EARLY ONLINE RELEASE

Note: This article was posted on the Archives Web site as an Early Online Release. Early Online Release articles have been peer reviewed, copyedited, and reviewed by the authors. Additional changes or corrections may appear in these articles when they appear in a future print issue of the Archives. Early Online Release articles are citable by using the Digital Object Identifier (DOI), a unique number given to every article. The DOI will typically appear at the end of the abstract.

The DOI for this manuscript is doi: 10.5858/arpa.2021-0044-OA

The final published version of this manuscript will replace the Early Online Release version at the above DOI once it is available.
A Physician-in-the-Loop Approach by Means of Machine Learning for the Diagnosis of Lymphocytosis in the Clinical Laboratory

Laura Bigorra, MSc; Iciar Larriba, MD; Ricardo Gutiérrez-Gallego, PhD

Context.—The goal of the lymphocytosis diagnosis approach is its classification into benign or neoplastic categories. Nevertheless, a nonnegligible percentage of laboratories fail in that classification.

Objective.—To design and develop a machine learning model by using objective data from the DxH 800 analyzer, including cell population data, leukocyte and absolute lymphoid counts, hemoglobin concentration, and platelet counts, besides age and sex, with classification purposes for lymphocytosis diagnosis.

Design.—A total of 1565 samples were included from 10 different lymphoid categories grouped into 4 diagnostic categories: normal controls (458), benign causes of lymphocytosis (567), neoplastic lymphocytosis (399), and spurious causes of lymphocytosis (141). The data set was distributed in a 60-20-20 scheme for training, testing, and validation stages. Six machine learning models were built and compared, and the selection of the final model was based on the minimum generalization error and 10-fold cross validation accuracy.

Results.—The selected neural network classifier rendered a global 10-class classification validation accuracy corresponding to 89.9%, which, considering the aforementioned 4 diagnostic categories, presented a diagnostic impact accuracy corresponding to 95.8%. Finally, a prospective proof of concept was performed with 100 new cases with a global diagnostic accuracy corresponding to 91%.

Conclusions.—The proposed machine learning model was feasible, with a high benefit-cost ratio, as the results were obtained within the complete blood count with differential. Finally, the diagnostic impact with high accuracies in both model validation and proof of concept encourages exploration of the model for real-world application on a daily basis.

(Arch Pathol Lab Med. doi: 10.5858/arpa.2021-0044-OA)

Both absolute lymphoid counts (ALCs) above $5 \times 10^9/\mu$L and/or presence of the atypical lymphoid flag trigger the analysis of the peripheral blood smear to classify the alteration as benign or malignant from the assessment of morphologic features. The consensus for morphologic diagnosis, and subsequently the International Council for Standardization in Hematology (ICSH), established that atypical lymphocytes should be further divided into “suspect reactive” (reactive lymphocytes), “suspect neoplastic” (abnormal lymphocytes), and “uncertain nature” categories.

The correct morphologic identification of lymphoproliferative neoplasms with peripheral expression is of utmost importance to trigger a subsequent clinical decision. Nevertheless, the assessment of the correct performance evaluated through external quality assessment (EQA) schemes reveals that the mean correct diagnosis of reactive cases is performed by 75% of participants, while that for neoplastic assignments is below 70%.

Modern analyzers, such as the Beckman Coulter DxH 800 (Miami, Florida), record cell population data (CPD) parameters of the leukocyte subpopulations by using impedance, conductivity, and scatter measures with the VCS (volumetric-conductivity-scatter) technology. These parameters have specific mean and standard deviation values for normal cells and any variation point to a particular pathology.

Machine learning (ML) is a tool that uses information from different sources, finding patterns from training data and allowing predictions to be made with classification purposes. Building on prior experiences using ML and allowing predictions to be made with classification purposes, the current work increases the diagnostic complexity by combining (1) 3 benign causes of lymphocytosis (viral infection, persistent polyclonal B-cell lymphocytosis, and post-splenectomy lymphocytosis), (2) 4 different types of neoplastic lymphocytosis (chronic lymphocytic leukemia, monoclonal B-cell lymphocytosis, CLL-like, and peripheral expression of follicular lymphoma) and (3) 2 spurious categories with lymphocytosis due to erythrolysis resistance (anaemia and liver disease)—which
together with the normal controls make a total of 10 different categories.

The goal was to develop a CPD-based ML model as an aid to the lymphocytosis diagnosis approach and evaluate its feasibility, standardization, and benefit-cost ratio. Moreover, as it is expected that this ML model would be integrated into the general routine workflow, a proof of concept (POC) was also designed to evaluate its daily-basis performance.

**MATERIALS AND METHODS**

The Figure summarizes the different phases of the ML approach following a scheme previously proposed for health care ML models.\(^{18}\) Data analysis and modelling were performed by using Python (3.6) and the Scikit-Learn (0.18.1) library.\(^{14,19,20}\)

**Problem Selection**

The focus was the classification of samples with lymphocytosis (ALC \(>3.5 \times 10^9/\mu L\)) and/or the presence of an atypical lymphoid flag into the categories recommended by the ICSH:\(^{5}\) (1) benign, (2) neoplastic, or (3) spurious.

**Data Collection**

**Human Specimens.**—A total of 1565 samples were included corresponding to (1) 458 normal controls; (2) categories with lymphocytosis including benign causes such as 346 cases of VI, 107 cases of expansion of polyclonal B cells (with or without lymphocytosis) with binucleated lymphocytes, and 114 cases involving post-splenectomy patients; (3) neoplastic cases including 191 CLL, 90 MBL, CLL-type, and cases with peripheral blood expression of SMZL \((n = 87)\) and FL \((31\) cases from 15 patients); and (4) spurious causes due to erythrocyte-hemolysis resistance in patients with anemia \((n = 109)\) or liver disease \((n = 32)\). Data use authorization was obtained from the institutional board. The treatment of data was carried out in accordance with the requirements of the legislation issued for protection of the data and privacy of the patients.

Hairy cell leukaemia and mantle cell lymphoma B-cell neoplasms were not included in the present study. The former because hairy cells are misinterpreted as monocytes in the DxH 800, and previous studies have resulted in specific CPD flags for the identification of hairy cells within the abnormal monocyte cluster.\(^{21}\) Mantle cell lymphoma B-cell neoplasms were disregarded because information on cyclin D1 was not available in the suspected cases for the diagnosis confirmation.

Venous blood samples were collected into tubes containing EDTA as anticoagulant and the analyses were performed within 6 hours after collection. The diagnostic confirmation and classification of the CLL, MBL, SMZL and FL samples was based on flow cytometric immunophenotype, following the World Health Organization (WHO) consensus guideline and according to the CLL-scoring system.\(^{22,23}\) The B-cell panel included CD19/CD5/CD23/CD10/CD105/CD79b/CD20/K/L/FMC7/CD38. The PPBL cases showed an expansion of B lymphocytes \((CD19^{+})\) that were CD5\(^{-}\), CD23\(^{+}\) with representation of both \(\kappa\) and \(\lambda\) light chains.\(^{24}\) Flow cytometric analysis was performed with a BD FACS Calibur cytometer (Becton-Dickinson, Franklin Lakes, New Jersey). Mouse anti-human monoclonal antibodies (DAKO Corporation, Carpin teria, California) were used as the primary antibodies, while anti-human \(\kappa\) and \(\lambda\) light-chain polyclonal antibodies were used for immunoglobulin light-chain expression analysis (Thermo Fisher Scientific Waltham, Massachusetts); all stainings were performed with appropriate controls. Viral infections were confirmed by chemiluminescence immunooassays (Liaison, DiaSorin, Italy), with positivity for Epstein-Barr or cytomegalovirus immunoglobulin M antibodies being the inclusion criteria.

**Features.**—The included features corresponded to population parameters (age and sex), hematoletic parameters including leukocyte counts and ALC, hemoglobin concentration and platelet counts, and lymphoid CPD, using a Beckman Coulter DxH 800 instrument with software version 3.9.0.31. Table 1 details the clinical characteristics of the patients, the percentage of patients with lymphocytosis, and the analyzer flags. For adults, lymphocytosis was considered when ALC was above \(3.5 \times 10^9/\mu L\), while for children the upper limit was considered by taking into account the different normal ranges by age.\(^{10}\)

All groups were composed of adult samples, with the exception of the VI group with 53% of the samples corresponding to a pediatric population (younger than 18 years) owing to the high specific age-related frequency of VIs due to Epstein-Barr virus in adolescents and young adults.\(^{10}\)

The DxH 800 provides CPD corresponding to cell volume, conductivity, and 5 scatter measurements, yielding the mean and standard deviation (SD) values for all 7 of these data, for a total of 14 CPDs.

Cell population data corresponding to scatter measurements are obtained by sensors placed at different angles around the flow cell collecting the scattered light of interest. Such scatter angles corresponded to 5.1\(^{\circ}\) (low angle light scatter, LALS), 10\(^{\circ}\) to 20\(^{\circ}\) (lower median angle light scatter, LMALS), 20\(^{\circ}\) to 42\(^{\circ}\) (upper median angle light scatter, UMALS), a channel that is the sum of the UMALS and LMALS regions (median-angle light scatter), and the axial light loss at 0\(^{\circ}\).\(^{11}\)

**Data Set Distribution.**—The 1565 cases from the total data set were distributed into the training, testing, and validation data sets in a 60–20–20 scheme. A total of 1278 cases (81.7%) were randomly assigned into the model-building data set (training-testing), and the remaining 287 cases (18.3%) were randomly assigned to the validation data set, used exclusively during the validation phase.\(^{13}\) The model-building data set was used for model tuning, training-testing, evaluation, and selection.\(^{18,20}\) (Figure).

Subsequently the model-building data set was randomly divided into training and testing data sets, controlled by setting the random state equal to zero.\(^{20}\) Of the 1278 cases from the model-building data set, 958 (75.0%) were included into the training data set, and the remaining 320 cases (25.0%) were included into the testing data set. The training-testing data sets represented 61.2% \((n = 958)\) and 20.4% \((n = 320)\) of the total data sets \((n = 1565)\), respectively.

**Machine Learning Development**

All the ML processes were performed by using 6 different classification algorithms: decision trees (DTs), random forests (RFs), naïve Bayes classifier (NBC), k-nearest neighbor (KNN), neural networks (NNs), and support vector machines (SVMs).\(^{19}\) To build the KNN, SVM, and NN models, the features were previously standardized by using the Scikit-Learn MinMaxScaler transformer.\(^{24}\)

**Model Tuning.**—The selection of the best combination of parameters for each ML algorithm was performed with the Scikit-Learn GridSearchCV, using 10-fold cross (10-FC) validation.\(^{14,19}\) Table 2 details the parameters and the range evaluated for each classifier. For the remaining parameters, default values as described in the Scikit-Learn library were used.

**Model Training-Testing.**—Training-testing accuracies and the variance among them were obtained for each of the 6 ML algorithms to analyze model overfitting and its generalization errors.\(^{18}\)

**Model Evaluation: 10-Fold Cross Validation.**—Model evaluation was performed through 10-FC validation. The model-building data set was randomly split into 10 different subsets (folds); in each fold the model was trained with 9 of the subsets and then evaluated with the holdout subset.\(^{14,20}\)

**Model Selection.**—The selection of the final model for the validation and performance analysis was based on the minimum generalization error from the training-testing analysis and the 10-FC validation accuracy.

**Model Validation and Diagnosis Impact**

The performance of the chosen model was evaluated by using the validation data set \((n = 287)\) by means of the global accuracy, Machine Learning Approach for Lymphocytosis Diagnosis—Bigorra et al
Different phases of the ML approach: problem selection, data collection, ML development (model tuning, training-testing, evaluation, and selection), and finally model validation. Lastly, the diagnosis impact on the 4 different categories is also detailed. Abbreviations: CLL, chronic lymphocytic leukemia; FL, follicular lymphoma; MBL, monoclonal B-cell lymphocytosis; ML, machine learning; NC, normal control; NN, neural networks; PPBL, persistent polyclonal B-cell lymphocytosis; RBC, red blood cell; SMZL, splenic marginal zone lymphomas; VI, viral infection.
Table 1. Population, Hematologic Parameters, and Analyzer Flags Observed for the 10 Different Categories

| Category                              | Age, Mean (SD), y | Sex Ratio, M:F | WBC, Mean (SD), 10^3/μL | ALC, Mean (SD), 10^3/μL | Hb, Mean (SD), g/dL | PLT, Mean (SD), 10^3/μL | Lymphocytosis, % | Atypical Lymphocytes | High-Event Rate in the Differential |
|---------------------------------------|------------------|----------------|-------------------------|-------------------------|---------------------|------------------------|------------------|----------------------|-------------------------------|
| Controls                              |                  |                |                         |                         |                     |                        |                  |                      |                               |
| NC (n = 458)                          | 51.9 (9.8)       | 1:1            | 6.2 (1.1)               | 1.9 (0.5)               | 14.2 (1.1)          | 181.0 (24.8)          | -                | -                    |                               |
| Benign lymphocytosis                  |                  |                |                         |                         |                     |                        |                  |                      |                               |
| VI (n = 346)                          | 22.2 (12.3)      | 1.1:1          | 10.7 (4.2)              | 6.7 (3.5)               | 13.6 (1.4)          | 200.0 (55.4)          | 82.9             | 98.0                 |                               |
| PPBL (n = 107)                        | 47.9 (10.6)      | 1:4            | 9.5 (2.7)               | 4.4 (1.4)               | 13.9 (1.1)          | 201.4 (69.7)          | 69.2             | 86.9                 |                               |
| Postsplenectomy (n = 114)             | 52.8 (13.6)      | 1.4:1          | 12.4 (2.4)              | 5.6 (0.9)               | 14.4 (1.4)          | 323.1 (76.5)          | 100.0            | 14.9                 |                               |
| Group mean                            | 33.2 (12.2)      |                | 10.8 (3.6)              | 6.0 (2.6)               | 13.8 (1.3)          | 225.0 (62.3)          | 83.8             | 79.2                 |                               |
| Neoplastic lymphocytosis              |                  |                |                         |                         |                     |                        |                  |                      |                               |
| CLL (n = 191)                         | 69.8 (10.7)      | 2:1            | 33.4 (34.5)             | 27.7 (32.5)             | 13.4 (1.8)          | 173.9 (59.8)          | 100.0            | 97.9                 |                               |
| MBL CLL-like (n = 90)                 | 69.2 (11.6)      | 1:1.3          | 9.6 (2.7)               | 4.8 (1.4)               | 13.8 (1.4)          | 192.5 (58.2)          | 81.1             | 73.3                 |                               |
| SMZL (n = 87)                         | 72.1 (10.8)      | 1:1.1          | 17.3 (20.3)             | 12.5 (19.6)             | 13.1 (1.8)          | 192.2 (67.8)          | 88.5             | 79.3                 |                               |
| FL (n = 31)                           | 61.0 (15.8)      | 1:2            | 33.8 (42.8)             | 28.4 (41.1)             | 13.0 (1.7)          | 144.9 (55.3)          | 83.9             | 87.1                 |                               |
| Group mean                            | 69.5 (11.3)      |                | 24.5 (24.9)             | 19.3 (23.3)             | 13.4 (1.7)          | 179.8 (60.9)          | 92.0             | 87.5                 |                               |
| Spurious lymphocytosis due to erythrocyte-lysis resistance | | | | | | | | | |
| Anemia (n = 109)                      | 50.1 (20.0)      | 1:1.1          | 8.7 (2.3)               | 4.2 (0.7)               | 10.5 (1.9)          | 322.0 (129.9)         | 89.9             | 54.1                 | 98.17                        |
| Liver disease (n = 32)                | 67.3 (15.7)      | 1:3.1          | 4.7 (3.8)               | 1.8 (0.7)               | 11.4 (2.8)          | 243.3 (130.4)         | 46.9             | 25.0                 | 87.50                        |
| Group mean                            | 54.0 (19.0)      |                | 7.8 (2.7)               | 3.6 (0.7)               | 10.7 (2.1)          | 304.2 (130.0)         | 80.1             | 47.5                 | 95.7                         |

Abbreviations: ALC, absolute lymphoid count; CLL, chronic lymphocytic leukemia; FL, follicular lymphoma; Hb, hemoglobin; NC, normal control; MBL CLL-like, monoclonal B-cell lymphocytosis; PLT, platelet; PPBL, persistent polyclonal B-cell lymphocytosis; SD, standard deviation; SMZL, splenic marginal zone lymphomas; VI, virally infected; WBC, white blood cells.

*a The categories were also grouped into benign, neoplastic, and spurious causes of lymphocytosis, and weighted mean and SD for all cases in each group are also detailed. The percentage of cases showing lymphocytosis is displayed. For adults, lymphocytosis was considered when ALC was above 3.5 × 10^3/μL, while for children the upper limit was considered, taking into account the different normal ranges by age.
approximately the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) using a limited amount of computer memory.

Abbreviations: SD, standard deviation; lbfgs, limited memory BFGS

is an optimization algorithm in the family of quasi-Newton methods that approximates the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) using a limited amount of computer memory.

The recommended diagnostic categories (normal controls, benign, neoplastic, and spurious) were obtained. The confusion matrices for the grouped categories and individual classes were detailed.

Table 2. Results for Model Tuning, Building-Testing, and Evaluation Phases

| Algorithm            | Parameter           | Model Tuning | Model Building-Testing | Model Evaluation |
|----------------------|---------------------|--------------|------------------------|------------------|
|                      | Selected            | Training     | Testing                | 10-Fold Cross    |
|                      |                     | Accuracy, %  | Accuracy, %            | Validation, %    |
|                      |                     |              |                        | SD               |
| Decision tree        | Maximum depth       | 7            | 92.0                   | 14.0             |
| k-nearest neighbor   | K                   | 1            | 100                    | 25.0             |
| Naive Bayes classifier | -                  |              | 86.0                   | 2.0              |
| Random forest        | Number of grown trees | 25       | 100.0                   | 14.0             |
| Support vector machines | Kernel            | Linear     | 95.3                   | 10               |
|                      | C                   | 100          | 85.3                   |                  |
|                      | Gamma               | 0.001        |                        |                  |
| Neural networks      | Hidden layer size   | 100          | 95.0                   | 8.0              |
|                      | Solver              | lbfgs        | 87.0                   |                  |
|                      | Alpha               | 0.1          |                        |                  |

Table 2 also displays the tuning parameters for each classifier and the obtained classification accuracies for the training, testing, and 10-fold cross validation. The variance reflects the difference between the training and testing accuracy. The SD column reflects the variance between the different training testing splits for the 10-fold cross validation.

Table 3 details the confusion matrix of the validation data set into the 4 aforementioned diagnostic categories. The rows display the true diagnosis and the columns the predicted diagnosis by the algorithm. Noteworthy, both normal controls and postsplenectomy patients had TPRs corresponding to 100%. Moreover, all but 5 of the categories presented TPRs above 75.0%. Among benign categories, following postsplenectomy patients, VI variance (ANOVA) with post hoc Tukey honest significance test revealed no statistically significant differences between the obtained 10-FC validation accuracies for the NN, SVM, and RF classifiers. ANOVA was applied as a Gaussian distribution and corroborated by the Shapiro-Wilk test.

Model Selection.—The selected model was the NN classifier, as it presented the highest 10-FC validation accuracy. Even if no statistical differences were found among NN, SVM, and RF classifiers, the former had the lower variance or generalization error (8.0% versus 10%–14%).

Model Validation and Diagnosis Impact

The ML model based on NN rendered classification accuracies of 95.0% for the training set, 87.0% for the testing set, and 83.8% for the 10-FC validation (Table 2). For the validation, 287 new samples were used, and the obtained performance metrics corresponded to a global accuracy of 89.9%, precision of 91.0%, recall of 90.0%, and an F-1 score of 90.0%. The distribution of the validation-set cases grouped into the recommended diagnostic categories (normal controls, benign, neoplastic, and spurious) yielded a global weighted accuracy of 95.8%.

Table 3 displays the confusion matrix of the validation data set into the 4 aforementioned diagnostic categories. The rows display the true diagnosis and the columns the predicted diagnosis by the algorithm. All normal controls were correctly classified within the normal control category, while 93.0%, 96.0%, and 92.3% of benign, neoplastic, and spurious lymphocytosis were correctly classified by the NN model. The overall misclassification was below 5.0%.

Finally, Table 4 presents a detailed confusion matrix with distribution of the 287 cases of the validation data set into the 10 different categories. The values over the diagonal in bold correspond to the true-positive rates (TPRs) for each cell class. The absolute number of cases is shown between parentheses.

Prospectively, a total of 100 cases with lymphocytosis and/or the presence of the atypical lymphoid flag were selected as a POC data set, to evaluate the model functioning beyond the development phase. The morphologic diagnosis was performed by 2 independent observers (1 clinical pathologist and 1 hematologist) as the ground truth. Cell morphology was evaluated with a microscope. The morphologic classification of the cases in the POC data set was performed according to ICSH recommendations for the standardization of peripheral blood cell morphologic features. A comparison between observers and ML diagnosis was performed. Finally, diagnostic category accuracy and weighted global accuracy were also obtained for the POC evaluation.

RESULTS

Machine Learning Development

Model Tuning.—The chosen parameters are detailed in Table 2. No tuning was performed for the NBC and standard settings were used.

Model Training-Testing.—Table 2 also displays the obtained classification accuracies for the training, testing, and 10-FC validation of each particular model. The variance (model overfitting) reflects the difference between training and testing accuracies.

Overall, with the exception of the NBC, the algorithms provided training accuracies superior to 90.0%. Testing results revealed the higher accuracies for NN, RF, and SVM classifiers, respectively, with values above 85%. Finally, the variance revealed that KNN, DTs, and RFs presented the highest overfitting of the training data, with values between 14% and 25%.

Model Evaluation: 10-Fold Cross Validation.—The best model after 10-FC validation corresponded to the NN classifier, with an accuracy of 83.8%, followed by the SVM (83.0%) and RF (79.5%) classifiers (Table 2).
showed a TPR corresponding to 95.1%, while PPBL showed lower outcome (77.8%), with misclassification with SMZL and normal control categories. Among neoplastic categories, CLL and MBL CLL-like presented higher TPRs corresponding to 88.6% and 76.5%, respectively. On the other hand, with the exception of SMZL cases, when misclassification occurred it only took place within the neoplastic category. Finally, the outcome for cases from spurious lymphocytosis showed a 95.0% TPR for anemia with erythrocyte lysis-resistance cases, while the lower outcome for the patients with liver disease (66.7%) implies mostly misclassification into the other spurious category.

For a better analysis, from the 287 cases of the validation set, 86 were normal controls, and the 201 remaining were distributed as follows: 75 corresponded to neoplastic lymphocytosis, 100 to a benign lymphocytosis, and 26 to spurious lymphocytosis. The error analysis revealed that normal control and SMZL categories included most of the incorrect cases from other categories. Still, only 5 of the 201 pathologic cases (2.5%) were incorrectly classified in the normal control category and a total of 5 of the 212 pathologic cases (2.5%) were incorrectly classified in the other spurious lymphocytosis. The higher misclassification rates reflected a total of 3 PPBL cases (16.7%) and 1 liver disease case (16.7%) incorrectly classified as SMZL.

### Table 4. Confusion Matrix of the Neural Network Classification Over the Validation Set

| Predicted Category | NC | VI | Post splenectomy | PPBL | CLL | MBL CLL-like | SMZL | FL | Liver Disease | Anemia |
|--------------------|----|----|------------------|------|-----|-------------|------|----|---------------|--------|
| True category      |    |    |                  |      |     |             |      |    |               |        |
| Normal controls    | 100.0 (86) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) |
| Benign             | 3.3 (2) | 95.1 (58) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 1.6 (1) | 0.0 (0) | 0.0 (0) |
| Neoplastic         | 0.0 (0) | 0.0 (0) | 100.0 (21) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) |
| Spurious           | 5.6 (1) | 0.0 (0) | 0.0 (0) | 77.8 (14) | 0.0 (0) | 0.0 (0) | 16.7 (3) | 0.0 (0) | 0.0 (0) | 0.0 (0) |
| CLL                | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 88.6 (31) | 2.9 (1) | 8.6 (3) | 0.0 (0) | 0.0 (0) | 0.0 (0) |
| MBL CLL-like       | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 11.8 (2) | 76.3 (13) | 11.8 (2) | 0.0 (0) | 0.0 (0) | 0.0 (0) |
| SMZL               | 5.9 (1) | 5.9 (1) | 5.9 (1) | 0.0 (0) | 17.6 (3) | 11.8 (2) | 52.9 (9) | 0.0 (0) | 0.0 (0) | 0.0 (0) |
| FL                 | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 16.7 (1) | 33.3 (2) | 0.0 (0) | 50.0 (3) | 0.0 (0) | 0.0 (0) |
| Liver disease      | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 16.7 (1) | 0.0 (0) | 66.7 (4) | 16.7 (1) |
| Anemia             | 5.0 (1) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 0.0 (0) | 95.0 (19) |        |

Abbreviations: CLL, chronic lymphocytic leukemia; FL, follicular lymphoma; NC, normal control; MBL CLL-like, monoclonal B-cell lymphocytosis; PPBL, persistent polyclonal B-cell lymphocytosis; SMZL, splenic marginal zone lymphomas; VI, viral infection.

*a The values are expressed as percentages, with respect to the total of cases corresponding to the true category per row, with the absolute number of cases between parentheses. The values over the diagonal in bold correspond to the true-positive rates for each cell class.

### Proof of Concept

From the total 100 cases in the POC data set, 36 were diagnosed as benign, 46 as neoplastic, and 18 as spurious lymphocytosis, respectively. The comparison between the observers’ consensus diagnosis and that obtained for the ML model presented an overall concordance of 76%. Considering the diagnostic categories, Table 5 shows the obtained accuracy for each. True positive rates were above 88.0% in the 3 categories and the weighted global accuracy corresponded to 91.0%.

### DISCUSSION

The diagnostic scenario for B-cell neoplasms has changed significantly from the Rappaport classification to the last WHO classification of tumors.22,26 The latter established a starting point, where the diagnosis of lymphoid malignancies is based on objective, standardized, and reproducible markers established by consensus of a panel of experts. However, some entities remain within a grey zone owing to morphologic or immunophenotypic overlap.26

Although stand-alone morphologic classification has some limitations, it still serves as the starting point for the diagnosis approach and triggers other ancillary confirmatory techniques. Our proposed ML model is also based on morphologic features, as CPD are automated data from the physical properties of the different leukocyte subpopula-
tions, using a combination of impedance (for cell size), conductivity (for nuclear density, granularity, and nucleucytoplasm ratio), and scatter measures (for structure, shape, and reflectivity on the cell). Nevertheless, these measures are objective, reproducible, and importantly, standardizable.

The ultimate aim of this work was to evaluate the diagnostic impact of an ML model on the lymphocytosis diagnosis approach into the different categories. For this purpose, we evaluated the performance of the models, but also their feasibility for their actual use in the clinical laboratory, for which the input data must be standardizable and subject to rigorous quality controls (QCs) in a quality regulation framework.

In terms of performance, the ultimate goal in the fine-tuning of an ML model is the achievement of the Bayes optimal error, generally equated to human-level performance. Considering the results obtained from EQA schemes as a measure of human-level performance, they were comparable with those achieved by our ML model in the detection of VI (64.3%–86% versus 95.1%) and neoplastic CLL and FL (CLL: 77%–93.1% versus 88.6%; FL: 56.7%–57.9% versus 50%) categories. Error analysis of our ML model highlighted that misclassification among the categories was as expected given the morphologic overlap amongst them. Such similarities included those observed among CLL and MBL CLL-like categories, as the differentiation amongst them is based on the size of the clonal population, and those observed between FL and SMZL with CLL. The overlap between FL and CLL could be explained by the atypical CLL morphology (associated with trisomy 12) with more than 15% of cells with cleaved nucleus resembling follicular lymphoma cells, and the presence of lymphoplasmacytic morphology, as found in SMZL. Similar misclassification rates were also found for FL as CLL in the results of EQA schemes. Besides, the resembling morphologic patterns of PPBL and SMZL explain the observed overlap between these 2 categories.

Of note, it is important to highlight that 100% of MBL CLL-like cases were classified within the neoplastic category, even with mean ALC as low as 4.8 x 10⁹/L and with 26.7% of the cases without the presence of the atypical lymphoid flag.

Secondly, we evaluated certain aspects of the feasibility of the model. Our proposed ML model uses numerical data from the Beckman Coulter DxC 800 analyzer. Although CPD have no specific QC yet, the VCS technology that provides them in the analyzer has already been used for other reported parameters, and hence subjected to QC, making possible its inclusion with specific analytical studies and proper software modifications, as in the case of the Monocyte Distribution Width, previously evaluated and proper software modifications, as in the case of the model. Our proposed ML model uses numerical data

The model validation revealed global, weighted, validation accuracy for the different diagnostic categories, corresponding to 95.8%, and a 91.0% accuracy in the POC. An accuracy above 90% is remarkable in view of the difficult diagnostic problem achieving, or even outperforming, the human-level performance reported by the EQA schemes. The main limitations affecting any ML model are related to data and interpretability/explainability of the model. In our model, data limitations could affect the mantle cell lymphoma category because of unavailability of the reference standard for its confirmation, but also because of the type of features included to feed the algorithm, which are based on morphologic parameters, with known grey areas, especially to differentiate among B-cell neoplasms. As usual, this model also was built with data from a single institution and the exchange of data between several institutions is mandatory for further implementation. Regarding model interpretability, as Aristotle noted: “the ability to explain how results are produced can be less important than the ability to produce such results and empirically verify their accuracy.” Thus, the empirical verification of an ML model presupposes a real-world prospective evaluation to assess its performance, using truly independent data sets to capture the variety of the real-world conditions. Moreover, an important part of the explainability of a model goes through a data transparency policy, as ML models involve not only algorithms but also data, and the same algorithm could lead to different outcomes when trained with different data. For such reasons, detailed information on the characteristics of the included groups was fully detailed for the readership to determine if the data set was appropriate.

To conclude, the proposed ML model not only achieves comparable human-level performance, but also meets specific requirements for its feasibility, such as high-quality labelled data in terms of reference standard (flow cytometry), and standardized and objective features following calibration and QC procedures. Of note, the application of the model has a high benefit-cost ratio, as the data required are obtained from the complete blood count with differential. The diagnostic impact with high accuracies in model validation and the performed POC provide encouraging possibilities for taking the model to the next level and training it with more categories and cases, and with a multicentric perspective, for its deployment for real-world application on a daily basis. Importantly, the application of ML requires a structure of regulatory and legal frameworks to ensure safe delivery of medical agents or other medical devices, being an artificial-intelligence–based technology defined as “Software as a Medical Device” by both the FDA and the International Medical Device Regulators Forum. Hopefully, after proper verification and regulatory process, an ulceror version of this model could serve as a diagnostic support tool, which could be integrated into the laboratory workflow, where the specialist’s responsibility would be to manage the extracted information by the ML model in the clinical context of the patient as a physician-in-the-loop or in symbiosis with ML models, in the path to achieving high-performance medicine.
