Open Journal of
Clinical & Medical
Case Reports

Volume 9 (2023)
Issue 27

Research Article ISSN: 2379-1039

Application of specialized machine learning for the prediction of Brucellosis disease

MH Tito*; M Arifuzzaman; MHE Jannat; A Nasrin; M Asaduzzaman; MM Hossain; SM Maruf; Afzal Haq Asif

***Corresponding Author: MH Tito**
Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.
Email: mokammel.17asvm014@bsmrstu.edu.bd

Abstract

Brucellosis is a bacterial disease that affects both humans and animals and is widely prevalent in many countries. In this study, we present a machine learning application for the prediction of brucellosis disease. The application utilizes various algorithms to analyze patient data and make predictions about the likelihood of infection. The results demonstrate the effectiveness of the machine learning approach in accurately predicting brucellosis disease and can help in early diagnosis and treatment. The study highlights the potential of machine learnings in improving disease diagnosis and management and can be a valuable resource for healthcare professionals in their efforts to control and eliminate brucellosis. The Radial Basis Function (RBF) is found with the highest Kappa Statistics, indicating that it has a high level of accuracy in its predictions. The attribute ranking was accomplished, and the results are summarized as: Average Wind Speed > Non-Pasteurized Dairy Product Ratio > Rural ratio > year > Month > male ratio > Season.

Keywords: Brucellosis; Cattle disease; Machine learning; J48 Tree; Apriori; SVM.

Introduction

Brucellosis is a bacterial disease that has been affecting humans and livestock for centuries, with over 500,000 new cases reported annually worldwide. The disease is caused by contact with infected animals, consumption of contaminated animal products, or inhaling airborne agents, and is a major public health hazard due to the increase of animal industries and urbanization [1]. The traditional methods of predicting and controlling brucellosis have been limited in their accuracy and efficacy, leading researchers to explore alternative solutions [3,4]. This is where machine learning comes in. Machine learning is a field of artificial intelligence that uses algorithms and statistical models to analyze data and make predictions. This technology has the potential to revolutionize the way we approach the prediction and control of diseases like brucellosis [5,6]. Machine learning (ML) techniques found to be beneficial and helpful in different branches of engineering. In this article, we have explored the use of different machine learnings for predicting brucellosis disease. We are examining different machine learning algorithms and their effectiveness in predicting brucellosis outbreaks, and how this information can be used to minimize its impact on public health. By using machine learning, we can analyze vast amounts of data in a matter of short time, improving the accuracy and efficiency of predicting brucellosis outbreaks. Machine learning can be applied in various stages of brucellosis prediction, including data preparation, model selection, and evaluation of the results. The algorithms used in machine learning can be trained using historical data on brucellosis outbreaks, allowing them to make predictions based on patterns and trends in the data. This information can be used to create early warning systems and intervention strategies to prevent the spread of brucellosis. The objectives of this study are to analyze data and identify the most important factors responsible for the outbreak of brucellosis [7,8].

Data collection

The data utilized in this research analysis was obtained from H. Bagheri [2]. The research focused on analyzing 109 datasets of Qazvin province (Iran), with data collected on a monthly basis. The study specifically examined time series data on brucellosis, using various covariates such as Livestock Ratio, Average Monthly Temperature, Non-Pasteurized Dairy Product Ratio, Minimum Temperature, Season, Average Wind Speed, Total Monthly Precipitation, Contact Ratio, Monthly Wind Speed, Rural ratio, cases, Average Age, Maximum Monthly Temperature, male ratio, Month, Year in Qazvin province. Add variable and targeted output. According to national rules, the clinical and epidemiological data of the patients were entered online into the Health Surveillance System [14,15].

Machine Learning Models

In this paper we have utilized total of four different ML's to explore, explain and model the behavior of brucellosis disease with respect to several input parameters. The ML's are selected bases on their suitability and acceptability are given below:

Radial basis function (RBF)

Radial basis function (RBF) is a powerful mathematical tool that has proven successful in various fields, including brucellosis. Its ability to accurately model complex data can predict an animal's likelihood of infection based on multiple features, such as age, weight, and vaccination status. RBF also identifies critical factors that contribute to the spread of brucellosis, informing prevention strategies. With high accuracy and reliability, RBF effectively diagnoses brucellosis in animals, detecting unique patterns in large datasets and even mild symptoms. RBF can be integrated with other diagnostic tools to provide a comprehensive evaluation. However, RBF can be computationally intensive and less effective for less commonly affected species. Biased data used for training may produce false-negative results, and diverse data is necessary for regional accuracy. RBF also requires regular updates and retraining to remain relevant. Despite these limitations, RBF is a valuable tool for understanding and mitigating brucellosis in animal populations [10].

Random Forest (RF)

Random Forest is a highly effective machine learning algorithm used in detecting brucellosis, a bacterial disease that affects both humans and animals. The algorithm is based on decision trees and combines the results of multiple trees to reduce over fitting. Random Forest considers various parameters like symptoms, medical history, and test results to predict the probability of the disease. Early detection of brucellosis is crucial, and the use of Random Forest has helped in the treatment and control of the disease. One of the significant advantages of Random Forest is its flexibility in analyzing complex datasets. It is highly accurate, reduces the potential for bias, and requires minimal tuning. However, the algorithm may not be suitable for small datasets or those with a limited number of variables. It requires significant processing power and time to run, and the results can be difficult to interpret. It is highly sensitive to outliers and impacted by missing data, which can reduce the accuracy of the findings. Additionally, it is not suitable for real-time analysis, making it challenging to use in clinical settings where rapid results are required [12].

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a popular machine learning algorithm used extensively in veterinary medicine to predict and diagnose various diseases in animals. One of the diseases that SVM has been used to detect is brucellosis, a bacterial disease commonly found in cattle. SVM creates a boundary between healthy and infected animals based on their clinical signs and symptoms, referred to as the support vector, which effectively differentiates between the two groups. SVM's use in diagnosing brucellosis in cattle has proven to be efficient and accurate, providing reliable results and reducing the risk of misdiagnosis. SVM has several advantages, including its ability to handle high dimensional and imbalanced datasets, good generalization performance, and clear boundary lines. However, SVM is computationally intensive, sensitive to outliers and noise in the data, requires careful parameter tuning, and may struggle with non-linear relationships and over fitting. Despite its limitations, SVM remains a versatile and powerful tool for brucellosis research and classification tasks [9].

J48 Trees

J48 Trees are an essential tool for predicting and classifying data in machine learning and data mining. In the case of brucellosis, J48 Trees can accurately predict the likelihood of an individual being infected based on factors such as age, occupation, and exposure to contaminated food. This information can prioritize individuals for testing and treatment, thus controlling the spread of the disease. The use of J48 Trees in brucellosis classification provides an effective and efficient method for predicting the presence of the disease in livestock populations. They are simple to understand and interpret, enabling farmers and veterinarians to use them with ease. Their adaptability means that they can be used in different applications, making them useful for tracking and managing outbreaks. However, despite their advantages, J48 Trees have limitations. They may not always accurately identify all cases of brucellosis, leading to false negatives. Technical expertise and specialized software are required for implementation, making them complex to use. The accuracy of J48 Trees depends on the quality of the data, leading to inaccurate predictions. They can be time-consuming and resource-intensive to train and validate, making them less accessible for some farmers and veterinarians. Finally, the accuracy of J48 Trees may vary depending on the population being analyzed, leading to potential biases in disease prediction [16].

Result & Discussion

The results from the four different ML's are tabulated and presented in Table 1

The Radial Basis Function, Random Forrest, Support Vector Machine, and J48 Tree machine learning methods' model evaluation criteria are contrasted in this table. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MARE), Kappa statistic, and Time (in seconds) are some of the evaluation criteria that are employed. Based on our experiment the results are summarized in the following table:

Table 1: Results from the four different ML's.

Model	**Evaluation Criteria**				
	RMSE	***MAE***	***MARE***	***KS***	**Processing *Time (s)***
Radial Basis Function	0.4178	0.3632	0.7301	0.5531	0.05
Random Forrest	0.4486	0.3944	0.7928	0.3634	0.03
Support Vector Machine	0.4907	0.2407	0.4840	0.5185	0.02
J 48 Tree	0.5512	0.3711	0.7460	0.3297	0.03
Radial Basis Function	0.4178	0.3632	0.7301	0.5531	0.05

The best performing model according to the evaluation criteria is the Support Vector Machine with the lowest RMSE value of 0.4907, the lowest MAE value of 0.2407, and the lowest MARE value of 48.4017%. The lowest RMSE and MAE values indicate that the model has the lowest deviation from the actual values. However, the Radial Basis Function model has a lower MARE value of 73.019% compared to the Support Vector Machine, which indicates that it has a lower error rate in terms of percentage. The Random Forrest model has the highest MARE value of 79.2856% among the four models, which means it has a higher error rate. The J48 Tree model has the highest RMSE value of 0.5512, which indicates it has the largest deviation from the actual values.

Overall, the Support Vector Machine model has the best performance according to the evaluation criteria, but the final choice of model should be based on the specific requirements and objectives of the project.

The reliability or inter-rater agreement between two raters is measured by the kappa statistic. The range of Kappa's value is -1 to 1, with -1 denoting no agreement and 1 denoting complete agreement. A value of 0 means that the agreement is equal to what would be expected by chance. A Kappa value of 0.5531 indicates a moderate agreement between the raters, and a value of 0.3634 indicates a fair agreement. A value of 0.5185 indicates moderate to substantial agreement, and a value of 0.3297 indicates fair agreement. The table compares four machine learning models (Radial Basis Function, Random Forest, Support Vector Machine, and J48 Tree) based on their processing time (in seconds) listed in the second column. The Support Vector Machine is the fastest, with a time of 0.02 seconds, followed by Random Forest (0.03 seconds), and J48 Tree and Radial Basis Function (both 0.03 and 0.05 seconds, respectively).

Results from H Bagheri 2020 by using different ML's are on the following table:

Table 2: Results from H Bagheri using different ML's.

Model	Evaluation Criteria		
	RMSE	*MAE*	*MARE*
Radial Basis Function	8.46	7.49	31%
Random Forrest	9.25	7.64	33%
Support Vector Machine	8.21	6.58	29%

From the above result, it is seen that our results are better in many fields than in H Bagheri (2020).

Apriori Algorithm

Apriori is a popular algorithm in data mining and machine learning that is used for frequent itemset mining and association rule learning. It can be applied to datasets related to Brucellosis to identify frequent patterns and associations between variables that may be useful for understanding the disease and designing prevention or control measures. The Apriori algorithm could be used to identify frequent combinations of risk factors, such as age, sex, occupation, and location, that are associated with a higher incidence of the disease. Additionally, the Apriori algorithm could be used in a research setting to identify potential risk factors or novel associations between variables and Brucellosis [13].

The following results are found after using Apriori algorithm:

Finding 5 selected best rules found as

1. Year ='(-inf-2012]' 36 ==> Non-Pasteurized Dairy Product Ratio='(-inf-7.125]' 36 <conf:(1)> lift:(1.24) lev:(0.06) [7] conv:(7)

2. Male ratio='(-inf-2.4]' Non-Pasteurized Dairy Product Ratio='(-inf-7.125]' 62 ==> Rural ratio='(-inf-4.203]' 59 <conf:(0.95)> lift:(1.08) lev:(0.04) [4] conv:(1.87)

3. Male ratio='(-inf-2.4]' Non-Pasteurized Dairy Product Ratio='(-inf-7.125]' Cases =Low 36 ==> Rural ratio='(-inf-4.203]' 34 <conf:(0.94)> lift:(1.07) lev:(0.02) [2] conv:(1.44)

4. Male ratio='(-inf-2.4]' Cases =High 35 ==> Rural ratio='(-inf-4.203]' 33 <conf:(0.94)> lift:(1.07) lev:(0.02) [2] conv:(1.4)

5. Male ratio='(-inf-2.4]' Cases =Low 43 ==> Rural ratio='(-inf-4.203]' 40 <conf:(0.93)> lift:(1.06) lev:(0.02) [2] conv:(1.29) high cases associated

Based on the Apriori rules, it can be seen that there are several strong associations between different variables and cases of brucellosis. The first rule states that there is a strong association between year of (-inf-2012) and non-pasteurized dairy product ratio of (-inf-7.125), with a confidence of 1. The lift value of 1.24 indicates that there is a strong association between the two variables, while the leverage value of 0.06 suggests that the relationship is weak. The conviction value of 7 suggests that the presence of a year of (-inf-2012) increases the likelihood of the non-pasteurized dairy product ratio being in the given range.

Based on the above rules findings, it can be concluded that there is a strong association between male ratio of (-inf-2.4), rural ratio of (-inf-4.203), year of (-inf-2012), and non-pasteurized dairy product ratio of (-inf-7.125) with cases of brucellosis. It is recommended to focus on these variables and populations for preventing and controlling the spread of brucellosis.

Ranked attributes

Based on the use of ML (Apriori) the attribute ranking was accomplished and the results are found as: Average Wind Speed > Non-Pasteurized Dairy Product Ratio > Rural ratio > year > Month > male ratio > Season.

Average Wind Speed: This attribute has been determined to be the most important or influential based on the ranking analysis. It likely holds a significant impact on the outcome or variable being studied.

Non-Pasteurized Dairy Product Ratio: This attribute is ranked second, indicating that it has a relatively high importance in relation to the studied variable. It suggests that the ratio of consumption of unpasteurized dairy products plays a notable role in the analysis.

Rural Ratio: This attribute is ranked third, suggesting that the ratio of rural areas or the presence of rural characteristics is considered significant in the analysis.

Year: This attribute holds a relatively lower ranking, indicating that it is of lesser importance compared to the preceding attributes. It suggests that the year in which the data was collected or the temporal aspect has some influence but is not as significant as the previous attributes.

Month: Similar to the year attribute, the month attribute holds a lower ranking, suggesting that the specific month during which the data was collected may have some impact but is not as influential as the previous attributes.

Male Ratio: This attribute is ranked sixth, indicating that the ratio of males in the population or specific context is considered to have a lesser impact on the studied variable compared to the attributes ranked higher.

Season: This attribute holds the lowest ranking among the listed attributes. It suggests that the season in which the data was collected or the seasonal factor is considered to have the least influence on the analyzed variable compared to the other attributes.

Derived tree from J48

In the (Figure 1) the determining factor for classifying cases is the average wind speed. If it's below 1.6, the average age is considered. A low classification occurs if the average age is under 40.41, otherwise, the year is responsible. When the average wind speed is above 1.6, the rural ratio determines the classification. If it's less than 1.2, the cases are low, otherwise, the month is responsible. If the year is before 2010, the cases are low. If the year is after 2010, the monthly wind speed determines the classification. A low classification occurs if the monthly wind speed is above 1, or if the male ratio (when the month is below 2) is above 2. When the average wind speed is below 2, the cases are low, otherwise, they are considered high.

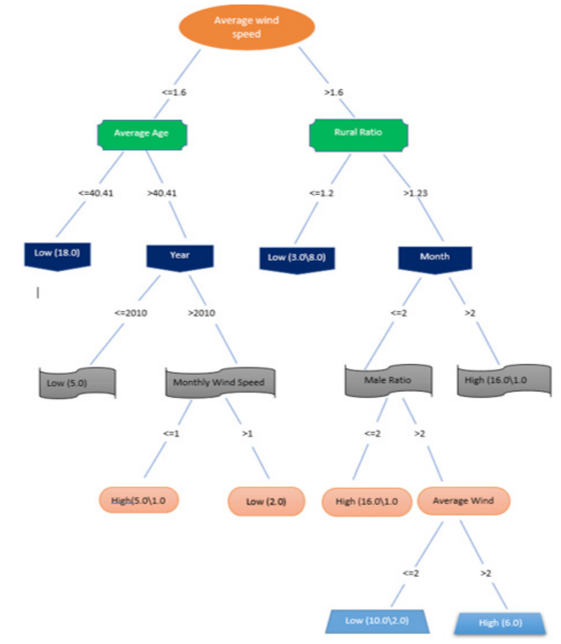

Figure 1: Tree from J48.

Conclusion

In conclusion, machine learning has enormous potential for the prediction and control of brucellosis disease. By using algorithms and statistical models to analyze large amounts of data, machine learning can help improve the accuracy and efficiency of predicting outbreaks of brucellosis. This information can be used to develop early warning systems, intervention strategies, and preventive measures to minimize the impact of brucellosis on public health.

The low incidence of disease cases has a robust correlation with three variables: the Male Ratio, the Non-Pasteurized Dairy Product Ratio, and the Rural Ratio, with a confidence of 94%.

A substantial relationship exists between a high incidence of disease and two factors, the Male Ratio and the Rural Ratio, boasting a confidence level of 94%.

Radial Basis Function had the highest Kappa Statistics, indicating that it has a high level of accuracy in its predictions.

The results of our model are better than those reported by H. Bagheri in 2020 as evidenced by a lower RMSE and MAE value. This suggests that our model produces more accurate predictions. For instance, our model's Support Vector Machine value of 0.4907 is significantly lower than the value of 8.21 reported by H. Bagheri, demonstrating the superiority of our model in terms of prediction accuracy.

In summary, the application of machine learning in the prediction of brucellosis disease has the potential to greatly enhance our ability to control and prevent the spread of this important public health hazard. With further research and development, we can work towards creating a world where brucellosis is no longer a threat to human and animal health.

References

1. Samaha H, Al-Rowaily M, Khoudair RM, Ashour HM. Multicenter Study of Brucellosis in Egypt," Emerg Infect Dis. 2008; 14: 1916–1918.

2. H. Bagheri et al., "Forecasting the monthly incidence rate of brucellosis in west of Iran using time series and data mining from 2010 to 2019," PLOS ONE. 2020; 15: e0232910.

3. Hassan MR, Mamun AA, Hossain MI, Arifuzzaman M. "Moisture Damage Modeling in Lime and Chemically Modified Asphalt at Nanolevel Using Ensemble Computational Intelligence," Computational Intelligence and Neuroscience, vol. 2018; p. e7525789.

4. Arifuzzaman M, Gazder U, Alam MS, Sirin O, Mamun AA. "Modelling of Asphalt's Adhesive Behaviour Using Classification and Regression Tree (CART) Analysis," Computational Intelligence and Neuroscience, vol. 2019; e3183050.

5. Arifuzzaman M. "Advanced ANN Prediction of Moisture Damage in CNT Modified Asphalt Binder," Journal of Soft Computing in Civil Engineering. 2017; 1: 1–11.

6. Rushd S, Hafsa N, Al-Faiad M, Arifuzzaman M. "Modeling the Settling Velocity of a Sphere in Newtonian and Non-Newtonian Fluids with Machine-Learning Algorithms," Symmetry. 2021; 13: 1.

7. Arifuzzaman M, Hassan MD R. "Moisture Damage Prediction of Polymer Modified Asphalt Binder Using Support Vector Regression," Journal of Computational and Theoretical Nanoscience. 2014; 11; 2221–2227.

8. Arifuzzaman M, Islam M, Hossain M, Tito MH, Anwar M, et al. "Application of AI on moisture damage of modified asphalt binders,". 2021; 307–311.

9. Wu C-H, Ho J-M, Lee DT. "Travel-time prediction with support vector regression," IEEE Transactions on Intelligent Transportation Systems. 2004; 5: 276–281.

10. Tapak L, Shirmohammadi-Khorram N, Hamidi O, Maryanaji J. "Predicting the Frequency of Human Brucellosis using Climatic Indices by Three Data Mining Techniques of Radial Basis Function, Multilayer Perceptron and Nearest Neighbor: A Comparative Study," Iranian Journal of Epidemiology. 2018; 14: 153–165.

11. Wu H, et al. "Time series analysis of weekly influenza-like illness rate using a one-year period of factors in random forest regression," BioScience Trends. 2017; 11: 292–296.

12. M. R. Segal, "Machine Learning Benchmarks and Random Forest Regression,". 2004.

13. T. Soman and P. O. Bobbie, "Classification of arrhythmia using machine learning techniques.," WSEAS Transactions on computers. 2005; 4: 548–552.

14. Eini P, Keramat F, Hasanzadehhoseinabadi M. "Epidemiologic, Clinical and Laboratory Findings of Patients with Brucellosis in Hamadan, West of Iran," Journal of Research in Health Sciences. 2012; 12: 2.

15. Zeinali M, Shirzadi M, Haj Rasooliha H. "National guideline for Brucellosis control," Iran Ministry of Health. 2011.

16. Kumar D, Sharma AK, Bajaj R, Pawar L. “Feature Optimized Machine Learning Framework for Unbalanced Bioassays,” in Cognitive Behavior and Human Computer Interaction Based on Machine Learning Algorithm, John Wiley & Sons, Ltd. 2021; 167–178.

Manuscript Information: Received: June 22, 2023; Accepted: August 09, 2023; Published: August 15, 2023

Authors Information: MH Tito[1]*; M Arifuzzaman[2]; MHE Jannat[1]; A Nasrin[3]; M Asaduzzaman[4]; MM Hossain[1]; SM Maruf[1]; Afzal Haq Asif[5]

[1]Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

[2]King Faisal University, Saudi Arabia.

[3]Combined Military Hospital, Bangladesh.

[4]National Heart Foundation Hospital & Research Institute, Bangladesh.

[5]Department of Pharmacy, College of Clinical Pharmacy, King Faisal University, Al-Ahsa 31982, Saudi Arabia.

Citation: Tito MH, Arifuzzaman M, Jannat MHE, Nasrin A, Asaduzzaman M, Hossain MM, Maruf SM, Asif AH. Application of specialized machine learning for the prediction of Brucellosis disease. Open J Clin Med Case Rep. 2023; 2091.

About the Journal: Open Journal of Clinical and Medical Case Reports is an international, open access, peer reviewed Journal focusing exclusively on case reports covering all areas of clinical & medical sciences.
Visit the journal website at www.jclinmedcasereports.com
For reprints and other information, contact info@jclinmedcasereports.com