J. Mantas et al. (Eds.)

doi:10.3233/SHTI230483

Machine Learning for Diagnosis and Screening of Chronic Lymphocytic Leukemia Using Routine Complete Blood Count (CBC) Results

Regina PADMANABHAN [a,1], Yousra EL ALAOUI [a], Adel ELOMRI [a], Marwa K. QARAQE [a], Halima EL OMRI [b], Ruba YASIN TAHA [b]
[a] *College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar*
[b] *Medical Oncology-Hematology Department, National Centre for Cancer Care and Research (NCCCR), Hamad Medical Corporation (HMC), Doha, Qatar*

Abstract. The comprehensive epidemiology and global disease burdens reported recently suggest that chronic lymphocytic leukemia (CLL) constitutes 25-30% of leukemias thus being the most common leukemia subtype. However, there is an insufficient presence of artificial intelligence (AI)-based techniques for CLL diagnosis. The novelty of this study is in the investigation of data-driven techniques to leverage the intricate CLL-related immune dysfunctions reflected in routine complete blood count (CBC) alone. We used statistical inferences, four feature selection methods, and multistage hyperparameter tuning to build robust classifiers. With respective accuracies of 97.05%, 97.63%, and 98.62% for Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), and XGboost (XGb)-based models, CBC-driven AI methods promise timely medical care and improved patient outcome with lesser resource usage and related cost.

Keywords. Chronic lymphocytic leukemia, routine blood test, machine learning

1. Introduction

CLL is a blood cancer that involves the progressive accumulation of aberrant lymphocytes in the bone marrow leading to immune dysfunction. As per the facts and figures so far, the burden of CLL is higher in males, elderly adults, and countries with high socio-demographic indexes [1]. In fact, the wide range and various morphologies of CLL cells make accurate CLL identification a nontrivial task for hematopathologists. In addition, approximately 60% of CLL patients are asymptomatic at the time of diagnosis, while symptomatic patients present with vague signs and symptoms. Presently, hospitals worldwide are undergoing a digital revamping to accommodate the smart options provided by artificial intelligence (AI) breakthroughs. For instance, smart diagnostic gadgets such as CellaVision (digital image analyzer) and Morphogo (AI-based bone marrow image analyzer), has been introduced in a hospital setting to

[1] Corresponding Author: Regina Padmanabhan, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar; Email: rpadmanabhan@hbku.edu.qa.

streamline the workflow and effectively reduce the turnaround time involved in hematological disease diagnostic pathway [2]. Thus far, the current CLL diagnosis techniques are mainly based on blood smear images, flow cytometry and immunophenotyping of circulating B lymphocytes. Machine learning (ML) techniques such as Support Vector Machine (SVM) and K-nearest neighbor (KNN) algorithms were used to identify CLL by image color segmentation [3]. Similarly, the least absolute shrinkage and selection operator (LASSO) regression on gene expression datasets was used to determine the six most prevalent diagnostic biomarkers for CLL [4]. While the aforementioned applications achieved very good performances, using imaging techniques for disease classification and abnormality detection is deemed time-consuming and computationally expensive. Furthermore, many AI methods reported so far in the context of CLL mainly investigate disease prognosis after CLL diagnosis [5]. Hence, reviews point out that the applicability of AI in the diagnosis of CLL is the least explored areas in hematology management, wherein further research is essential [6]. To bridge this gap, we demonstrate in our paper the potential of a ML algorithm to leverage the power of simple blood count test outcome in classifying patients with chronic lymphocytic leukemia (CLL).

2. Methods

Dataset for CLL diagnosis: Study data included a total of 682 CBCs records, 88 confirmed CLL (label 1) and 594 control or benign (label 0) records collected from the National Center for Cancer Care and Research (NCCCR) at Hamad Medical Corporation (HMC) in Doha, Qatar, during 2016 to 2020 (IRB approved). The first column of Table 1 shows the 20 candidate parameters of routine CBC extracted from hospital records.

Building a robust CLL diagnostic model: First, we identified relevant CBC parameters that represent CLL characteristics the most and at the same time allow discrimination of CLL from benign cases. As shown in Table 1, we looked at the group-wise mean and standard deviation of each CBC parameter, p-value, correlation coefficient and feature ranks of all the parameters to find the final feature set. The last column in Table 1 shows the final 11 features selected. We selected the top three models (LR, LDA, and XGB) out of 8 candidate models (KNN, SVM, LR, DT, GNB, RF, XGB, QDA) based on the accuracy of classification using 5-fold cross-validation (Table 2).

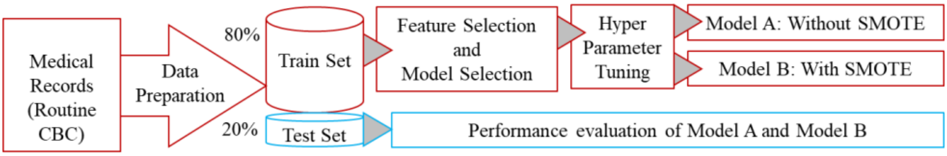
Medical Records (Routine CBC)
Data Preparation
80%
20%
Train Set
Test Set
Feature Selection and Model Selection
Hyper Parameter Tuning
Model A: Without SMOTE
Model B: With SMOTE
Performance evaluation of Model A and Model B

Figure 1. Workflow diagram used to build the proposed CLL diagnostic model. Only training data (label 0:473, label 1:71) were used for feature selection and model building and then the model was validated on two unseen data sets, a balanced set (labels 0,1=17, 17) and an imbalanced set (labels 0,1=121, 17).

Next, we used grid search and hyper opt parameter search algorithms in Scikit-learn (Python) to finalize the model parameters of the selected XGboost model. To improve accuracy and overcome the drawbacks of an imbalanced dataset, data augmentation (SMOTE-Synthetic Minority Over-sampling Technique) was added to derive Model B (Figure 1). Finally, set aside test sets were used to evaluate model performance.

Table 1. Group-wise statistical values and feature ranks for CBC parameters, where CO: Correlation with Outcome. We exploited feature selection characteristics of chi-square (robustness), mutual information (entropy reduction), extra tree classifier (random selection), and XGboost classifier (gradient boosting) to derive cumulative feature rank (FR).

Feature	**Control** **(μ (std.))**	**CLL** **(μ (std.))**	***p*-value**	**CO**	**FR**
WBC (white blood cell)	6.44 (1.6)	53.3 (78.1)	<0.05	**0.49**	1
RBC (red blood cell)	4.84 (0.6)	4.48 (0.94)	<0.05	-0.17	-
Hgb (hemoglobin)	13.13 (1.5)	12.5 (2.3)	<0.05	-0.12	-
HCT (hematocrit)	39.9 (4.2)	38.6 (6.5)	**0.028**	-0.094	-
MCV (mean corpuscular volume)	83.1 (7.7)	87.2 (9.2)	<0.05	0.17	-
MCH (mean corpuscular hemoglobin)	27.4 (3.1)	28.3 (3.87)	**0.014**	0.1	-
MCHC (MCH concentration)	32.9 (1.3)	32.5 (1.7)	0.013	-0.11	-
RDW-CV (red cell distribution width)	14.3 (2.3)	15.1 (2.5)	<0.05	0.11	-
Platelets	270.0 (74.0)	179.7 (91.4)	<0.05	**-0.37**	7
MPV (mean platelet volume)	10.3 (1.2)	9.69 (1.7)	<0.05	-0.16	-
ANC (absolute neutrophil count)	3.61 (1.3)	4.8 (3.3)	<0.05	**0.24**	6
Lymphocyte count	2.1 (0.3)	42.7 (67.1)	<0.05	**0.49**	2
Monocyte count	0.51 (0.17)	3.98 (13.4)	<0.05	**0.24**	9
Eosinophil count	0.17 (0.15)	0.29 (0.32)	<0.05	**0.21**	11
Basophil count	0.04 (0.03)	0.11 (0.09)	<0.05	**0.44**	10
Neutrophil %	54.86 (10.0)	20.2 (16.8)	<0.05	**-0.72**	3
Lymphocyte %	33.6 (9.1)	72.4 (18.0)	<0.05	**0.77**	4
Monocyte %	8.03 (2.1)	5.60 (5.84)	<0.05	**-0.27**	5
Eosinophil %	2.78 (2.19)	1.18 (1.4)	<0.05	**-0.25**	-
Basophil %	0.72 (0.3)	0.34 (0.3)	<0.05	**-0.35**	8

3. Results and Discussion

The challenges posed by missing values (0.8%) in the train set and small sample size (71 -CLL) on the model building were tackled by using KNN imputation and data augmentation, respectively. As shown in Table 2, even though there is only marginal improvement in the accuracy, recall (sensitivity) has improved by 5.88% with SMOTE.

Table 2. Results of model selection trials when top three (QDA, LR, XGb) models were further analyzed with 20 and 11 CBC features. As there is no considerable gain with additional features, XGboost model with 11 features were selected as final model (*learning_rate*=0.12, *booster*=gbtree, *number of estimators*=150). Performance with and without data augmentation is also given.

	Model Selection		
Model	Performance (20 Features)	Performance (11 Features)	
Quadratic Discriminant Analysis (QDA)	95.71%	97.05%	
Logistic Regression (LR)	98.53%	97.63%	
XGboost (XGb)	98.71%	98.62%	
Performance of XGboost model (11 features) with and without SMOTE for various test sets			
Metrics	Model A validated on balanced test set (label 1: 17, label 0: 17)	Model B validated on balanced test set (label 1: 17, label 0: 17)	Model B, imbalanced test set (label 1: 17, label 0: 121)
Accuracy	97.05%	97.05%	98.55%
FI-Score	96.96%	97.14%	94.44%
Recall	94.12%	100%	100%
AUC	100%	100%	100%

Hence, we hypothesize that Model B will remain robust for a diverse set of datasets. Clinical implications of early leukemia diagnosis includes (i) lesser treatment cost and drug toxicity (ii) improved treatment response and survival rate, (iii) chance of cure or management of disease with the use of easily available first-generation drugs, and (iv) amble time for chromosomal analysis to device personalized treatment. Even though further confirmatory cytomorphological and cytogenetic analysis are required, preliminary screening with proposed smart assistive devices ensures that only very likely suspects are sent for such more invasive tests. Thus, reducing congestion in the hematological department due to over-referral from general hospitals, overall hospital burden and patient anxieties. As all models exhibited acceptable performance, we foresee the use of the proposed kind of practice-changing smart clinical assistive devices in reducing delay in leukemia diagnosis.

4. Conclusions

We show that our rigorous feature selection predicated on both statistical inferences and data-driven learning outcomes along with the multistage hyperparameter tuning approach (grid search and hyperopt) has resulted in a high-performance model (Accuracy=98.55%). These novel results with simple CBC test will accelerate the integration of AI-enabled smart diagnostic devices in hematological disease diagnosis. The unavailability of extensive clinical data may limit the capacity of this model to capture all biological variabilities of CLL and hence questions the generalizability of this model. Hence, robustness of the proposed model will be evaluated further by end of 2023 using upcoming CBC test results from NCCCR, Qatar.

Acknowledgement

This work was supported by the National Priorities Research Program -Standard (NPRPS) Twelfth (12th) Cycle grant# NPRP12S-0219-190108, from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work, and are solely the responsibility of the author[s].

References

[1] Yao Y, Lin X, Li F, Jin J, Wang H. The global burden and attributable risk factors of chronic lymphocytic leukemia in 204 countries and territories from 1990 to 2019: analysis based on the global burden of disease study 2019. BioMedical Engineering OnLine. 2022;21(1):1-22.

[2] Lin E, Fuda F, Luu HS, Cox AM, Fang F, Feng J, et al., editors. Digital pathology and artificial intelligence as the next chapter in diagnostic hematopathology. Seminars in Diagnostic Pathology; 2023: Elsevier.

[3] Mohammed EA, Far BH, Mohamed MM, Naugler C, editors. Application of support vector machine and k-means clustering algorithms for robust chronic lymphocytic leukemia color cell segmentation. 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services 2013. IEEE.

[4] Zhu Y, Gan X, Qin R, Lin Z. Identification of Six Diagnostic Biomarkers for Chronic Lymphocytic Leukemia Based on Machine Learning Algorithms. Journal of Oncology. 2022;2022.

[5] Agius R, Brieghel C, Andersen MA, Pearson AT, et al. Machine learning can identify newly diagnosed patients with CLL at high risk of infection. Nature communications. 2020;11(1):363.

[6] El Alaoui Y, Elomri A, Qaraqe M, Padmanabhan R, Yasin Taha R, El Omri H, et al. A review of artificial intelligence applications in hematology management: Current practices and future prospects. Journal of Medical Internet Research. 2022;24(7):e36490.