Letters in Animal Biology

Journal homepage: www.liabjournal.com

A comparative study of ensemble machine learning algorithms for brucellosis disease prediction

Mokammel Hossain Tito,[1]* Md. Arifuzzaman,[2] Most Hoor E. Jannat,[1] Md. Siddiqur Rahman,[3] Sayra Tasnin Sharmy,[3] Alifa Nasrin,[4] Md. Asaduzzaman,[5] Md. Ashrafuzzaman,[6] Dipok Biswas Prince,[1] Afzal Haq Asif [2]

[1] *Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh*
[2] *King Faisal University. Saudi Arabia*
[3] *Bangladesh Agricultural University, Bangladesh*
[4] *Combined Military Hospital, Bangladesh*
[5] *National Heart Foundation Hospital & Research Institute, Bangladesh*
[6] *Islamic University Bangladesh*

Article info

Received: 02 September 2023
Received in revised form: 01 October 2023
Accepted: 02 October 2023
Published online: 03 October 2023

Keyword

Ensemble machine learning
AdaBoostM1
Vote
Bagging
LogitBoost
Brucellosis

* *Corresponding author:*
M.H. Tito
Email: mokammel.17asvm014@bsmrstu.edu.bd

Reviewed by:
Hasnain Idrees (*University of Agriculture, Fasilabad, Pakistan*)

Abstract

Brucellosis, caused by *Brucella spp*., is a global public health concern, particularly in underdeveloped regions. Cattle, predominantly infected with *B. abortus*, encounter reproductive challenges, reduced productivity, and fertility issues. Effective control measures, including serological tests like iELISA (indirect Enzyme-linked Immunosorbent Assay) are vital. This research harnesses machine learning techniques, encompassing AdaBoostM1, Vote, Bagging, and LogitBoost, to forecast Brucella infection in cattle, utilizing comprehensive data sourced from Qazvin, Iran. Detailed model descriptions are provided, highlighting AdaBoostM1 as the optimal choice, boasting a robust 75% correlation, low RMSE (Route Mean Square Error), MAE (Mean Absolute Error), and a commendable Kappa Statistics score of 0.4965. Ensemble machine learning demonstrates significant potential in Brucellosis prediction, adept at handling intricate datasets, and enhancing predictive accuracy. AdaBoostM1 stands out as the preferred model, offering valuable insights for Brucellosis prediction and contributing to the enhancement of disease control strategies.

1. Introduction

Brucellosis is a type of zoonotic illness caused by gram-negative bacteria known as Brucella spp. This bacterium can infect various animal species, including humans, and is highly contagious and widespread worldwide, especially in low to moderate-income underdeveloped countries. Although developed countries have successfully eradicated it, it remains a significant public health risk in underdeveloped countries(Arifuzzaman et al. 2021; Rahman et al. 2014). The primary cause of *Brucella* infection in cattle is *B. abortus*, and it can also be caused by *B. melitensis* and occasionally by *B. suis*. The infection typically affects the reproductive organs of cattle, resulting in placentitis followed by abortion, which can lead to a loss of productivity and reproduction, chronic metritis, and decreased fertility rates. Infected animals can also shed bacteria through their excretions, which is a significant source of transmission to other susceptible hosts (Atallah and Al-Mousa 2019; Rahman et al. 2014). To control the spread of brucellosis, it is crucial to perform regular serological tests on animals, such as the indirect enzyme-linked immunosorbent assay (iELISA). At present machine learning technique is a very popular technique to find out the risk factors (Tito et al. 2023) which are more responsible for spreading of disease. In clinical settings, machine-learning algorithms such as data mining techniques (Arifuzzaman et al. 2021) are being used as a valuable tool for risk assessment and clinical decision-making (Ballesteros-Ricaurte et al. 2022; Rushd et al. 2021). The objective of this paper is to find out the best technique for predicting brucella disease. In this paper we have used AdaBoostM1, Vote, Bagging, and LogitBoost to predict Brucella infection in cattle.

2. Methodology

The information used in this study's analysis came from an open source data (Bagheri et al. 2020). The original study concentrated on the analysis of 109 sets of data set from the Iranian region of Qazvin, where data were gathered on monthly basis. The research focused on time series data on brucellosis and used a variety of covariates, including rural ratio, non-pasteurized dairy products ratio, males ratio, average age, contact ratio, livestock ratio, climatic parameters including (average temperature, minimum and maximum monthly temperature, monthly precipitation, monthly wind speed, and average wind speed), month, season, and year in the province of Qazvin. Clinical and epidemiological information about the patients was entered into the Health Surveillance System online in accordance with government regulations (Budgaga et al. 2016; Rushd et al. 2021).

2.1 Description of machine learning models

2.1.1 AdaBoostM1

AdaBoostM1 is a Machine Learning) (ML) algorithm that combines weak classifiers to form a strong classifier for classification and prediction tasks. It iteratively trains classifiers, focusing on misclassified samples and assigns greater weights to difficult ones (Freund and Schapire 1997). It can predict the risk of Brucella infection in cattle based on various risk factors, handling high-dimensional and noisy data while identifying informative factors. However, drawbacks include requiring a large amount of training data for optimal performance, sensitivity to outliers or imbalanced data, and difficulty in interpreting underlying relationships (Darabi et al. 2019).

2.1.2 Vote

Vote" is an ensemble learning method in machine learning that combines multiple classifiers and aggregates their predictions to produce a final prediction. In the context of predicting Brucella infection in cattle, "Vote" can be used to combine the predictions of multiple classification models, such as decision trees, logistic regression, and support vector machines (Vapnik 2000). The main advantage of "Vote" is that it can improve the accuracy of the final prediction by combining the strengths of different classifiers. However, "Vote" can also suffer from the same limitations as other ensemble learning methods, such as overfitting to the training data, and it may not always produce the best possible results. Therefore, it is important to carefully evaluate the performance of "Vote" and other machine learning models in predicting Brucella infection in cattle, and to consider the strengths and limitations of each method when developing effective control programs (Hossain et al. 2021).

2.1.3 Bagging

Bagging, or Bootstrap Aggregating, is a machine learning technique that involves creating multiple subsets of a dataset and training a base learning algorithm on each subset. The predictions made by the base learning algorithms are then combined to make a final prediction, which is typically more accurate and less prone to overfitting than a prediction made by a single algorithm (Breiman 1996). Bagging can be useful for predicting Brucella infection in cattle, as it allows for the identification of the most important risk factors associated with the disease. Some potential advantages of bagging include increased accuracy, reduced overfitting, and the ability to handle large datasets. However, some potential disadvantages of bagging include increased computational requirements and the possibility of reduced interpretability due to the use of multiple algorithms. Overall, bagging can be a useful machine learning technique for predicting Brucella infection in cattle, but it is important to carefully consider its pros and cons before using it in practice (Budgaga et al. 2016; Hossain et al. 2021).

2.1.4 LogitBoost

LogitBoost is a machine learning algorithm that combines boosting and logistic regression techniques to build a predictive model. Boosting is a method that sequentially trains weak classifiers and combines their results to form a stronger classifier. Logistic regression, on the other hand, is a statistical model that estimates the probability of a binary outcome. The combination of these techniques in LogitBoost results in an algorithm that is well-suited for binary classification tasks, such as predicting the presence of Brucella infection in cattle (Friedman 2001). The main advantage of LogitBoost is its ability to handle high-dimensional data with complex interactions between features. It can also handle missing data and outliers. However, a potential disadvantage of LogitBoost is its tendency to overfit the training data, leading to poor generalization to new data. Additionally, the interpretation of the model can be challenging, as it is based on a combination of many weak classifiers. Despite these challenges, LogitBoost can be a useful tool for predicting Brucella infection in cattle and identifying the most important risk factors associated with the disease (Mahdavi and Aziz 2020).

2.2 Mechanism of Ensemble Learning

In Fig. 1 a predictive ensemble model was developed using four different techniques, namely AdaBoostM1, Vote, Bagging, and LogitBoost. By combining the predictions of these models, it is possible to accurately predict the number of cases (Tapak et al. 2018). This approach enables a more robust prediction by leveraging the strengths of multiple models, which can help to overcome the limitations of individual models. With this ensemble model, it is expected that the accuracy of predictions will be higher than using any one of the individual models alone (Hossain et al. 2021).

3. Result and Discussion

Table 1 presents the evaluation results of four machine learning models used for predicting brucellosis disease in cattle. The models are assessed based on various evaluation criteria, including Correlation Coefficient (CC), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE), Kappa Statistics, and the computational time required for training (in seconds).

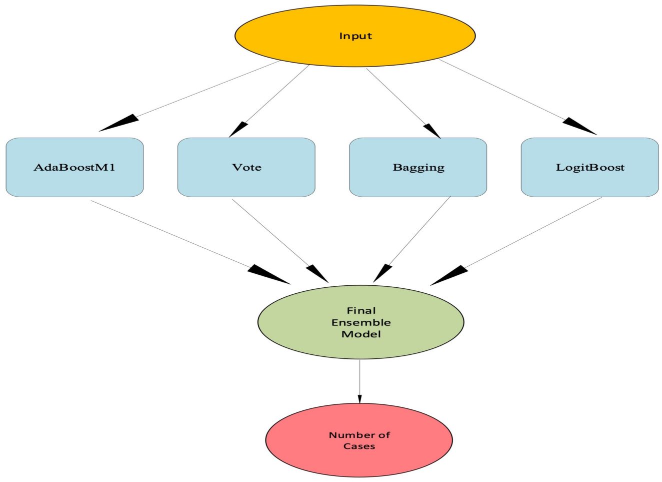

Fig. 1 Mechanism of Ensemble Learning

The AdaBoostM1 model achieved a correlation coefficient (CC) of 75%, indicating a strong correlation between predicted and actual values. The RMSE is 0.4376, signifying relatively small prediction errors. The MAE is 0.3185, representing the average absolute error. The RAE is 64.0429%, indicating that, on average, predictions are within this percentage of the actual values. The Kappa Statistics score is 0.4965, indicating good agreement with actual outcomes. The model required 0.08 seconds for training.

The Vote model achieved a CC of 68.5185%, showing a moderate correlation between predicted and actual values. The RMSE is 0.4466, indicating slightly larger prediction errors compared to AdaBoostM1. The MAE is 0.3926, representing the average absolute error. The RAE is 78.9408%, suggesting somewhat higher errors on average. The Kappa Statistics score is 0.3616, indicating moderate agreement with actual outcomes. Notably, this model required no additional training time beyond model creation (0 seconds).

Bagging achieved a CC of 70.3704%, indicating a moderately strong correlation. The RMSE is 0.4462, similar to Vote. The MAE is 0.375, signifying the average absolute error. The RAE is 75.3945%, indicating average errors within this percentage. The Kappa Statistics score is 0.4041, indicating reasonable agreement with actual outcomes. The model required 0.06 seconds for training.

LogitBoost achieved a CC of 69.4444%, showing a moderate correlation. The RMSE is 0.4871, indicating relatively larger prediction errors compared to AdaBoostM1 and Bagging. The MAE is 0.3191, representing the average absolute error. The RAE is 64.1537%, indicating that predictions are within this percentage of the actual values on average. The Kappa Statistics score is 0.383, indicating moderate agreement with actual outcomes. However, this model required the longest training time at 0.37 seconds.

Based on the evaluation criteria and considering the overall performance, AdaBoostM1 appears to be the most suitable model for predicting brucellosis disease in cattle. It exhibits the highest correlation, the lowest RMSE and MAE, and a good Kappa Statistics score. Although it has a slightly longer training time compared to some models, its accuracy and agreement with actual outcomes justify its selection for predicting Brucellosis.

The performance of four models – AdaBoostM1, Vote, Bagging, and Logitboost is illustrated in Fig. 2. In the figure, the orange bar represents correctly classified instances, the yellow bar represents incorrectly classified instances, and the green bar represents the length of successful completion of the task, which is considered in second place. Among all the correctly classified instances, AdaBoostM1 achieves the highest accuracy at 75%. On the other hand, Vote has the lowest accuracy of 68.52%. Despite its low accuracy, Vote requires the least amount of time. Conversely, Logitboost, which takes the longest time, does not yield satisfactory results. Considering the entire model, AdaBoostM1 emerges as the best option. It achieves higher accuracy with minimal errors and provides results within a short timeframe.

Ensemble machine learning represents a ground-breaking approach in the field of veterinary medicine (Alqaissi et al. 2022). To the best of our knowledge, this marks the inaugural utilization of ensemble learning within the veterinary field. The inspiration for this concept stemmed from our observations in the identification of human infectious diseases (Alqaissi et al.

Table 1 Summary the performance of four Ensemble machine learning models

Model	Evaluation criteria					
	CC	RMSE	MAE	RAE	Kappa Statistics	Time (s)
AdaBoostM1	75%	0.4376	0.3185	64.04%	0.4965	0.08
Vote	68.52%	0.4466	0.3926	78.94%	0.3616	0
Bagging	70.37%	0.4462	0.375	75.39%	0.4041	0.06
LogitBoost	69.44%	0.4871	0.3191	64.15%	0.383	0.37

CC = Correlation Coefficient; RMSE = Root Mean Square Error; MAE = Mean Absolute Error; RAE = Relative Absolute Error

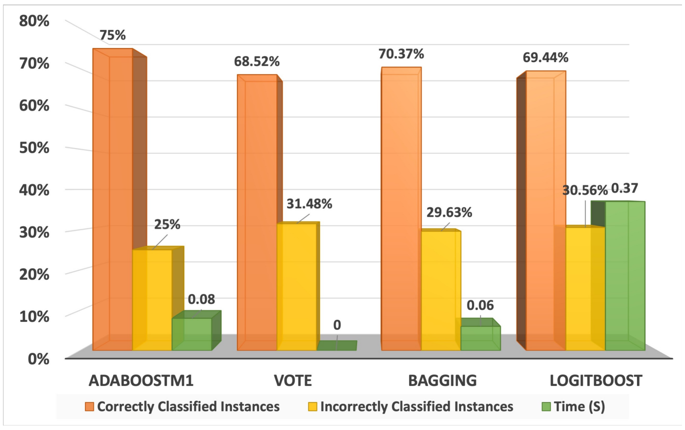

***Fig. 2** Graphical illustration depicting the effectiveness of various methods*

2022; Santangelo et al. 2023), kidney disease (Nikhila 2021), heart disease (Alqahtani et al. 2022) , diabetes (Abnoosian et al. 2023), skin disease (Verma et al. 2020), and pneumonia (Kundu et al. 2021).

4. Conclusions

Ensemble machine learning techniques can be useful tools for predicting the risk of brucellosis infection in cattle and identifying the most important risk factors associated with the disease. These techniques have the potential to handle high-dimensional and noisy data, produce more accurate predictions, and reduce overfitting compared to single models. In conclusion, from the models which we have used AdaBoostM1 is the recommended model for predicting Brucellosis disease.

5. Recommendation and future prospects

Although this dataset has potential for further improvement, its use is currently limited due to certain constraints. Machine learning can be applied for predicting different kinds of animal disease such as Lumpy Skin Disease, Rabies, Anthrax, Foot and Mouth Disease, etc.

Declarations

Funding: None

Conflict of interest: None

Ethical approval: Not applicable

Acknowledgements: None

References

Abnoosian K, Farnoosh R, Behzadi MH. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics 24(1): 337. https://doi.org/10.1186/s12859-023-05465-z

Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. (2022). Cardiovascular disease detection using ensemble learning. Computational Intelligence and Neuroscience e5267498. https://doi.org/10.1155/2022/5267498

Alqaissi EY, Alotaibi FS, Ramzan MS. (2022). Modern machine-learning predictive models for diagnosing infectious diseases. Computational and Mathematical Methods in Medicine e6902321. https://doi.org/10.1155/2022/6902321

Arifuzzaman M, Islam M, Hossain M, Tito MH, Anwar M, Al Fuhaid A. (2021). Application of AI on moisture damage of modified asphalt binders. 4th Smart Cities Symposium (SCS 2021) 307-311. https://doi.org/10.1049/icp.2022.0361

Atallah R, Al-Mousa A. (2019). Heart disease detection using machine learning majority voting ensemble method. 2nd International Conference on New Trends in Computing Sciences (ICTCS), Amman, Jordan. Pp. 1-6. https://doi.org/10.1109/ICTCS.2019.8923053

Bagheri H, Tapak L, Karami M, Hosseinkhani Z, Najari H, Karimi S, Cheraghi Z. (2020a). Forecasting the monthly incidence rate of brucellosis in west of Iran using time series and data mining from 2010 to 2019. PLOS ONE 15(5): e0232910. https://doi.org/10.1371/journal.pone.0232910

Ballesteros-Ricaurte JA, Fabregat R, Carrillo-Ramos A, Parra C, Pulido-Medellín MO. (2022). Systematic literature review of models used in the epidemiological analysis of bovine infectious diseases. Electronics 11(15): 2463. https://doi.org/10.3390/electronics11152463

Breiman L. (1996). Bagging predictors. Machine Learning 24(2): 123-140. https://doi.org/10.1007/BF00058655

Budgaga W, Malensek M, Pallickara S, Harvey N, Breidt FJ, Pallickara S. (2016). Predictive analytics using statistical, learning, and ensemble methods to support real-time exploration of discrete event simulations. Future Generation Computer Systems 56: 360-374. https://doi.org/10.1016/j.future.2015.06.013

Darabi H, Choubin B, Rahmati O, Torabi Haghighi A, Pradhan B, Klove B. (2019). Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. Journal of Hydrology 569: 142-154. https://doi.org/10.1016/j.jhydrol.2018.12.002

Freund Y, Schapire RE. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1): 119-139. https://doi.org/10.1006/jcss.1997.1504

Friedman JH. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5): 1189-1232.

Hossain M, Haque MS, Arifuzzaman M, Hossain SMZ. (2021a). Artificial neural network based system for distorted image recognition. 3rd Smart Cities Symposium (SCS 2020) 503-508. https://doi.org/10.1049/icp.2021.0852

Kundu R, Das R, Geem ZW, Han GT, Sarkar R. (2021). Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PLOS ONE 16(9): e0256630. https://doi.org/10.1371/journal.pone.0256630

Mahdavi A, Aziz J. (2020). Estimation of semiarid forest canopy cover using optimal field sampling and satellite data with machine learning algorithms. Journal of the Indian Society of Remote Sensing 48(4): 575-583. https://doi.org/10.1007/s12524-020-01102-x

Nikhila. (2021). Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm. 2021 International Conference on

Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India. Pp. 476–480. https://doi.org/10.1109/ICCCIS51004.2021.9397144

Rahman S, Sarker AS, Melzer F, Sprague LD, Neubauer H. (2014). The prevalence of Brucella abortus DNA in seropositive bovine sera in Bangladesh. African Journal of Microbiology 8(48): 3856-3860. https://doi.org/10.5897/AJMR2014.6031

Rushd S, Hafsa N, Al-Faiad M, Arifuzzaman M. (2021). Modeling the settling velocity of a sphere in newtonian and non-newtonian fluids with machine-learning algorithms. Symmetry 13(1): 71. https://doi.org/10.3390/sym13010071

Santangelo OE, Gentile V, Pizzo S, Giordano D, Cedrone F. (2023). Machine learning and prediction of infectious diseases: A systematic review. Machine Learning and Knowledge Extraction 5(1): 175-198. https://doi.org/10.3390/make5010013

Tapak L, Shirmohammadi-Khorram N, Hamidi O, Maryanaji Z. (2018). Predicting the frequency of human brucellosis using climatic indices by three data mining techniques of radial basis function, multilayer perceptron and nearest neighbour. A comparative study. Iranian Journal of Epidemiology 14: 153-165.

Tito M, Arifuzzaman M, Jannat M, Nasrin A, Asaduzzaman M, Hossain M, Maruf S, Asif AH. (2023). Application of specialized machine learning for the prediction of Brucellosis disease. Open Journal of Clinical and Medical Reports 9(27): 2091.

Vapnik VN. (2000). The nature of statistical learning theory. Springer, New York. https://doi.org/10.1007/978-1-4757-3264-1

Verma AK, Pal S, Kumar S. (2020). Prediction of skin disease using ensemble data mining techniques and feature selection method - A comparative study. Applied Biochemistry and Biotechnology 190(2): 341-359. https://doi.org/10.1007/s12010-019-03093-z

Citation

Tito MH, Arifuzzaman M, Jannat MHE, Rahman MS, Sharmy ST, Nasrin A, Asaduzzaman M, Ashrafuzzaman M, Prince DB, Asif AH. (2023). A comparative study of ensemble machine learning algorithms for brucellosis disease prediction. Letters in Animal Biology 03(2): 23 – 27.