SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 5, NO. 1, APRIL 2022
DOI: 10.35377/saucis.05.01.1094043

Research Article

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

Özgül ÖZER[1], Nazlı ARDA[2]
[1]Corresponding 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

Received 27 March 2022; Revised ; Accepted 4 April 2022; Published online 30 April 2022

Abstract

Celiac disease; is an autoimmune digestive system disease characterized by chronic intestinal inflammation and villus antrophy 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 clinicial 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 accurary 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. Nutritionals can not be absorbed sufficiently as the result of villus antrophy [1]. While the symptoms of the disease are similar to many other diseases and these symptoms are diffirent 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 useage of serological tests has also been suggested for the follow up of patient dietary compliance. Antibodies againts 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 × 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 clinicial 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 usefull tool for applying machine learning algoriths for beginners without coding backround. 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 clinicical studies assesing the accuracy of celiac serological tests. 80 Studies including sensitivity, specificity, positive and negative predictive values were included. We processed and analyzed the datas 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 transfered to Knime workflow with CSV reader node. The input table is splited 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.

File Table - 17:7 - CSV Reader

File Edit Hilite Navigation View

Table "default" - Rows: 79 Spec - Columns: 4 Properties Flow Variables

Row ID	Sensitivity	Specificity	PPV	NPV
Row0	87,1	94,1	95,2	84,4
Row1	91	97	91	97
Row2	93	98,2	93,9	98,5
Row3	95,7	94,3	95,7	94,4
Row4	100	97,7	100	98
Row5	90,1	75	7,7	99,7
Row6	90,1	87,1	27,3	99,4
Row7	89,5	87,5	21,4	99,5
Row8	95,2	93	93,7	96,8
Row9	90,9	90,9	28,6	99,6
Row10	97,9	92,5	89,4	99
Row11	86,8	42,9	3,57	99,3
Row12	98,82	100	95,35	100
Row13	98	38	57	96
Row14	90	54	98,7	12,5
Row15	85	59	98,1	14,2
Row16	100	61	100	98,7
Row17	83	30	95,2	9,5
Row18	92	94	89	96
Row19	76	95	84	92
Row20	76	85	83	79
Row21	76	68	65	79
Row22	95	66	71	94
Row23	91	96,8	91,2	96,8
Row24	98,4	100	95,5	100
Row25	100	95	100	98
Row26	100	81	100	93
Row27	89,3	87,1	90,4	93,4
Row28	90	90	61	98
Row29	96	82	79	97
Row30	99	74	89	96
Row31	99	68	92	95
Row32	99	62	93	94
Row33	100	51	95	92

File Table - 17:7 - CSV Reader

File Edit Hilite Navigation View

Table "default" - Rows: 79 Spec - Columns: 4 Properties Flow Variables

Row ID	Sensitivity	Specificity	PPV	NPV
Row36	100	19	98	88
Row37	100	21	98	88
Row38	97,4	93,3	90,3	98,2
Row39	97,4	93,8	83,3	100
Row40	95	93	93	95
Row41	95,3	74,2	91	85,2
Row42	100	94	100	94
Row43	99,5	98,3	99,6	98,1
Row44	100	95,7	100	95,9
Row45	99,5	42,9	42,9	99,5
Row46	99,5	73,3	84,6	98,9
Row47	99,1	82,6	73,1	99,5
Row48	98,2	87	66,3	99,5
Row49	100	83,5	82	96,9
Row50	99	79	78,9	99,1
Row51	100	100	100	100
Row52	100	100	100	100
Row53	100	93,7	100	94,4
Row54	100	100	100	100
Row55	94	95,8	93,1	91,6
Row56	64,5	95,3	83,3	88
Row57	87	73	84	77
Row58	96,9	91	94,5	97,2
Row59	97,04	90,24	88,1	97,62
Row60	80,7	96,9	66	98,5
Row61	81,9	90,6	65,9	95,8
Row62	89,2	90,6	76,3	96,1
Row63	95,2	75,2	85,7	90,8
Row64	87	66	58	95
Row65	98,5	84,4	98,2	86,8

Figure 1 Accuracy parameters of serological tests

File Table - 18:1 - CSV Reader

File Edit Hilite Navigation View

Table "default" - Rows: 81 Spec - Columns: 4 Properties Flow Variables

Row ID	Sensitivity	Specificity	PPV	NPV
Row66	94,2	76,9	92,9	70,7
Row67	93	29	42	89
Row68	93	13	25	86
Row69	99,1	60	94,7	90,5
Row70	99,1	25	80	90,5
Row71	90	93,3	70	98,1
Row72	90	50	40	92,8
Row73	96	89	97	87
Row74	98	91	92	98
Row75	93	53	73	86
Row76	98	79	79	98
Row77	94	52	87	74
Row78	92	80	67	96
Row79	95	93	50	99,5
Row80	87	82	86	83

Figure 2 Accuracy parameters of serological tests

2.2. Applying Machine Learning Algorithms

4 different machine learning algorithms are used after partitioning. Decision tree learner, naives 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.

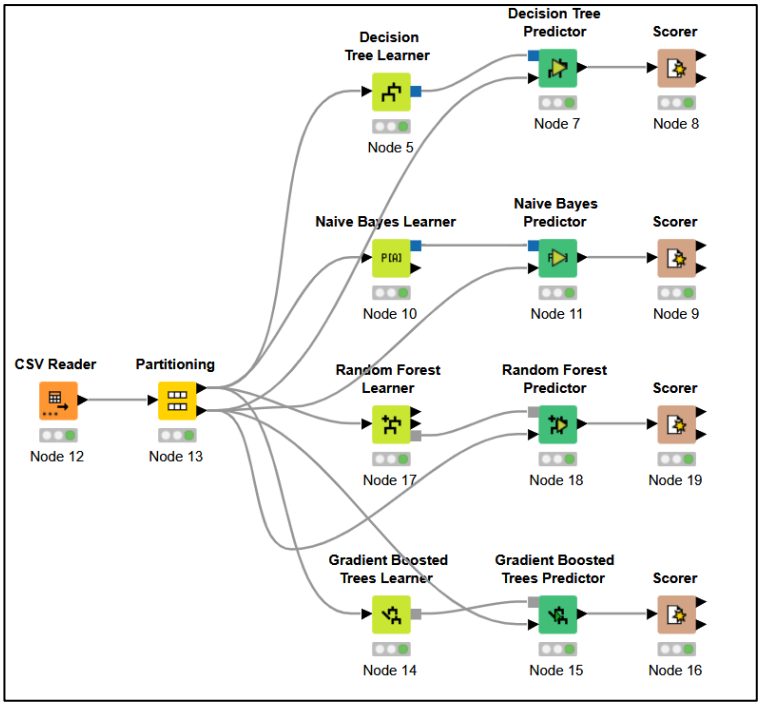

Figure 3 Knime workflow

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 navie 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.

Name	Value
Cohen's kappa	0.8790525785318326
#False	9
#Correct	71
Error	0.1125
Accuracy	0.8875
knime.workspace	C:\Users\ozgul\knime-workspace

Figure 4 Accuracy statistics for decision tree preditor node

Name	Value
Cohen's kappa	0.6703862660944205
#False	24
#Correct	56
Error	0.3
Accuracy	0.7
knime.workspace	C:\Users\ozgul\knime-workspace

Figure 5 Accuracy statistics for naive bayes predictor node

Name	Value
Cohen's kappa	1.0
#False	0
#Correct	80
Error	0.0
Accuracy	1.0
knime.workspace	C:\Users\ozgul\knime-workspace

Figure 6 Accuracy statistics for random forest predictor node

Name	Value
Cohen's kappa	0.6871280394490733
#False	23
#Correct	57
Error	0.2875
Accuracy	0.7125
knime.workspace	C:\Users\ozgul\knime-workspace

Figure 7 Accuracy statistics for gradient boosted trees predictor

4. Conclusion

Data mining approches have been successfully applied to different practical problems not only in clinical medicine but also in epidemiological studies and meta-analysis. These approches 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 overfittig and the small number of sample size. Decision tree predictor which provides %88,7 accuracy can be used as a effective machine learning algorithm for predicting potential covariates for missing values in meta analyses.

References

[1] D. Schuppan, "Current concepts of celiac disease pathogenesis," *Gastroenterology*, vol. 119, pp. 234–242, 2000.

[2] S. Lohi et al, "Increasing prevalence of coeliac disease over time," *Alimentary Pharmacology &Therapeutics*, vol. 26, no. 9, pp. 1217-25, 2005.

[3] M. Parizade, Y. Bujanover, B. Weiss V., Nachmias and B. Shainberg, "Performance of Serology Assays for Diagnosing Celiac Disease in a Clinical Setting," *Clinical and Vaccine Immunology*, vol. 16, pp. 1576–1582, 2009.

[4] A. Fasano and C. Catassi, "Current approaches to diagnosis and treatment of celiac disease: An evolving spectrum", *Gastroenterology*, vol. 120, no. 3, pp. 636-51, 2001.

[5] A. Marlou and A.D. Leffler, "Serum Markers in the Clinical Management of Celiac Disease," *Digestive Disease*, vol. 33, pp. 236–243, 2015.

[6] D. Basso et al. "A new indirect chemiluminescent immunoassay to measure Anti-Tissue Transglutaminase antibodies," *J Pediatr Gastroenterol Nutr*., vol. 43, pp. 613-8, 2006.

[7] P. Eusebi, "Diagnostic Accuracy Measures," *Cerebrovascular Diseases*, vol.36, pp. 267–272, 2013.

[8] A. Hoyer and A. Zapf, "Studies for the Evaluation of Diagnostic Tests," *Deutsches Ärzteblatt International*, vol. 18, pp. 555–60, 2021.

[9] O. Rozenberg, A. Lerner, A. Pacht, M. Grinberg, D. Reginashvili, C. Henig and M. Barak, "A new algorithm for the diagnosis of celiac disease," *Cellular & Molecular Immunology*, vol. 8, pp. 146–149, 2011.

[10] Z. Obermeyer and J. E. Emanuel, "Predicting the Future-Big Data, Machine Learning, and Clinical Medicine," *N Engl J Med*, vol. 375, no.13, pp. 1216–1219, 2016.

[11] M. Saken, M. Y. Banzragch and N. Yumusak, "Impact of image segmentation techniques on celiac disease classification using scale invariant texture descriptors for standard flexible endoscopic systems," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 29, pp. 598 – 615, 2021.

[12] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *International Journal of Medical Informatics*, vol.77, pp.81–97, 2008.

[13] Y. Long, L. Wang and M. Sun, "Structure Extension of Tree-Augmented Naive Bayes," *Entropy*, vol. 21, pp.721, 2019.

[14] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg and O. Lyashevska, "Predictive analytics with gradient boosting in clinical medicine," *Annals of Translational Medicine*, vol. 7, no.7, pp.152, 2019.

[15] M. Song, H. Jung, S. Lee, D. Kim, and M. Ahn, "Diagnostic Classification and Biomarker Identification of Alzheimer's Disease with Random Forest Algorithm," *Brain Sciences*, vol. 11, no. 4, pp. 453, 2021.

[16] A. W. Warr, "Scientific workflow systems: Pipeline Pilot and KNIME," *J Comput Aided Mol Des.*, vol. 26, no.4, pp. 801–804, 2012.

[17] Y. Yuan and R. Little, "Meta-Analysis of Studies with Missing Data," *Biometrics*, vol.65, pp. 487-496, 2009.

[18] J. M. Schauer, K. Diaz, T.D. Pigott and J. Lee, "Exploratory Analyses for Missing Data in Meta-Analyses and Meta-Regression: A Tutorial," *Alcohol and Alcoholism*, vol. 57, pp. *35-36*, 2022.