Asian Journal of Agriculture and Biology
https://doi.org/10.35495/ajab.2025.017

Original Article

Brucellosis risk factors in dairy cattle: A machine learning approach to safeguarding human health

SM Azizul Karim Hussaini[1], Farhan Ibne Siddique[2], Mokammel Hossain Tito[8], Abdullah Al Mamun[1], Md Arifuzzaman[3], Afzal Haq Asif[4], Shahzad Khan[5], Muhammad Shahzad Chohan[5], Samar Sindi[6], Talha Yusuf [7], Md. Siddiqur Rahman[1]*

[1]Department of Medicine, Faculty of Veterinary Science, Bangladesh Agricultural University, Bangladesh
[2]Armed Forces Medical College, Dhaka Cantonment, Dhaka, Bangladesh
[3]Department of Civil & Environmental Engineering, King Faisal University, Saudi Arabia
[4]Department of Pharmacy Practice, College of Clinical Pharmacy, King Faisal University, Saudi Arabia
[5]Department of Biomedical Sciences, College of Clinical Pharmacy, King Faisal University, Saudi Arabia
[6]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Saudi Arabia
[7]Department of Computer Sciences, FCIT, King Abdulaziz University, Saudi Arabia
[8]Department of Medicine, Gazipur Agricultural University, Bangladesh

**Corresponding author's email:* siddique.medicine@bau.edu.bd

Received: 23 January 2025 / Accepted: 15 April 2025 / Published Online: 04 May 2025

Abstract

Brucellosis is a highly contagious zoonotic disease caused by the bacterium Brucella spp. that distresses mutually animals and humans, especially in underdeveloped countries with poor control programs. In adult cattle, the disease affects mainly the reproductive organs, thus causing major losses in production and reproduction, such as abortion and reduced fertility. This study involves the collection of 460 blood samples from dairy farms, which were analysed for brucellosis infection using the Rose Bengal Test (RBT). Additionally, data on the animals' history, including placenta (retained), repeat breeding, their age, abortion, and lastly calving, were also recorded. To address the problem of class imbalance between the positive and negative classes, a technique, known as Synthetic Minority Over-sampling Technique (SMOTE) was applied in the research work. A total of five algorithms were used in this paper among them multilayer perceptron (MLP) and weekadeeplearning4j showed the best results for the prediction of brucellosis having 93.59% and 93.94% accuracy, respectively. Besides, risk factors are ranked based on their importance as ordered as retained placenta > repeat breeding > calving > abortion > age, and three association rules are made to understand the correlation of the factors for occurring the disease. By applying this study, early diagnosis of the disease could be possible to mitigate the economic losses.

Keywords: Brucellosis, Zoonotic disease, Dairy cattle, Machine learning, Risk factors, Public health

How to cite this article:

Hussaini SMAK, Siddique FI, Tito MH, Al Mamun A, Arifuzzaman M, Asif AH, Khan S, Chohan MS, Sindi S, Yusuf T and Rahman MS. Brucellosis risk factors in dairy cattle: A machine learning approach to safeguarding human health. Asian J. Agric. Biol. 2025: 2025017. DOI: https://doi.org/10.35495/ajab.2025.017

Introduction

Brucellosis, with approximately 500,000 new cases recorded yearly for human disease, ranks among the most prevalent zoonoses globally. Despite this, it is still mostly ignored outside of a few industrialized countries where it has been completely eliminated (O'Callaghan, 2020). The Mediterranean, the Middle East, parts of Africa, Asia, and Latin America are among the regions where it is common and endemic (Godfroid et al., 2010). Abortions, infertility, and decreased milk production are among the reproductive issues caused by brucellosis in cattle, which significantly harms farmers and the agricultural sector financially (Franc et al., 2018). Humans can contract it by eating infected animal products or coming into close contact with ill animals. It mostly affects livestock, including as cattle, goats, sheep, and pigs (WOAH, 2018). Gram-negative intracellular bacteria called Brucella species are the cause of it (Akhvlediani et al., 2017; Ducrotoy et al., 2018). Most of the twelve species of Brucella that are now known to exist have the ability to infect both humans and other animals (Asmare et al., 2013). Brucella suis is less prevalent than B. melitensis, however Brucella abortus is the most common cause of Brucella infection in cattle (Mick et al., 2014). Serious health, financial, and livelihood effects result from the disease's frequent disregard in low- and moderate-income nations with poor control methods (McDermott et al., 2013). This could be the outcome of the disease remaining mislabeled as an unusual reproductive condition (Alamian and Dadar, 2020; Ducrotoy et al., 2017). Numerous countries are obstructed by bovine brucellosis, which presents a significant threat to public health and affects cattle (Yu and Nielsen, 2010; Khurana et al., 2021).

The infection causes placentitis and abortion when it pinpoints in the adult cattle's reproductive organs. This vanguards to losses in reproduction and production, including diminished fertility, chronic metritis, and decreased milk production (Franc et al., 2018; Kothalawala et al., 2017). Nevertheless, most infected animals undergo only one abortion throughout their lifetime and tolerate the infection for the duration of their lives (Godfroid et al., 2010). Cows that are not pregnant or who have had an initial abortion do not exhibit any indications of the illness. Bacteria may be defecated by animals in their waste, which is considered a crucial mechanism for the transmission of infection among susceptible hosts (Jamil et al., 2020; Hosein et al., 2018). Therefore, using serological tests to frequently check animals for brucellosis would support identify diseased animals, provide efficient control measures, and reduce the disease's spread (Gwida et al., 2016). In a developing country like Bangladesh, where research facilities are insufficient, serological testing will be extremely difficult and expensive besides in the Mymensingh district, the infectious disease was responsible for 48,436,400 taka (605455 US dollars) in total losses every year (Ahmed et al., 2018).

Notwithstanding the existence of a regulator program and targeted vaccines for wildlife, seroprevalence studies show that brucellosis is endemic in both humans and animals (Hosein et al., 2018; Abdelbaset et al., 2018). A brief study indicated that the seroprevalence of brucellosis in sheep and cattle was 16.3% and 16.7%, respectively (Selim et al., 2019). The motives for the persistence of brucellosis remain ambiguous despite efforts to control it. The implementation of effective strategic control activities may be stalled by insufficient epidemiological data about the seroprevalence of Brucella disease and its linked risk features (Eltholth et al., 2017; Gwida et al., 2016).

Machine learning algorithms are now being used in healthcare settings for the purpose of hazard factor evaluation and clinical decision-making. (Tito et al., 2023; Mburu et al., 2018; Selim et al., 2021). In humans, cardiovascular disease, kidney disease, diabetes, respiratory diseases are now diagnosed with the help of machine learning (Mathur et al., 2020; Ahsan et al., 2022). In the past ten years, data mining methodologies have been employed in veterinary epidemiological studies. Recent predictions of Brucella infection in cattle have utilised data mining techniques, including decision trees, random forests, support vector machines, and multivariate adaptive regression (Sharmy et al., 2024; Tito et al., 2024a; Rahman et al., 2024). To the best of our knowledge, no previous research has explored the potential of deep learning, ensemble learning, and classifier models (functional, lazy, tree) in predicting Brucella infection in cattle. Therefore, we set out to use machine learning to predict brucellosis in Bangladeshi dairy cattle by identifying the key risk factors and their correlation.

https://doi.org/10.35495/ajab.2025.017

Material and Methods

Data collection

The dataset is obtained from the Military Farm (MF) and the Central Cattle Breeding Dairy Farm (CCBDF), both located in Savar, as seen in Figure (1a). A total of 460 blood samples were collected Figure (1b) from dairy cows and sera were separated and stored at 4° C. Minimum sample size was 210 according to sample size calculation with 16.3% seroprevalence in Bangladesh (Selim et al., 2019). Furthermore, a questionnaire was created to gather animal data in accordance with the findings of the literature survey including repeat breeding, retained placenta, abortion, age, and calving identified as important features. Then the blood sample was tested through Rose Bengal Test (RBT) and confirmed the diagnosis Figure (1c) in this study.

Statistical analysis

Figure 2 shows the distributions of categorical variables such as age, calves, abortion, repeated breeding, retained placenta, and RBT that are connected to livestock or veterinary data. The distributions reveal imbalances across categories. These trends suggest that certain conditions or demographics are either rare or skewed in the dataset, which could impact subsequent analyses. Machine learning is able to handle such kind of data.

Pre-processing

SMOTE (Synthetic Minority Over-sampling Technique), introduced by (Chawla et al., 2002), is one of the most popular oversampling techniques for handling class imbalance problems in datasets. Instead of duplicating existing minority class instances, SMOTE synthesizes new ones by interpolating between existing minority instances and their nearest neighbors. It enriches the feature space with meaningful synthetic examples, which, in effect, prevents overfitting a common problem of simple oversampling techniques. SMOTE, by systematically generating balanced training data, enhances the classifier's ability to learn patterns from both the majority and minority classes Figure 3 and Figure 4.

The SMOTE algorithm identifies a user-specified number of k nearest neighbors for each minority class instance based on a distance measure, for example, Euclidean distance. Synthetic samples are then generated by interpolating between the feature vectors of a minority instance and its selected neighbors, based on the formula:

$$x_{new} = x_i + \lambda \cdot (x_j - x_i) \quad \text{.....(i)}$$

where x_i is the original instance, x_j is a neighbor, and λ a random number between 0 and 1. This is repeated until the target class balance is reached. Only then will SMOTE reduce the class imbalance ratio, which would otherwise bias the classifier toward the majority class, while preserving the feature space distribution. It will ensure better sensitivity, more excellent generalization, and more balanced decision boundaries for the classifier to be applied in several domains, such as medical diagnosis.

Model selection and validation

All the models are selected based on their previous performance in the field of disease prediction. A 10-fold cross-validation was conducted for each model, utilizing 66% of the data for training and 44% for validation. The comprehensive framework of our study is depicted in Figure 5.

https://doi.org/10.35495/ajab.2025.017

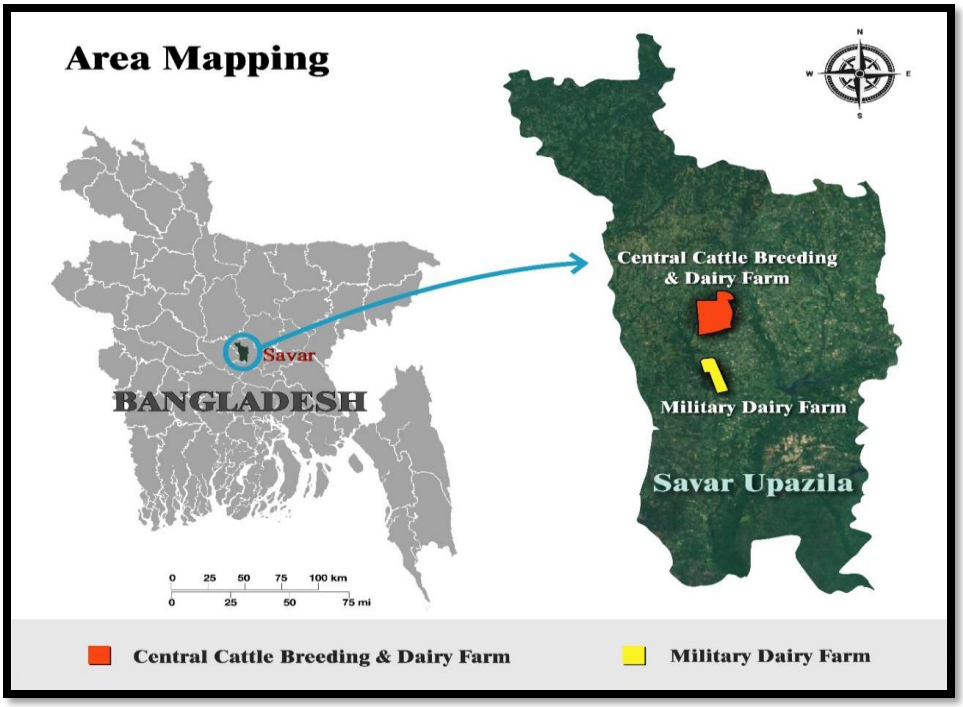

Figure-(1a). Study area is Savar (Military Farm and Central Cattle Breeding and Dairy Farm) for collecting data

Figure-(1b). Collection of blood for laboratory test.

https://doi.org/10.35495/ajab.2025.017

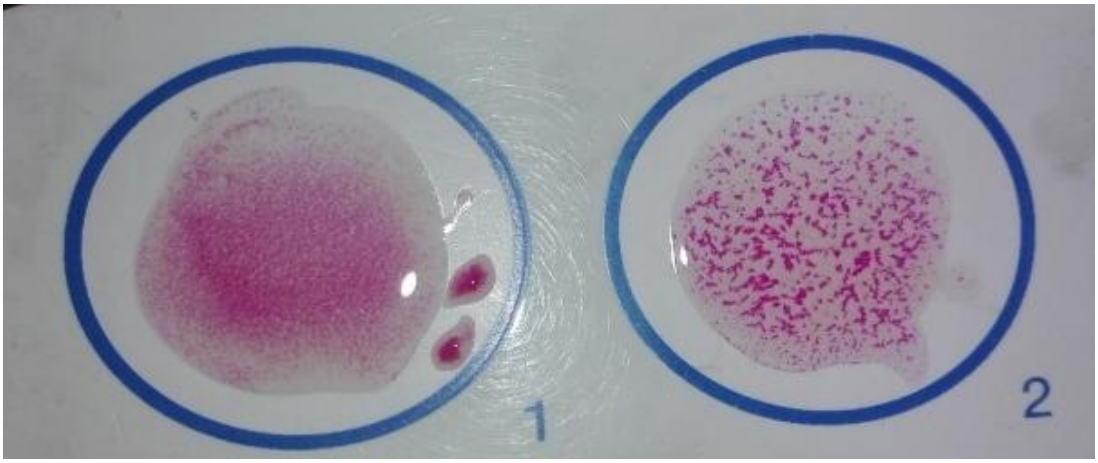

Figure-(1c). Rose Bengal Test (RBT) agglutinated sera indicating Brucella positive (2).

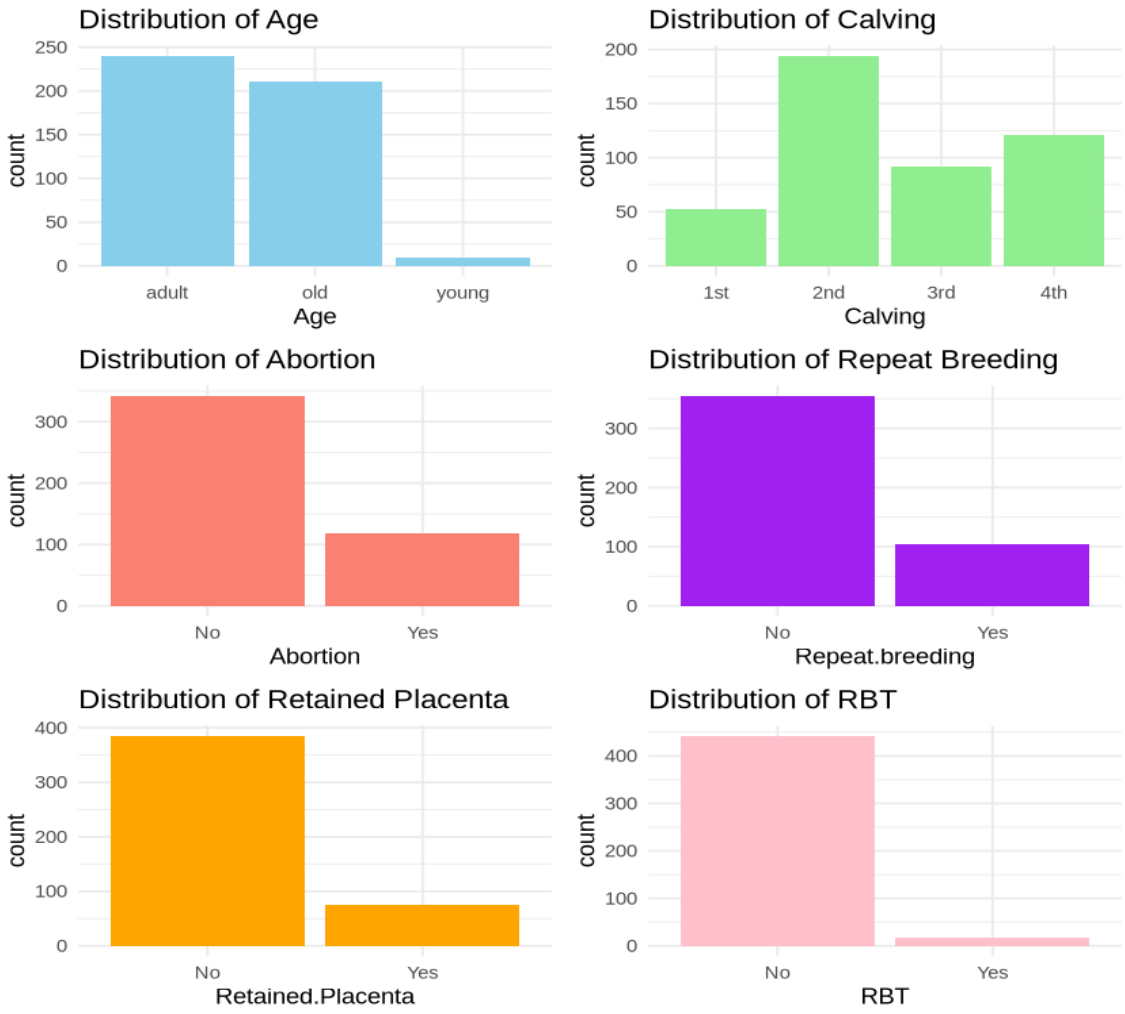

Figure-2. Histograms of the considered attributes.

https://doi.org/10.35495/ajab.2025.017

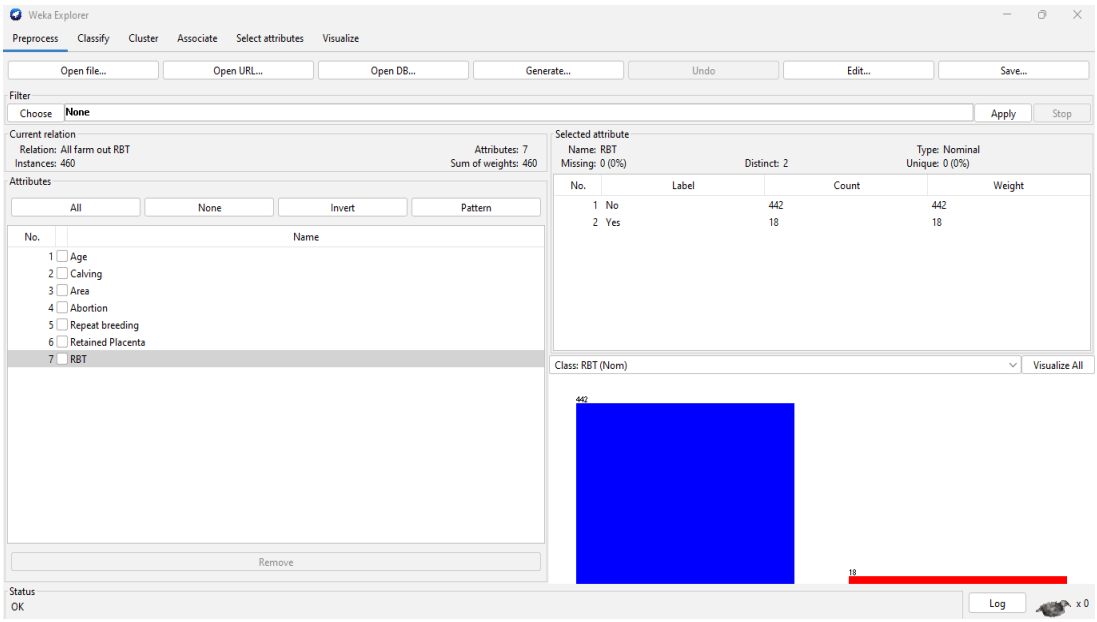

Figure-3. Imbalanced dataset before using SMOTE.

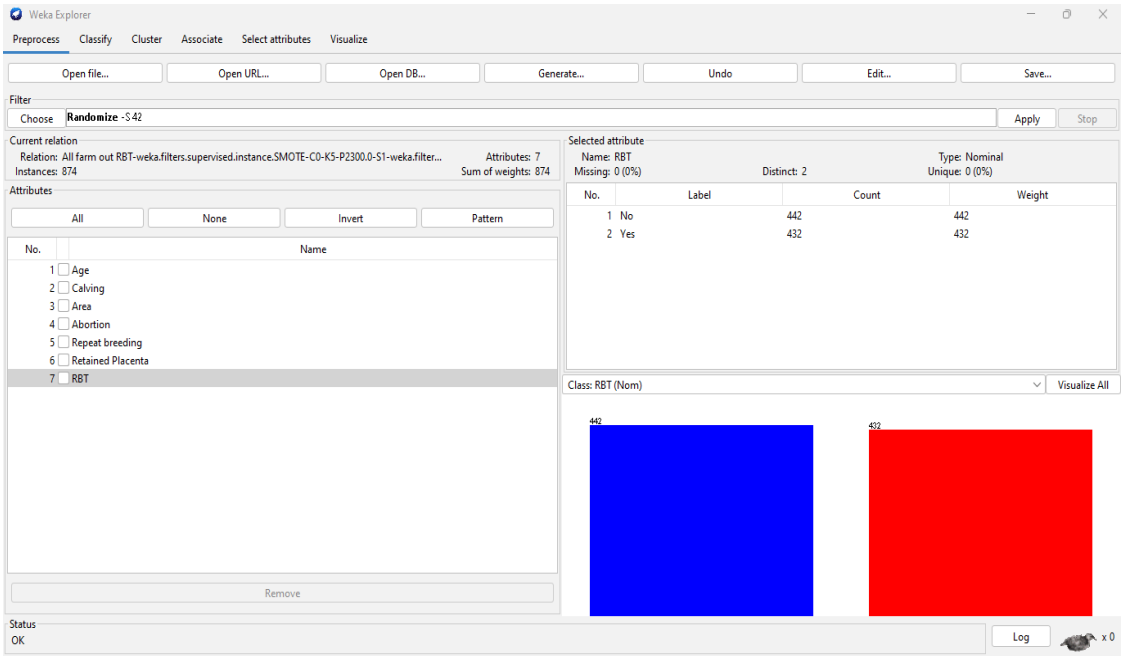

Figure-4. Balanced dataset after using SMOTE.

https://doi.org/10.35495/ajab.2025.017

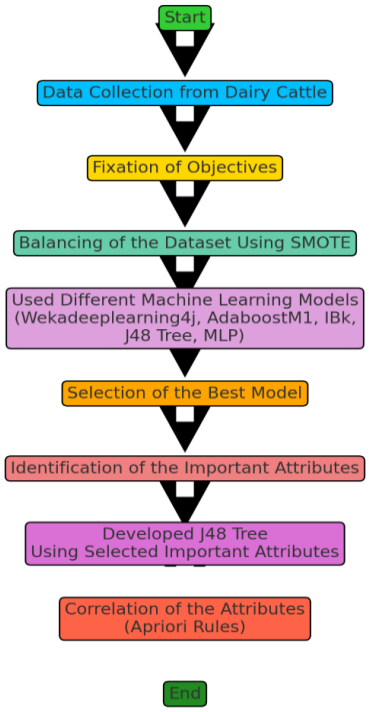

Figure-5. Overall workflow of the research

Description for different machine learning models

This paper employs machine learning algorithms, to be exact five, with the corresponding discussion provided below:

Multilayer Perceptron (MLP)

The feed forward artificial neural network model known as a multilayer perceptron (MLP) converts sets of input data into a collection of suitable outputs. It is among the neural network architectures that are most commonly used to diagnose diseases (Bishop, 1992; Tito et al., 2024b; Ripley, 1994). It is made up of a network of nodes that are layered. A typical MLP network consists of an input layer for external inputs, one or more hidden layers, and an output layer that provides classification results, with processing nodes usually numbering at least three, and sometimes more. (Figure 6).

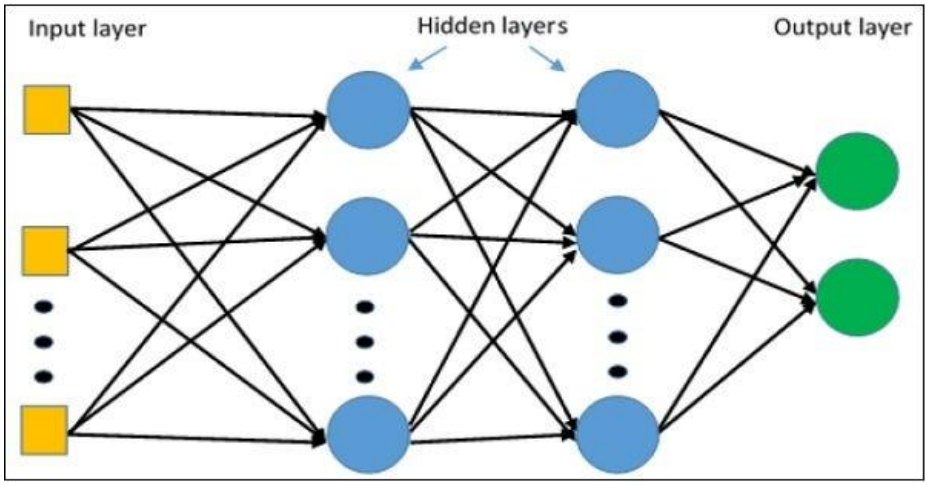

Figure-6. Workflow of Multilayer perceptron (MLP)

https://doi.org/10.35495/ajab.2025.017

WekaDeeplearning4j

WekaDeeplearning4j is an advanced integration of deep learning capabilities into the Weka machine-learning environment using the Deeplearning4j library. That is, the main purpose of this integration will let practitioners and researchers perform complicated predictive modeling, mainly disease prediction and analysis, by availing the deep neural network capabilities (Lang et al., 2019). It can be tightly integrated with Weka's feature-selection and data-preprocessing suite to ensure the best possible representation of inputs for neural networks. Moreover, its ability to deal with imbalanced datasets, common in disease-related data, increases its usefulness in identifying minority class instances, such as those representing rare disease occurrences. WekaDeeplearning4j has been applied in many research areas starting from epidemiology to genomics and clinical diagnostics, exhibiting capability in improving predictive accuracy for supporting data-driven decision-making within the healthcare systems (Lang et al., 2019; Mohd Radzi et al., 2022).

Here are some key concepts and governing equations we presented that assist the current WekaDeeplearning4j model:

Feedforward Neural Networks: The output of a neuron in a fully attached layer is processed as:

$$y = f\left(\sum_{f=1}^{n} w_i x_i + b\right) \ldots\ldots\ldots\ldots\ldots \textbf{(ii)}$$

where (wi) are the weights, (xi) represent the inputs, (b) act as the bias, and (f) is performing as the activation function (e.g., ReLU, sigmoid).

Convolutional Neural Networks (CNNs): The final output from a convolutional layer is shown by:

$$y_{i,j,k} = f\left(\sum_{m=1}^{M}\sum_{n=1}^{N}[\![x_{i+m=1,j+n=1} \cdot w_{m,n,k} + b_k)]\!]\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textbf{(iii)}$$

where (x) is the input, (w) are representing as the convolutional filters, (b) is similar to the bias, and (f) is behaving as the activation function. The indices (i,j,k) talk about the depth and spatial dimensions of the respective output feature map.

Recurrent Neural Networks (RNNs): The concealed state (h_t) in an RNN is reorganized as:

$$h_t = f(W_h h_{t=1} + W_x x_t + b) \ldots\ldots\ldots\ldots\ldots\ldots \textbf{(iv)}$$

where (W_x) and (W_h) are the weight matrices, (x_t) represents the input at a time step (t), and (f) is performing as an activation function.

Long Short-Term Memory (LSTM): LSTM units have more complex equations to handle long-term dependencies:

$$i_t = \sigma(W_i x_t + U_i h_{t=1} + b_i)$$
$$f_t = \sigma(W_f x_t + U_f h_{t=1} + b_f)$$
$$o_t = \sigma(W_o x_t + U_o h_{t=1} + b_o)$$
$$\tilde{c} = tanh(W_c x_t + U_c h_{t=1} + b_c)$$
$$c_t = f_t \odot c_{t=1} + i_t \odot \tilde{c}_t$$
$$h_t = o_t \odot \tanh(c_t)$$

J48 Tree

According to Pham in 2017, the most potent machine learning techniques for disease diagnosis and prediction is C4.5, is instigated by the J48 decision tree algorithm (Pham et al., 2017). This algorithm produces a decision tree by reiteratively sectionalization datasets based on the attributes values, and it also optimizes the assortment criteria in order to augment information gain. This has built it very appropriate for investigation in multifaceted medical datasets (Quinlan and Rivest, 1989). The author stated that J48 is decent at defining the risk factors and patterns connected with diseases by mapping attribute-based decision rules into a tree types of structure (as shown in Figure 7) to allow intuitive visualization of the decision process. Straightforwardness at the level of model output in medicinal research is of extreme consequence, as results simple enough for clinical decision-making will be provided (Quinlan and Rivest, 1989; Singh et al., 2021; Zou et al., 2018).

Some fundamental equations and perceptions behind the J48 decision tree algorithm are:

Information Gain: The J48 algorithm uses information gain to decide which attribute to fragment on at each node in the corresponding tree. Information gain is also grounded on the notion of entropy from the established information theory. The entropy (H(S)) of a set (S) is given by:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \ldots\ldots\ldots\ldots\ldots\ldots \textbf{(v)}$$

where (p_i) is the proportion of examples in class (i).

https://doi.org/10.35495/ajab.2025.017

Gain Ratio: To address the bias of information gain towards attributes with many values, J48 uses the gain ratio. The gain ratio is defined as:

$$\textbf{Gain Ratio}(A) = \frac{\textbf{Information Gain}(A)}{\textbf{Split Information}\ (A)}$$

where the split information is:

$$\textbf{Split Information}\ (A) = -\sum_{v \in \textbf{Values}(A)} \frac{|S_v|}{|S|} \log_2 \left(\frac{|S_v|}{|S|}\right) \text{ ..... (vi)}$$

Here, (S_v) is the subset of (S) for which attribute (A) has the value (v).

Pruning: Experts use pruning to make the decision tree smaller and then to stop it from getting too complicated. One way for doing this is called "reduced error pruning." Basically, it gets rid of parts of the tree that is not help in making better predictions when it is tested on a separate set of data. The error rate (E) of a subtree can be determined as:

$$E = \frac{\textbf{Number of misclassified instances}}{\textbf{Total number of instances}}$$

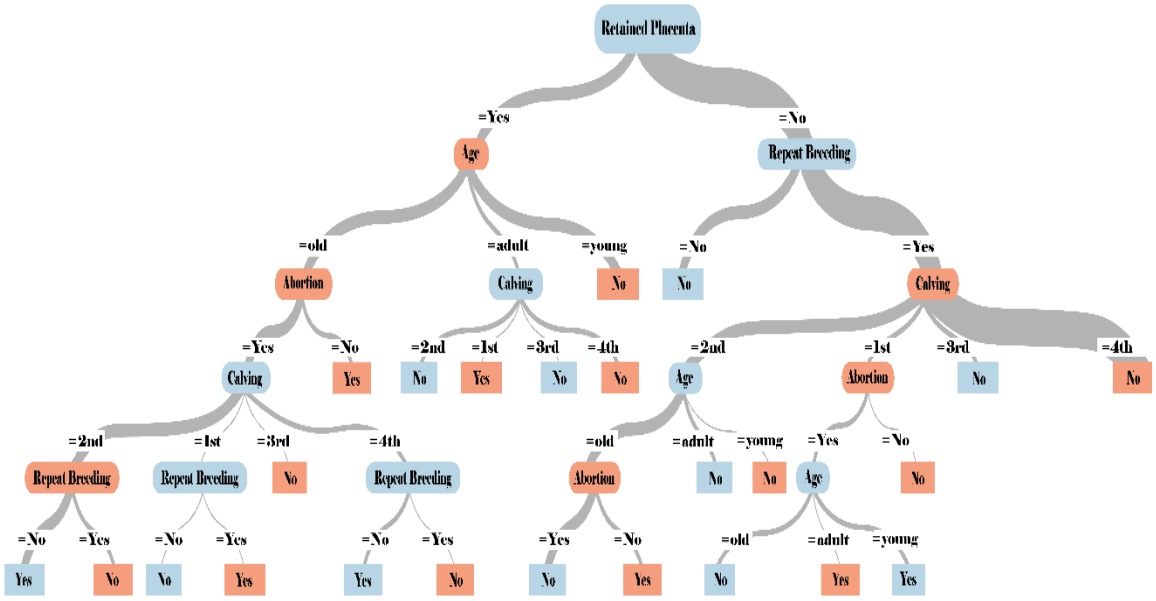

Figure-7. J48 tree underlying working pattern

AdaboostM1

Adaptive Boosting (AdaboostM1), is a broadly recognised ensemble learning technique which enhances their level of performance for weak classifiers by concentrating on misclassified instances via sequential iterations. This special adaptability makes AdaboostM1 predominantly effective for disease diagnosis and prediction problems, which often have imbalanced data and where the identification of intermittent cases is imperative. It has been proven to be productive in the prediction of diseases such as diabetes, plague, and cancer, as it increases model sensitivity and detects faint patterns related to mainly clinical features (Abdeldjouad et al., 2020; Eyupoglu and Karakuş, 2024; Potenciano et al., 2016).

Some fundamental equations and notions behind the AdaboostM1 algorithm are described below.

Weight Initialization: Initially, each training instance is assigned an equal weight:

$$w_i^{(1)} = \frac{1}{N} \text{ ..... (vii)}$$

where (N) is the total number of training instances.

Weak Learner Training: In each iteration (t), a weak learner (h_t(x)) is being trained from the weighted training data.

https://doi.org/10.35495/ajab.2025.017

Error Calculation: The weighted error (epsilon$_t$) of the weak learner is also calculated as:

$$\epsilon_t = \sum_{i=1}^{N} w_i^{(t)} I\left(y_i \neq h_t(x_i)\right) \text{.....(viii)}$$

where (1) is the indicator function that is 1 if the prediction is incorrect and 0 otherwise.

Alpha Calculation: The weight of the weak learner's vote, (alpha_t), is processed as:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \text{..... (ix)}$$

Weight Update: The weights of the training instances are restructured to focus more on the misclassified occurrences:

$$w_i^{(t+1)} = w_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right) \text{..... (x)}$$

The weights are then normalized so that they can be sum equal to 1.

Final Hypothesis: The final robust classifier (H(x)) is a weighted majority vote of the (T) weak classifiers:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \text{..... (xi)}$$

K-Nearest Neighbors

The IBk algorithm is a type of the k-Nearest Neighbors (k-NN) algorithm, which has widely been utilized in medical diagnostics and disease prediction problems (Musa et al., 2024). The algorithm acts by judging the feature set of a patient with alike cases in a dataset to construct predictions based on proximity in feature space. The non-parametric nature of the algorithm itself makes it flexible and adaptive to the datasets deprived of any predefined distributions thus, the icon tool for analyzing dissimilar medical data, where disease dynamics may diverge extensively from one population to other cases (John et al., 2024).

Some vital equations and perceptions behind the k-NN algorithm are described below.

Distance Metric: The K-NN algorithm bank on a distance metric to find the nearest neighbors. The most conjoint distance metric is the Euclidean distance, defined as:

$$\text{dist}(\mathbf{x}, \mathbf{z}) = \sqrt{\sum_{i=1}^{d} (x_i - z_i)^2} \text{..... (xii)}$$

where (\mathbf{x}) and (\mathbf{z}) are two points in a (d)-dimensional space.

Minkowski Distance: A more wide-ranging form of distance metric is the Minkowski distance, which includes the Euclidean distance as a special case:

$$\text{dist}(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{d} |x_i - z_i|^p\right)^{\frac{1}{p}} \text{..... (xiii)}$$

For (p = 1), it enhances the Manhattan distance, and for (p = 2), it becomes the Euclidean distance.

Classification Rule: For a specified test point (\mathbf{x}), the K-NN algorithm allots the supreme common label among its (k) nearest neighbors. Formally, if (S_x) indicates the set of the (k) nearest neighbors of (\mathbf{x}), predicted label (\hat{y}) is:

$$\hat{y} = \text{mode}(\{y_i : \mathbf{x}_i \in S_x\}) \text{..... (xiv)}$$

where (y_i) is the label of the (i)-th neighbor.

Weighted K-NN: In particular variations, the neighbors are weighted by their distance to the test point, giving nearer neighbors more persuade. The weight (w_i) for the (i)-th neighbor maybe written by:

$$w_i = \frac{1}{\text{dist}(\mathbf{x}, \mathbf{x}_i)} \text{..... (xv)}$$

The predicted label is then established by the weighted majority vote.

Results

Table 1 summarizes the performance metrics of five machine learning models. The models were evaluated using key indicators, including accuracy, kappa statistic, precision, recall, and F-measure, among others. In case of the accuracy of the five machine learning models WekaDeeplearning4j achieved the highest accuracy at 93.9359%, closely followed by MLP with 93.5927% and IBk with 93.135%. The J48 tree model demonstrates slightly lower accuracy at 92.3341%, while AdaboostM1 has the lowest accuracy among the models at 86.3844%. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), reveals notable differences. MLP has the lowest MAE at 0.1029, followed closely by IBk at 0.1077, indicating high precision. In contrast, AdaboostM1 has the highest MAE at 0.2208, showing a larger deviation from true values. Similarly, for RMSE, MLP achieves the lowest value at 0.2317, with IBk close behind at 0.2393, while AdaboostM1 again exhibits the highest error at 0.3174 (Figure 8).

https://doi.org/10.35495/ajab.2025.017

Table-1. Algorithm performance assessed by several assessment metrics

Parameter	**Machine learning models**				
	MLP	**WekaDeeplearning4j**	**J48 tree**	**AdaboostM1**	**IBk**
Accuracy	93.5927 %	93.9359 %	92.3341 %	86.3844 %	93.135 %
Kappa statistic	0.872	0.8788	0.8469	0.728	0.8629
Mean absolute error	0.1029	0.1315	0.118	0.2208	0.1077
Root mean squared error	0.2317	0.2436	0.2546	0.3174	0.2393
Relative absolute error	20.5738 %	26.2941 %	23.5954 %	44.1624 %	21.542 %
Root relative squared error	46.341 %	48.7267 %	50.9194 %	63.4799 %	47.8624 %
TP Rate	0.936	0.939	0.923	0.864	0.931
FP Rate	0.063	0.060	0.075	0.135	0.067
Precision	0.941	0.943	0.931	0.868	0.937
Recall	0.936	0.939	0.923	0.864	0.931
F-Measure	0.936	0.939	0.923	0.864	0.931
MCC	0.877	0.882	0.854	0.732	0.869
ROC Area	0.969	0.963	0.945	0.926	0.961
PRC Area	0.968	0.954	0.937	0.920	0.956
Required Time (s)	0.47	2.19	0.01	0.02	0.01
TP Rate: Total positive rate; FP Rate: False positive rate; MCC: Matthews correlation coefficient; ROC Area: Receiver operating characteristic Area; PRC Area: Precision-Recall Curve Area					

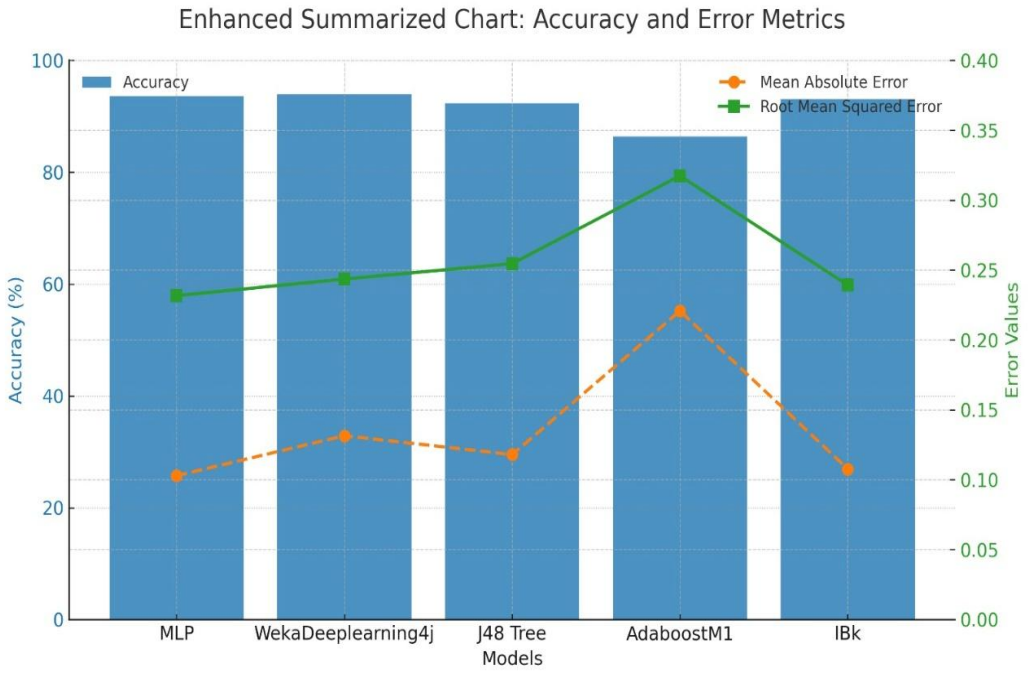

Figure-8. Comparison of accuracy and different error rates.

https://doi.org/10.35495/ajab.2025.017

In Figure 9, considering precision, WekaDeeplearning4j leads at 0.943, slightly ahead of MLP (0.941) and IBk (0.937), while J48 Tree and AdaboostM1 score 0.931 and 0.868, respectively. The recall values follow a similar trend, with WekaDeeplearning4j at 0.939, MLP at 0.936, IBk at 0.931, J48 Tree at 0.923, and AdaboostM1 at 0.864. F-Measure values mirror the recall, reinforcing the overall performance hierarchy. Finally, the kappa statistic, which measures agreement, is highest for WekaDeeplearning4j (0.8788) and MLP (0.872), with IBk (0.8629) and J48 Tree (0.8469) performing well, and AdaboostM1 scoring the lowest at 0.728. Based on these evaluation criteria the accuracy was identified.

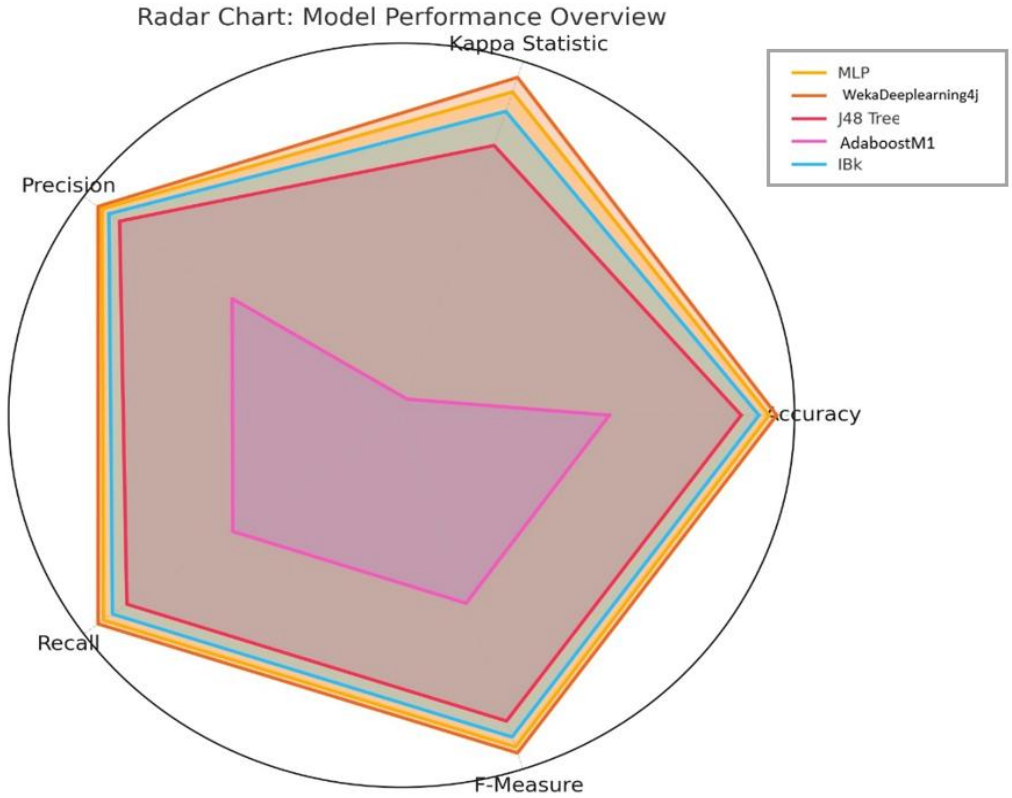

Figure-9. The radar chart illustrates the comparative evaluation of key metrics across various machine learning algorithms.

A total of five machine learning algorithms including MLP, WekaDeeplearning4j, J48 Tree, AdaboostM1, and IBk are used for predicting brucellosis in dairy cattle. Based on the different evaluation metrics WekaDeeplearning4j and MLP were (Abunna, 2018) found as the best algorithms to prediction brucellosis in dairy cattle.

Table-2. Ranked attributes

Name of attributes	**Weight of the attributes**
Repeat breeding	0.1251
Retained Placenta	0.2678
Calving	0.1046
Abortion	0.1031
Age	0.0884

Important risk factors are ordered as Retained Placenta > Repeat breeding > Calving > Abortion > Age

https://doi.org/10.35495/ajab.2025.017

Association rules (Apriori rules)

Based on the important risk factors we have found the following correlation of the factor for occurring brucellosis. The identified rules are:

1. Age=adult, Repeat breeding=No, Abortion=No, Retained Placenta=No ==> RBT=No <conf:(0.99)> lift:(1.97) lev:(0.1) [85] conv:(43.5)
2. Abortion=No, Repeat breeding=No, Retained Placenta=No ==> RBT=No <conf:(1)> lift:(1.97) lev:(0.15) [132] conv:(66.97)
3. Age=old, Retained Placenta=Yes ==> RBT=Yes <conf:(0.91)> lift:(1.83) lev:(0.12) [108] conv:(4.98)

Discussion

Brucellosis is a significant zoonotic disease that severely lowers livestock industry output while causing chronic and impairing health problems in humans. It is an endemic disease of thousands of years in Bangladesh. This research aims to predict brucellosis using various machine learning models on animal data, while identifying noteworthy risk features and their correlations related to the disease in Bangladeshi dairy cattle, to support a precise framework for applicable diagnosis and prevention-control schemes. The primary finding of the current study indicated that MLP and weekadeelearning4j exhibited the most effective performance in predicting brucellosis in dairy livestock. The MLP and weekadeelearning4j machine-learning algorithms are potent and resilient methods utilised in animal health research to evaluate risk variables (Michalak and Giacobini, 2023). The mathematical features of other models, which depend on sample size and data structure, may be the cause of their comparatively lower performance when compared to MLP and Weekadeelearninh4j (Nemes et al., 2009). Besides, among the risk factors we ranked the important risk factors retained placenta, repeat breeding, calving, abortion, age which are aligned with (Abunna, 2018)**.** Association rule mining is a data mining technique used extensively in the diagnosis field to discover hidden patterns, correlations, and associations among various variables within large datasets. Regarding disease prediction, association rules can be applied to establish relations among various symptoms, diagnosis parameters in order to achieve early-stage diagnosis and prevention (Abbasi et al., 2020). Our association rules revealed the underlying correlation of the attributes for occurring the disease. In the first rule, we found in case of adult cow which has no abortion, no repeat breeding and no retention of placental history then the animal was disease free and the rule shows it is 99% confident. In case of second one, it shows, when an animal has no abortion, neither repeat breeding and nor retention of placenta history, it is likely to be disease negative and the rule is 100% confident. In third rule, the apriori rule showed when the animal is old having the placental problem then it has 91% chances to be Brucella affected.

If this strategy is successfully implemented at the field level, it will be feasible to diagnose the disease quickly and affordably. Additionally, if steps are taken to prevent or reduce the risk factors, it will improve the economic condition of farmers and ultimately the nation's economy. The quantity of the dataset, factors taken into account, and geographic location could all affect the study's outcome.

Brucellosis can significantly impact a country's economy by affecting livestock production, trade, and public health. The successful implementation of this approach can contribute to the country's economic stability by improving livestock health, reducing production losses, and facilitating trade. For example, in Brazil, the total estimated losses attributed to bovine brucellosis were approximately R$ 892 million (Santos et al., 2013). By effectively controlling brucellosis, countries can reduce economic losses and improve their overall economic performance.

The Savar district is the sole source of the data, and there are very few samples. Area can have an impact on animal health as well as forecast results, and low data may indicate less accuracy.

Conclusion

A total of five machine learning algorithms we used, and two algorithms were found for predicting brucellosis including MLP and weekadeeplearning4j having 93.59% and 93.94% accuracy, respectively. Besides, found the most important factors which are Retained Placenta > Repeat breeding > Calving > Abortion > Age. Also, we identified three apriori rules to see the underlying correlation of the attributes for occurring the disease. By applying these techniques in the field level it will be very helpful to identify the disease early without laboratory and prevention of the

disease. Several initiatives can be taken to prevent or mitigate the risk factors associated with brucellosis. These include vaccination of animals, testing and culling of infected animals, and maintaining proper hygiene and biosecurity measures on farms. By employing these initiatives, farmers can safeguard their livestock and livelihoods from the economic impact of brucellosis with the help of expert. It can help in improving farmers' income and livelihoods anticipated to production losses, including abortions, reduced milk yields, and trade restrictions.

Future Recommendations

This dataset consists of small number of data, in the upcoming days more data can be used for more reliability of the disease prediction.

Acknowledgements

Located at Farmgate, Dhaka the Department of Livestock Services has well-funded the research under the LDDP Research and Innovation Sub-project, Livestock and Dairy Development Project-Transmission dynamics of brucellosis and abortion risk factors in large dairy herds in Bangladesh (project code: RP-C-01-10).

Disclaimer: None.
Conflict of Interest: None.
Source of Funding: None.

Ethical Approval Statement

The study protocol was peer reviewed and approved by the Ethical Review Committee of appropriate authority. Animal research was approved by the Faculty of Veterinary Science of Bangladesh Agricultural University.

Data Availability Statement

The dataset is accessible on the GITHUB website: https://github.com/Mokammel-14/Brucellosis-Dataset

Contribution of Authors

Hussaini SMAK & Rahman MS: Data collection, designing the experiment and preparing the initial draft of manuscript.
Hussaini SMAK, Siddique FI and Sindi S: Reviewed literature, critiqued and edited the final draft of manuscript.
Asif AH, Yusuf T & Khan S: Reviewed literature and prepared the initial draft of manuscript.
Tito MH, Arifuzzaman M, Asif AH & Chohan MS: Formal analysis and prepared the initial draft of manuscript.
Rahman MS, Al Mamun A, Tito MH & Arifuzzaman M: Conceived ideas and designed the experiment.

References

Abbasi M, Karimipour F and Gholipour S, 2020. Detection of the Association Rules of the Occurrence of Brucellosis in Humans Using Spatial Data Mining. Depict. Health. 11: Article 1.

Abdelbaset AE, Abushahba MFN, Hamed MI and Rawy MS, 2018. Sero-diagnosis of brucellosis in sheep and humans in Assiut and El-Minya governorates, Egypt. Int. J. Vet. Sci. Med. 6: S63–S67.

Abdeldjouad FZ, Brahami M and Matta N, 2020. A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques. In: Jmaiel M, Mokhtari M, Abdulrazak B, Aloulou H, Kallel S (Eds.), The Impact of Digital Technologies on Public Health in Developed and Developing Countries. Springer Int. Publ. pp. 299–306.

Abunna F, 2018. Assessment of major reproductive health problems, their effect on reproductive performances and association with brucellosis in dairy cows in Bishoftu town, Ethiopia. J. Dairy Vet. Anim. Res. 7: Article 1.

Ahmed BS, Osmani MG, Rahman AKMA, Hasan MM, Maruf AA, Karim MF, Karim SMA, Asaduzzaman M, Hasan MR, Rahman MM and Rahman MS, 2018. Economic impact of caprine and ovine brucellosis in Mymensingh district, Bangladesh. Bangladesh J. Vet. Med. 16: 193–203.

Ahsan MM, Luna SA and Siddique Z, 2022. Machine-learning-based disease diagnosis: A comprehensive review. Front. Healthcare 10: 541. https://doi.org/10.3390/healthcare10030541.

Akhvlediani T, Bautista CT, Garuchava N, Sanodze L, Kokaia N, Malania L, Chitadze N, Sidamonidze K, Rivard RG, Hepburn MJ, Nikolich MP, Imnadze P and Trapaidze N, 2017. Epidemiological and Clinical Features

of Brucellosis in the Country of Georgia. PLoS ONE. 12: e0170376.

Alamian S and Dadar M, 2020. Brucella melitensis infection in dog: A critical issue in the control of brucellosis in ruminant farms. Comp. Immunol. Microbiol. Infect. Dis. 73: 101554.

Asmare K, Sibhat B, Molla W, Ayelet G, Shiferaw J, Martin AD, Skjerve E and Godfroid J, 2013. The status of bovine brucellosis in Ethiopia with special emphasis on exotic and crossbred cattle in dairy and breeding farms. Acta Trop. 126: 186–192.

Bishop C, 1992. Neural networks and their diagnostic applications. Rev. Sci. Instrum. 63: 4772–4774.

Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP, 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16: 321–357.

Ducrotoy M, Bertu WJ, Matope G, Cadmus S, Conde-Álvarez R, Gusi AM, Welburn S, Ocholi R, Blasco JM and Moriyón I, 2017. Brucellosis in Sub-Saharan Africa: Current challenges for management, diagnosis and control. Acta Trop. 165: 179–193.

Ducrotoy MJ, Muñoz PM, Conde-Álvarez R, Blasco JM and Moriyón I, 2018. A systematic review of current immunological tests for the diagnosis of cattle brucellosis. Prev. Vet. Med. 151: 57–72.

Eltholth MM, Hegazy YM, El-Tras WF, Bruce M and Rushton J, 2017. Temporal Analysis and Costs of Ruminant Brucellosis Control Programme in Egypt Between 1999 and 2011. Transbound. Emerg. Dis. 64: 1191–1199.

Eyupoglu C and Karakuş O, 2024. Novel CAD Diagnosis Method Based on Search, PCA, and AdaBoostM1 Techniques. J. Clin. Med. 13: Article 10.

Franc KA, Krecek RC, Häsler BN and Arenas-Gamboa AM, 2018. Brucellosis remains a neglected disease in the developing world: A call for interdisciplinary action. BMC Public Health. 18: 125.

Godfroid J, Nielsen K and Saegerman C, 2010. Diagnosis of Brucellosis in Livestock and Wildlife. Croat. Med. J. 51: 296–305.

Gwida M, El-Ashker M, Melzer F, El-Diasty M, El-Beskawy M and Neubauer H, 2016. Use of serology and real-time PCR to control an outbreak of bovine brucellosis at a dairy cattle farm in the Nile Delta region, Egypt. Ir. Vet. J. 69: Article 3.

Hosein HI, Zaki HM, Safwat NM, Menshawy AMS, Rouby S, Mahrous A and Madkour BE, 2018. Evaluation of the General Organization of Veterinary Services control program of animal brucellosis in Egypt: An outbreak investigation of brucellosis in buffalo. Vet. World. 11: 748–757.

Jamil T, Melzer F, Saqib M, Shahzad A, Khan Kasi K, Hammad Hussain M, Rashid I, Tahir U, Khan I, Haleem Tayyab M, Ullah S, Mohsin M, Mansoor MK, Schwarz S and Neubauer H, 2020. Serological and Molecular Detection of Bovine Brucellosis at Institutional Livestock Farms in Punjab, Pakistan. Int. J. Environ. Res. Public Health. 17: Article 4.

John TJL-S, Kanwar O, Abidi E, Nekidy WE and Piechowski-Jozwiak B, 2024. Towards artificial intelligence-based disease prediction algorithms that comprehensively leverage and continuously learn from real-world clinical tabular data systems. PLoS Digit. Health. 3: e0000589.

Khurana SK, Sehrawat A, Tiwari R, Prasad M, Gulati B, Shabbir MZ, Chhabra R, Karthik K, Patel SK, Pathak M, Iqbal Yatoo M, Gupta VK, Dhama K, Sah R and Chaicumpa W, 2021. Bovine brucellosis – a comprehensive review. Vet. Q. 41: 61–88.

Kothalawala KAC, Makita K, Kothalawala H, Jiffry AM, Kubota S and Kono H, 2017. Association of farmers' socio-economics with bovine brucellosis epidemiology in the dry zone of Sri Lanka. Preventive Veterinary Medicine. 147: 117–123.

Lang S, Bravo-Marquez F, Beckham C, Hall M and Frank E, 2019. WekaDeeplearning4j: A deep learning package for Weka based on Deeplearning4j. Knowl.-Based Syst. 178: 48–50.

Mathur P, Srivastava S, Xu X and Mehta JL, 2020. Artificial intelligence, machine learning, and cardiovascular disease. Clin. Med. Insights Cardiol. 14: 1179546820927404. https://doi.org/10.1177/1179546820927404.

Mburu JW, Kingwara L, Ester M and Andrew N, 2018. Use of classification and regression tree (CART), to identify hemoglobin A1C (HbA1C) cut-off thresholds predictive of poor tuberculosis treatment outcomes and

associated risk factors. J. Clin. Tuberc. Other Mycobact. Dis. 11: 10–16.

McDermott J, Grace D and Zinsstag J, 2013. Economics of brucellosis impact and control in low-income countries. Revue Scientifique et Technique. 32(1):249–261.

Michalak K and Giacobini M, 2023. Evolutionary Neural Networks for Livestock Disease Prevention. Proc. Companion Conf. Genet. Evol. Comput. pp. 395–398.

Mick V, Carrou GL, Corde Y, Game Y, Jay M and Garin-Bastuji B, 2014. Brucella melitensis in France: Persistence in Wildlife and Probable Spillover from Alpine Ibex to Domestic Animals. PLoS ONE. 9: e94168.

Mohd Radzi SF, Hassan MS and Mohd Radzi MAH, 2022. Comparison of classification algorithms for predicting autistic spectrum disorder using WEKA modeler. BMC Med. Inform. Decis. Mak. 22: 306.

Musa M, Mohammed AU, Musa S and Jabaka LM, 2024. A K-Nearest Neighbor (KNN)-Algorithm in poultry diseases monitoring system. IJSGS. 10(3): Article 696. https://doi.org/10.57233/ijsgs.v10i3.696

Nemes S, Jonasson JM, Genell A and Steineck G, 2009. Bias in odds ratios by logistic regression modelling and sample size. BMC Med Res Methodol. 9:56. https://doi.org/10.1186/1471-2288-9-56

O'Callaghan D, 2020. Human brucellosis: Recent advances and future challenges. Infect. Dis. Poverty. 9: Article 4.

Pham BT, Tien Bui D and Prakash I, 2017. Landslide susceptibility assessment using bagging ensemble-based alternating decision trees, logistic regression, and J48 decision trees methods: A comparative study. Geotech. Geol. Eng. 35: 2597–2611.

Potenciano V, Abad-Grau MM, Alcina A and Matesanz F, 2016. A comparison of genomic profiles of complex diseases under different models. BMC Med. Genomics. 9: Article 3.

Quinlan JR and Rivest RL, 1989. Inferring decision trees using the minimum description length principle. Inf. Comput. 80:227–248. https://doi.org/10.1016/0890-5401(89)90010-2).

Rahman MH, Akther S, Haque MN, Ali MZ, Zihadi MAH and Rahman MZ, 2024. Ovine brucellosis in Bangladesh: Seroprevalence and its associated risk factors. Journal of Research in Veterinary Sciences. 3(1): 28. https://doi.org/10.5455/JRVS.20240519053238

Ripley BD, 1994. Neural networks: A review from a statistical perspective: Comment. Stat. Sci. 9: 45–48.

Santos RL, Martins TM, Borges ÁM and Paixão TA, 2013. Economic losses due to bovine brucellosis in Brazil. Pesq. Vet. Bras. 33: 759–764. https://doi.org/10.1590/S0100-736X2013000600012.

Selim A, Attia K, Ramadan E, Hafez YM and Salman A, 2019. Seroprevalence and molecular characterization of Brucella species in naturally infected cattle and sheep. Prev. Vet. Med. 171: 104756.

Selim A, Megahed A, Kandeel S, Alanazi AD and Almohammed HI, 2021. Determination of seroprevalence of contagious caprine pleuropneumonia and associated risk factors in goats and sheep using classification and regression tree. Animals. 11: Article 4.

Sharmy ST, Yeasmin F, Ahmed A, Tito MH, Rahman MS, Ehsan MA and Rahman AKMA, 2024. Identification of *Brucella* spp. in aborted fetuses by guinea pig inoculation. Bangladesh J. Vet. Med. 22: 1–5. https://doi.org/10.33109/bjvmjj2024fam1

Singh A, Mehta JC, Anand D, Nath P, Pandey B and Khamparia A, 2021. An intelligent hybrid approach for hepatitis disease diagnosis: Combining enhanced k-means clustering and improved ensemble learning. Expert Syst. 38: e12526.

Tito MH, Arifuzzaman M, Jannat MHE, Rahman MS, Sharmy ST, Nasrin A, Asaduzzaman M, Ashrafuzzaman M, Prince DB and Asif AH, 2023. A comparative study of ensemble machine learning algorithms for brucellosis disease prediction: Detection of brucellosis using artificial intelligence. Lett. Anim. Biol. 3: 23–27.

Tito MH, Jannat MHE, Afrose M, Ahmed SMJ, Maruf SM, Hossain MA, Samani S, Mira RJ, Saha B, Jihad AI and Das TK, 2024a. Deciphering foot and mouth disease predictive modeling: uncovering attribute correlations and risk factors with advanced machine learning. *Vet. Sci. Res. Rev.* 10(2): 58-71.

https://dx.doi.org/10.17582/journal.vsrr/2024/10.2.58.71

Tito MH, Rahman MS, Jannat MHE, Sharmy ST, Nasrin A and Asaduzzaman M, 2024b. Prediction of brucellosis disease with ensemble machine learning application. IET Conference Proceedings 2023, 163–166. [Online]. Available at: https://doi.org/10.1049/icp.2024.0918.

WOAH, 2018. https://www.woah.org/app/uploads/2021/03/3-01-04-brucellosis-1.pdf

Yu WL and Nielsen K, 2010. Review of detection of Brucella sp. by polymerase chain reaction. Croat. Med. J. 51: 306–313.

Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H, 2018. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 9: Article 1.