JOURNAL OF
CONTEMPORARY MEDICINE

DOI:10.16899/jcm.1243103
J Contemp Med 2023;13(3):459-468

Original Article / Orijinal Araştırma

Predicting of Bacteremia in Patients with Acute Brucellosis Using Machine Learning Methods

Akut Brusellozlu Hastalarda Bakteriyeminin Makine Öğrenmesi Yöntemleri Kullanılarak Tahmin Edilmesi

Mehmet Çelik[1], Mehmet Reşat Ceylan[1], Deniz Altındağ[2], Nevin Güler Dinçer[3], Sait Can Yücebaş[4], Sevil Alkan[5]

[1]University of Harran Faculty of Medicine, Department of İnfectious Diseases and Clinical Microbiology, Şanlıurfa, Turkey
[2]Cizre Dr. Selahattin Cizrelioğlu State Hospital, Şırnak, Turkey
[3]University of Muğla Sıtkı Koçman Faculty of Science, Department of Statistics, Muğla, Turkey
[4]Çanakkale Onsekiz Mart University, Fculty of Engineering, Department of Computer Engineering, Çanakkale, Turkey
[5]Department of Infectious Diseases and Clinical Microbiology, Çanakkale Onsekiz Mart University, Faculty of Medicine, Çanakkale, Turkey

Abstract

Aim: Accurate and early diagnosis of brucellosis is crucial to slow the spread of the disease and provide rapid treatment to patients. The aim of this study was to develop a machine learning-based predictive model for the diagnosis of bacteremia in brucellosis patients based on some hematological and biochemical markers without the need for blood culture and bone marrow culture and to investigate the importance of this method in predicting bacteremia in brucellosis.

Material and Method: In this study, 162 patients over 18 years of age diagnosed with brucellosis were included and the patients were divided into two groups according to bacteremia status. Data were collected retrospectively and analyzed by machine learning. Twenty demographic, hematological, and biochemical laboratory parameters and 30 classifiers were used to predict bacteremia in brucellosis. The classifiers were developed using Python programming language. To assess the classification performance of the methods used, Accuracy (ACC), f-measure (F), and ROC under area (AROC) criteria were utilized. All classification methods were executed with a 15-fold cross-validation test set selection method. The feature importance method was used to select the most discriminative features for the classification of blood culture positivity.

Results: Extratree classifier with "entropy" criterion (ETC1) showed the best predictive performance with ACC values ranging between 0.5 and 1.00, F values between 0.53 and 1, and AROC values between 0.62 and 1. The neutrophil percentage, the lymphocyte percentage, the eosinophil percentage, alanine aminotransferase, and C-reactive protein values were determined as the most distinguishing features with scores of 0.723, 1.000, 0.920, 0.869, and 0.769, respectively.

Conclusion: This study showed that the ETC1 classifier may be helpful in determining bacteremia in brucellosis patients, and that elevated lymphocytes, alanine aminotransferase and C-reactive protein and low neutrophils and eosinophils may indicate bacteremic brucellosis.

Keywords: Brucellosis, brucella, machine learning methods, classification, bacteremia

Öz

Amaç: Brusellozun doğru ve erken teşhisi, hastalığın yayılımını yavaşlatmak ve hastalara hızlı tedavi sağlamak için çok önemlidir. Bu çalışmanın amacı, bruselloz hastalarında bakteriyemi tanısı için kan kültürü ve kemik iliğine kültürüne ihtiyaç duymadan bazı hematolojik ve biyokimyasal belirteçlere dayalı makina öğrenmesi temelli bir prediktif model geliştirmek ve bu yöntemin brusellozda bakteriyemiyi öngörmedeki önemini araştırmaktı.

Gereç ve Yöntem: Bu çalışmaya bruselloz tanısı konulan 18 yaş üstü 162 hasta dahil edilmiş olup, hastalar bakteriyemi durumuna göre iki gruba ayrıldı. Hastaların verileri retrospektif olarak toplandı ve makine öğrenmesi yöntemiyle analiz edildi. Brusellozda bakteriyemiyi tahmin etmek için yirmi demografik, hematolojik ve biyokimyasal laboratuvar parametresi ve 30 sınıflandırıcı kullanıldı. Sınıflandırıcılar Python programlama dili kullanılarak geliştirildi. Kullanılan yöntemlerin sınıflandırma performansını değerlendirmek için Doğruluk (ACC), f-ölçütü (F) ve alan altında ROC (AROC) ölçütleri kullanıldı. Tüm sınıflandırma yöntemleri 15 kat çapraz doğrulama test seti seçim yöntemi ile gerçekleştirildi. Kan kültürü pozitifliğinin sınıflandırılmasında en ayırt edici özelliklerin seçilmesi için özellik önemi yöntemi kullanıldı.

Bulgular: "Entropi" ölçütlü ekstratree sınıflandırıcı (ETC1), 0,5 ile 1,00 arasında değişen Acc değerleri, 0,53 ile 1 arasında değişen F değerleri ve 0,62 ile 1 arasında değişen AROC değerleri ile en iyi tahmin performansını gösterdi. Nötrofil yüzdesi, lenfosit yüzdesi, eozinofil yüzdesi, alanin aminotransferaz ve C-reaktif protein değerleri sırasıyla 0,723, 1,000, 0,920, 0,869 ve 0,769 skorlarıyla en ayırt edici özellikler olarak belirlendi.

Sonuç: Bu çalışma, ETC1 sınıflandırıcısının bruselloz hastalarında bakteriyemiyi belirlemede yardımcı olabileceğini, lenfosit, alanin aminotransferaz ve C-reaktif protein yüksekliğinin; nötrofil ve eozinofil düşüklüğünün bakteremik brusellozu gösterebileceğini göstermiştir.

Anahtar Kelimeler: Bruselloz, brusella, makine öğrenme yöntemleri, sınıflandırma, bakteriyemi

Corresponding *(İletişim)*: Mehmet ÇELİK, Karşıyaka district, 498th street, Tema Yeşilvadi buildings, A-block, Floor: 6, No: 12 Haliliye/Şanlıurfa, Turkey
E-mail *(E-posta)*: dr.mcelik12@gmail.com
Received *(Geliş Tarihi)*: 27.01.2023 **Accepted *(Kabul Tarihi)*:** 02.04.2023

INTRODUCTION

Brucellosis is a globally common zoonotic disease caused by Brucella spp. a Gram-negative intracellular bacterium. [1] It is endemic in many countries in Northern and Eastern Africa, Central Asia, India, Central and South America, and Mediterranean countries in Europe and the Middle East. [2] According to the World Health Organization (WHO), approximately 500,000 new brucellosis cases are reported annually. However, the true number of cases is higher than the reported number of cases.[3] Transmission of brucellosis is mostly due to the consumption of unpasteurized milk/dairy products in endemic countries and occupational exposure in developed countries.[4]

Symptoms and signs such as fever, sweating, fatigue, and osteoarthritis are frequently seen in brucellosis, and more serious conditions may occur in different organs.[4-6] Because the clinical presentation of brucellosis is variable and non-specific, confirmation of the diagnosis with laboratory tests is essential for providing appropriate treatment to the patient. Diagnosis of human brucellosis requires laboratory tests involving nucleic-acid amplification assays, serology, and culture. Bone marrow and blood culture are the gold-standard diagnostic tests.[3] The rate of blood culture positivity (bacteremia) in brucellosis varies between 15-90%. Especially in acute brucellosis, culture positivity rates are usually higher. [7] However, the results of these tests are delayed. The aim of this study is to predict bacteremia in acute brucellosis based on some hematological and biochemical markers of brucellosis patients without the need for blood culture and bone marrow culture. For this purpose, classification methods, one of the machine learning methods, were used in this study.

MATERIALS AND METHOD

The organization of this study is as follows:

Data Collection

Data were collected retrospectively in this study and 162 patients with a diagnosis of brucellosis were included in the study.

Patients over the age of 18 who were diagnosed with acute brucellosis and admitted to the Infectious Diseases and Clinical Microbiology outpatient clinic of Harran University Hospital between 2018 and 2020 were included in the study.

Hematologic and biochemical laboratory results and age/gender information of these patients were obtained from the hospital information management system.

Brucellosis definition: The criteria used for the diagnosis of brucellosis are growth in the culture of Brucella spp. in blood or other body fluids and together with clinical symptoms such as fever, sweating, chills, joint-muscle pain, headache, and weakness, being of serum Brucella tube agglutination titer equal to or greater than 1/160 or being of at least a four-fold titer increase in the serum sample taken at two-week intervals. The presence of clinical symptoms and signs for less than 2 months was considered acute brucellosis.[4]

Hematological and biochemical parameters: From the hematological examinations of the patients included in the study at the time of application; white blood cell (WBC), hemoglobin (HGB), hematocrit (HCT), platelet (PLT), neutrophil (NEUT), neutrophil % (NEUT%), lymphocyte (LYMP), lymphocyte % (LYMP%), monocytes % (MO%), eosinophil % (EOZ%), from biochemical tests; creatinine (CRE), aspartate aminotransferase (AST), alanine aminotransferase (ALT), total bilirubin (T.BIL), direct bilirubin (D.BIL), lactate dehydrogenase (LDH), FER, C reactive protein (CRP) results were evaluated.

Ethics Considerations

This study was supported by the Clinical Research Ethics Committee of Harran University with the number 22.10.21 on May 23, 2021. All procedures in the study were performed in accordance with the World Medical Association Declaration of Helsinki.

Data Preprocessing

Missing values were completed by using KNNImputer.[8] which is one of the Scikit-learn classes. KNNImputer is based on finding k neighbors nearest to the instance involving the missing values by using a distance measure (generally Euclidian distance). The missing values are completed by taking the arithmetic means of the relating values of the k-nearest neighbor.

Statistical Analyses

Statistical analysis was performed by using SPSS 21 package program (IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp). Continuous variables were presented as mean ±standard deviation. Categorical variables were shown as frequencies and percentages. The normality of continuous variables was tested by using the One-Sample Kolmogorov-Smirnov test. Two independent samples t-test was utilized for normally distributed variables and the Mann-Whitney U test for non-normally distributed variables to test whether the difference between the parameters of bacteremic and non-bacteremic patients was statistically significant. the p-value is lower than 0.05 and was considered statistically significant in all statistical tests.

Classification

In this study, classification which is one of the popular machine learning methods was used to predict the relationship between hematological and biochemical features and bacteremia. Classification is a process of predicting a function (f) between the features (X) and the labels (C) as f:X→C in the labeled data set.[9] The main objective of the classification is to assign the instances to a predefined class according to the features. Classification is performed in two main steps training and testing. In the training step, a classifier or

function (f) is predicted based on the relationship between the features (X) and the classes (C). The test step is to evaluate the classification performance of the predicted classifier by using various evaluation criteria. The original data set firstly is divided into two distinct subsets as training and test sets. There exist various test set selection techniques, especially Holdout and cross-validation. To determine the classification method, and provide the best prediction results, this study uses a cross-validation test set selection technique.[10]

Classification methods can be divided into two main categories as base and ensemble classifiers. The base classifier is based on predicting a single classifier for the classification problem. In this study, K-Nearest Neighbor (KNN), three Support Vector Machines (SVM) classifiers, Gaussian Naïve Bayes (GNB), Decision Tree Classifier (DTC), and Logistic Regression (LR) were used from the base classifier category. Ensemble classification methods are based on combining several base classifiers such as KNN, SVM, Bayes, etc. to improve the prediction performance. These methods can be divided into three main categories as bagging, boosting, and stacking.[11-14]

KNN classifier is based on finding the k nearest instances to new instance to be classified by using a distance function. For this objective, the distances of new instance to all instances in the data set are calculated. The distances are sorted as ascending and k instances nearest to new instance are found. The class of new instance is predicted by majority vote method.

SVM classifier is based on finding optimal hyperplane which maximizes the margin between the different classes. For the binary class and linearly separable classification problems, the process of SVM can be briefly described as follows. Let (x_i, y_i) be the training data set, where x_i is the ith input, consisting of p features and $y_i \epsilon \{-1, +1\}$ is the class label corresponding to of ith input. The separating line(for binary class problems) to be found can be written as follows:

$$w^T x + b = 0$$

Where w indicates the normal of the line and b is the bias. Support vectors are utilized to find the parameter of the line.

The SVM tries to find the hyperplane, which makes the margin ($\frac{2}{\|w\|}$) maximum. This problem is equivalent with the following optimization problem:

$$MinJ(w, b, \lambda) = \frac{1}{2\|w\|^2} - \sum_{j=1}^{k} \lambda_j [y_j(wx + b) - 1] \quad (2)$$

If the first derivative of the objective function given in Eq. (2) is taken separately with respect to w and b and set to 0, required equations for finding w and b are obtained. For two-class classification problems that are not linearly separable, the feature space is first transformed into a linearly separable space using a kernel function. Then, the objective function given in Eq. (2) is tried to be minimized.

Bayes Classification: Bayes classification is based on estimating probability of belonging of a new instance to given a class by using following equation:

$$P(C_j|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C_j) * P(C_j)}{P(\boldsymbol{X})} \; j = 1,2, \ldots, c \quad (3)$$

Where c is the number of class, $P(C_j)$ is the probability of class C_j, $P(\boldsymbol{X}|C_j)$ is the conditional probability that the instance is X, given that the class is C_j and lastly $P(C_j|\boldsymbol{X})$ is the conditional probability that the class is C_j, given that the instance is X. The instance, X, is assigned to the class with the highest probability of $P(C_j|\boldsymbol{X})$.

Logistic Regression: Similar to the Bayesian classification method, logistic regression estimates the probabilities of classes for a given X instance. For the binary outcome classification problems (such as the presence or absence of a disease), the probability is estimated as follows:

$$P(Y = 1|\boldsymbol{X}) = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}} \quad (4)$$

Where, $P(Y = 1)$ is the probability of presence of interested outcome, such as a disease. The probability of absence of interested outcome is estimated by using following equation:

$$P(Y = 0) = 1 - P(Y = 1) \quad (5)$$

In fact, logistic regression estimates the model parameters ($\boldsymbol{\beta}$) given in Eq. (4). For this objective, Eq. (5) is modified as follows:

$$\log\left(\frac{P(Y = 1|\boldsymbol{X})}{1 - P(Y = 1|\boldsymbol{X})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (6)$$

Decision Trees (DT): DT is based on partitioning of the feature space into homogenous subsets, recursively. As a result of DT classification process, a tree-like structure consisting of a root node, multiple internal nodes, terminal nodes and branches is obtained. The root and internal nodes correspond to a feature in the data set. The terminal nodes include class labels. To construct the tree structure, some evaluation criteria such as information entropy, gain ratio and Gini index are used (Ahybridensemblemethodforpulsarcandidateclassification--AstrophysicsandSpaceScience.pdf). These criteria rank the features according to its contribution to classification performance. The root node is the feature that contributes the most to classification success. In other words, it is the most distinguishing feature. The root node is splitted into the branches according to its categories or values. The root node (internal node) is determined for each branch again.

This process is repeated until a terminal node is reached. As can be understood from here, internal nodes are determined by using the data set which includes the instances providing specified property of its parent node, while root node is determined by using whole data set.

Ensemble Classification Methods: Ensemble classification methods are based on combining of several base classification methods to improve the prediction performance. These methods can be divided into three main categories as bagging, boosting, and stacking (**Figure 2**). General working principle of the ensemble methods can be summarized as follows. In bagging, firstly, N samples, each of with n dimensions are created by using simple random sampling (with replacement) method from original training set (**Figure 2a**). Each sample is trained via a base classifier such as KNN, SVM, Bayes etc. simultaneously. The prediction results are aggregated by using some methods such as weighted average or majority voting. The base classifiers execute independently of each other in bagging. The main idea behind boosting method is to obtain a strong classifier by combining weak classifiers (**Figure 2b**). In this method, a single sample with n dimension is constituted at the beginning of the classification process. The selected base classifier is applied to the sample and misclassified instances are identified. In the next step, a new sample is created by assigning higher weights to misclassified instances. Base classifier is applied to new sample and misclassified instances are determined again. The weights of misclassified instances are increased. This process is repeated until the predetermined number of repetitions or the desired training error is reached. As a result of the process, a high-performance classifier is obtained by combining the weak classifiers. As can be understood, the boosting is a method working sequentially while the bagging is method working parallelly. The bagging and boosting have in common is that they both use the same base classifier during the classification process. The stacking ensemble method (**Figure 1**) directly works on the original training set without creating sub-samples. The training set is learned by using different types of base classifiers and classification results are combined by using a meta-classifier.

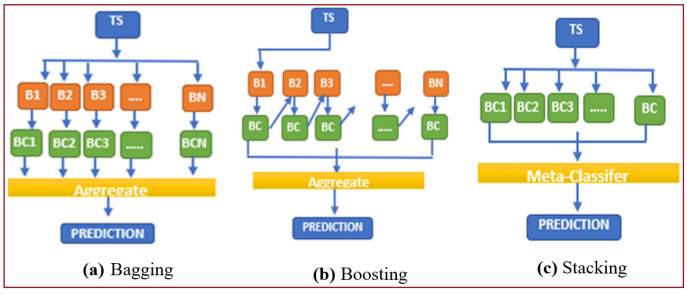

Figure 1. Ensemble Methods

In this study, 7 base classifiers and 23 ensemble classifiers are used. The reason of using numerous classifiers is to identify the predictive model best reflecting the relationship between the hematological and biochemical markers and the bacteremia.

Evaluation Criteria

A confusion matrix is utilized to compare and evaluate the performance of classification methods. The confusion matrix is given as in **Figure 2** for our study.

		Predicted Class	
		Nonbacteremic	Bacteremic
Actual Class	Nonbacteremic	True Negatives (TN)	False Positives (FP)
	Bacteremic	False Negatives (FN)	True Positives (TP)

Figure 2. Confusion Matrix

In **Figure 2**, TN is the number of participants who are classified as non-bacteremic while non-bacteremic in actual (correctly classified), FP is the number of participants who are classified as bacteremic while non-bacteremic in actual (incorrectly classified), FN is the number of participants who are classified as a non-bacteremic while is bacteremic in actual (incorrectly classified) and lastly TP is the number of participants who are classified as a bacteremic while is bacteremic in actual. Some evaluation criteria obtained by using confusion matrix can be given as follows [15]:

$$Accuracy\ (Acc) = \frac{TP + TN}{TN + FP + FN + TP} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$_{F}-Measure\ (F) = \frac{2 * Precision * Recall}{Precision + recall} \quad (4)$$

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \quad (5)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN} \quad (6)$$

AROC is also used to evaluate the performance of the classification methods. AROC refers to the area under the curve obtained by plotting the TPR against the FPR The closer the ACC, Precision, Recall, F-Measure, AROC, and TPR values of a classification method are to 1, the higher the classification success. It is desirable that the FPR be close to 0. In this study, ACC, F, and AROC performance metrics are used to compare the classification methods used.

Test Set Selection

Classification consists of two main steps as training and testing. In training step, classification model is predicted by utilizing the relationship between independent and dependent (class) variables. Test step includes evaluating the performance of predicted classification model. Data

set firstly should be divided into two distinct subsets as training set and test set to perform these steps. Thus, test set selection is important subject in the classification. Two methods have been widely used as Hold-out and k-fold cross-validation for this objective. In the Hold-out method, training percentage firstly is determined and the instances in the determined percentage of the data set constitute the training set, the remaining part the test set. In the k-fold cross validation method, data set is firstly divided into k subsets. Hold-out method is repeated ask times such that each time a subset is selected as test set and remaining k-1 subset sets as training set. **Figure 3** illustrates k-5-folds cross validation method.

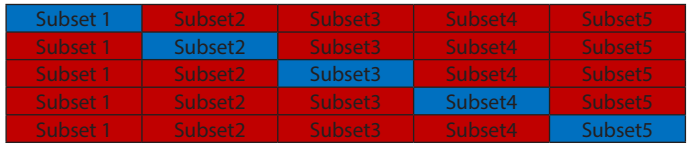

Figure 3. 5-fold cross validation

In **Figure 3**, subsets with blue color indicate the test sets, subsets with red color the training sets. According to **Figure 3**, Subset1 is selected as test set, remaining part (Subset2 + Subset3+ Subset 4+ Subset 5) training in the first execution of the algorithm. In the second execution, Subset 2 is selected as test set, remaining part training (Subset1+ Subset3+ Subset4+ Subset5) set. This procedure is repeated until each subset is the test set once.

Experimental Setup

This section consists of three subsections. First subsection gives the brief information about the statistical properties of the data set. In Subset 2, methods having the highest classification performance based on the cross-validation are determined. Subsect 3 provides the results of feature importance.

RESULTS

Statistical Properties of Data Set

The data set contained a percentage of 2.38% missing values. Missing values were predicted using the KNNImputer method. This study aims to predict the blood culture positivity by using the classification from machine learning algorithms. For this objective, we collected 162 patients' data with diagnosing of acute brucellosis, 54.9% (n=89) of whom are in blood culture negativity group (labelled as 0), 45.1% (n=73) in blood culture positivity group (labelled as 1), 54.9% (n=89) female and 45.1% (n=73) male. The patients in the blood culture negativity group of 41.6% (n=37) were female, 58.4 % (n=52) male, the patients in the blood culture positivity group of 49.3% (n=36) were female, 50.1% (n=37) male. 20 features relating to these participants were studied. The mean± standard deviation of the features is given in **Table 1**.

Table 1. Descriptive Statistics

Features	Overall	0 (n=89)	1 (n=73)	p
Age	39.13±14.76	40.28±15.43	37.72±13.99	0.36
WBC	7067.21±2889.26	7756.22±2756.38	6227.17±2842.24	0.00
HGB	13.06±1.66	13.19±1.57	12.91±1.76	0.28
HCT	38.97±5.10	39.07±5.52	38.85±4.55	0.79
NEUT	4052.30±2451.43	4725.87±2567.68	3231.10±2033.54	0.56
NEUT %	54.11±12.89	58.67±12.15	48.52±11.56	0.00
LYMP	2343.98±913.38	2192.40±801.17	2528.77±1008.96	0.02
LYMP %	34.89±12.06	30.21±10.90	40.59±10.95	0.00
MO %	9.19±2.85	8.91±3.00	9.53±2.64	0.16
EOZ %	1.54±1.61	2.02±1.85	0.94±0.95	0.00
PLT	245.51±81.33	249.06±74.17	241.18±90.05	0.54
CRE	0.78±0.27	0.74±0.15	0.82±0.37	0.16
AST	37.69±44.39	28.20±26.17	49.25±57.64	0.00
ALT	35.93±32.59	27.99±28.11	45.62±35.17	0.00
T. BİL	0.64±0.36	0.59±0.33	0.69±0.40	0.02
D.BİL	0.28±0.16	0.26±0.13	0.31±0.18	0.02
LDH	308.81±192.93	268.92±86.77	357.45±263.93	0.00
FER	306.69±383.61	238.11±269.88	390.30±476.60	0.00
CRP	3.08±3.81	2.57±4.10	3.69±3.35	0.00

*WBC: White blood cell, HGB: Hemoglobin, HCT: Hematocrit, NEUT: Neutrophil, NEUT%: Neutrophil %, LYMP: Lymphocyte, LYMP %: Lymphocyte %, MO %: Monocytes %, EOZ %: Eosinophil %, PLT: Platelet, AST: Aspartate aminotransferase, ALT: Alanine aminotransferase, T.BİL: Total bilirubin, D.BİL: Direct bilirubin, LDH: Lactate dehydrogenase, FER: Ferritin, CRP: C-reactive protein

As can be seen in **Table 1**, the mean of WBC, NEUT%, and EOZ% values are significantly higher in the non-bacteremic class. The mean of LYMP, LYMP%, AST, ALT, T.BİL, D.BİL, LDH, FER and CRP are significantly higher than in the bacteremic class. **Figure 4** denotes the box plot of the features for each class.

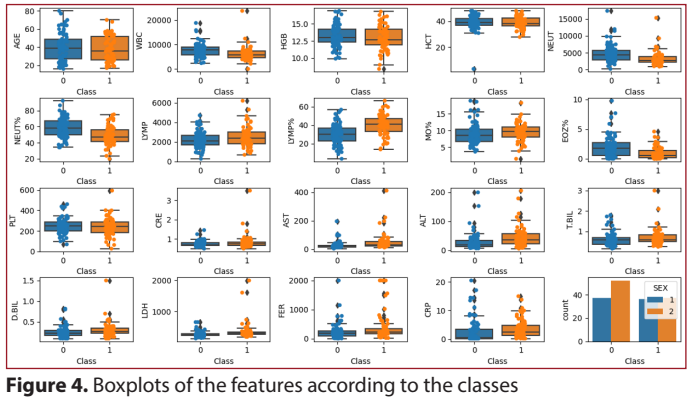

Figure 4. Boxplots of the features according to the classes

Boxplot provides visualization of the distribution of the data. The lines in the middle of the boxes correspond to the median. The start and the finish lines of the boxes represent the first (Q_1) and third (Q_3) quartiles. The difference between two horizontal lines (whiskers) is a measure of heterogeneity in data and is generally calculated as [Q_1-1.5x (Q_3-Q_1) Q_3+1.5x(Q_3-Q_1)]. The instances outside of the horizontal lines represent the outliers. According to this, the medians of Age, WBC, HGB, HCT, NEUT, NEUT%, EOZ%, and PLT are higher in the non-bacteremic class, while the medians of the other features in

the bacteremic class. When the outliers are not considered, the spread of age, WBC, HCT, NEUT, NEUT%, MO%, and EOZ% are higher in the non-bacteremic class, the spread of HGB, LYMP, LYMP%, PLT, CRE, AST, and ALT are higher in the bacteremic class.

Model Development

We utilized the Python environment and Scikit-learn library for executing the classification methods. 30 classifiers, 7 of which are based, 23 of which are ensemble were included in this study. The classification methods, their abbreviations, and forms of usage were given in **Appendix 1**. The cross-validation test set selection method was used, and the number of folds was selected as 15. ACC, F, and AROC were calculated for all folds and classification methods. **Figure 5** shows the box plot of Acc, F, and AROC values obtained from 15 folds.

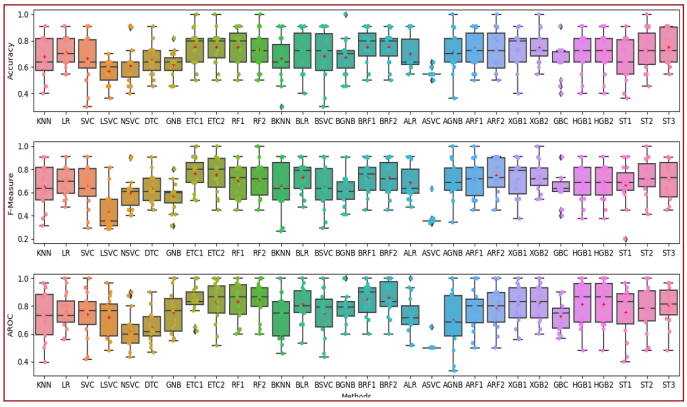

Figure 5. Boxplot of 15- fold cross validation Acc, F and AROC of classification methods

Table 2 gives the arithmetic means and 95% confidence interval of performance metrics for all classification methods, separately.

From **Figure 5** and **Table 2**, the highest Acc values were obtained from ETC1, ETC2, RF1, BRF1, BRF2, XGB2, and ST3, the highest F value from ETC1, and the highest AROC value from BF2. The lowest Acc, F, and AROC values were obtained from the ASCV. From these results, it can be said that ETC1, ETC2, RF1, BRF1, BRF2, XGB2, and ST3 generally provided good classification performance, while ASVC had the worst performance. Acc values ranged between 0.4 and 1, F values between 0.375 and 1, and AROC values between 0.48 and 1 in these methods. However, it is observed that the ETC1 is more successful in classification when evaluating three performance metrics, simultaneously.

3.3 Feature Importance

In this subsection, it was determined which features are the most distinctive in predicting bacteremia in brucellosis. For this objective, the feature importance method was used. This method was executed together with ETC1 which was found as the best method in the classification of bacteremia. All feature scores were normalized into a range of 0-1, with a minimum score of 0 and a maximum score of 1. **Figure 6** gives the bar plots of the feature scores obtained.

Table 2. Mean and confidence interval of the performance metrics.

Methods	Mean [95% Confidence Interval]		
	Acc	F	AROC
KNN	0.68 [0.58 0.77]	0.65 [0.54 0.76]	0.73 [0.63 0.83]
LR	0.72 [0.65 0.78]	0.70 [0.62 0.77]	0.76 [0.69 0.83]
SVC	0.66 [0.56 0.76]	0.65 [0.55 0.76]	0.74 [0.65 0.83]
LSVC	0.57 [0.51 0.63]	0.43 [0.34 0.52]	0.72 [0.64 0.80]
NSVC	0.61 [0.53 0.69]	0.60 [0.51 0.68]	0.63 [0.55 0.70]
DTC	0.65 [0.59 0.72]	0.64 [0.57 0.71]	0.65 [0.58 0.71]
GNB	0.62 [0.56 0.68]	0.56 [0.49 0.63]	0.76 [0.67 0.84]
ETC1	0.75 [0.68 0.83]	0.78 [0.69 0.85]	0.84 [0.78 0.90]
ETC2	0.75 [0.68 0.83]	0.75 [0.66 0.84]	0.83 [0.75 0.92]
RF1	0.75 [0.68 0.83]	0.70 [0.62 0.79]	0.83 [0.75 0.91]
RF2	0.73 [0.65 0.82]	0.71 [0.62 0.80]	0.85 [0.78 0.92]
BKNN	0.66 [0.57 0.76]	0.66 [0.54 0.78]	0.73 [0.65 0.82]
BLR	0.73 [0.64 0.82]	0.73 [0.66 0.81]	0.81 [0.74 0.88]
BSVC	0.68 [0.58 0.80]	0.65 [0.54 0.75]	0.75 [0.66 0.84]
BGNB	0.67 [0.60 0.75]	0.61 [0.54 0.68]	0.79 [0.72 0.85]
BRF1	0.75 [0.68 0.83]	0.73 [0.65 0.81]	0.85 [0.78 0.92]
BRF2	0.75 [0.68 0.82]	0.73 [0.65 0.81]	0.86 [0.79 0.93]
ALR	0.70 [0.63 0.78]	0.69 [0.60 0.77]	0.74 [0. 66 0.81]
ASVC	0.56 [0.54 0.57]	0.37 [0.33 0.41]	0.51 [0.49 0.53]
AGNB	0.71 [0.62 0.80]	0.69 [0.60 0.79]	0.70 [0.58 0.82]
ARF1	0.74 [0.66 0.82]	0.72 [0.63 0.81]	0.77 [0.68 0.85]
ARF2	0.73 [0.65 0.82]	0.75 [0.66 0.84]	0.79 [0.71 0.87]
XGB1	0.73 [0.65 0.82]	0.72 [0.63 0.81]	0.81 [0.73 0.89]
XGB2	0.75 [0.68 0.82]	0.73 [0.66 0.81]	0.82 [0.75 0.89]
GBC	0.64 [0.57 0.71]	0.63 [0.56 0.70]	0.73 [0.67 0.79]
HGB1	0.71 [0.62 0.79]	0.69 [0.60 0.78]	0.82 [0.73 0.91]
HGB2	0.71 [0.62 0.79]	0.69 [0.60 0.78]	0.82 [0.73 0.91]
ST1	0.64 [0.55 0.73]	0.67 [0.57 0.76]	0.75 [0.65 0.85]
ST2	0.73 [0.64 0.82]	0.72 [0.63 0.81]	0.78 [0.68 0.86]
ST3	0.75 [0.68 0.83]	0.71 [0.62 0.81]	0.80 [0.73 0.88]

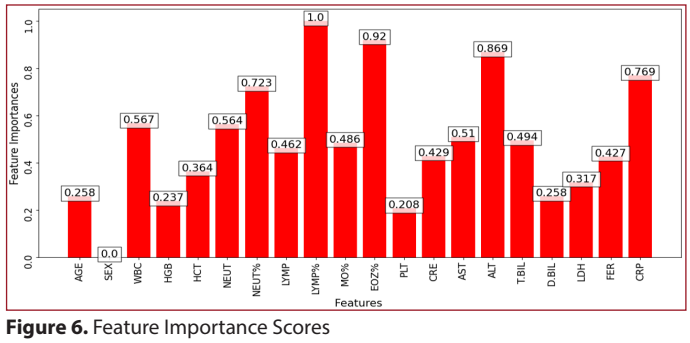

Figure 6. Feature Importance Scores

When examining **Figure 4**, the most distinctive features are NEUT % (0.723), LYMP % (1.000), EOZ% (0.920), ALT (0.869), and CRP (0.769). Besides, the WBC, NEUT, and AST have also moderate importance.

Lastly, in this section, the ETC1 classification method was applied to the data sets including all features and only important features, separately and the confusion matrixes given in **Figure 7**. were obtained.

Appendix1: Classification methods, their abbreviations and form of usage

Classification Methods	Abbreviation	Type	Form of Usage
K-Nearest Neighbor	**Knn**	Base	n_neighbors=10
Logistic Regression	LR	Base	solver='**lbfgs**'
	SVC	Base	decision_function_shape='ovo'
Support Vector Machines	LSVC	Base	LinearSVC()
	NSVC	Base	NuSVC()
Decision Trees	DTC	Base	max_depth=20, random_state=42
Gaussian Naïve **Bayes**	GNB	Base	GaussianNB()
Extra Trees	ETC1	Ensemble	criterion ="**entropy**", max_depth=20, bootstrap=**True**
	ETC2	Ensemble	criterion ="**gini**", max_depth=20, bootstrap=**True**
Random Forest	**Rf1**	Ensemble	criterion='**entropy**', max_depth=20, max_samples=20
	Rf2	Ensemble	criterion='**gini**', max_depth=20, max_samples=20
Bagging	B**Knn**	Ensemble	base_estimator=KNeighborsClassifier(n_neighbors=10), max_samples=0.7,max_features=0.7, n_estimators=20
	BLR	Ensemble	base_estimator=LogisticRegression(solver='**lbfgs**'), max_samples=0.7,max_features=0.7, n_estimators=20
	BSVC	Ensemble	base_estimator=svm.SVC(decision_function_shape='ovo'), max_samples=0.7, max_features = 0.7, n_estimators=20
	BGNB	Ensemble	base_estimator=GaussianNB(), max_samples=0.7,max_features=0.7, n_estimators=50
	B**Rf1**	Ensemble	base_estimator=RandomForestClassifier(criterion='**gini**', max_depth=200, max_samples=50), max_samples=0.7,max_features=0.7, n_estimators=20
	B**Rf2**	Ensemble	base_estimator=RandomForestClassifier(criterion='**entropy**', max_depth=200, max_samples=50), max_samples=0.7,max_features=0.7, n_estimators=20
Adaboost	ALR	Ensemble	base_estimator=LogisticRegression(solver='**lbfgs**'), algorithm="**SAMME**", n_estimators = 100, random_state=**None**
	ASVC	Ensemble	base_estimator=svm.SVC(decision_function_shape='ovo'), algorithm="**SAMME**", n_estimators = 100,random_state=**None**
	AGNB	Ensemble	base_estimator=GaussianNB(), algorithm="**SAMME**", n_estimators = 100,random_state=**None**
	A**Rf1**	Ensemble	base_estimator=RandomForestClassifier(criterion='**gini**', max_depth=50, max_samples=20), algorithm="**SAMME**", n_estimators = 100,random_state=**None**
	A**Rf2**	Ensemble	(base_estimator=RandomForestClassifier(criterion='**entropy**', max_depth=50, max_samples=20),algorithm="**SAMME**", n_estimators = 100,random_state=**None**
XGBoost	XGB1	Ensemble	base_score=0.5,learning_rate=0.2, n_estimators=100, objective='**binary:logistic**', tree_method='**exact**', booster='**gbtree**'
	XGB2	Ensemble	base_score=0.5,learning_rate=0.2, n_estimators=100, objective='**binary:logistic**', tree_method='**exact**',booster='**gblinear**'
Gradient Boosting	GBC	Ensemble	(n_estimators=100, learning_rate=0.2, max_depth=50, random_state=0
Histogram-Based Gradient Boosting	HGB1	Ensemble	loss='**log_loss**'
	HGB2	Ensemble	loss='**binary_crossentropy**'
	ST1	Ensemble	level0.append(('**rf1**', RandomForestClassifier(criterion='**entropy**', max_depth=20, max_samples=20))) level0.append(('**rf2**',RandomForestClassifier(criterion='**gini**', max_depth=20, max_samples=20))) level0.append(('**df1**', DecisionTreeClassifier(max_depth=20, random_state=42))) level1 = ExtraTreesClassifier(criterion ="**gini**", max_depth=20, bootstrap=**True**) s1 = StackingClassifier(estimators=level0, final_estimator=level1, cv=5)
	ST2	Ensemble	level0.append(('**knn**', KNeighborsClassifier(n_neighbors=10))) level0.append(('**cart**', DecisionTreeClassifier(max_depth=20, random_state=42))) level0.append(('**bayes**', GaussianNB())) level1 = RandomForestClassifier(criterion='**gini**', max_depth=100, max_samples=20) s1 = StackingClassifier(estimators=level0, final_estimator=level1, cv=5)
	ST3	Ensemble	level0.append(('**knn**', KNeighborsClassifier(n_neighbors=10))) level0.append(('**cart**', DecisionTreeClassifier(max_depth=20, random_state=42))) level0.append(('**bayes**', GaussianNB())) level1 = RandomForestClassifier(criterion='**gini**', max_depth=100, max_samples=20) s1 = StackingClassifier(estimators=level0, final_estimator=level1, cv=5)

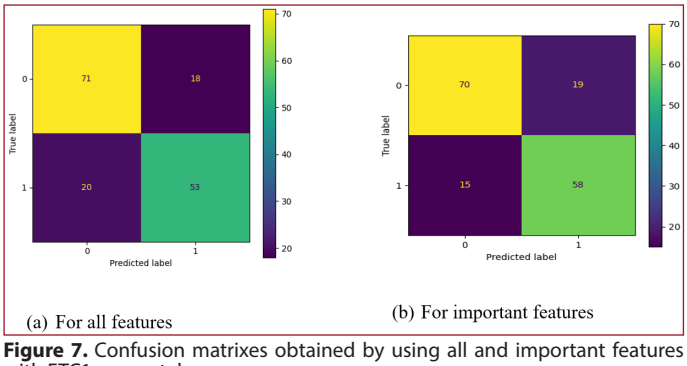

Figure 7. Confusion matrixes obtained by using all and important features with ETC1, separately

As can be seen in **Figure 7**, ETC1 correctly classified 71 of 89 (80%) instances in the non-bacteremic group and 53 of 73 (73%) instances in the bacteremic group when all features were used in the classification. When only important features were considered, the Acc value was found as 0.79. Besides, ETC1 correctly classified 70 of 89 (79%) instances in non-bacteremic and 58 of 73 (79%) instances in bacteremic. As can be understood from this result, the TP rate was increased for the bacteremic group in the second case.

DISCUSSION

Brucellosis is one of the most dangerous zoonotic diseases. It causes significant clinical conditions in humans and leads to a significant loss of productivity in the livestock industry.[16] A definitive diagnosis of brucellosis is the isolation of bacterium from blood, bone marrow or body fluids, and other tissues.[17,18] The presence of bacteremia is important in brucellosis. When serology is negative due to various factors, the diagnosis of brucellosis is supported by a positive culture. In addition, the presence of bacteremia may be important for treatment change in experimental protocols and appears to provide an increased risk for relapse of the disease. The presence of bacteremia is synonymous with the development of secondary seeding and focal complications.[7,19] There may be clinical and laboratory differences in brucellosis patients with and without bacteremia. Kaduna et al.[7] found that AST, ALT elevation, and leukopenia were to be higher in bacteremic patients than in non-bacteremic patients. In the study of Qie et al.[20] thrombocytopenia and CRP elevation were found to be higher in bacteremic patients. In a study conducted on pediatric patients with a diagnosis of brucellosis, high CRP, ALT, and AST levels were found to be important markers for blood culture positivity drawing a conclusion that lower hemoglobin, iron, and vitamin D levels and higher leukocyte, CRP, and ferritin levels were associated with blood culture positivity rate. In these studies, the statistical characteristics of the laboratory findings of the patients were generally emphasized to identify important biomarkers in distinguishing between bacteremic and non-bacteremic patients.[21,22] Recently, classification, which is one of the machine learning methods has been widely exploited to diagnose a disease and to determine important features for diagnosing the disease. The correct and early diagnosis of brucellosis is very crucial. The definitive diagnostic test is the blood culture, but it is time-consuming. Therefore, we aimed to investigate whether some hematological and biochemical parameters are useful in predicting bacteremia with the help of the machine learning method, which is one of the artificial intelligence applications.

Some studies about this subject can be summarized as follows: Chicco and Jurman[23] used the RF classification method to diagnose hepatitis C diseases and to determine the most diagnostic features for hepatitis C. They found that RF provided good performance for diagnosing hepatitis C and AST and ALT levels were diagnostic features. Chicco and Oneto[24] applied nine classification methods to a dataset of electronic health records, consisting of 364 patients and 29 features to predict septic shock. As a result of this study, they observed that the NB classifier had the highest accuracy value, and creatinine, Glasgow coma scale, mean arterial pressure, and initial procalcitonin were the most diagnostic features to predict septic shock.[24] Xiong et al.[25] employed RF, SVM, and LR classification methods to predict the severity of illness of COVID-19 patients at the time of hospital admission and to identify the most important features in distinguishing severe COVID-19 patients. The dataset used in this study consists of 23 features and a total of 287 patients, 36.6% of whom were severe cases and 63.4% of whom were non-severe cases. They concluded that RF yielded the best performance and chest-CT, neutrophil to lymphocyte ratio, lactate dehydrogenase, and D-dimer were important features. Kou et al.[26] proposed a feature representation algorithm to identify the pathogenicity of the influenza B virus. In the study, firstly, 67 RF classifiers were used to determine the informative features. Then, the classification performances of RF, SVM, NB, and KNN were compared based on the optimal features set, and lastly, the RF classifier was selected for pathogenicity identification of IBV according to evaluation criteria. Herein we aimed to predict the classification of bacteremia in patients with acute brucellosis based on some hematological and biochemical markers. Besides, it investigated the most important hematological and biochemical features in predicting bacteremia. The main objective of this study is to decide faster whether the patients are bacteremic or not by identifying the important features indicating the existence of bacteremia. To our best knowledge, this study is the first study conducted for this topic.

To achieve this objective, a dataset consisting of 162 patients with a diagnosis of acute brucellosis, 89 (54.9%) of whom has non-bacteremic, 73 (45.1%) bacteremic, and 20 features including age, sex, and 18 hematological and biochemical markers were collected retrospectively. 30 classification methods, 7 of which were base classifiers, and 23 of which were ensemble classifiers were applied to the collected bacteremia data set. Firstly, statistical

characteristics of the features used were examined according to a class of bacteremia. The mean of WBC, NEUT% and EOZ% values were significantly higher in the non-bacteremic group. The mean of LYMP, LYMP%, AST, ALT, T.BIL, D.BIL, LDH, FER, and CRP levels were significantly higher than in the bacteremic group. In the second step, the classification process had been performed for each method separately and the method providing highest classification performance was determined according to three performance metrics. According to means of the performance metrics, it was decided that ETC1 had the highest classification performance. The means of ACC, F, and AROC values were found as 0.75, 0.78, and 0.84 for ETC1, respectively. To determine the most distinguishing features in the classification of bacteremia, feature importance was used. The normalized feature importance scores were found as 0.723, 1.000, 0.93, 0.869, and 0.769 for NEUT %, LYMP %, EOZ %, ALT, and CRP, respectively. It concluded that the most important feature was the LYMP %. Besides, it was observed that the WBC, NEUT, and AST had also moderate importance. Lastly, ETC1 was applied to the data sets including all features and only important features separately, the results were evaluated by utilizing the confusion matrix. When considering all features simultaneously, ETC1 correctly classified 71 of 89 (80%) instances in the non-bacteremic group and 53 of 73 (73%) instances in the bacteremic group. When the ETC1 was executed by considering only the important features, 70 of 89(79%) instances in the non-bacteremic group and 58 of 73(79%) instances in the bacteremic group were correctly classified. As a result of the study, it was observed that the high levels of LYMP %, ALT, and CRP and low levels of NEUT% and EOZ% can indicate bacteremia in brusellosis.

Limitations

This study is a single center, had limited number of patients and retrospective design.

CONCLUSIONS

The definitive diagnosis of brucellosis is the isolation of Brucella spp. in blood or bone marrow culture. However, despite technological developments, the growth of bacteria in culture and identification after growth is time-consuming. For all that, the levels of some laboratory biomarkers may differ in bacteremic and non-bacteremic patients, and we used the machine learning algotritms to predict bacteremia in brucellosis. Our results showed that the ETC1 classifier can be used as a predictive tool for bacteremia in brucellosis patients based on hematological and biochemical parameters. The feature importance method was used for determining the most distinguishing features of bacteremia. It is concluded that the most important feature was the LYMP% and that the WBC, NEUT, and AST have also moderate importance, and that high levels of LYMP %, ALT and CRP, and low levels of NEUT %, and EOZ % are parameters that can predict bacteremia.

ETHICAL DECLARATIONS

Ethics Committee Approval: This study was supported by the Clinical Research Ethics Committee of Harran University with the number 22.10.21 on May 23, 2021.

Informed Consent: Because the study was designed retrospectively, no written informed consent form was obtained from patients.

Referee Evaluation Process: Externally peer-reviewed.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Author Contributions: All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Akhvlediani T, Bautista CT, Garuchava N, Sanodze L, Kokaia N, Malania L, et al. Epidemiological and clinical features of brucellosis in the country of Georgia. PLoS One 2017;12:e0170376.
2. Bahmani N, Bahmani A. A review of brucellosis in the Middle East and control of animal brucellosis in an Iranian experience. Reviews in Medical Microbiology 2022;33(1):e63-e69.
3. Yagupsky P, Morata P, Colmenero JD. Laboratory diagnosis of human brucellosis. Clinical Microbiology Reviews 2020;33(1):e00073-19.
4. Buzgan T, Karahocagil MK, Irmak H, et al. Clinical manifestations and complications in 1028 cases of brucellosis: a retrospective evaluation and review of the literatüre. Int J Infect Dis 2010;14(6):e469-478.
5. Moosazadeh M, Nikaeen R, Abedi G, Kheradmand M, Safiri S. Epidemiological and clinical features of people with Malta fever in iran: a systematic review and meta-analysis. Osong Public Health and Research Perspectives 2016;7(3):157–67.
6. Zheng R, Xie S, Lu X, et al. A systematic review and meta-analysis of epidemiology and clinical manifestations of human brucellosis in China. BioMed research international 2018;2018:Article ID 5712920.
7. Kadanali A, Ozden K, Altoparlak U, Erturk A, Parlak M. Bacteremic and nonbacteremic brucellosis: clinical and laboratory observations. Infection 2009;37(1):67-9.
8. Choudhury A, Kosorok MR. Missing data imputation for classication problems. arXiv:2002.10709 2020;1-27.
9. Bailly A. Time Series Classification Algorithms with Applications in Remote Sensing. General Mathematics [math.GM]. Université Rennes 2, 2018. English.
10. Shahub S, Upasham S, Ganguly A, Prasad S. Machine learning guided electrochemical sensors for passive sweat cortisol detection. Sens Bio-Sens Res 2022;38:1-11.
11. Breiman L. Bagging predictors. Machine learning 1996;24(2):123-40.
12. Freund Y. Boosting a weak learning algorithm by majority. Information and computation 1995;121(2):256-85.
13. Wolpert DH. Stacked generalization. Neural networks 1992;5(2):241-59.
14. 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.
15. Safdari R, Deghatipour A, Gholamzadeh M, Maghooli K. Applying data mining techniques to classify patients with suspected hepatitis C virus infection. Intelligent Medicine 2022;21:24.
16. Megahed A, Kandeel S, Alshaya DS, et al. A comparison of logistic regression and classification tree to assess brucellosis associated risk factors in dairy cattle. Preventive Veterinary Medicine 2022;203:105664.

17. Al Dahouk S, Tomaso H, Nöckler K, Neubauer H, Frangoulidis D. Laboratory-based diagnosis of brucellosis – a review of the literature. Part I: techniques for direct detection and identification of Brucella spp. Clin Lab 2003;49(9–10):487–505.
18. Al Dahouk S, Nöckler K. Implications of laboratory diagnosis on brucellosis therapy. Exp Rev Anti-infective Therapy 2011;9(7):833-45.
19. Pappas G, Papadimitriou P. Challenges in Brucella bacteraemia. Int J Antimicrobial Agents 2007;30:29-31.
20. Qie C, Cui J, Liu Y, Li Y, Wu H, Mi Y. Epidemiological and clinical characteristics of bacteremic brucellosis. J Int Med Res 2020;48(7):1-7.
21. Özdem S, Tanır G, Öz FN, et al. Bacteremic and Nonbacteremic Brucellosis in Children in Turkey. J Tropic Pediatr 2022;68(1): 114.
22. Kara SS, Cayir Y. Predictors of blood culture positivity in pediatric brucellosis. J Coll Physicians Surg Pak 2019;29(07):665-70.
23. Chicco D, Jurman G. An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis. IEEE Access, 2021; 9:24485-98.
24. Chicco D, Oneto L. Data analytics and clinical feature ranking of medical records of patients with sepsis. BioData Mining 2021;14(12):1-22.
25. Xiong Y, Ma Y, Ruan L, Li D, Lu C, Huang L. Comparing different machine learning techniques for predicting COVID-19 severetity. Infectious Diseases of Poverty 2022;11(1):1-9.
26. Kou Z, Fan X, Li J, Shao Z, Qiang X. Using amino acid features to identify the pathogenicity of influenza B virüs. Infect Dis Poverty 2022;11(1):1-13.