Comparison of Different Machine Learning Algorithms to Predict the Diagnostic Accuracy Parameters of Celiac Serological Tests

Özgül ÖZER1, Nazlı ARDA2

1Corresponding Author; Institute of Graduate Studies in Science, Istanbul University Turkey; ozgul.ozer@ogr.iu.edu.tr
2 Department of Molecular Biology and Genetics, Faculty of Science, Istanbul University, Turkey; narda@istanbul.edu.tr

Abstract

Celiac disease; is an autoimmune digestive system disease characterized by chronic intestinal inflammation and villus atrophy affecting genetically predisposed individuals. Diagnosis is based on serological tests and small bowel biopsy. Because of the diversity in the clinical features of the disease, various patient profile and the non-standardized serological tests, it is difficult to diagnose the celiac disease. Sensitivity, specificity, positive and negative predictive values are important parameters for the accuracy of the tests and they are missing in some clinical studies. It is difficult do standardize the tests with these missing values for clinicians. The aim of this study is to train different machine learning algorithms and to test their performance in prediction of the diagnostic accuracy parameters of celiac serological tests. Decision trees are effective machine learning algorithms for predicting potential covariates with %88.7 accuracy.

Keywords: machine learning, diagnostic test accuracy, CAD diagnosis of celiac disease, celiac serological tests

1. Introduction

Celiac disease (CD) is the inflammation of the small intestine caused by dietary gluten in genetically predisposed individuals. The incidence of the disease is %1 in most countries. Patients are required to follow a gluten-free diet life-long. Nutritional can not be absorbed sufficiently as the result of villus atrophy [1]. While the symptoms of the disease are similar to many other diseases and these symptoms are different in each individual, it is very difficult to diagnose. %80-90 of the patients are still under-diagnosed while only %10 of the patients know that they have celiac disease [2]. In a serologic screening research, involving more than 17,000 Italian schoolchildren, the ratio of individuals who know their disease to those who do not know is 1/7 [3,4].

Although the patient profile with celiac disease may be variable, serological tests are a cheap and non-invasive method for clinicians to identify the disease. The usage of serological tests has also been suggested for the follow up of patient dietary compliance. Antibodies against to gluten proteins in the foods and to structural proteins in intestinal mucosa (endomisium, reticulin, transglutaminase) are the targets of the tests. In 1960s, it is found that the gliadin componds in wheat are involved in the pathogenesis of the disease. Anti-gliadin antibodies (AGA) are the first autoantibodies used in the diagnosis of celiac disease and then anti-endomisium (EMA) antibodies began to be used in the diagnosis at 1980s. Endomisium is a structural protein of intestinal tissue.

It is not recommended to use Anti-endomisium antibodies in patients with mild bowel lesions (Marsh 3A) and children under 2 years of age. In 1990s, the role of the is tissue transglutamase (tTG) enzyme in celiac pathogenesis is well understood and tTG antibody tests are became very popular at diagnosis. [5]. We can use Anti-gliadin antibodies (AGA) for screening aims while anti-tissue transglutaminase (dTG) and anti-endomysium (EMA) autoantibodies are giving better results at diagnosis and patient follow-up [6].

Diagnostic test accuracy determines if the test identify the target situation accurately. There are some parameters like sensitivity, specificity, likelihood ratios, Youden's index which tells us the diagnostic
accuracy of tests. These parameters can be calculated from $2 \times 2$ contingency table that includes the number of true-positive, true-negative, false-positive, and true-positive test results. Sensitivity is the ratio of individuals correctly identified with target situation. A test with 100% sensitivity means all diseased individuals are correctly identified. There are no false negatives. These parameters differ between analysis. Specificity and sensitivity of some assays are lower than expected in some clinical applications [7,8,9].

Machine learning is extensively applied in the field of medical informatics, including gene and protein structure prediction, genome analysis, drug discovery, text mining and image processing. There are limited number of studies about the prediction of diagnostic test accuracy parameters using machine learning algorithms [10,11].

Machine learning workflows are complex and difficult to understand since the accuracy of the algorithms is distinct from each other. Decision trees provide high classification accuracy and can be used in different areas of medical decision making. Simple decisions are used for prediction consecutively in decision tree algorithms. Bayesian classifier is also one of the most useful and effective predictive data mining method. Naive Bayes models uses the method of maximum likelihood for parameter estimation in practical applications. A family of algorithms based on a common principle are used for training instead of a single algorithm [12,13]. Random forests have been successfully used in classification, regression and clustering tasks. Boosting is also a flexible nonlinear regression procedure that helps improving the accuracy of trees [14,15].

KNIME Platform is a very useful tool for applying machine learning algorithms for beginners without coding background. Procedures like clicks, drags, and drops can be followed easily. This paper describes the overall process of applying different machine learning algorithms via the KNIME analytics platform in a simple way [16].

2. Material and Method

2.1. Dataset and Data Preprocessing

The Pubmed database was searched (January 2000 - January 2022) for clinical studies assessing the accuracy of celiac serological tests. 80 Studies including sensitivity, specificity, positive and negative predictive values were included. We processed and analyzed the data using the Konstanz Information Miner (KNIME) analytics platform. The procedures to install KNIME extensions were followed. After installing Knime extensions, we created the Knime workflow. Datasets are transferred to Knime workflow with CSV reader node. The input table is split into two partitions (%70 train dataset, %30 test dataset) with partitioning node as shown in Figure 1-2. Sensitivity was designed as target value since there is a correlation between sensitivity and the other values.
2.2. Applying Machine Learning Algorithms

4 different machine learning algorithms are used after partitioning. Decision tree learner, Naïves Bayes learner, random forest learner, gradient boosted trees learner nodes are trained with training datasets while predictors nodes made predictions with test datasets. Scorer nodes calculated and represent the accuracy statistics as shown in Figure 3.
3. Results

Accuracy values of sensitivity predictions are; %88 for decision tree predictor, %70 for naive bayes predictor, %100 for random forest predictor and %71 for gradient boosted trees predictor as shown in Figure 4-7.

Decision tree predictor node provided highest Cohen’s kappa value with 0.87 while naive bayes predictor node the lowest value with 0.67. Decision tree predictor provided the lowest error rate with 0.1 while naive bayes predictor calculated the highest error value.
Data mining approaches have been successfully applied to different practical problems not only in clinical medicine but also in epidemiological studies and meta-analysis. These approaches can offer predictions for missing parameters which are in fact not ignorable in meta-analyses and systemic reviews. Machine Learning algorithms can highlight the gaps in the evidence based medicine by predicting potential covariates [17,18].

%100 accuracy of random forest predictor in this study, can be explained with overfitting and the small number of sample size. Decision tree predictor which provides %88.7 accuracy can be used as an effective machine learning algorithm for predicting potential covariates for missing values in meta analyses.
References