Robust Detection of Chronic Lymphocytic Leukemia with Support Vector Machines and Flow Cytometry

Barış Boral[1*]

1 Department of Immunology, Ankara Oncology Training and Research Hospital, Ankara, Türkiye

Abstract

Aim: Our aim is to build a precise automatic tool for the diagnosis of CLL with the help of machine learning algorithms and flow cytometry immunophenotypic data.

Methods: We run experiments with two machine learning methods. First one is decision tree which was previously used in other similar works and second one is support vector machines which is considered to be a more robust classification method.

Results: Among the 40 CLL patients from the test set, the model correctly predicts 38 of them and among the 20 other B-CLPD patients, the model predicts 18 of them correctly. Its sensitivity, which is the fraction of true positive predictions among all positive samples, is 95% (38/40).

Conclusion: The model achieves very high accuracies on our leave out test set. This model can be a useful tool for automatic CLL diagnosis.

Keywords: *CLL, flow cytometry, machine learning*

1. Introduction

Chronic lymphocytic leukemia (CLL) is one of the most common types of adult leukemia in Western countries. It occurs more during or after middle age compared to childhood[1, 2]. Its diagnosis is based on parameters from blood counts, differential counts, a blood smear, and immunophenotyping[2].

Flow cytometry is used to assist the diagnosis and monitoring of malignant hematopoietic myeloid and lymphoid tumors. Furthermore, they are the most informative tests to confirm a diagnosis of CLL. In the diagnosis of CLL, CD5, CD19, CD20, CD23, and surface or cytoplasmic kappa and lambda light chains are regarded as essential markers. However, there are difficulties in differentiating the CLL diagnosis from the other B-cell chronic lymphoproliferative disorders (B-CLPD), because those markers can be seen in other B-CLPD as well[3]. For example, the CD5 expression can also be seen in other lymphoid malignancies, such as mantle cell lymphoma and the expression of CD23 can be observed in marginal zone lymphoma. In other words, those markers are not specific to CLL. Because of that, other markers, CD10, CD43, CD79b, CD81, FMC7 and CD200 can be useful in discriminating the CLL diagnosis from the other B-CLPD[3]. Scoring systems are also available to differentiate the CLL from the other B-CLPD and the most frequently used one is Matutes score[4]. The Matutes score system looks at the positivity of CD23 and CD5 and the absence or poor presence of CD79b (or CD22), FMC7 and SmIg. It converts those observations to a 0-5 numerical score. If the score ends up greater than 3, the subject is classified as a CLL patient [4]. However, this scoring system does not perfectly differentiate diagnosis between CLL and other B-CLPD [3]. B-CLPD neoplasms are challenging to diagnose due to their overlapping clinical features as mentioned before[3]. Machine learning algorithms have become useful tools in classification tasks. They learn from data and remove the need for hand-designed heuristics. As they find applications in many domains, healthcare providers also recognize their capabilities and use medical clinical decision algorithms based on a set of decision rules to improve diagnosis with reduced cost[5]. A machine learning approach to differentiate the CLL diagnosis from the other B-CLPD may reduce the cost, speed-up the diagnosis, and may even improve the accuracy of the correct diagnosis. Previous studies propose a decision tree in differential diagnosis of lymphoproliferative diseases[6].

In this study, we aim to develop an easy, precise, automatic tool for the diagnosis of CLL with the help of machine learning methods learned from flow cytometry immunophenotypic data.

* Corresponding Author: Barış Boral
e-mail: boralbaris@gmail.com
Received: 14.08.2023, Accepted: 22.08.2023, Available Online Date: 31.08.2023
Cite this article as: Boral B. Robust Detection of Chronic Lymphocytic Leukemia with Support Vector Machines and Flow Cytometry. J Cukurova Anesth Surg. 2023; 6(2): 324-6. doi: 10.36516/jocass.1342711
Copyright © 2023 This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CC-BY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

2. Materials and methods

2.1. Study Groups

The data of patients diagnosed with B lymphoproliferative disease between 2020-2023 were retrospectively analyzed. A total of 152 patients' data were analyzed. Among the analyzed patients, 108 patients were diagnosed with CLL and 44 patients with other B-CLPD. Among the patients diagnosed with other B-CLPD, 18 patients were diagnosed with mantle cell lymphoma, 14 patients with marginal zone lymphoma, 4 patients with splenic marginal zone lymphoma, 5 patients with hairy cell lymphoma, 2 patients with follicular lymphoma, and one patient with burkitt lymphoma.

The diagnosis of B-CLPD was made according to the most recent revision of the WHO classification of lymphoid neoplasms, released in 2022, based on clinical data and morphological, immunophenotypic, and genetic criteria[4].

2.2. Flow Cytometry

Peripheral blood samples were taken into tubes containing EDTA. Samples were prepared and analyzed with a flow cytometer immediately after collection. Then, immunophenotypic analysis was performed using monoclonal antibodies; CD19, CD5, CD23, CD81, FMC7, CD22, CD43, CD103, CD11c, CD123, HLA-DR, CD10, CD38, CD25, CD200, CD79b and CD20.

2.3. Machine Learning Set-up

We run experiments with two machine learning methods. First one is decision tree which was previously used in other similar works[6] and second one is support vector machines which is considered to be a robust classification method. Decision tree algorithm considers each attribute of the data to pick the best one that will result in branched out nodes with the most purity. We use the Gini index to measure the purity of the nodes. If a node is pure, that means all samples that fall into that node share the same class. They are considered unstable classifiers as new samples in the training set may cause the structure of the decision tree to change drastically.

On the other hand, support vector machines find a decision boundary between the positive and negative classes with the largest margin. Since the margin between the negative and positive samples are maximized, they are considered to be more stable and robust.

We use CLL as our positive class and other B-CLPD as our negative class for both of these methods and learn a model for this binary classification problem. Our dataset as mentioned in the Study Groups section includes 108 CLL patients and 44 other B-CLPD patients. We use 40 patients from CLL group and 20 other B-CLPD patients for our test set in which we evaluate our models. The rest of the data is used for the training. For each of these data points, we consider the 17 immunophenotypic attributes mentioned in the Flow Cytometry subsection.

For the decision tree and support vector machine algorithms, we use Scientific computing tools (version 1.2.3, SciPy.org) for Python (version 2.7.16, Python.org).

3. Results

In this section, we report our prediction results of the decision tree and support vector machine models. Firstly, the decision tree method serves as a baseline. It achieves 93.3% accuracy on the test set. Its confusion matrix is shown in Table 1. Among the 40 CLL patients from the test set, the model correctly predicts 38 of them and among the 20 other B-CLPD patients, the model predicts 18 of them correctly. Its sensitivity, which is the fraction of true positive predictions among all positive samples, is 95% (38/40). Its specificity, which is the fraction of true negative predictions among all negative samples, is 90% (18/20).

Next, we evaluate the support vector machine model which is considered to be a more powerful and robust method. It achieves 98.3% accuracy. Table 2 shows the confusion matrix of the model. The support vector machine model classifies all 20 other B-CLPD patients correctly and among the CLL patients the model correctly predicts 39 out of the 40 patients. The sensitivity of this model is 100% (39/39) and specificity is 95.23% (20/21).

Note that both models perfectly classify the training data they learn from as shown in Table 1 and 2.

Table 1

Confusion matrix of decision tree method on the training and test set.

		Train Dataset		Test Dataset	
		Ground Truth		Ground Truth	
		CLL	Other B-CLPD	CLL	Other B-CLPD
Predictions	CLL	68	0	38	2
	Other B-CLPD	0	24	2	18

Chronic lymphocytic leukemia (CLL), Other B-cell chronic lymphoproliferative disorders (B-CLPD).

Table 2

Confusion matrix of support vector machine method on the training and test set

		Train Dataset		Test Dataset	
		Ground Truth		Ground Truth	
		CLL	Other B-CLPD	CLL	Other B-CLPD
Predictions	CLL	68	0	39	0
	Other B-CLPD	0	24	1	20

Chronic lymphocytic leukemia (CLL), Other B-cell chronic lymphoproliferative disorders (B-CLPD).

4. Discussion

Flow cytometry tests are very important diagnostic tools in B-CLPD, especially in CLL[3]. The Matutes score that was mentioned in the introduction has been used for more than 20 years[4]. However, detecting CLL is not a solved task due to the ambiguous immunophenotypes[7]. To overcome this challenge Vergnolle et al[5] has developed a decision tree that enables the differentiation of CLL from non-CLL cases. Özdemir et. al.[8] also created a similar decision tree with sensitivity of 97.78% and specificity of 93.33%. We also set a decision tree model which achieved sensitivity of 95% and specificity of 90%. The results may differ because different datasets are used but these results serve as a baseline for our comparison. The support vector machine from our experiments has sensitivity of 100% and specificity of 95.23% and achieves better results than the decision tree baseline.

Decision tree models are interpretable. They are easily built and computationally efficient. However, they are simple models and considered to be not stable. In this work, we show that with support vector machine models, better results can be obtained. Those results show that automatic tools can be considered for the diagnosis of CLL.

Interpreting flow cytometry tests is a difficult task that requires experts in the field. Performing these tests by unauthorized persons may result in misdiagnoses and incorrect or unnecessary drug use. With these newly developed methods, it is aimed to prevent such errors.

In this study, we collect a dataset to learn a classification model of CLL and other B-CLPD classes using immunophenotyping with flow cytometry. We train a machine learning model specifically a Support Vector Machine which is a robust classification method. The model achieves very high accuracies on our leave out test set which shows that it can be a useful tool for automatic CLL diagnosis.

Acknowledgements

None.

Statement of ethics

This was a retrospective and single-center study which was approved by the Ankara Oncology Training and Research Hospital local Ethics Committee and was conducted in accordance with the Declaration of Helsinki. (AEŞH-EK1-2023-472)

Conflict of interest statement

The authors declare that they have no financial conflict of interest with regard to the content of this report.

Funding source

The authors received no financial support for the research, authorship, and/or publication of this article.

Author contributions

Author read and approved the final manuscript.

References

1.Mato A, Jahnke J, Li P, et al. Real-world treatment and outcomes among older adults with chronic lymphocytic leukemia before the novel agents era. Haematologica. 2018; 103(10): 462-5.
https://doi.org/10.3324/haematol.2017.185868
2.Hallek M. Chronic lymphocytic leukemia: 2020 update on diagnosis, risk stratification and treatment. Am J Hematol. 2019; 94(11): 1266-87.
https://doi.org/10.1002/ajh.25595
3.Alaggio R, Amador C, Anagnostopoulos I, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms [published correction appears in Leukemia. 2023 Jul 19;:]. Leukemia. 2022;36(7):1720-48.
https://doi.org/10.1038/s41375-022-01620-2
4.Matutes E, Owusu-Ankomah K, Morilla R, et al. The immunological profile of B-cell disorders and proposal of a scoring system for the diagnosis of CLL. Leukemia. 1994; 8(10): 1640-5.
5.Vergnolle I, Ceccomarini T, Canali A, et al. Use of a hybrid intelligence decision tree to identify mature B-cell neoplasms. Cytometry B Clin Cytom. 2023; 10.1002/cyto.b.22136.
https://doi.org/10.1002/cyto.b.22136
6.Moraes LO, Pedreira CE, Barrena S, et al. A decision-tree approach for the differential diagnosis of chronic lymphoid leukemias and peripheral B-cell lymphomas. Comput Methods Programs Biomed. 2019; 178: 85-90.
https://doi.org/10.1016/j.cmpb.2019.06.014
7.Frater JL, McCarron KF, Hammel JP, et al. Typical and atypical chronic lymphocytic leukemia differ clinically and immunophenotypically. Am J Clin Pathol. 2001; 116(5): 655-64.
https://doi.org/10.1309/7Q1J-1AA8-DU4Q-PVLQ
8.Ozdemir ZN, Falay M, Parmaksiz A, et al. A novel differential diagnosis algorithm for chronic lymphocytic leukemia using immunophenotyping with flow cytometry. Hematol Transfus Cell Ther. 2023; 45(2): 176-181.
https://doi.org/10.1016/j.htct.2021.08.012