Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

Prediction of brucellosis based on hematological biomarkers via ensemble classification methods

Ahmet Sahin[a,*], Mehmet Celik[b], Mehmet Resat Ceylan[b], Deniz Altındag[c], Esra Gurbuz[d], Nevin Guler Dincer[e], Sevil Alkan[f]

[a] *Dr. Ersin Arslan Training and Research Hospital, Department of Infectious Diseases and Clinical Microbiology, Gaziantep, Türkiye*
[b] *Harran University, Faculty of Medicine, Department of Infectious Diseases and Clinical Microbiology, Şanlıurfa, Türkiye*
[c] *Cizre Dr. Selahattin Cizrelioğlu State Hospital, Department of Infectious Diseases and Clinical Microbiology, Şırnak, Türkiye*
[d] *University of Health Sciences, Van Training and Research Hospital, Department of Infectious Diseases and Clinical Microbiology, Van, Türkiye*
[e] *Muğla Sıtkı Koçman University, Faculty of Science, Department of Statistics, Muğla, Türkiye*
[f] *Çanakkale Onsekiz Mart University, Faculty of Medicine, Department of Infectious Diseases and Clinical Microbiology, Çanakkale, Türkiye*

ARTICLE INFO

Keywords:
Brucellosis
Ensemble classification methods
Machine learning
Permutation importance
Zoonotic disease

Received: Oct 01, 2023
Accepted: Nov 30, 2023
Available Online: 27.12.2023

DOI:
10.5455/annalsmedres.2023.10.273

Abstract

Aim: Some hematological changes are frequently observed in the clinical course of brucellosis. This study aimed to predict the diagnosis of brucellosis based on some hematological biomarkers with the help of ensemble classification methods.

Materials and Methods: A total of 23 ensemble classification methods, including 10 bagging, 9 boosting, and 4 stacking approaches were applied to the brucellosis data set. Each subject in the brucellosis data set contains 13 features, including age, gender, and 10 hematological variables.

Results: This study included a total of 257 participants [173 (67.3%) brucellosis patients and 84 (32.7%) healthy controls]. The mean values of white blood cells (WBC), hemoglobin (HGB), neutrophil (NEUT), neutrophil/lymphocytes (NEUT/LYMP), and monocytes/lymphocytes (MO/LYMP) of brucellosis patients were found to be significantly lower than those of healthy controls. Random Forest with Gini criterion (RF2) was selected to be the best fit model with a mean accuracy of 0.728. HGB (mean score = 0.1814), age (0.1311), NEUT/LYMP (0.0938), WBC (0.0817) and mean platelet volume (MPV) (0.0815) were determined as most diagnostic parameters in brucellosis.

Conclusion: The lower levels of HGB, WBC, and NEUT/LYMP and higher levels of age and MPV may be important indicators for the diagnosis of brucellosis.

Introduction

Brucellosis is an important zoonotic disease caused by Gram-negative coccobacillus of the Brucella genus [1]. It is widespread in the world, and it is prevalent in South America, East Mediterranean countries, and possibly Sub-Saharan Africa [2, 3]. Infection is usually transmitted through the consumption of contaminated food products of animal origin, such as unpasteurized milk and milk products [4]. Symptoms and signs such as fever, fatigue sweating, and osteoarticular involvement are common in brucellosis cases. Brucellosis is clinically categorized as subclinical, acute, subacute, or chronic forms [5]. Microbiological diagnosis of human brucellosis is based on three different modalities: serological methods, microbiological cultures, and nucleic acid amplification tests [6]. With advances in computer processing and artificial intelligence (AI), attempts have been made to develop intelligent tools that can learn and automate decision support without the need for human input, including the diagnosis of infection [7]. Machine Learning (ML), a branch of AI, is a scientific discipline that deals with the development, implementation, and analysis of automated methods that allow a machine to evolve through a learning process. With the rapid development of computer technology, ML has been widely applied in medicine [8, 9]. In the literature, it is seen that ML methods such as Artificial Neural Networks (ANNs), Decision Trees (DTs), and Support Vector Machines (SVMs) are used [10, 11]. However, using a single technique may not always detect high accuracy. Each technique used in classification has advantages and disadvantages. Ensemble classification method is a popular

*Corresponding author:
Email address: ahmet27sahin@hotmail.com (Ahmet Sahin)

paradigm that leverages the strengths of individual classifiers and mitigates their weaknesses. Ensemble classification method consist of combining more than one single classifier depending on a certain combination rule [11, 12]. This study aimed to predict brucellosis with the help of ensemble classification methods by using some hematological biomarkers in the absence of serological and microbiological tests.

Materials and Methods

The objective of this study is to develop a predictive model, based on ensemble classification models and some hematological biomarkers, to predict brucellosis disease. In this section, brief information was given about the data and methods used to achieve the objective determined. Approval was obtained from Harran University Clinical Research Ethics Committee before the study (Date and Number: 18.10.2022-173789).

Data and data preprocessing

Study protocol

This study was designed as a case-control study. The data set consisted of patients over the age of 18 years with brucellosis who were admitted to the Infectious Diseases and Clinical Microbiology outpatient clinic of a state hospital between 2018 and 2022. A control group of healthy volunteers without any symptoms or underlying disease was also included in the study. Information about the hematologic laboratory results, age, and gender of patients with brucellosis and healthy volunteers were obtained from the hospital information management system.

Definition of Brucellosis

The criteria used for the diagnosis of brucellosis are growth in the culture of *Brucella* spp. in blood or other body fluids cultures (BACT/ALERT 3D/60 bioMérieux, France) and together with clinical symptoms such as fever, chills, joint-muscle pain, headache, sweating, and weakness, or serum *Brucella* tube agglutination (Wright test, Biomedica, Canada) titer equal to or greater than 1/160 or at least a four-fold titer increase in the serum sample taken at two-week intervals. Brucellosis was categorized into three groups according to the duration of clinical signs and symptoms acute brucellosis (0-2 months), subacute brucellosis (2-12 months), and chronic brucellosis ($>$ 12 months) [5].

Laboratory tests of the patients included in the study and the healthy control group were evaluated at the time of admission to the hospital. These tests included white blood cell (WBC), hemoglobin (HGB), neutrophil (NEUT), lymphocyte (LYMP), monocyte (MO), platelet (PLT), NEUT/LYMP, PLT/LYMP, MO/LYMP and mean platelet volume (MPV).

Data preprocessing

There were some missing values and outliers in the dataset. The missing values are imputed by using Scikit-Learn's KNNImputer [13]. The outliers were determined by calculating Median Absolute Deviation (MAD) statistic and removed from the dataset. MAD was calculated as follows:

$$MAD_i = median(\mid X_{ij} - median(X_i) \mid)\ i = 1, 2..., p \quad (1)$$

Where X_{ij} is the i. feature of j. subject, median (X_i) is the median of i. feature and p is the number of features. k is usually is set as 1.4826 [14]. X_{ij} values outside the interval defined as bellow were determined as outliers:

$$[median(X_i) - 3 * MAD_i, median(X_i) + 3 * MAD_i] \quad (2)$$

Ensemble classification

Classification is the task of assigning a class label to new instance based on the relationship between a categorical class variable (Y) and features vectors (X=$\{X_1, X_2, \ldots, X_p\}$):

$$y_{n+1} = \int (X_{n+1}) \quad (3)$$

Where, X_{n+1} is feature vector of new instance, f is a classifier function predicted by a classification method [15]. There exist numerous classification methods used for predicting the $\int$function such as Naïve Bayes (NB), K-nearest Neighbor (KNN), Support Vector Machines (SVM), Adaboost, XGBoost, Random Forest (RF).But these classification methods can be collected into two main titles as individual and ensemble. In individual classification, single type of classification model is applied to the training set and thus a single model is predicted to explain the relationship between the variables. Ensemble classification is based on combining the prediction results of a set of individual classifiers using some methods such as weighted voting, majority voting or weighted averaging. Major advantages of ensemble classification are to solve the problem of high bias and variance mostly faced in the individual classification and to increase the classification accuracy [16]. Ensemble classification methods can be divided into three categories including bagging, boosting, and stacking [17-20].

Bagging: Firstly, constructs N number of training sets from original data set using simple random sampling (with replacement). An individual classifier is predicted for each training set and the prediction results are aggregated to obtain the final predictions. In this approach, all individual classifiers are predicted in parallel. RF is most known bagging ensemble classification method.

Boosting: Builds the training sets step by step. In the first step, a single training set is produced from original dataset. All instances have equal weight of being selected to the first training set. Individual classifier is predicted for the produced training set and misclassified instances are identified. In the second step, a new training set is produced by increasing the weight of misclassified instances. The individual classifier is predicted by learning the second training sets and misclassified instances are re-identified. The third training set is produced by increasing the weights of misclassified instances. This process is repeated until the desired training error is achieved.

Stacking: Two or more different types of individual classifiers are applied to original data set, separately and the prediction results obtained are combined using a meta classifier.

Performance evaluation criteria used in classification

To compare performance of the classification methods, accuracy (Acc), precision, recall, F1-measure (F), Kappa statistics, Area under receiver operating characteristic (AROC) etc. criteria are generally used. The most of these criteria are calculated based on confusion matrix [21].

True Positive (TP) and True Negative (TN) refer to the number of cases correctly classified in positive class and negative class, respectively. False Positive (FP) indicates the number of cases classified as positive when true class is negative, and False Negative (FN) indicates the number of cases classified as negative when true class is positive. Based on these definitions, some criteria are calculated as follows:

$$A_{cc} = \frac{TP + TN}{TN + FP + FN + TP} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F - Measure\,(F) = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

$$TruePositiveRate\,(TPR) = \frac{TP}{TP + FN} \tag{8}$$

$$FalsePositiveRate\,(FPR) = \frac{FP}{FP + TN} \tag{9}$$

All criteria take values between 0 and 1.The closer the value is to 1, the more successful the classification.

Statistical analysis

All statistical analyses were carried out by using SPSS 22 software. The variables that follow normal distribution were presented as mean±standard deviation (MS) and the variables that do not follow normal distribution as median [interquartile range=Third Quartile-First Quartile] (MI) (Table 1 and Table 2) . Besides %95 confidence interval (CI) was calculated for the variables with normal distribution, fences (Fe) were calculated for the variables not followed normal distribution. The normality of the variables was tested based on One Sample Kolmogorov Smirnov test. To compare the mean of two continuous variables with normal distribution, Student T-test was performed. Mann-Whitney U test was used for the variables that did not satisfy the assumption of normal distribution. In order to determine whether there is a significant difference between the means of two or more variables, One-Way ANOVA was applied when the normal distribution assumption was satisfied, and the Kruskall-Wallis test was applied when it was not satisfied. The correlation between normally distributed variables was calculated according to Pearson coefficient, and the correlation between non-normally distributed variables was calculated according to Spearman's Rho coefficient. p- values less than or equal to 0.05 were considered statistically significant.

Results

This study intends to predict brucellosis disease based on some hematological and demographic variables. To achieve this objective, a total of 23 ensemble classification methods, including 10 bagging, 9 boosting, and 4 stacking approaches were applied to the brucellosis data set. Each subject in the brucellosis data set contains 13 features, including age, gender, class (Healthy control or patients with brucellosis), and 10 hematological variables. In this section, the statistical properties of the subjects firstly were investigated. In the next step, the classifica-

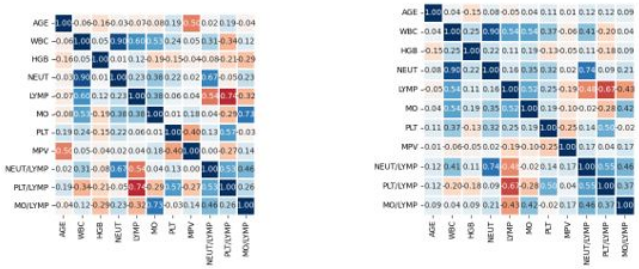

Figure 1. Correlation heatmaps for control and brucellosis classes.

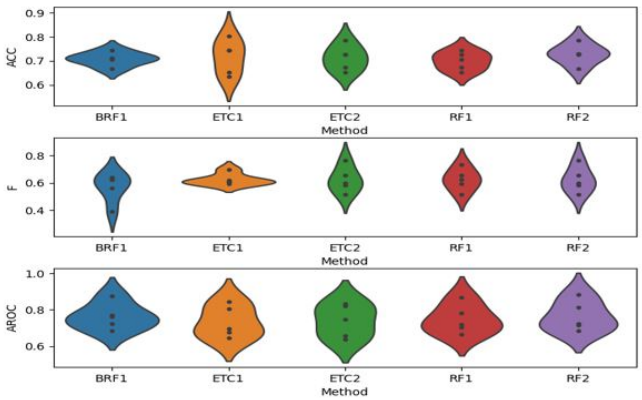

Figure 2. Violin plot of the Acc, F and AROC distributions for top five methods.

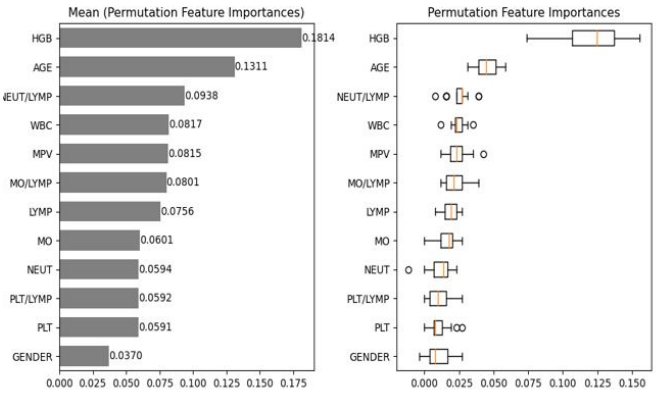

Figure 3. Feature importance scores for RF2 method.

Table 1. . Demographic and hematological properties of the overall participants.

Variables	Statistic	Control (0)	Brucellosis (1)	p
Age	MS / MI CI/Fe	32[47-28] [-0.5 75.5]	40.61±14.02 [38.50 42.71]	0.22
WBC	MS / MI CI/Fe	7167.26±1425.11 [6857.99 7476.53]	6702.72±1701.60 [6447.36 6958.08]	**0.02**
HGB	MS / MI CI/Fe	15 [15.60-13.23] [9.68 19.16]	12.90 [14.10-12.20] [9.35 16.95]	**0.00**
NEUT	MS / MI CI/Fe	4026.33±1069.77 [3794.18 4258.49]	3571.91±1297.67 [3377.17 3766.65]	**0.01**
LYMP	MS / MI CI/Fe	2316.69±541.75 [2199.12 2434.26]	2421.10±663.90 [2321.47 2520.73]	0.21
MO	MS / MI CI/Fe	581.75±191.60 [540.17 623.33]	524.40[665-417] [45 1037]	0.26
PLT	MS / MI CI/Fe	248.64±47.17 [238.41 258.88]	246[277.50-208.50] [105 381]	0.74
MPV	MS / MI CI/Fe	9.77±1.07 [9.53 10.00]	9.90[10.50-9.30] [7.5 12.3]	0.09
NEUT/LYMP	MS / MI CI/Fe	1.81±0.56 [1.69 1.83]	1.52[1.90-1.06] [-0.2 3.16]	**0.00**
PLT/LYMP	MS / MI CI/Fe	0.11±0.03 [0.11 0.12]	0.104[0.13-0.08] [0.005 0.205]	0.29
MO/LYMP	MS / MI CI/Fe	0.26±0.08 [0.24 0.28]	0.22[0.27-0.18] [0.05 0.41]	**0.00**

WBC: White blood cell, HGB: Hemoglobin, NEUT: Neutrophil, LYMP: Lymphocyte, MO: Monocytes, PLT: Platelet, MPV: Mean platelet volume, NEUT/ LYMP: Neutrophil/ Lymphocyte, PLT/ LYMP: Platelet/ Lymphocyte, MO LYMP: Monocytes/ Lymphocyte.

tion performance of the methods used in this study was compared and the best-fit model was selected according to three evaluation criteria. The classification experiments were performed by using Python 3.9.0 with some Python libraries such as Sklearn, Numpy, Pandas, Seaborn, and Matplotlib.

Results of data preprocessing and statistical analysis

A total of 305 participants, 96 (31.5%) healthy controls and 209 (68.5%) patients diagnosed with brucellosis, were included in this study. Of the patients diagnosed with brucellosis, 80 (38.3%) had acute, 75 (35.9%) subacute, and 54 (25.8%) had chronic brucellosis sub-diagnosis. In terms of gender, 131 (43%) of the patients were female and 174 (57%) were male. The mean age of all patients was 40.13±14.86 years.

When outliers were removed from the dataset, a total of 257 patients were included in this study, including 84 (32.7%) healthy controls and 173 (67.3%) brucellosis patients. Of the brucellosis patients, 63 (36.4%) had acute brucellosis, 69 (39.9%) subacute brucellosis and 41 (23.7%) chronic brucellosis sub diagnosis. In terms of gender, 110 (42.8%) of the patients were female and 147 (57.2%) were male. Among healthy controls, 58 (69%) were female and 26 (31%) were male. The mean age of the patients included in the study was 39.84±14.21 years, 38.27±14.56 years in healthy controls and 40.61±14.02 years in brucellosis patients.

Other statistical characteristics of the cases are given as mean ± standard deviation in Table 1. Independent sample t-test and Mann-Whitney U test were used to determine whether there was a statistically significant difference between the brucellosis patients and the control group in terms of age and hematologic variables.

According to Table 1, there was no significant difference between the control and brucellosis groups in terms of mean age and LYMP, MO, PLT, MPV, and PLT/LYMP values. Mean values of WBC, HGB, NEUT NEUT/LYMP, and MO/LYMP were significantly lower in brucellosis patients compared to the control group ($p<0.05$). Detailed information about brucellosis patients is given in Table 2. One-Way ANOVA and Kruskal-Wallis test were used to statistically compare the mean or median values of the brucellosis groups.

According to Table 2, no statistically significant difference was found between the brucellosis subgroups in terms of age and hematologic variables. Figure 1 shows the correlations between variables for control and brucellosis classes separately.

Some of the results obtained from Figure 1 can be summarized as follows. In the control group, there was a moderate and positive correlation between age and MPV values, and a moderate and negative correlation between MPV and PLT values. There was also a moderate and positive correlation between WBC and LYMP, WBC and MO, and

Table 2. Demographic and hematological properties of patients with brucellosis.

Variables	Statistic	Acute	Subacute	Chronic	p
Age	MS / MI CI/Fe	38.08±13.54 [34.67 41.49]	42.67±14.61 [39.16 46.18]	41.02±13.42 [36.79 45.26]	0.17
WBC	MS / MI CI/Fe	6696.98±1818.30 [6239.05 7154.92]	6714.64±1525.33 [6348.21 7081.06]	6691.46±1835.79 [6112.02 7270.91]	0.99
HGB	MS / MI CI/Fe	12.90 [14.20 11.90] [8.45 17.65]	13.12±1.42 [12.78 13.46]	12.70[12.40 14.10] [9.85 16.65]	0.80
NEUT	MS / MI CI/Fe	3535.87±1393.37 [3184.96 3886.79]	3536.52±1149.29 [3260.43 3812.61]	3686.83±1403.04 [3243.98 4129.68]	0.81
LYMP	MS / MI CI/Fe	2389.21±685.35 [2216.60 2561.81]	2526.67±613.30 [2379.34 2674.00]	2292.44±700.01 [2071.49 2513.39]	0.18
MO	MS / MI CI/Fe	557.87±199.49 [507.63 608.11]	560.58±162.13 [521.64 599.53]	537.50±180.44 [480.55 594.45]	0.79
PLT	MS / MI CI/Fe	243.30±62.86 [227.47 259.13]	250.59±61.06 [235.93 265.26]	243.95±51.30 [227.76 260.14]	0.75
MPV	MS / MI CI/Fe	9.99±0.83 [9.78 10.20]	9.84±0.87 [9.63 10.04]	10.20±0.88 [9.93 10.48]	0.10
NEUT/LYMP	MS / MI CI/Fe	1.57±0.70 [1.39 1.74]	1.46±0.53 [1.32 1.58]	1.73±0.73 [1.50 1.96]	0.10
PLT/LYMP	MS / MI CI/Fe	0.11±0.04 [0.10 0.12]	0.10±0.03 [0.10 0.11]	0.11±0.03 [0.10 0.12]	0.36
MO/LYMP	MS / MI CI/Fe	0.22±0.10 [0.20 0.25]	0.21[0.17 0.26] [0.04 0.40]	0.22[0.18 0.31] [-0.02 0.51]	0.38

WBC: White blood cell, HGB: Hemoglobin, NEUT: Neutrophil, LYMP: Lymphocyte, MO: Monocytes, PLT: Platelet, MPV: Mean platelet volume, NEUT/ LYMP: Neutrophil/ Lymphocyte, PLT/ LYMP: Platelet/ Lymphocyte, MO LYMP: Monocytes/ Lymphocyte.

NEUT and MO values, and a high and positive correlation between WBC and NEUT values in both groups.

Performance analysis of ensemble classification methods

A total of 23 ensemble-based classification methods were used to predict the diagnosis of brucellosis. The classification methods used are given in Appendix 1. 5-fold cross-validation method was used to test the validity of the methods. Acc, F, and AROC were selected to evaluate the performance of the methods. Table 3 gives the mean and median values of the evaluation criteria obtained from each fold. Figure 2 shows the violin plot of Acc, F, and AROC values for top five methods.

The results shown in bold in Table 3 show the best five values of the evaluation criteria according to the mean and median values. As can be seen from Table 3 and Figure 2, the mean Acc values ranged from 0.596 to 0.728, the mean F values from 0.402 to 0.641, and the mean AROC values from 0.5 to 0.765. The median Acc value was between 0.577 and 0.745, the median F value was between 0.402 and 0.655, and the median AROC value was between 0.5 and 0.763. When all criteria were evaluated simultaneously, the top five methods with the highest classification performance were BRF1, ETC1, ETC2, RF1, and RF2. However, RF2 was selected as the most appropriate method because it was among the top five most successful models for all criteria except the median of the AROC values.

To further investigate the class-based performance of RF2, RF2 was rerun for the entire dataset without selecting a test set. RF2 correctly classified 57 out of 84 (68%) subjects in the control group and 165 out of 173 (95%) in the brucellosis group. However, 27 subjects (32%) in the control group were misclassified as brucellosis, and 8 subjects (0.05%) in the brucellosis group were misclassified as control. Based on these results, we concluded that RF2 was more successful in predicting brucellosis patients compared to the control group. Finally, the permutation importance method was used to identify the main features in predicting the diagnosis of brucellosis [22]. The obtained feature scores were given in Figure 3.

According to Figure 3, the top five most discriminative parameters in predicting the diagnosis of brucellosis were detected as; HGB value (with an average score of 0.1814), age (with an average score of 0.1311), NEUT/LYMP value (with an average score of 0.0938), WBC value (with an average score of 0.0817 and MPV value with an average score of 0.0815).

Discussion

Brucellosis, as it is a systemic disease, can cause changes in inflammatory parameters [23]. Hematological complications due to brucellosis are common. This may be related to the bacteria's tropism in the reticuloendothelial system. Changes in hematological parameters are mostly observed

Table 3. Mean and median values of evaluation criteria.

Method	Acc		F		AROC	
	Mean	Median	Mean	Median	Mean	Median
AGNB	0.670	0.673	0.594	0.634	0.619	0.643
ALR	0.689	0.692	0.584	0.639	0.681	**0.745**
ARF1	**0.712**	**0.712**	0.617	0.635	0.670	0.632
ARF2	**0.728**	**0.712**	0.603	**0.655**	0.633	0.691
ASVC	0.673	0.673	0.402	0.402	0.500	0.500
BGNB	0.701	0.692	**0.629**	**0.650**	0.744	0.689
BKNN	0.642	0.647	0.440	0.430	0.578	0.585
BLR	0.693	0.673	0.586	**0.641**	**0.753**	**0.763**
BRF1	0.708	0.712	0.570	0.625	**0.764**	**0.763**
BRF2	0.704	0.692	0.563	**0.641**	**0.761**	**0.739**
BSVC	0.673	0.673	0.402	0.402	0.591	0.615
ETC1	**0.717**	**0.745**	**0.626**	0.610	0.734	0.697
ETC2	**0.712**	**0.725**	**0.624**	0.601	0.739	**0.746**
GBC	0.603	0.577	0.567	0.571	0.654	0.666
HGB	0.627	0.686	0.516	0.502	0.670	0.632
RF1	0.701	0.706	**0.625**	0.623	**0.748**	0.721
RF2	**0.728**	**0.731**	**0.641**	**0.655**	**0.765**	0.724
ST1	0.642	0.647	0.539	0.510	0.600	0.621
ST2	0.697	0.673	0.577	0.617	0.690	0.676
ST3	0.689	0.686	0.586	0.639	0.663	0.666
ST4	0.596	0.608	0.528	0.561	0.554	0.574
XGB1	0.635	0.686	0.543	0.558	0.684	0.663
XGB2	0.669	0.667	0.477	0.453	0.668	0.654

Acc: Accuracy, F: F1-measure, AROC: Area under receiver operating characteristic.

in brucellosis patients. Therefore, this study aimed to investigate which hematological parameters can be used in predicting bacteremia with the help of ML methods. To our knowledge, the current study is the first study that investigates hematological parameters in predicting bacteremia with the help of ML methods.

Hematologically, leukopenia, lymphomonocytosis, and mild anemia are frequently seen in brucellosis [24, 25]. Olt et al. found a significant relationship between HGB, NEUT/LYMP ratio, and brucellosis [26]. In Iran, Akya et al. showed that CRP, WBC, and NEUT counts can be used as biomarkers in the preliminary diagnosis of brucellosis [23]. In another study, statistically significant differences were found in HGB, thrombocyte, MPV, and NEUT/LYMP ratio in brucellosis patients, while it was stated that MPV and NEUT/LYMP ratio could be used as inflammation markers in brucellosis [27]. In our study, there were a total of 257 patients with 84 healthy controls and 173 brucellosis patients (48 patients were excluded from the study due to outliers). The mean levels of WBC, HGB, NEUT NEUT/LYMP, and MO/LYMP of brucellosis patients were significantly lower than healthy controls ($p<0.05$). Low levels of NEUT and HGB are common laboratory findings in brucellosis. Also, LYMP was found to be higher in the brucellosis group compared to the control group, although it was not statistically significant ($p=0.21$). According to these results, we think that WBC, HGB, NEUT, NEUT/LYMP, and MO/LYMP levels may be important biomarkers in predicting brucellosis.

ML methods have been widely used in the field of medicine and have achieved good results in many diseases' diagnosis, and risk assessments in recent years [28-31].It was reported that the ensemble classification method, which is a type of ML, can be used in the early diagnosis and classification of some infectious diseases. Brinati et al. reported that the RF model is the best classifier method in their study where they investigated different ML methods to detect COVID-19 from routine blood tests [32]. Chicco and Jurman used the RF classification method to diagnose hepatitis C and to determine the most diagnostic features for hepatitis C. They found that RF provided the best performance for diagnosing hepatitis C and ALT and AST were diagnostic features [33]. In a study conducted in China, the eXtreme Gradient Boosting (XGBoost) model, it was determined as the most suitable method for determining the incidence and predicting hand, foot, and mouth disease [34]. In a study conducted in China, a comparison was made between the autoregressive integrated moving average (ARIMA) model and the XGBoost model to determine which method is more suitable for estimating the occurrence of brucellosis. As a result, the XGBoost model was found to be more suitable for predicting human brucellosis cases in China [35]. In our study, 23 ensemble-based classification methods were used to classify patients with brucellosis and to predict brucellosis disease. Acc, F, and AROC criteria were employed to compare the classification performance of the methods used. The mean Acc values were found to be between 0.596 and 0.728, mean F values between 0.402 and 0.641, and mean AROC values between 0.5 and 0.765. The median Acc value was between 0.577 and 0.745, the median F value between 0.402 and 0.655, and the median AROC between 0.5 and 0.763. When evaluating all results simultaneously, it was detected that the five methods with the highest classification performance were BRF1, ETC1, ETC2, RF1, and RF2. RF2 was found to be the most suitable model as it had the highest (mean Acc value 0.728, mean F value = 0.641, mean AROC value=0.765) and provided better classification results for all criteria. When RF was re-executed with the whole data set without using the test set, it was seen that RF2 correctly classified 95% of the brucellosis patients and 68% of subjects in the control class. According to the result of the permutation importance method, HGB was found to be the most important parameter with a mean score of 0.1814 for predicting brucellosis. Age and NEUT/LYMP, WBC, and MPV levels were also determined as important parameters. From these results, it was concluded that the lower values of HGB, WBC, and NEUT/LYMP and average age and high levels of MPV may be important indicators for diagnosing brucellosis.

Limitations

The major limitation of our study is its retrospective nature and small sample size. Larger studies are needed on this subject.

Conclusion

In conclusion the RF2 ensemble classification model provided the promise results for predicting or diagnosing brucellosis patients. HGB was found to be the most effective biomarker in the prediction the brucellosis. The

other important biomarkers were identified to as older age, NEUT/LYMP, WBC, and MPV.

Ethical approval

Approval was received for this study from Harran University Clinical Research Ethics Committee (Date and Number: 18.10.2022-173789).

References

1. Bagheri Nejad R, Krecek RC, et al. Brucellosis in the Middle East: Current situation and a pathway forward. PLoS Negl Trop Dis. 2020;14(5):e0008071. doi:10.1371/journal.pntd.0008071.
2. Pappas G, Papadimitriou P, Akritidis N, et al. The new global map of human brucellosis. Lancet Infect Dis. 2006;6(2):91-9. doi:10.1016/s1473-3099(06)70382-6.
3. Golshani M, Buozari S. A review of Brucellosis in Iran: Epidemiology, Risk Factors, Diagnosis, Control, and Prevention. Iran Biomed J. 2017;21(6):349-59. doi:10.18869/acadpub.ibj.21.6.349.
4. Yagupsky P, Baron EJ. Laboratory exposures to brucellae and implications for bioterrorism. Emerg Infect Dis. 2005;11(8):1180-5. doi:10.3201/eid1108.041197.
5. Buzgan T, Karahocagil MK, Irmak H, et al. Clinical manifestations and complications in 1028 cases of brucellosis: a retrospective evaluation and review of the literature. Int J Infect Dis. 2010;14(6):e469-78. doi:10.1016/j.ijid.2009.06.031.
6. Yagupsky P, Morata P, Colmenero JD. Laboratory Diagnosis of Human Brucellosis. Clin Microbiol Rev. 2019;33(1). doi:10.1128/cmr.00073-19.
7. Rawson TM, Hernandez B, Moore LSP, et al. Supervised machine learning for the prediction of infection on admission to hospital: a prospective observational cohort study. J Antimicrob Chemother. 2019;74(4):1108-15. doi:10.1093/jac/dky514.
8. Qiu J, Wu Q, Ding G, et al. A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing. 2016;2016(1):67. doi:10.1186/s13634-016-0355-x.
9. Hamida S, el Gannour O, Cherradi B, et al. Optimization of Machine Learning Algorithms Hyper-Parameters for Improving the Prediction of Patients Infected with COVID-19. 2020.
10. Esfandiari N, Babavalian MR, Moghadam A-ME, Tabar VK. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications. 2014;41(9):4434-63.
11. Hosni M, Abnane I, Idri A, et al. Reviewing ensemble classification methods in breast cancer. Comput Methods Programs Biomed. 2019;177:89-112. doi:10.1016/j.cmpb.2019.05.019.
12. Chandra A, Yao X. Ensemble Learning Using Multi-Objective Evolutionary Algorithms. Journal of Mathematical Modelling and Algorithms. 2006;5(4):417-45. doi:10.1007/s10852-005-9020-3.
13. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520-5. doi:10.1093/bioinformatics/17.6.520.
14. Leys C, Ley C, Klein O, et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology. 2013;49:764–6. doi:10.1016/j.jesp.2013.03.013.
15. Ren Y, Zhang L, Suganthan PN. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. IEEE Computational Intelligence Magazine. 2016;11:41-53.
16. Priasni T, Oswari T. Comparative study of standalone classifier and ensemble classifier. TELKOMNIKA (Telecommunication Computing Electronics and Control). 2021;19:1747. doi:10.12928/telkomnika.v19i5.19508.
17. Breiman L. Bagging predictors. Machine learning. 1996;24:123-40.
18. Freund Y. Boosting a Weak Learning Algorithm by Majority. Information and Computation. 1995;121(2):256-85. doi:https://doi.org/10.1006/inco.1995.1136.
19. Wolpert D. Stacked Generalization. Neural Networks. 1992;5:241-59. doi:10.1016/S0893-6080(05)80023-1.
20. Shahhosseini M, Hu G, Pham H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. Machine Learning with Applications. 2022;7:100251. doi:10.1016/j.mlwa.2022.100251.
21. Özkan Y, Demirarslan M, Suner A. Effect of data preprocessing on ensemble learning for classification in disease diagnosis. Communications in Statistics-Simulation and Computation. 2022:1-21.
22. Breiman L. Random Forests. Machine Learning. 2001;45:5-32. doi:10.1023/A:1010950718922.
23. Akya A, Bozorgomid A, Ghadiri K, et al. Usefulness of Blood Parameters for Preliminary Diagnosis of Brucellosis. J Blood Med. 2020;11:107-13. doi:10.2147/jbm.S245513.
24. Çelik M, Ceylan M, Altındağ D, et al. Diagnostic significance of hematological parameters in brucellosis. Journal of Clinical Medicine of Kazakhstan. 2023;20:50-5. doi:10.23950/jcmk/12929.
25. El-Koumi MAM, Afify MM, Al-Zahrani SHM. A prospective study of brucellosis in children: relative frequency of pancytopenia. Iran J Pediatr. 2014;24(2):155-60.
26. Olt S, Ergenç H, Açıkgöz SB. Predictive contribution of neutrophil/lymphocyte ratio in diagnosis of brucellosis. Biomed Res Int. 2015;2015:210502. doi:10.1155/2015/210502.
27. Bozdemir Ş E, Altıntop YA, Uytun S, et al. Diagnostic role of mean platelet volume and neutrophil to lymphocyte ratio in childhood brucellosis. Korean J Intern Med. 2017;32(6):1075-81. doi:10.3904/kjim.2016.092.
28. Yan Y, Chen C, Liu Y, et al. Application of Machine Learning for the Prediction of Etiological Types of Classic Fever of Unknown Origin. Front Public Health. 2021;9:800549. doi:10.3389/fpubh.2021.800549.
29. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia. 2019;60(10):2037-47. doi:10.1111/epi.16333.
30. Allen A, Mataraso S, Siefkas A, et al. A Racially Unbiased, Machine Learning Approach to Prediction of Mortality: Algorithm Development Study. JMIR Public Health Surveill. 2020;6(4):e22400. doi:10.2196/22400.
31. Jo YT, Joo SW, Shon SH, et al. Diagnosing schizophrenia with network analysis and a machine learning method. Int J Methods Psychiatr Res. 2020;29(1):e1818. doi:10.1002/mpr.1818.
32. Brinati D, Campagner A, Ferrari D, et al. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. J Med Syst. 2020;44(8):135. doi:10.1007/s10916-020-01597-4.
33. Chicco D, Jurman G. An Ensemble Learning Approach for Enhanced Classification of Patients With Hepatitis and Cirrhosis. IEEE Access. 2021;PP:1-. doi:10.1109/ACCESS.2021.3057196.
34. Meng D, Xu J, Zhao J. Analysis and prediction of hand, foot and mouth disease incidence in China using Random Forest and XGBoost. PLoS One. 2021;16(12):e0261629. doi:10.1371/journal.pone.0261629.
35. Alim M, Ye GH, Guan P, et al. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. BMJ Open. 2020;10(12):e039676. doi:10.1136/bmjopen-2020-039676.