Mathematical Biosciences and Engineering

MBE, 22(5): 1140–1158.
DOI: 10.3934/mbe.2025041
Received: 30 October 2024
Revised: 05 February 2025
Accepted: 10 February 2025
Published: 08 April 2025

https://www.aimspress.com/journal/MBE

Research article

Artificial neural networks to predict the presence of Neosporosis in cattle

Javier Antonio Ballesteros-Ricaurte[1,2,3,*], Ramon Fabregat[3], Angela Carrillo-Ramos[4], Carlos Parra[4] and Andrés Moreno[4]

[1] Escuela de Ingeniería de Sistemas y Computación, Universidad Pedagógica y Tecnológica de Colombia, Tunja 150003, Colombia
[2] Doctorado en Ingeniería, Pontificia Universidad Javeriana, Bogotá 110231, Colombia
[3] Broadband Communications and Distributed Systems, University of Girona, Girona 17007, Spain
[4] Departamento de Ingeniería de Sistemas, Pontificia Universidad Javeriana, Bogotá 110231, Colombia

* **Correspondence:** Email: javier.ballesteros@uptc.edu.co; Tel: +603184001387.

Abstract: The prediction of bovine infectious diseases is a constant challenge as generally, only laboratory data is available not allowing the study of their relationship with each disease's risk factors. The diseases neosporosis and bovine viral diarrhea, which are present in Colombia, the United States, Mexico, Brazil, and Argentina, cause reproductive problems in cattle and generate economic losses for ranchers. Although there are mathematical models that can evaluate which cattle are susceptible to these diseases, these provide limited information, maintaining the need for a model that provides information on both transmission and mechanisms for controlling the disease. In this article, a machine learning model is presented that combines laboratory data with risk factors in a neural network to predict the presence of bovine neosporosis. The proposed model was implemented with data from previous studies conducted in the municipality of Sotaquirá, Boyacá, Colombia, and obtained an accuracy of 94% in predicting the presence of the disease. It can be concluded that incorporating laboratory data into machine learning algorithms improves the prediction of the presence of these diseases. Furthermore, the proposed system not only predicts but also provides useful information for clinical decision-making, making it a valuable tool in the veterinary field.

Keywords: machine learning; bovine; infectious diseases; event; prediction; metrics

1. Introduction

Bovine neosporosis, caused by the protozoan Neospora caninum, is one of the worldwide leading infectious diseases responsible for abortions in cattle [1]. This disease causes significant economic losses for producers due to reduced fertility, decreased milk production, and high costs associated with control measures and replacement of affected animals. The risk factors have been evaluated in previous studies, observations, and surveys conducted with the personnel in charge of cattle [2]. These surveys provide information about administered vaccines, symptoms exhibited by cattle, and their interactions with other animals, among other aspects. Traditional methods for diagnosing bovine neosporosis include serological tests such as ELISA, which, although widely used, have limitations in terms of accessibility, cost, and response time. Also, conducting these studies presents some challenges: the costs are high for small and medium-sized cattle associations, and the periodic serum sampling induces stress in cattle, leading to reduced milk production [2].

In this project, bovine neosporosis has been chosen because epidemiological studies are expensive for cattle farmers, and the research group in Veterinary Medicine and Zootechnics, GIDIMEVETZ at the Pedagogical and Technological University of Colombia (UPTC) has available data collected from studies conducted to determine the presence of this disease in the Boyacá department in Colombia [3]. Although epidemiological studies have been conducted in different countries to better understand this disease, and its presence has been reported in Colombia, its epidemiological distribution across the country's regions is still unknown [2].

To determine the occurrence of bovine neosporosis, it is necessary to have mechanisms that utilize available data, such as health records, demographic data (e.g., age, breed), and previous medical history (i.e., diagnoses, laboratory tests, and risk factors). The availability of this data, along with advances in computer hardware [4] as well as machine and deep learning algorithms, make it possible to develop systems to predict the presence of the disease with high accuracy and in a short time [5].

Machine learning algorithms have been utilized on basic patient data in human medicine for aspects such as disease prediction [4,5], treatment organization [6], medical diagnosis [7], and medication selection [8]. They have also been applied in public health decision-making and logistics for locating healthcare service providers [9]. In recent years, machine learning has been successfully used in the detection of various infectious diseases in animals and humans. Currently, a disease called lumpy skin disease is spreading very rapidly among cattle and water buffalos. In [10], authors aimed to predict whether cattle in a specific geographic location currently have or may develop lumpy skin cancer in the future. They applied various machine learning algorithms to the nodular skin disease dataset, comparing their accuracy in predicting the disease. Among all the applied algorithms, the random forest algorithm outperformed the others, achieving the highest accuracy of 97.7%. In [11], machine learning techniques such as support vector machine (SVM), gradient boosting, and random forest were used to diagnose lumpy skin disease. The aim of that study was to create a diagnostic tool that is accurate and effective for the early identification and treatment of this infectious disease. In [12], the objective was to implement a machine learning approach to predict maternal and environmental factors associated with infectious APM (abortions and perinatal mortalities). Production type (dairy/beef), gestation length, and season were successfully predicted by the ensemble, with modest predictive power ranging from 63% to 73%. In addition to the predictive accuracy of individual variables, the MLP (machine learning pipeline) hierarchically identified predictive drivers of environmental/maternal characteristics associated with APM. In [13], machine learning algorithms,

including artificial neural networks and random forests, were applied to predict bovine leukosis virus seropositivity in dairy cattle, demonstrating the utility of these techniques in veterinary epidemiology.

Despite advances in the use of machine learning tools such as artificial neural networks (ANN) to predict bovine infectious diseases, this remains an underexplored topic [1]. The main challenges include incomplete and outdated information [14]; in addition, trends in bovine infectious diseases are not well-known. The absence of models that integrate multiple sources, including clinical, management, and environmental data, restricts the ability to produce accurate and generalizable predictions. Furthermore, result interpretation remains a challenge, which hinders the adoption of these tools by veterinarians and farmers.

The objective of this study is to design a predictive model for the presence of infectious diseases on different farms using the occurrence data on bovine neosporosis, risk factor data, contextual variables, and clinical laboratory data from previous studies. Predicting infectious diseases such as neosporosis in cattle presents significant challenges, particularly in rural environments where the availability and quality of the data may be limited. This is due to the lack of systematic and standardized records on many livestock farms. Despite these limitations, advances in machine learning, such as the use of neural networks, offer a promising framework for maximizing the value of available data.

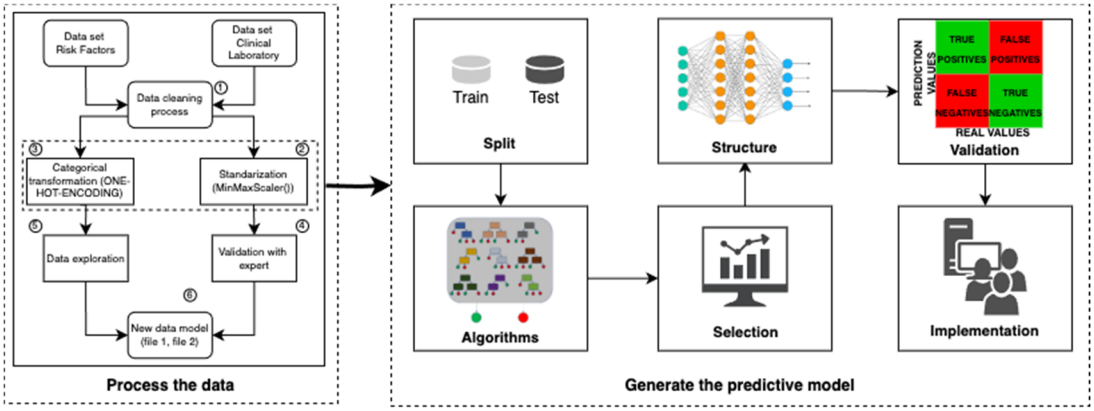

Figure 1. General process for developing the predictive model.

Figure 1 illustrates the general process for developing the proposed predictive model. This model consists of subprocesses for processing data and the predictive model generation. The purpose of the model is to reduce the reporting time of the presence of the disease, seeking to minimize social costs and economic losses caused by these infectious diseases, taking advantage of the available information and laying the foundations for future improvements in data collection and standardization, thus allowing its applicability in contexts with limited resources.

2. Prior study to process the data

To build machine learning models, it is essential to conduct a preliminary data study, which includes exploring and analyzing the available dataset to gain a comprehensive understanding of its features, identify patterns, detect outliers, and uncover relationships between variables.

Data preprocessing is one of the most critical steps in any machine learning application [15]. It

involves a set of techniques and procedures to manipulate and transform raw data into a more useful and suitable format for analysis, interpretation, or application. Relevant features are extracted from the raw data and standardized to optimize performance. Ensuring that the machine learning algorithm performs well with both the training dataset and new data is crucial.

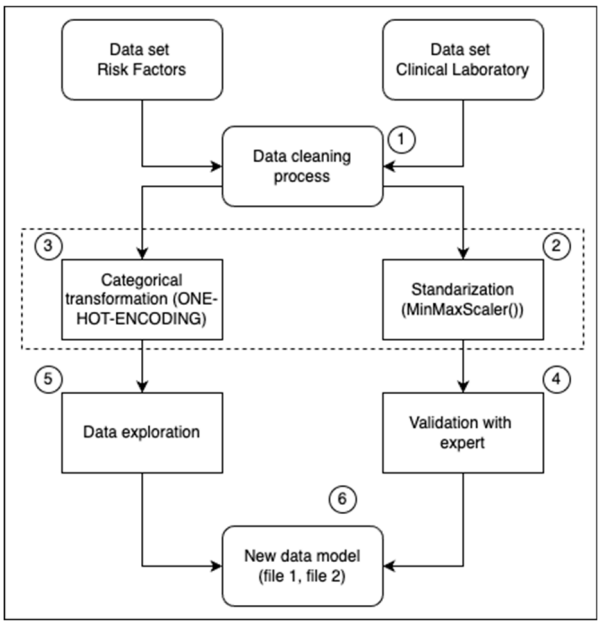

Figure 2. Flowchart illustrating data processing.

Data preprocessing involves several stages, as depicted in Figure 2. It is important to note that these stages are not always followed linearly. In this project, preprocessing is applied to both datasets, and each stage is explained below.

2.1. Data collection

In the dairy section of the Boyacá Department, data is generated through various processes using both internal and external sources of information. There is no standardized format for recording this data, meaning each stakeholder provides it in different formats. Furthermore, information is lost due to the lack of proper and comprehensive record-keeping on farms. Farmers record information about cattle in notebooks, resulting in limited available data.

For this project, the data from the cattle population used for analysis and characterization has the following properties:

- This population includes Ayrshire, Holstein, Jersey, and Normando breeds, with ages ranging from 2 to 4 years.
- The reference population comprises female cattle from various cattle farms specializing in milk production in an extensive farming system, with an average production rate of 21.2 L/cow/day [16].
- Data on farm abortions, weak calves, and embryonic deaths were obtained through surveys conducted on the farms where serum samples were collected (from studies registered by the research group). Individual cattle data was also collected.
- The following criteria were met: authorization was requested from the farm owners, who

expressed their consent to participate in the study; the privacy and confidentiality of participant information were guaranteed; the owners' data and the identification of Neospora caninum-seropositive animals remained completely anonymous.

All data were stored in an excel file to be used for model validation. Two data sources were considered for training and validating the prediction model:

a) Data from previous studies: The dataset used for this study was collected from farms in specific regions of Colombia, with a total of 1000 bovines. The GIDIMEVETZ research group at UPTC has conducted several studies on bovine neosporosis [16], obtaining clinical laboratory data and identifying risk factors for the disease. While representative of these areas, this dataset may not capture the variability in environmental and management conditions present in other regions. This data was analyzed to understand and evaluate the relevant variables for the modeling process. Additionally, it was used to train and validate the proposed prediction model.
b) Data from sample collection: 10 cattle farms with a total of 460 cattle in the municipality of Toca (Boyacá) were selected to collect new serum samples, conduct surveys with the farm managers and individuals familiar with the daily cattle behavior and risk factors, and analyze disease trends.

Serological samples were taken randomly from female cattle within a specific age range across the dairy breeds considered. The samples were drawn directly from the vein using needles and stored in tubes. The tubes were labeled with the animal's identification number, age, farm name, and reproductive status, and the sera were then transported to the Clinical laboratory and stored at -4 ºC until analyzed.

Table 1. Description of variables in the clinical laboratory data file.

Clinical laboratory	Data type
CP (positive control): Used to confirm that the ELISA test is accurately detecting the antigen.	Float
CN (negative control): Used to verify that there is no nonspecific reactivity in the sample that could produce a false positive signal.	Float
Blank: Used to establish the baseline and detect any possible interference in the ELISA test.	Float
Do corr: Correlation in the ELISA test refers to the elimination of interference or background noise in the test sample and the test controls.	Float
Prom blank (average of the blank variable): The average value of the absorbance measured in the blank control in the ELISA test.	Float
Prom CP corr (positive control average): Positive control average value used as a reference to determine the presence and amount of antigen in the test sample.	Float
Prom CN corr (negative control average): The value used to determine the absence of antigen in the test sample.	Float
Rate Neospora: The ratio between the absorbance of the test sample and the absorbance of the positive control.	Float
Antigens: The amount of antigen present in the test sample.	Float
Result: From the processing of the samples with the ELISA kit, the result is obtained, which can be positive, negative, or suspicious.	Integer

In the clinical laboratory, samples were centrifuged at 1500 rpm for 10 min, and the resulting serum

was transferred to a storage tube. Samples were processed using the indirect ELISA technique [17] to detect antibodies against Neospora caninum, using the corresponding commercial kit. The kit instructions were followed to determine if a sample was positive or negative. It is worth noting that the centrifugation machines are not connected to a computer, instead they print a report. Clinical laboratory staff placed the samples in the machine, waited for the results, and manually entered them into an excel file. Each ELISA kit, corresponding to each disease, provides detailed instructions for processing the samples and interpreting the results in the clinical laboratory.

Table 2. Description of the variables in the risk factors file.

Risk factors	Data type
ID: Identification number of each bovine.	Integer
Age: Age of each bovine (years).	Integer
Breed: Term used to classify groups of cattle based on their physical characteristics. The breeds considered are Ayrshire, Holstein, Jersey, and Norman.	Integer
Abortion: If the female bovine has suffered a pregnancy loss before the normal term, whether spontaneous or induced.	Boolean
Repetition: Lack of conception or repeated reproductive failures in a female bovine over several mating seasons.	Boolean
Non-carrying: Absence of pregnancy after mating or artificial insemination. It can be caused by various reasons.	Boolean
Dystocia: Difficulty in childbirth (can be caused by various factors).	Boolean
Weak calf: One that is born with low weight and shows signs of weakness and decreased body temperature after birth.	Boolean
Embryonic death: Loss of the embryo in the early stages of development. It can be caused by various factors.	Boolean
Natural breeding: A male bovine is used on the farm for animal reproduction.	Boolean
Insemination: Artificial insemination is used on the farm for animal reproduction.	Boolean
Drinkers: Cattle drink from a container where water is poured or from a well. These waterers can be for only the cattle on a farm, or they can be shared with other farms. When they share drinking fountains, the value is taken to be true, otherwise it is false.	Boolean
Pastures: Area of fenced field covered with grass that farms distribute to rotate cattle. They are for the cattle on the farm, although sometimes they are shared with the cattle on other farms. The true value is given if the paddocks are shared.	Boolean
Corral: An isolated place that some farmers have on their farms to take sick cattle and thus not infect other cattle.	Boolean
Milking type: Can be manual or automatic. In the manual case, farmers or farm managers carry out the milking process and often do not follow good practices. In the automatic case, they use machines to milk.	Boolean
Answer: Whether the bovine has the disease or not (in this case the disease bovine neosporosis is considered).	Boolean

After visiting the farms, conducting surveys with farmers or managers, collecting serum samples, and taking them to the clinical laboratory for processing, data were organized into two files as shown

in Tables 1 and 2. The clinical laboratory data file contains 10 variables obtained from the sample processing in the clinical laboratory. The risk factors data file contains 14 variables collected from the surveys, covering both demographic characteristics and livestock management practices, which have a significant impact on susceptibility to the disease studied.

Risk factors were selected through a structured process that combined expert analysis, literature review, and machine learning techniques. Variables such as abortions, dystocia, and shared pastures were chosen based on previous studies that identified them as key determinants in the spread of infectious diseases in cattle. A contingency table and correlation analysis were used to identify significant relationships between risk variables and the presence of bovine neosporosis. Specific factors such as types of waterers and access to veterinary services were included due to their high prevalence in the livestock system of the Boyacá region, Colombia. SHAP (Shapley Additive Explanations) values helped confirm the importance of each variable in the predictions, validating the inclusion of the most influential factors such as abortions, breed, and dystocia. This data was stored in an excel file containing 1000 records. Each record was linked to an identifier that included the name of the bovine, age, breed, locality, and farm.

The results of the surveys were entered into an excel file. Survey responses help to identify the risk factors (abortion, repletion, non-pregnancy, dystocia, weak calves, and embryonic death) and contextual variables (natural mating, artificial insemination, watering sources, pastures, pens, and milk type) present on the farms. They also help to identify whether farm managers or owners use pens to separate sick cattle, have other animals on the farms, implement control measures, or utilize the services of specialized veterinary personnel.

2.2. Data exploration

The objective of exploring data from surveys, clinical laboratory tests, and farm observations regarding bovine infectious diseases is to understand the data structure. This process involves handling outliers, addressing missing data, removing extra spaces, and normalizing values. These processes are carried out using Python libraries and applied to both the clinical laboratory data and the risk factor data of the cattle. Due to the low number of records, it is not viable to eliminate records with missing values.

The dataset for bovine infectious disease prediction is divided into three subsets: basic characteristics (age, gender, breed), risk factors (artificial insemination, abortion, heat repeat, direct mating with a bull, non-pregnancy, dystocia, weak calves at birth, and embryonic death), and clinical laboratory tests (disease standard value, positive control, negative control, blank, ratio, and result). Each subset consists of several features that need adjustment in the process. For this, dimensionality reduction is used based on expert judgment [6], and techniques such as the correlation matrix between variables and contingency table are employed to obtain the features most related to the disease.

A correlation matrix was created for the clinical laboratory data to show the correlations between each pair of variables. Each row and column represents a different variable, and the values at the intersection of rows and columns indicate the degree of correlation between those two variables. As shown in Figure 3, correlation values range from -1 to 1, where positive values indicate positive correlation, and negative values indicate negative correlation.

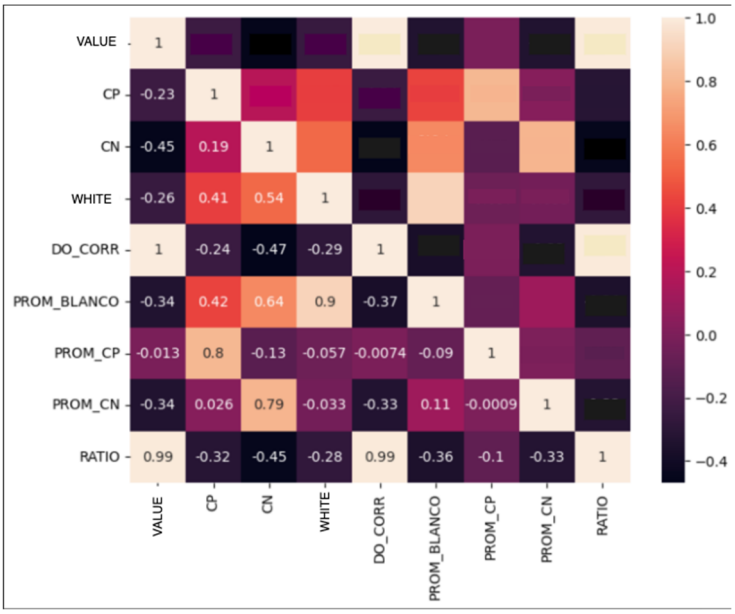

Figure 3. Correlation matrix for the clinical dataset.

The correlation matrix reveals the relationships between variables in the dataset helping to identify significant patterns and trends. A positive correlation indicates that as one variable increases, the other also increases. When analyzing the correlation between CP and average blank, it is visible that an increase in CP corresponds to an increase in average blank. A negative correlation means that as one variable increases, the other decreases. Analyzing the correlation between CP and the antigen, the results suggest that an increase in CP leads to a decrease in antigen levels.

To indicate the relationships in the risk factor dataset, a contingency table is used to summarize and analyze the relationship between two categorical variables [15]. To illustrate this process, consider the relationships between the variable breed and age with the response variable (presence of the disease). Table 3 displays the relationship between the breed variable and the presence of bovine neosporosis.

Table 3. Contingency table showing the correlation between breed and the presence of bovine neosporosis disease.

		Breed				
		Ayrshire	Holstein	Jersey	Normando	Total
Neospora	Negative	118 (49.2%)	143 (49.1%)	132 (62.3%)	157 (61.9%)	550
	Positive	122 (50.8%)	148 (50.9%)	80 (37.7%)	100 (38.1%)	450
	Total	240	291	212	257	1000

The results of the contingency tables are used to show the distribution of one categorical variable concerning another. Figure 4 illustrates that the Ayrshire and Holstein breeds are balanced in terms of the presence or absence of the disease, whereas the Jersey and Normando breeds exhibit a lower percentage of disease presence.

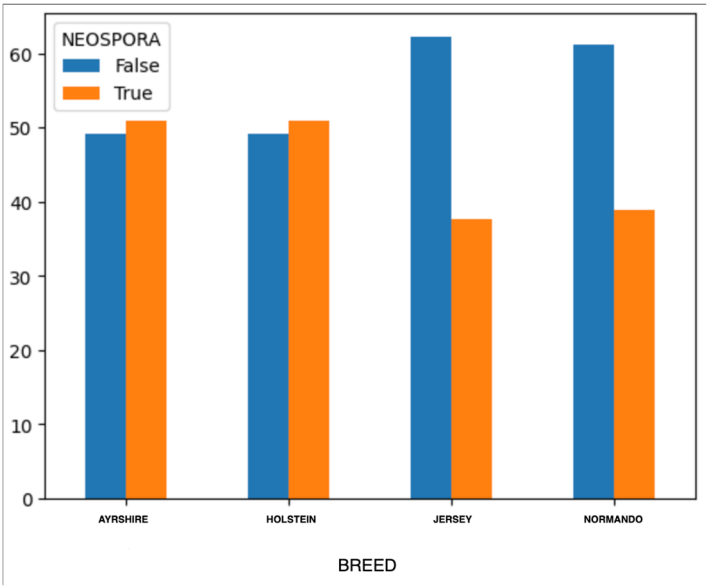

Figure 4. Visualization of the relationship between breed and presence of the disease.

The relationship between the age and the presence of bovine neosporosis is shown in Table 4 and Figure 5. The disease's presence is higher at ages 2 and 3 but less common at age 4.

Table 4. Contingency table showing the correlation between age and the presence of the bovine neosporosis disease.

		Breed			
		2–3 years	3–4 years	> 4 years	Total
Neospora	Negative	127 (58.3%)	313 (51.2%)	110 (64.3%)	550
	Positive	91 (41.7%)	298 (48.8%)	61 (35.7%)	450
	Total	218	611	171	1000

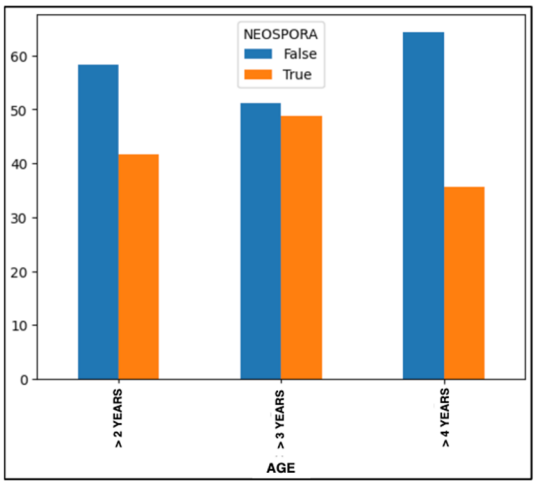

Figure 5. Visualization of the relationship between age and presence of the disease.

2.3. Feature selection

To improve model interpretability and ensure a better understanding of the features influencing the prediction of the presence of neosporosis in cattle, the SHAP (Shapley Additive Explanations) method was used, as described in [18]. SHAP values break down the model predictions into individual feature contributions, facilitating the identification of the most influential factors. In [19], the authors described how SHAP helps to identify important patterns in medical data and complements other selection methods.

Likewise, the study by Yu et al. [20] demonstrated how the strategic combination of machine learning methods with appropriate feature selection can improve the prediction of complex events, such as harmful algal blooms. This approach highlights the importance of identifying optimal combinations of relevant factors to improve the accuracy of predictive models. Inspired by this methodology, they incorporated SHAP to assess and validate the importance of specific features in the veterinary context.

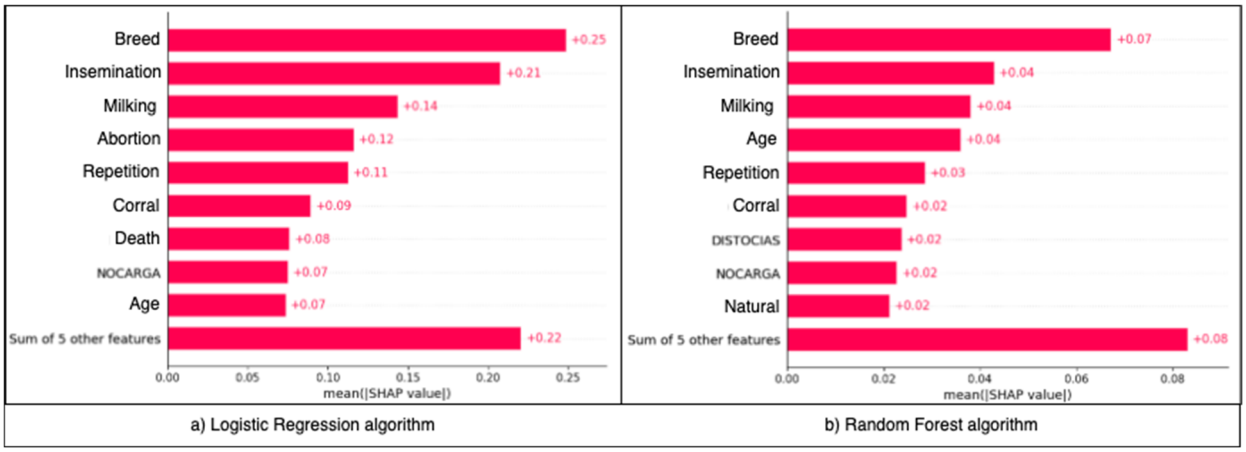

Figure 6. The most important variables using a) logistic regression and b) random forest algorithms.

The combined use of feature selection and explanation techniques not only strengthens the robustness of our model but also provides an interpretable and quantifiable understanding of its performance. Figure 6 presents a SHAP summary chart illustrating the most relevant features for the prediction of bovine neosporosis using logistic regression and random forest algorithms. In both cases, the top three features are breed, insemination, and milking type. The order of characteristics changes according to the algorithm's determined importance. These results indicate that the proposed features are important in predicting the disease's presence. Additionally, the SHAP values indicate that each variable positively contributes to the prediction.

In addition, SHAP values help identify key variables that influence each prediction, allowing veterinarians to understand exactly which factors are contributing to the diagnosis of each animal. For example, if the model predicts a high risk of neosporosis for a particular bovine, the SHAP chart can show that breed and milking type have a significant impact on the estimated probability. This allows veterinarians to justify and communicate specific decisions, such as implementing changes in animal management or intensifying preventative measures in more susceptible breeds.

2.4. Transformation of the variables

Categorical transformation, normalization, and encoding techniques were applied to ensure data comparability and to keep values within an appropriate range. For disease risk factors, which are categorical variables (such as age and breed), the one-hot encoding process was applied [15,21]. This process involves converting each category of a variable into a binary vector where each position represents a distinct category. For example, regarding the age variable with the categories 2 years, 3 years, and 4 years, one-hot encoding produces the results shown in Table 5.

Table 5. Example of age variable coding.

Observation	Age	2 years	3 years	4 years
1	2 years and 4 months	1	0	0
2	3 years and 8 months	0	1	0
3	4 years and 2 months	0	0	1

For the clinical laboratory data, the MinMaxScaler transformation [15] was applied to convert numerical variables into a specific range between 0 and 1. This process ensures comparability across variables while preserving the data's shape and distribution, enabling the model to adequately capture relationships and patterns in the data. By combining the results from the previous steps, a dataset is obtained, which integrates different variables related to the diseases. It is important to note that this process differs from what is done in other works, such as [4], as it incorporates multiple data types.

3. Predictive model generation

To generate the predictive model, it is necessary to analyze various machine learning algorithms and select the one that performs best according to evaluation metrics [22]. The dataset obtained in the previous section is used to select the algorithm. From the algorithm selection, the model is defined and connected to the data to create the predictive model for the disease. These steps and model creation are described below.

3.1. Splitting the data

Splitting the dataset is a standard operation in machine learning. At the very least, data is divided into two parts: a larger portion for training, and a smaller part for evaluating the trained model for validation. Typically, the division of data into these sets is done in 80/20 or 70/30 ratios [15]. The process of dividing the data into these two sets is generally done randomly after choosing the desired percentage for each set.

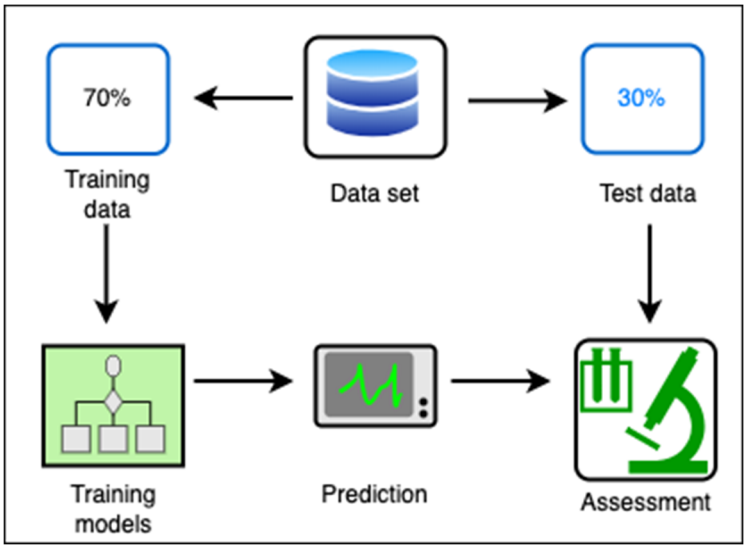

Figure 7. Splitting the dataset into training and testing.

In this study, the dataset consists of 1000 records. As shown in Figure 7, 70% was randomly selected to train the model (training data) and 30% was used to evaluate it (test data). The two datasets were also utilized in the process of algorithm selection and evaluation, including random forest, logistic regression, support vector machine, and artificial neural networks.

3.2. Comparison of machine learning algorithms

To select the machine learning algorithm that best fits the data, a comparative evaluation of the metric results is performed for random forests, logistic regression, support vector machines, and artificial neural networks, considering that they all perform well in the classification task. The confusion matrix is used to evaluate the algorithms [15], allowing for the construction of other evaluation metrics such as sensitivity, accuracy, precision, and F1-score. These metrics are used to determine if the results are similar or if there is a model that offers a better solution.

The process begins by defining the hyperparameters for each algorithm, along with their respective values. This process involves testing all possible parameter combinations, utilizing the GridSearchCV scikit-learn library [15] developed in Python. Each set of parameters is tested in 30 iterations. Table 6 presents the hyperparameters considered for each algorithm, as well as the best parameter combination for each algorithm using the training data. Although a set of possible values is provided to the algorithm, the solution might include other values discovered by the algorithm itself, as seen with the number of neurons in the artificial neural network algorithm.

Table 6. Hyperparameters for each algorithm.

Algorithm	Possibles parameters	Best selection hyperparameters
Random forest	Number of estimators: 10, 100, 1000 max_features: sqrt, log2	Number of estimators: 100 max_features: sqrt
Logistic regression	C: 100, 10, 1, 0.1, 0.01 Regularization: L1, L2 Solver: newton-cg, lbfgs, liblinear	C: 0.1 Regularization: L1 Solver: newton-cg
Support vector machine	C: 50, 10, 1, 0.1, 0.01 Regularization: L1, L2 Kernel: poly, rbf, sigmoid Gamma: scale	C: 0.1 Regularization: L1 Kernel: sigmoid Gamma: scale
Artificial neural network	Number of layers: 2–12 Number of neurons: 2, 4, 8, 16, 32, 64, 128, 256 Type of learning: 0.0001, 0.001, 0.01 Optimizing algorithm: Adam, stochastic gradient descent, adadelta Lot size: 1, 25, 50 Number of epochs: 20, 40, 100, 200, 300	Number of layers: 6 Number of neurons: 128, 64, 32, 16, 8, 4, 2 Type of learning: 0.001 Optimizing algorithm: Adam Lot size: 1, 25, 50 Number of epochs: 100

After selecting the hyperparameters, a model is created for each algorithm. Each model is trained with the training dataset and validated with the validation dataset. For each experiment, a confusion matrix is obtained to determine the performance of each machine learning algorithm. The results of the four metrics used to evaluate each algorithm based on the confusion matrix are shown in Table 7.

Table 7. Metric results on the validation dataset.

Algorithm	Accuracy	Precision	Sensitivity	F1
Logistic regression	0.44	0.37	0.40	0.38
Random forest	0.54	0.42	0.39	0.40
Support vector machine	0.49	0.61	0.43	0.54
Artificial neural network	0.82	0.88	0.94	0.91

Considering the results, the best algorithm for the given data is artificial neural networks. This algorithm outperformed all reference models, achieving an accuracy of 82% for bovine neosporosis. Its outstanding sensitivity (94%) demonstrates an exceptional ability to identify positive cases. The high F1-score (91%) confirms its balance between accuracy and sensitivity, even in the presence of imbalanced data.

The ANN efficiently handles nonlinear relationships, capturing interactions between risk factors and clinical outcomes. The deep structure (four hidden layers) and optimization with Adam allowed for adequate model tuning. The use of early stopping and cross-validation minimized overfitting. The ANN took full advantage of the 24 features of the dataset, while models such as logistic regression

and SVM are more limited in this aspect.

The accuracy metric confirms the predictive ability of the model, while the sensitivity metric results measure the model's tendency to predict positive values more accurately than negative ones.

3.3. Artificial neural network structure

The results shown in Table 6 suggest that a six-layer network with a variable number of neurons, which can be 2, 4, 8, 16, 32, 64, or 128, would be optimal. In addition, the use of Adam as an optimizing algorithm is recommended due to its ability to adapt to different learning rates for each parameter.

Considering these results, it is proposed that the neural network is of the feed forward type and that it has a wide and deep structure as proposed in [4]. The feed forward design is chosen to process the 24 input features, of which 5 derive from the clinical laboratory data and the remaining 19 from the risk factors. The wide component is responsible for memorizing the interactions of the considered features. The deep component is a neural network made up of six dense layers (hidden layers of the neural network), which use a varied number of neurons and different activation functions to process and transmit information through the network. The input layer integrates all 24 features, followed by four hidden layers, each with 32, 16, 8, and 4 neurons, respectively, a configuration determined from the results obtained from the hyperparameters and multiple tests. Figure 8 shows the internal structure of the proposed network.

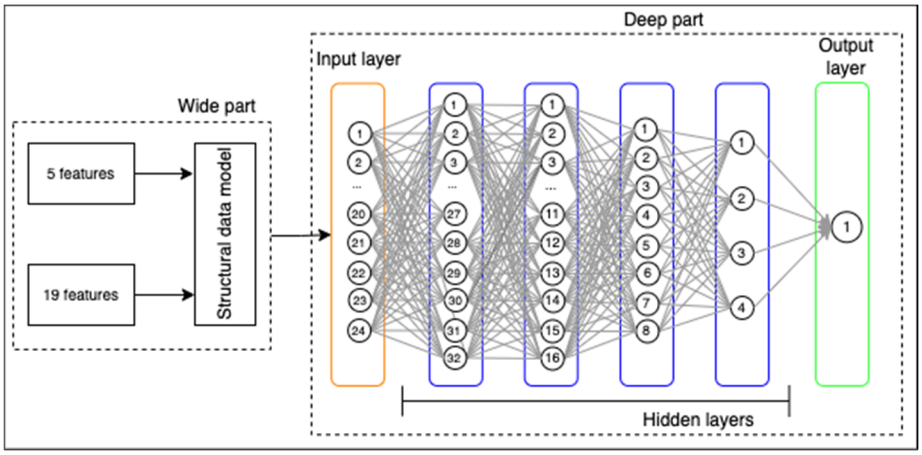

Figure 8. Internal structure of the feed forward neural network.

The implementation of this network is performed using a sequential model of dense layers, where each neuron in a layer is connected to all neurons in the next layer. In the input layer, the tanh (hyperbolic tangent) activation function is used to normalize the input values, facilitating the initial learning of the network. The four hidden layers use the ReLU (rectified lineal unit) activation due to its simplicity and impact on model performance. In the output layer, the activation function is sigmoid (for binary classification problems, as the target variable result is true or false). The architecture of the network is implemented using Keras, a Python library [23] that facilitates the construction and training of neural network models.

To compile the neural network, the Adam optimizer is used, enabling the network to continue learning. The learning rate (lr) is the factor by which the weights of the connections between neurons

are adjusted during the training process. Setting this value to lr = 0.001 avoids manual adjustment. The binary cross entropy function [15] is used as the loss function, representing the difference between expected and predicted outcomes. Accuracy is used as the performance metric.

The next step is training the model using the fit function in Keras. This function takes parameters such as the training dataset, the percentage of data used as the validation set, the number of epochs (training iterations over the training dataset), and the batch size. To achieve good performance of the neural network, it is important to avoid overfitting. The early stopping method is used to detect overfitting. It identifies the epoch at which the network begins to overfit and stops training when this occurs [15]. In this case, the early stopping is configured with the following parameters: min_delta (0.0001); number of epochs with no improvement after which training will be stopped (10).

The accuracy obtained with this model for bovine neosporosis is 94% with 95% confidence, and the precision of the model varies between 92.67% and 95.33%. In the proposed predictive model, the early stopping function halted the training at epoch 44. Figure 9 shows the loss function curve of both the training set and the test set: it decreases and converges at various points in the graph.

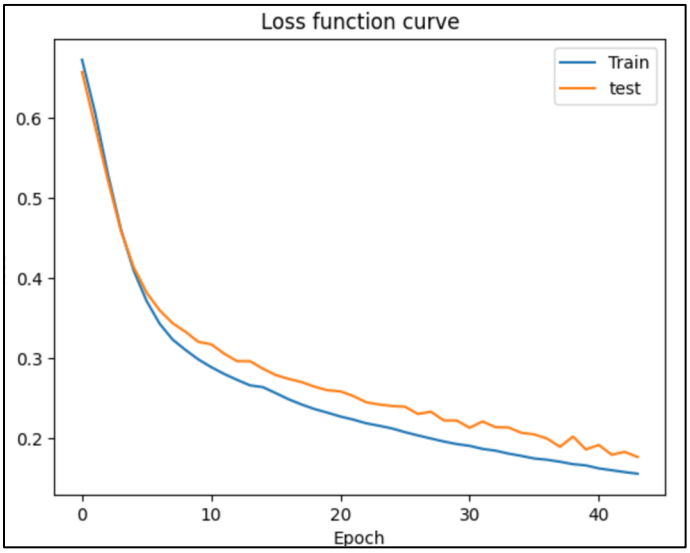

Figure 9. Loss function curve of the feed forward neural network.

3.4. Model validation

Once the neural network training is completed, the model is validated using the test set (X_test), which has not been seen by the network during training. This set is loaded into the model.predict() function, generating predictions that are compared with the real values to measure the model's performance. To quantify the accuracy of the predictions, metrics such as mean absolute error and the mean squared error are used, which evaluate the discrepancy between the predicted and observed values. The model achieved a precision of 97.36% and an accuracy of 96.12%, indicating a high level of reliability in predicting the presence of neosporosis in cattle.

The validation process was conducted with a new dataset consisting of 460 cattle from farms located in the municipality of Toca. The animal dataset was selected with the authorization of the cattle farms during the years 2020 and 2021, by allowing access to their farms and taking serum samples to

process them in the clinical laboratory using ELISA kits.

The steps outlined in the data processing section (conducting surveys with cattle ranchers or farm managers to gather information about the cattle and processing laboratory test results) were followed. After processing the data, the dataset was prepared for use in the predictive model, and predictions were obtained from the trained model.

To validate the results, accuracy, precision, sensitivity, and F1-score metrics were used. From the confusion matrix (Figure 10) the obtained values were: precision of 97.86%, accuracy of 97%, sensitivity of 96.21%, and F1-score of 98.9%. These results indicate that the model is capable of adequately predicting positive cases found on the farms.

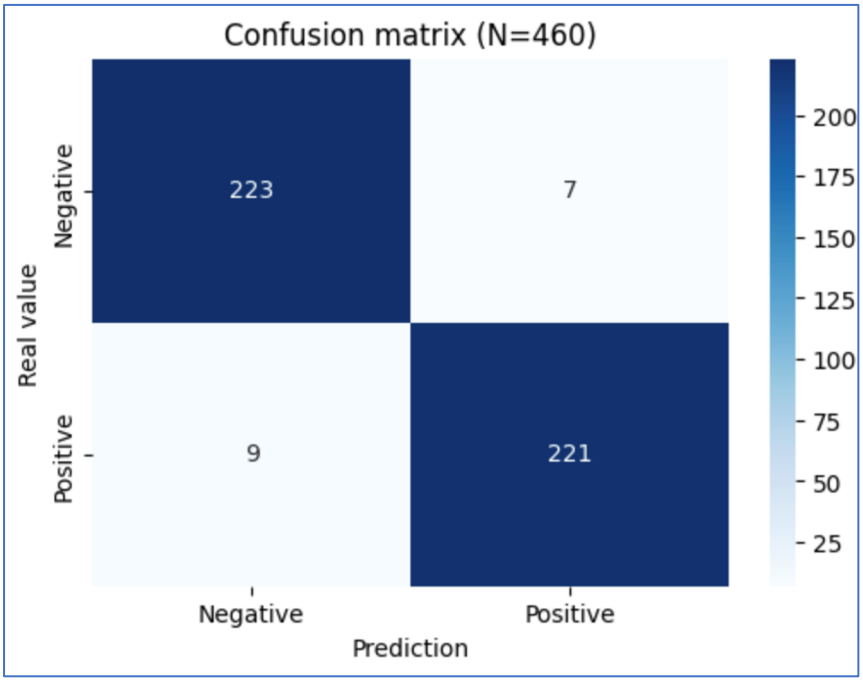

Figure 10. Confusion matrix.

On the new dataset, the model identified 262 positive cases of the disease. These prediction results were compared with the results of clinical laboratory test outcomes and farm diagnoses, information that was available to the veterinarians who took the serum samples from the cattle. The model showed a 10% increase in positive cases compared to the cases recorded by the veterinarians.

Although this study demonstrates the potential of the model using available datasets, the importance of implementing cross-validation techniques and incorporating independent validation sets in future work is recognized. These practices, widely recommended in the literature [20], would further improve the robustness and generalization of the model, especially in different contexts and livestock populations.

3.5. Implementation

The process of implementing this neural network prediction model aims to use the model trained with new datasets that meet the established criteria. This avoids the need to train the model every time.

Once the neural network prediction model has been validated, it needs to be saved before being used to preserve its current state and avoid having to adjust the parameters repeatedly.

4. Conclusions

This article presents the construction of a predictive model for the bovine neosporosis infectious disease. Data obtained in the clinical laboratory, alongside data on risk factors and contextual information gathered through surveys related to bovine infectious diseases, was used. With guidance from veterinary doctors of the GIDIMEVETZ research group, we validated the necessary parameters.

We used and compared various machine learning algorithms to determine the most suitable one. After training, validating, and executing the predictive model and comparing the results of the metrics obtained by different algorithms, we found that the proposed neural network yielded the best results in predicting disease onset. The artificial neural network-based model demonstrated significant superiority over the reference models (logistic regression, SVM, and random forest) in all aspects evaluated. This highlights the importance of using advanced machine learning architectures for complex problems, especially when dealing with datasets with multiple interrelated and nonlinear factors.

Considering the literature review on the use of machine learning to predict bovine infectious diseases, the predictive model was developed. Based on the information generated to support the decision-making of veterinarians, it can be stated that the impact of this model is positive for the dairy ecosystem of the Department of Boyacá in Colombia, as for the information generated to aid veterinary decision-making.

It is important to highlight the model's limitations in predicting the two bovine infectious diseases within the dairy ecosystem of the Boyacá department. These limitations are related to the lack of data availability, the scarcity of information on the presence of diseases, the difficulty and cost of collecting data on farms, and the lack of access to data from organizations. Despite these limitations, the model performs well with both the training and the validation datasets.

This study presents a model to predict the presence of neosporosis in cattle, with an approach that could be scalable to other regions by incorporating specific data from different geographic and environmental contexts. A future research direction is to extend the scope of the model to include multiple infectious diseases, such as BVD or leptospirosis, by integrating additional data and developing machine learning architectures capable of addressing multi-objective problems. This expansion would not only broaden the applicability of the model but would also contribute to a better understanding of the interactions between shared and specific risk factors.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research is financed by the Boyacá Government through the scholarship obtained in the call number 733 of the Ministry of science technology and innovation (MinCiencias) aimed to the Formation of High-Performance Human Capital. The data, the sampling and processing of the samples in the laboratory is due to the work and advice of the GIDIMEVETZ Research Group of the UPTC.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. J. A. Ballesteros-Ricaurte, R. Fabregat, A. Carrillo-Ramos, C. Parra, M. O. Pulido-Medellín, Systematic literature review of models used in the epidemiological analysis of bovine infectious diseases, *Electronics*, **11** (2022), 2463. https://doi.org/10.3390/electronics11152463
2. G. M. Figueredo, *Serological Survey of Bovine Infectious Causes of Reproductive Disordes in Colombia*, Ph.D thesis, University of Parma, 2011.
3. M. O. Pulido-Medellín, D. García-Corredor, R. Andrade-Becerra, Seroprevalencia de Sarcocystis spp. en un hato lechero del municipio de Toca, Colombia, *Rev. Salud Anim.*, **35** (2013), 159–163.
4. B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, T. Do, et al., Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records, *Comput. Methods Programs Biomed.*, **182** (2019), 105055. https://doi.org/10.1016/j.cmpb.2019.105055
5. J. A. Cruz, D. S. Wishart, Applications of machine learning in cancer prediction and prognosis, *Cancer Inform.*, **2** (2006), 59–78. https://doi.org/10.1177/117693510600200030
6. R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: An unsupervised representation to predict the future of patients from the electronic health records, *Sci. Rep.*, **6** (2016), 26094. https://doi.org/10.1038/srep26094
7. T. Bhardwaj, P. Somvanshi, Machine learning toward infectious disease treatment, *Adv. Intell. Syst. Comput.*, **748** (2019), 683–693. https://doi.org/10.1007/978-981-13-0923-6_58
8. P. Sajda, Machine learning for detection and diagnosis of disease, *Annu. Rev. Biomed. Eng.*, **8** (2006), 537–565. https://doi.org/10.1146/annurev.bioeng.8.061505.095802
9 S. Oh, J. Cha, M. Ji, H. Kang, S. Kim, E. Heo, et al., Architecture design of healthcare software-as-a-service platform for cloud-based clinical decision support service, *Healthcare Inf. Res.*, **21** (2015), 102–110. https://doi.org/10.4258/hir.2015.21.2.102
10. N. Ujjwal, A. Singh, A. K. Jain, R. G. Tiwari, Exploiting machine learning for lumpy skin disease occurrence detection, in *10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, (2022), 1–6. https://doi.org/10.1109/ICRITO56286.2022.9964656
11. N. K. Krishna, R. M. Benjamin, N. Tejaswi, Machine learning diagnosis of lumpy skin disease in cattle herds, in *Proceedings of the 3rd International Conference on Applied Artificial Intelligence and Computing*, (2024), 426–431. https://doi.org/10.1109/ICAAIC60222.2024.10575895
12. G. Villa-Cox, H. van Loo, S. Speelman, S. Ribbens, J. Hooyberghs, B. Pardon, et al., Machine learning modeling to predict causes of infectious abortions and perinatal mortalities in cattle, *Theriogenology*, **226** (2024), 20–28. https://doi.org/10.1016/j.theriogenology.2024.05.041
13. A. A. Megahed, R. Bommineni, M. Short, K. N. Galvão, J. H. J. Bittar, Using supervised machine learning algorithms to predict bovine leukemia virus seropositivity in dairy cattle in Florida: A 10-year retrospective study, *Preventative Vet. Med.*, **235** (2025), 106387. https://doi.org/10.1016/j.prevetmed.2024.106387
14. S. Chae, S. Kwon, D. Lee, Predicting infectious disease using deep learning and big data, *Int. J. Environ. Res. Public Health*, **15** (2018), 1–19. https://doi.org/10.3390/ijerph15081596

15. S. Raschka, V. Mirjalili, *Python Machine Learning*, Second Edition, Packt Publishing Ltd, 2017.
16. M. O. Pulido-Medellín, D. J. García-Corredor, J. C. Vargas-Abella, Seroprevalencia de Neospora caninum en un Hato Lechero de Boyacá, Colombia, *Rev. Inv. Vet. Perú*, **27** (2016), 355–362. https://doi.org/10.15381/rivep.v27i2.11658
17. M. O. P. Medellín, A. D. Anaya, R. A. Becerra, Asociación entre variables reproductivas y anticuerpos anti Neospora caninum en bovinos lecheros de un municipio de Colombia, *Rev Mex. Cienc. Pecu.*, **8** (2017), 167–174. https://doi.org/10.22319/rmcp.v8i2.4439
18. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, (2017), 4768–4777.
19. S. Ahmed, M. S. Kaiser, M. S. Hossain, K. Andersson, A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions, *IEEE Access*, **13** (2024), 37370–37388. https://doi.org/10.1109/ACCESS.2024.3422319
20. P. Yu, R. Gao, D. Zhang, Z. P. Liu, Predicting coastal algal blooms with environmental factors by machine learning methods, *Ecol. Indic.*, **123** (2021), 107334. https://doi.org/10.1016/j.ecolind.2020.107334
21. W. Zhang, T. Du, J. Wang, Deep learning over multi-field categorical data, in *Advances in Information Retrieval (ECIR)*, (2016), 45–57. https://doi.org/10.1007/978-3-319-30671-1_4
22. B. Ivorra, D. Ngom, Á. M. Ramos, Be-CoDiS: A mathematical model to predict the risk of human diseases spread between countries—validation and application to the 2014–2015 Ebola virus disease epidemic, *Bull. Math. Biol.*, **77** (2015), 1668–1704. https://doi.org/10.1007/s11538-015-0100-x
23. P. Nickerson, P. Tighe, B. Shickel, P. Rashidi, Deep neural network architectures for forecasting analgesic response, in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, (2016), 2966–2969. https://doi.org/10.1109/EMBC.2016.7591352

AIMS Press