	131 (24.7)	168 (31.6)
CBC[a]				
WBC, $\times 10^3/\mu L$	7.8 (2.2-30.1)	2.4 (1.5-13.1)	9.5 (3.7-25.8)	2.4 (1.8-3.2)
Hemoglobin, g/dL	8.9 (7.9-10.0)	9.5 (8.2-10.9)	8.9 (7.8-10.1)	9.4 (8.2-10.9)
MCV, fL	95.8 (90.6-100.9)	89.4 (86.2-92.2)	85.3 (81.0-89.0)	90.6 (85.2-97.6)
Platelet count, $\times 10^3/\mu L$	56.0 (29.0-98.2)	27.0 (14.0-57.5)	66.0 (38.0-114.5)	72 (46.0-104.0)
Manual differential, peripheral blood[a]				
Blasts, %	27.0 (9.9-58.0)	55.0 (10.5-74.5)	48.8 (12.2-72.6)	0.0 (0.0-0.0)
Manual differential, bone marrow[a]				
Blasts, %	58.0 (34.0-77.4)	81.0 (71.8-84.0)	90.0 (84.3-94.0)	1.0 (0.6-1.8)
Flow cytometry, bone marrow[a]				
Blasts, %	50.0 (30.8-73.2)	84.0 (73.0-87.2)	82.0 (65.6-90.0)	0.8 (0.4-1.3)
Potential specimen quality indicators, present/not present[b]				
Hypocellular bone marrow biopsy	3/185[c]	0/32[d]	1/123[e]	25/137[c,d,e]
Less-than-optimal aspirate smears	40/160	3/29	28/103	25/143
Gross % blast underestimate by FC	22/173	3/29	16/114	NA

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; APL, acute promyelocytic leukemia; FC, flow cytometry, MCV, mean corpuscular volume; NA, not applicable.
[a]Data shown as median (25th to 75th percentiles).
[b]Potential quality indicators were compared between groups using χ^2 test and were significantly different only for hypocellular bone marrow biopsies ($P < .0001$). Fisher exact tests revealed that hypocellular bone marrow biopsy was more common for nonneoplastic cytopenias compared with
[c]AML/not APL ($P < .0001$),
[d]APL ($P = .017$), and
[e]ALL ($P < .0001$).

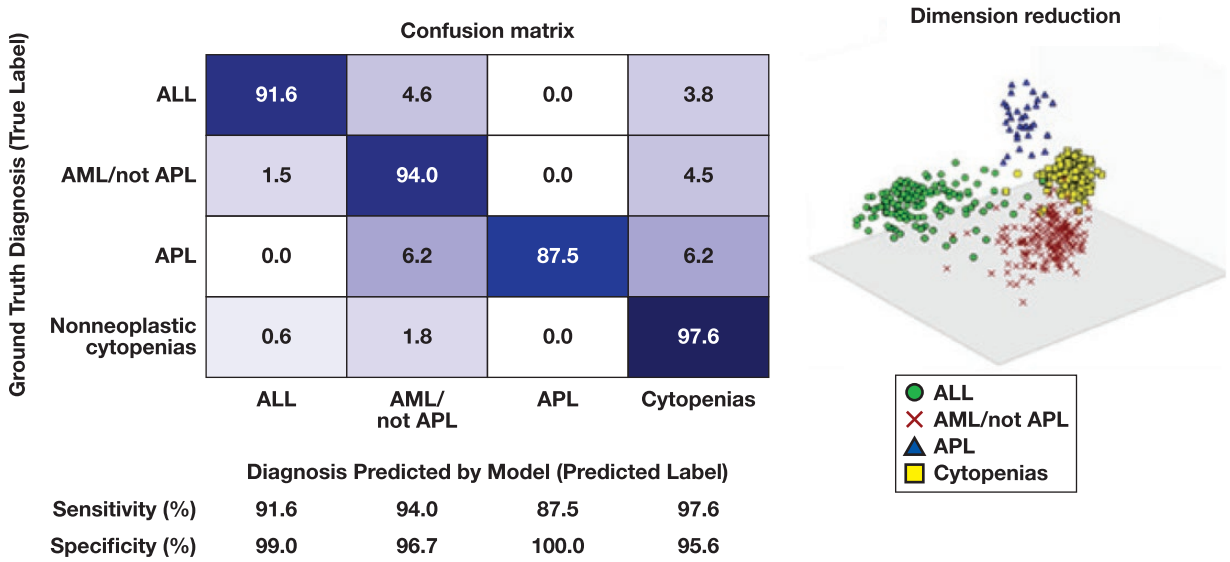

FIGURE 2 Performance of the machine learning (ML) model for classification of acute leukemias and distinction from nonneoplastic cytopenias. The ML model was trained to classify patients' flow cytometry data into 4 categories corresponding to the ground-truth diagnoses: acute promyelocytic leukemia (APL), acute myeloid leukemia (AML)/not APL, acute lymphoblastic leukemia (ALL), and nonneoplastic cytopenias (cytopenias). **A**, Classification accuracy of the final ML model for each category using the whole data set was depicted by a confusion matrix; sensitivity and specificity for each predicted category were also determined. **B**, Dimensionality reduction using principal component analysis was performed to depict the data output distribution of the model on a 3-dimensional plot; the ground-truth diagnosis was denoted with different icons.

acute leukemias misclassified as nonneoplastic cytopenias (AML/not APL, n = 9; ALL, n = 5; APL, n = 2).

The percentages of major hematolymphoid populations determined by manual FC data analysis were compared between the combined group of acute leukemias misclassified as nonneoplastic cytopenias and the 4 groups consisting of only correctly classified specimens (Supplemental Table S2). The acute leukemias misclassified as nonneoplastic cytopenias had lower proportions of blasts compared

with all 3 groups of correctly classified acute leukemias (median, 16% vs 84% [APL], 52% [AML/not APL], 84% [ALL]; $P < .05$ for all 3 comparisons). They also had higher proportions of lymphocytes compared with the correctly classified acute leukemias (median, 30% vs 7% [APL], 11% [AML/not APL], 6% [ALL]; $P < .05$ for all 3 comparisons).

No significant difference was found between the distribution for flow cytometers 1, 2, and 3 used for data acquisition for correctly classified (63.2%, 35.2%, 1.6%) and incorrectly classified specimens (64.5%, 35.5%, 0%).

Because rapid, accurate detection of APL is so important, observations about misclassification were specifically sought for this category. No specimens were misclassified as APL. No APL was misclassified as ALL, but 4 of 32 (12.5%) specimens were misclassified as either nonneoplastic cytopenias (n = 2) or AML/not APL (n = 2) TABLE 2. The 2 samples misclassified as nonneoplastic cytopenias were similar to the larger assembled group of acute leukemias misclassified as such because they had low proportions of blasts based on manual FC data analysis (ie, 8% and 18%, respectively). The other 2 that were misclassified had 70% and 87% blasts, respectively, and no obvious reason for misclassification as AML/not APL.

Among AML/not APL cases, 58 of 200 had monocytic differentiation (29%). There was no difference in the model's performance between those with or without monocytic differentiation. Two of 58 with monocytic differentiation and 10 of 142 without monocytic differentiation were misclassified (ACC, 96.6 vs 92.9%; $P = .5147$). Both AML/not APL specimens with monocytic differentiation were misclassified as nonneoplastic cytopenias, and 1 was similar to the overall larger assembled group of acute leukemias so misclassified because of a low proportion of blasts by manual FC data analysis (16%), despite 63% blasts on the aspirate smear differential. Neither had good reason to consider chronic myelomonocytic leukemia (CMML) as an alternative diagnosis. Among correctly classified AML/not APL with monocytic differentiation cases, only 1 with 20% blasts/promonocytes was difficult to distinguish from CMML, while another patient had been under observation for previously diagnosed CMML when he presented with 51% blasts in the bone marrow morphologically.

Three Potential Specimen Quality Indicators and Impact on ML Model Performance

A hypocellular bone marrow biopsy was more common for nonneoplastic cytopenias compared with all 3 acute leukemias TABLE 1, but no differences were found among the 4 categories based on "less than optimal aspirate smears" or, among the acute leukemia categories, based on "gross % blast underestimate by FC" with reference to the aspirate smear manual differential.

Acute leukemias misclassified as nonneoplastic cytopenias were more often associated with a "gross % blast underestimate by FC" compared with correctly classified APL, AML/not APL, and ALL (78.6% of the specimens vs 3.6%, 8.7%, and 9.2%, respectively; $P < .0001$ for all 3 comparisons). Both APL specimens misclassified as nonneoplastic cytopenias and the AML/not APL with monocytic differentiation case with a low proportion of blasts also so misclassified met criteria for "gross % blast underestimate by FC."

Abnormal Populations in Addition to Acute Leukemia

Abnormal B-cell populations other than acute leukemia were detected in 20 of 531 cases (3.8%). Only 1 of these specimens was among those misclassified by the ML model. Chronic lymphocytic leukemia/small lymphocytic lymphoma (38% of total events) and another B-cell–lineage lymphoproliferative disorder (53% of total events) were detected in 2 patients with AML/not APL. The latter case was misclassified as nonneoplastic cytopenias but was also noted to have only 9% blasts by FC. Light chain restricted B cells (≤12% of total events) suggesting monoclonal B-cell lymphocytosis were identified in 8 patients with AML/not APL, 2 with ALL, and 8 with nonneoplastic cytopenias.

DISCUSSION

Our ML model, developed for bone marrow specimens originally obtained to evaluate for potential new acute leukemia, demonstrated excellent performance (ACC, 94.2%; AUC, 99.5%) to rapidly classify FC data into 4 categories (APL, AML/not APL, ALL, and nonneoplastic cytopenias). The model achieved this accuracy using the complete FC panel for new acute leukemia but also

TABLE 2 Number of Specimens Misclassified by Machine Learning Model

Correct Category, No.	No. Misclassified (% of Diagnostic Category)	Incorrect Classifications
AML/not APL (n = 200)	12 (6)	ALL (n = 3)
		Nonneoplastic cytopenias (n = 9)
APL (n = 32)	4 (12.5)	AML/not APL (n = 2)
		Nonneoplastic cytopenias (n = 2)
ALL (n = 131)	11 (8.4)	—
B-cell lymphoblastic leukemia (n = 118)	9 (7.6)	AML/not APL (n = 4)
		Nonneoplastic cytopenias (n = 5)
T-cell lymphoblastic leukemia (n = 13)	2 (15.4)	AML/not APL (n = 2)
Nonneoplastic cytopenias (n = 168)	4 (2.4)	AML/not APL (n = 3)
		ALL (n = 1)

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; APL, acute promyelocytic leukemia.

HTTPS://DOI.ORG/10.1093/AJCP/AQAB148

demonstrated similar performance with only 3 FC parameters. The findings suggested that an ML approach could serve as an automated triage tool to rapidly identify and prioritize FC results for hematologic malignancies, including APL and other acute leukemias.

Computational and ML algorithms for FC data interpretation[17] and morphologic image analysis[18] proposed specifically for hematolymphoid disorders have been reviewed and include some designed to assist in acute leukemia diagnosis. The potential of morphologic image analysis for acute leukemia classification has been demonstrated.[19] Using FC data, 87% of 23 algorithms achieved greater than 97% ACC when challenged to distinguish AML from healthy donors.[20] Although that finding was published 8 years ago, adoption of computational tools in clinical laboratories has remained low.[7] Infinicyt software, supported by the EuroFlow Consortium, has provided the computational tools most commonly used for clinical FC testing.[7] Infinicyt offers partially automated gating for a small number of FC panels, including an acute leukemia orientation tube (ALOT), along with tools to determine whether populations match those within their database.[6,21] For ALOT analysis, Infinicyt recommends the subsequent panel for full characterization (ie, AML, T-ALL, or B-ALL panel) or indicates that a manual decision is needed. Despite multiple other methods,[22,23] population clustering or automated gating alone typically requires substantial human interaction to review populations before interpretation. However, methods to cluster or model FC data, followed by supervised ML (SML) classification, can provide more automated results for diagnostic category, residual disease status, or outcome prediction.[11,12,14,24] Our prior study used such an approach to identify specimens with residual AML and MDS.[14] We adapted those methods for the current study to detect and classify new acute leukemias.

Our method applied unsupervised representative learning with an integrated use of generative probabilistic GMM to provide an overall representation of individual patient FC data and the Fisher kernel method to express the degree of difference between patients. We then used a supervised SVM classifier to maximize discrimination between nonneoplastic cytopenias and 3 acute leukemia categories. In this paradigm, the discriminative power from the Fisher scoring–based phenotype vector enhanced the capacity of the linear SVM classifier. Different methods to model clinical FC data, followed by other SML classifiers, have been used to classify subtypes of mature B-cell neoplasms and healthy samples[11] and to detect B-cell neoplasms with good accuracy, identify specimens that need add-on markers, and show the potential to autoverify normal results.[12] Zhao et al[11] used self-organizing maps, followed by a deep convolutional neural network as the classifier. Ng et al[12] used FC data from their B-cell screening panel as input for uniform manifold approximation and projection, a method related to *t* distributed stochastic neighbor embedding, followed by a random forest classifier. Unlike our approach, the methods in those 2 studies used dimensionality reduction of FC data with inherent loss of information.

Evaluation of each FC parameter's contribution to model performance in this study demonstrated improvements up to a combination of 3 parameters (FSC-A, SSC-H, CD117) but no further significant improvements with more. There was no significant difference in performance when using the full set of parameters and any number of markers beyond the top 3, which implied that the learned SVM classifier tended to be robust and did not dramatically change or overfit because of different feature dimensions. The results illustrated the high contribution of light scatter properties and raised the possibility that a simplified screening method for hematologic malignancy based primarily on light scatter properties might be feasible; however, this hypothesis remains to be explored. Importantly, unlike a simplified screening method, the minimum number of markers needed for an ML model to classify data could not be the only consideration for panels intended to contribute to definitive FC interpretations. For example, the best 3-parameter combination in our study would be completely inadequate for a human analyst to render an interpretation. Optimization of such panels to be used with ML classification would need to take into account the data laboratory professionals need to catch errors, manage data classified into broad or indeterminate categories, and render timely, conclusive interpretations.

Given the importance of preventing errors for patient testing, we tried to identify features associated with risk for misclassification. We found a low proportion of blasts and evidence that suboptimal, hemodilute specimens contribute to misclassification of acute leukemias, including 2 of the 4 misclassified APL specimens, as nonneoplastic. Increasing the number of specimens to train the model could potentially improve performance, particularly for the underrepresented APL category. Augmenting model training with more acute leukemia cases with low blast counts and/or training the model to recognize hemodilute specimens would also have the potential to further improve performance. Otherwise, the numbers of specimens misclassified from one individual category into another were too low to identify commonalities to help understand why they were misclassified. Further understanding of additional factors that contribute to misclassification would be important before implementing an ML approach clinically. Although understanding how ML models learn to make predictions is difficult,[10] it is a growing area of research.[25]

Although our study demonstrated that an ML model could distinguish among 3 categories of acute leukemia involving bone marrow and nonneoplastic cytopenias, clinical laboratories encounter a much broader spectrum of hematologic malignancies[15] and specimen types. For example, 3.8% of our specimens harbored abnormal populations beyond acute leukemia, which our model was not trained to detect. We did not train the model to detect or classify myeloid neoplasms other than AML, and we excluded acute leukemias of ambiguous lineage. The model would need to be trained to classify more categories, but an indeterminate category could be used for less frequent conditions until sufficient specimens were accrued for additional specific categories. An ML model would also need to be trained for other specimen types (eg, blood, lymph nodes, body fluids) for which it would be used.

In this study, we did not evaluate the degree of standardization needed for the model to be implemented across multiple laboratories. For example, decision support tools available with Infinicyt rely on strictly standardized antibody panels, specimen processing, and instrument settings.[6] Two aspects of our

approach may permit sufficient flexibility for the model to be applied across institutions. First, preprocessing the FC data before model input includes compensation and normalization. First, z score normalization takes into account statistical characteristics (mean and SD) of the FC data and helps offset variations for light scatter and fluorescent intensities.[26] Second, we anticipate that the use of GMM to capture the phenotypic representation for each specimen in a probabilistic manner[27] may permit us to apply our approach to those distributions derived for similar parameters across institutions without extreme stringency related to antibody panels, specimen processing, and instrument settings. One of our major upcoming goals is to obtain a large multicenter database to further evaluate these potential solutions and determine the minimum level of standardization needed to successfully generalize our approach.

The ML model we developed demonstrated excellent performance to rapidly classify real-world FC data into 4 categories for patients evaluated for potential acute leukemia. It accomplished this success with substantially fewer FC markers than currently exist in our new acute leukemia panel. These results can be used to help design AI-based decision support tools to address the wider spectrum of conditions clinical laboratories encounter. Anticipating further development and multicenter studies to evaluate the generalizability of such an approach, we are optimistic that AI-assisted decision support will lead to greater efficiency and increase patient access to fast and accurate diagnoses for hematologic malignancies.

Acknowledgments: We acknowledge the technical and administrative assistance of Matthew Wild, Gwendolyn Illar, Ruth Bates, and Wendy Shallenberger in the Division of Hematopathology and the Clinical Flow Cytometry Laboratory at UPMC.

REFERENCES

1. Craig FE, Foon KA. Flow cytometric immunophenotyping for hematologic neoplasms. *Blood*. 2008;111:3941-3967.
2. Borowitz MJ, Wood BL, Devidas M, et al. Prognostic significance of minimal residual disease in high risk B-ALL: a report from Children's Oncology Group study AALL0232. *Blood*. 2015;126:964-971.
3. Lahuerta JJ, Paiva B, Vidriales MB, et al; GEM (Grupo Español de Mieloma)/PETHEMA (Programa para el Estudio de la Terapéutica en Hemopatías Malignas) Cooperative Study Group. Depth of response in multiple myeloma: a pooled analysis of three PETHEMA/GEM clinical trials. *J Clin Oncol*. 2017;35:2900-2910.
4. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol*. 2016;16:449-462.
5. Liu P, Liu S, Fang Y, et al. Recent advances in computer-assisted algorithms for cell subtype identification of cytometry data. *Front Cell Dev Biol*. 2020;8:234.
6. Pedreira CE, Costa ESD, Lecrevise Q, et al; EuroFlow. From big flow cytometry datasets to smart diagnostic strategies: the EuroFlow approach. *J Immunol Methods*. 2019;475:112631.
7. Cheung M, Campbell JJ, Whitby L, et al. Current trends in flow cytometry automated data analysis software [published online ahead of print February 19, 2021]. *Cytometry A*. doi:10.1002/cyto.a.24320.
8. Loken MR. Multidimensional data analysis in immunophenotyping. *Current Protocols in Cytometry*. 2001. doi:10.1002/0471142956.cy1004s00.
9. Shouval R, Fein JA, Savani B, et al. Machine learning and artificial intelligence in haematology. *Br J Haematol*. 2021;192:239-250.
10. Radakovich N, Nagy M, Nazha A. Machine learning in haematological malignancies. *Lancet Haematol*. 2020;7:e541-e550.
11. Zhao M, Mallesh N, Höllein A, et al. Hematologist-level classification of mature B-cell neoplasm using deep learning on multiparameter flow cytometry data. *Cytometry A*. 2020;97:1073-1080.
12. Ng DP, Zuromski LM. Augmented human intelligence and automated diagnosis in flow cytometry for hematologic malignancies. *Am J Clin Pathol*. 2021;155:597-605.
13. Reiter M, Diem M, Schumich A, et al; International Berlin-Frankfurt-Münster (iBFM)-FLOW-network and the AutoFLOW Consortium. Automated flow cytometric MRD assessment in childhood acute B-lymphoblastic leukemia using supervised machine learning. *Cytometry A*. 2019;95:966-975.
14. Ko BS, Wang YF, Li JL, et al. Clinically validated machine learning algorithm for detecting residual diseases with multicolor flow cytometry analysis in acute myeloid leukemia and myelodysplastic syndrome. *EBioMedicine*. 2018;37:91-100.
15. Swerdlow SH, Campo E, Harris NL, et al, eds. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. Rev. 4th ed. Lyon, France: IARC; 2017.
16. Hunt AM, Shallenberger W, Ten Eyck SP, et al. Use of internal control T-cell populations in the flow cytometric evaluation for T-cell neoplasms. *Cytometry B Clin Cytom*. 2016;90:404-414. doi:10.1002/cyto.b.21335.
17. Duetz C, Bachas C, Westers TM, et al. Computational analysis of flow cytometry data in hematological malignancies: future clinical practice? *Curr Opin Oncol*. 2020;32:162-169.
18. Nanaa A, Akkus Z, Lee WY, et al. Machine learning and augmented human intelligence use in histomorphology for haematolymphoid disorders. *Pathology*. 2021;53:400-407.
19. Reta C, Altamirano L, Gonzalez JA, et al. Segmentation and classification of bone marrow cells images using contextual information for medical diagnosis of acute leukemias. *PLoS One*. 2015;10:e0130805.
20. Aghaeepour N, Finak G, Hoos H, et al; FlowCAP Consortium; DREAM Consortium. Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*. 2013;10:228-238.
21. Lhermitte L, Mejstrikova E, van der Sluijs-Gelling AJ, et al. Automated database-guided expert-supervised orientation for immunophenotypic diagnosis and classification of acute leukemia. *Leukemia*. 2018;32:874-881.
22. Finak G, Frelinger J, Jiang W, et al. OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput Biol*. 2014;10:e1003806.
23. Liu X, Song W, Wong BY, et al. A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol*. 2019;20:297.
24. Rajwa B, Wallace PK, Griffiths EA, et al. Automated assessment of disease progression in acute myeloid leukemia by probabilistic analysis of flow cytometry data. *IEEE Trans Biomed Eng*. 2017;64:1089-1098.
25. Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ*. 2019;364:l886.
26. Lee G, Stoolman L, Scott C. Transfer learning for auto-gating of flow cytometry data. Paper presented at: Proceedings of ICML Workshop on Unsupervised and Transfer Learning. *Proc Machine Learning Res*. 2012;27:155-165. http://proceedings.mlr.press/v27/lee12a.html.
27. Baudry JP, Raftery AE, Celeux G, et al. Combining mixture components for clustering. *J Comput Graph Stat*. 2010;9:332-353.