Original Research

Journal of Advanced Veterinary Research
(2024) Volume 14, Issue 6, 975-981

Machine learning algorithms for clinical mastitis prediction in a dairy herd use automated milking system data

Dina N. Faris[1*], Ahmed M. Gad[2], Mahmoud S. El-Tarabany[3], Sherif I. Ramadan[4], Ghada G. Afifi[5], Eman A. Manaa[4]

[1]*Department of Animal Wealth Development (Biostatistics), Faculty of Veterinary Medicine, Benha University, Moshtohor, Toukh, 13736, Qalyubia, Egypt.*
[2]*Statistics Department, Faculty of Economics and Political Science, Cairo University, Giza, Egypt.*
[3]*Animal Wealth Development Department, Faculty of Veterinary Medicine, Zagazig University, El-Zeraa str. 114, Sharkia, Zagazig 44511, Egypt.*
[4]*Animal and poultry production, Department of Animal Wealth Development, Faculty of Veterinary Medicine, Benha University, Moshtohor, Toukh, 13736, Qalyubia, Egypt.*
[5]*Genetic Engineering and Biotechnology Research Institute, University of Sadat City, El-Monofya, Egypt.*

ARTICLE INFO

Recieved: 01 April 2024

Accepted: 22 May 2024

***Correspondence:**

Corresponding author: Dina N. Faris
E-mail address: dina.abdallah@fvtm.bu.edu.eg

Keywords:

Automatic milking system, Clinical Mastitis, Decision tree, Machine learning.

ABSTRACT

Bovine clinical Mastitis (CM) is the most important disease in the dairy industry affecting both animal welfare and farm profitability. Therefore, early and accurate detection of the disease is a valuable timely intervention. In this article, six different machine learning classification algorithms were compared to obtain a prediction model for early detection of the disease. These algorithms are Support Vector Machine, Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbor, Classification and Regression Decision Tree, and Random Forest. The algorithms are applied to the milk production of Holstein Friesian cows milked by an automated milking system using the dairy records and disease events. This includes 1493 cows with clinical Mastitis and 2387 healthy cows. The six models were evaluated based on five performance metrics criteria: accuracy, precision, recall, F1-score, and area under the curve (AUC). The accuracy rate ranged from 62% to 74%. The AUC is used to choose the best model. The Decision Tree algorithm and Gaussian Naïve Bayes scored the highest AUC of 71%. However, the Decision Tree algorithm is more stable with respect to other metrics (73% for accuracy and 64% for Precision, Recall, and F1-score). Hence, it can be considered the best predictive CM model with moderate accuracy. Out of the 15 input features, days in milk, age of the animal, lactation order, 305 days mature herd equivalent, and average daily milk yield were the only important features shared in establishing the Decision Tree model.

Introduction

Clinical Mastitis (CM) is the most frequent production disorder of dairy farming. It has a negative impact on both animal health and production efficiency (Ruegg, 2017). Annually, about 20–40% of all the herd lactating cows suffer one or more cases of CM (Hogeveen *et al*., 2019), which results in an average failure cost of $147 per cow per year, particularly contributed by milk production losses and culling, which represents 11% to 18% of the gross margin per cow per year (Cheng and Han, 2020). In addition to lower milk yield, the prevalence of CM results in a financial burden to the farmers. This is due to the fact that each CM case involves veterinary expenses, therapeutic costs, labor costs, non-saleable milk discard, premature culling loss, future reproductive problems, replacement costs, and death (Ghafoor and Sitkowska, 2021). For all these reasons, it becomes crucial to predict CM very early.

Technological innovation in the last decades has impacted numerous facets of the modern dairy industry. This encourages many farms to implement an automatic milking system (AMS), despite its higher economic initial cost (Bausewein *et al*., 2022). Reducing the workload and providing a more flexible work schedule (Vik *et al*., 2019) are not the only benefits of the AMS, but also a vast quantity of cow-level dairy data become available (King *et al*., 2018). However, till now there is a shortage in data mining and integration which implies that these data are not being utilized to their full potential. As a result, multiple dairy farming issues such as poor longevity, low performance, and health problems remain uncontrolled perfectly (Cockburn, 2020).

Advanced data analysis techniques such as machine learning (ML) methods may offer new advances in precision livestock management, involving critical disease detection and prediction, production management, and farming decision-making processes (Hossain *et al*., 2022). ML is a subfield of computer science that gives computers the ability to "learn" without being explicitly programmed. Generally, ML is suitable for handling large and high-dimensional datasets and prioritizes predictive accuracy over hypothesis-driven inference (Bi *et al*., 2019).

The early detection of cows with a high risk of health problems, such as lameness, clinical and subclinical mastitis, ketosis, and metritis, is crucial for dairy farms. It enhances and prevents the negative impacts of these disorders early (Zhou *et al*., 2022). The ML methods are progressively finding their way into the dairy industry in this regard. For example, Dhoble *et al*. (2019) combined ML and Cytometric fingerprinting for the early prediction of Bovine Mastitis through the evaluation of the microbiological milk quality. Also, the ML was applied for the detection of claw lesions in dairy farms (Volkmann *et al*., 2021), and for the prediction of the calving time in dairy cows using the behavioral and activity sensors data, a recurrent neural network ML algorithm was used (Keceli *et al*., 2020).

Unlike traditional statistical methods, ML models can analyze categorical data accurately and are insensitive to missing data (Fatima and Pasha, 2017). Also, they can deal with the complex, nonlinearity, and outliers problems of the data (Dong *et al*., 2022). However, the ML proves great potential for precision livestock farming, particularly in the domain of early disease prediction (Gokul Krishna*et al*., 2023). There are few literature use ML for CM prediction in dairy cows. At the same time, there are no studies that used ML for CM prediction in Egypt in dairy cows.

Therefore, the objective of this article is twofold. First, to establish six different supervised ML algorithms for classification and prediction of CM onset in Friesian female dairy cattle using automated milking system data. These methods are the Support Vector Machine (SVM), the logistic regression (LR), the Gaussian Naïve Bayes (NB), the K-nearest neighbor (KNN), the Classification and Regression Decision Tree (CART-DT), and the Random forest (RF). The second is to compare the accuracy, precision, F1-Score, Recall, and Area under the ROC curve (AUC) of the six models aiming to select the optimal model.

Materials and methods

In this article, we tried to build an ML predictive model of cow CM us-

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
ISSN: 2090-6277/2090-6269/ © 2011-2024 Journal of Advanced Veterinary Research. All rights reserved.

ing available data from an AMS database. A retrospective cross-sectional survey was conducted to collect a random sample of automatic-made records directly from a dairy farm. However, the raw data might not be sufficient to yield results immediately. Firstly, some data preparation is necessary before using the predictive model.

Ethical approval

The current work was approved by the Committee of Animal Care and Welfare, Benha University, Faculty of Veterinary Medicine, Egypt (BUFVTM:17-04-23).

Data collection

The original data was collected from a private commercial Egyptian dairy farm located at 80th Km of Cairo Alexandria desert road during the period extending from July 2016 to November 2019. A total of 3880 dairy records of Holstein Friesian cows (containing the dairy and some health information of each cow during its last production season) were selected randomly from different five units on the farm. Cows were housed in an open yard shaded with free stalls, the floor was lined with sand, and equipped with a cool spraying system in the summer thus relieving heat stress in the summer months. A total mixed ration method with computerized calculating systems was used that controls feeding portions according to the reproductive and productive demands of animals. All day long water was supplied freely to animals. Pre-milking and post-milking udder hygiene measures were practiced by dipping the teats in an iodine solution. The milking process was performed in a herringbone with a rapid exit automatic milking parlor three times a day using machine milking, and milk parameters for each cow were recorded in a computerized database. Detection of CM depends on the presence of clinical signs on the udder such as hotness, redness, swelling, painful reaction, and hardness of udder tissues, and then the infection is confirmed by the California Mastitis Test (CMT). 15 features were selected to study their impact on the prediction model of CM including the reproductive status (Pregnant, not pregnant, and bred), lactation order, age at the last season (years), average daily milk yield (DMY), total milk this lactation (TOTM), milk peak (MPEAK), 305 days mature herd equivalent (305 ME), days in milk (DIM), days open, calving season (summer extend from 21st March to 20th September and winter extend from 21st September to 20th march), lameness, abortion, metritis, milk fever, and retained placenta onset (all classified into yes and no). The output variable to be predicted was CM which was classified into (yes = Mastitis, and no = healthy).

Data pre-processing

Data pre-processing is an initial step of the ML including cleaning, scaling, transformation, and feature engineering to make the quality of data better for building a predictive model of optimal classification performance (Iliou *et al*., 2015).

The first pre-processing step was the data visualization which revealed no missing values in the data. The CM prevalence was 38.5%. Table 1 presents the mean and standard deviation (SD) of the independent numerical features divided according to CM positive or negative. Figure 1 shows the distribution of the independent categorical feature frequencies versus the CM.

The second step was identifying the independent and dependent features and labeling the categorical features (reproductive status, calving season, lameness, abortion, metritis, milk fever, retained placenta, and mastitis) using the Label-Encoder. Outliers were detected graphically as shown in Figure 2 by using a boxplot and statistically by using the interquartile Range Method (IQR). The IQR is defined as the difference between Q3 (the 75th percentile) and Q1 (the 25th percentile) and any value outside the range of [Q1 − 1.5 × IQR or Q3 + 1.5 × IQR] is considered to be an outlier (Li *et al*., 2021). The outliers have been replaced with percentile for the features; (lactation order, age at the last season, DMY, TOTM, MPEAK, 305ME, and days open). Figure 3 shows the boxplot after the outliers transformation.

Table 1. Summary statistics for the numerical independent features.

Features	Clinical Mastitis			
	Positive		Negative	
	Mean	SD	Mean	SD
Lactation order	2.7	1.6	2	1.4
Age at the last season	4.1	1.87	3.24	1.6
DMY	26.6	11	31.5	10.9
TOTM	8445.8	4548.2	6142.7	4810.4
MPEAK	43.8	9.9	41.16	9
305ME	10673.5	2305.8	11184.1	2060.4
DIM	287.9	156.6	213.5	167.1
Days open	211.8	149.4	161.1	142.7

SD: standard deviation; DMY: average daily milk yield; TOTM: total milk this lactation; MPEAK: milk peak; 305ME: 305 days mature herd equivalent; DIM: days in milk.

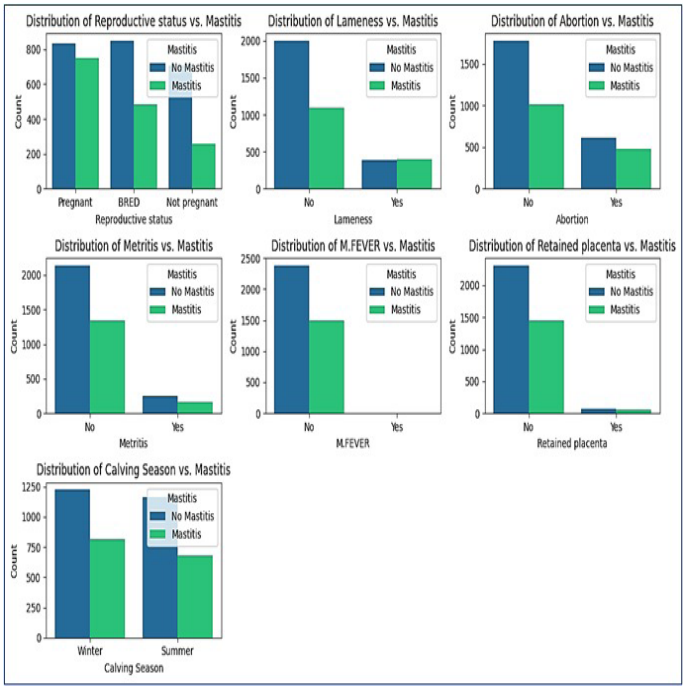

Fig. 1. The distribution of the independent categorical features frequencies versus the CM.

The Z-score standardization method for data scaling was used. The values for an attribute, A, were standardized based on the mean (μ) and the standard deviation (σ_A) of A. Avalue, *vi*, of A was normalized to v_i by computing:

$$\acute{v}_i = \frac{v_i - \mu}{\sigma_A}$$ (Han *et al*, 2012).

For feature extraction, we computed the autoencoders (AE) method. Figure 4 presents the plot of the AE model with no compression. The AE is an unsupervised neural network that consists of two linked parts: the encoder and the decoder. The encoder learns how to interpret the input and compresses the input into a latent representation called (the bottleneck layer), while the decoder takes the output of the encoder and tries to reconstruct the input again from the intermediate code (Ardelean *et al*., 2023).

Finally, data were randomly split into a training set (90% of the sample) for model training and a testing set (10% of the sample) for verification of the prediction performance. Models were prepared using 10-fold

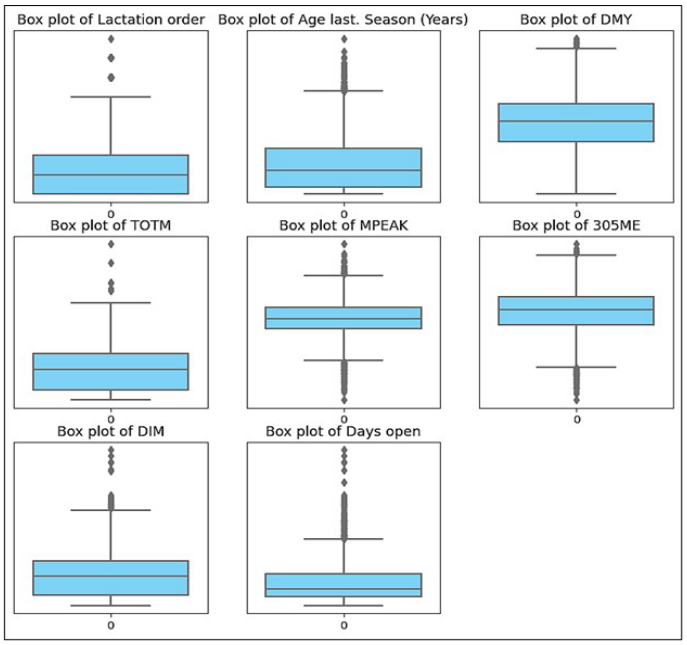

Fig. 2. Shows the boxplot for the numerical features containing outliers.

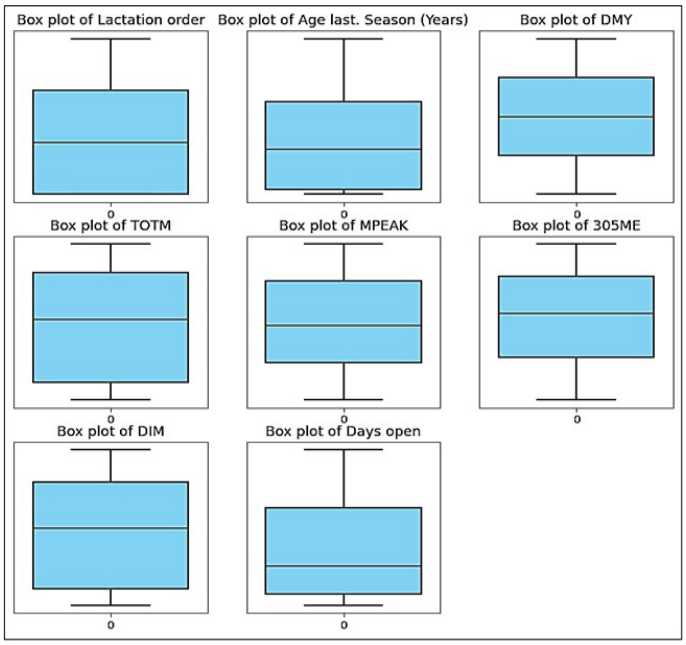

Fig. 3. Shows the boxplot for the numerical features after the outliers transformation.

cross-validation (CV) + a grid search method of hyperparameters optimization. The estimation of model accuracy was based on the average of the 10-fold repeated CV over the total number of the candidate values of each parameter in the grid search as shown in Table 2. The tuned best hyperparameters giving the best validation accuracy were then used to build the predictive algorithm.

ML algorithm

Six classification ML algorithms were selected and implemented to build CM prediction models, which were SVM, LR, Gaussian Naïve Bayes, KNN, CART-DT, and RF.

SVM Method

The general idea of the SVM as shown in Figure 5, was to obtain an optimal hyperplane that linearly separates the d-dimensional data perfectly into its two classes such that the marginal distance was maximum between the hyperplane and the support vectors; the nearest instance of each class to the hyperparameter (Schölkopf, 2003).

The best optimal hyperparameters for establishing the SVM model after fitting 10-folds for each of 54 candidate values of grid search (totaling 540 fits) were: C-parameter of 0.01, kernel function was linear, and a kernel coefficient gamma of 0.1.

Table 2. Grid search parameters and the final optimal parameters for each algorithm.

Algorithm	Grid-search parameters	Final optimal parameters	Validation accuracy (%)
SVM	C: [0.01, 0.05, 0.1, 0.5, 0.8, 1] Kernel: [Linear, RBF, sigmoid] Gamma: [0.1, 1, auto]	C: [0.01] Kernel: [Linear] Gamma: [0.1]	73
LR	Fit intercept: [True, False] penalty: [l1, l2] C: [0.1, 0.2, 0.3, 0.5, 1, 5, 10, 100] Solver: [lbfgs, liblinear, newton-cg, newton-cholesky, sag, saga]	Fit intercept: [True] penalty: [l2] C: [0.1] Solver: [lbfgs]	70
Gaussian NB	var_smoothing: np.logspace(0,-9, num=100)	var_smoothing: 1.0	69
KNN	n_neighbors: [1,2,3,5] weights: [uniform, distance] metric: [Euclidean, Manhattan, Hamming, Jaccard, Cosine]	n_neighbors: [1] weights:[uniform] metric:[Euclidean]	63
DT	criterion: [Gini, Entropy] max_depth: [3, 5, 7, 9, 11] min_samples_split: [2, 3, 5, 10] min_samples_leaf: [1, 2, 4]	criterion: [Gini] max_depth: [3] min_samples_split: [2] min_samples_leaf: [1]	71
RF	n_estimators: [50, 100, 200, 2000] criterion: [Gini, Entropy] max_depth: [3, 5, 7] min_samples_split: [2, 5, 7] min_samples_leaf: [1, 2, 4] max_features: [auto, sqrt, log2]	n_estimators: [50] criterion: [Gini] max_depth: [3] min_samples_split: [2] min_samples_leaf: [1] max_features: [auto]	70

SVM: Support vector machine, LR: Logistic regression, Gaussian NB: Gaussian Naïve Bayes, KNN: K-Nearest Neighbor, DT: Decision tree, RF: Random forest.

The LR

The LR is a classification-supervised ML algorithm developed for predicting a binary outcome for an event based on the previous observations of a data set (Nusinovici *et al*., 2020).

The LR model was given by the linear relation between the logit and the values of the explanatory variables as:

$$\boldsymbol{Log\left(\pi/(1-\pi)\right) = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k,}$$

where π is the outcome probability, Log($\pi/(1-\pi)$) is the log of odds, β_0 is the intercept, β_1 is the regression coefficient of the X_1 independent variable (predictor), and β_k is the regression coefficient for the X_k independent variable (Bender, 2009).

The best optimal hyperparameters for building the LR model after fitting 10-folds for each of 192 candidate values of grid search (totaling 1920 fits) were: C-parameter of 0.1, fit-intercept: True, regularization penalty function l2, and lbfgs solver.

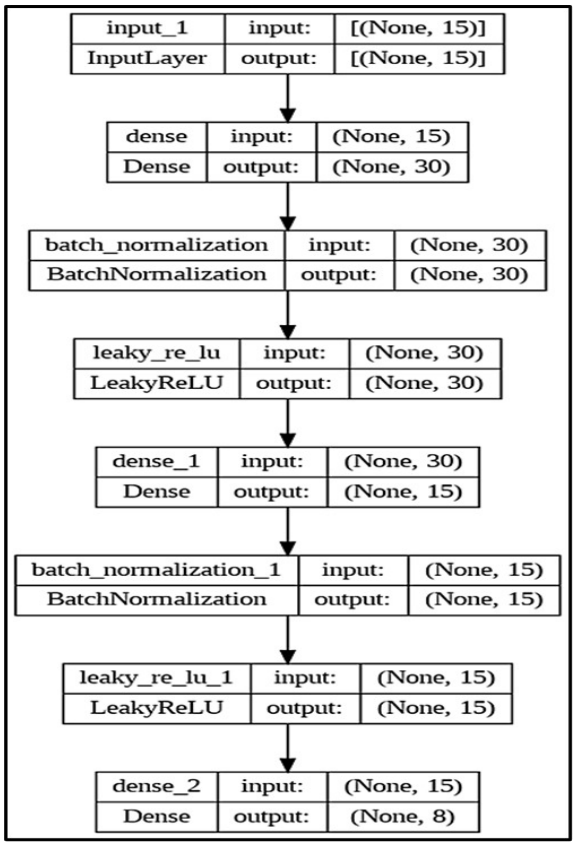

Fig. 4. Shows the AutoEncoder (AE) model plot with no compression for the feature extraction.

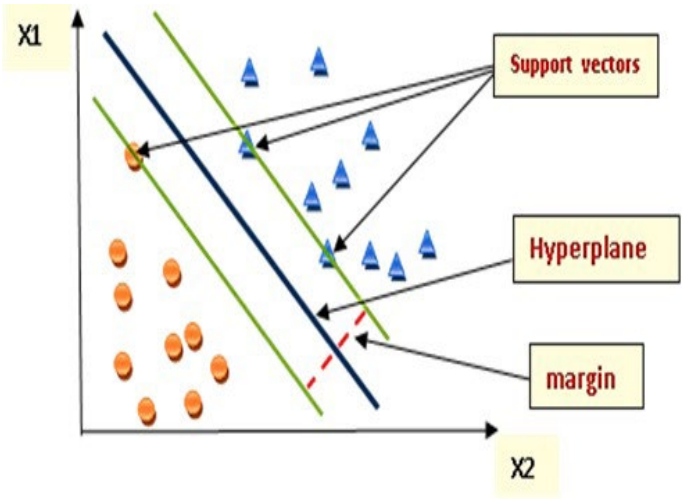

Fig. 5. The general idea of linear support vector machine (SVM).

The Gaussian NB

The NB is a simple but powerful probabilistic supervised ML classifier based on the Bayes theorem that greatly simplifies learning by assuming that features are independent given classes. Although independence is generally a poor assumption, in practice naive Bayes often compete well with more sophisticated classifiers (Lee *et al*., 2001).

Var_smoothing (in this case np. logspace starts from 0, ends at -9, and generates 100 samples) is the only hyperparameter to be tuned for the Gaussian NB algorithm. After fitting 10 folds for each of 100 candidate values of grid search (totaling 1000 fits), Var_smoothimg =1 was the best hyperparameter.

The KNN

One of the most popular supervised ML classification algorithms is the distance-based algorithm. It is based on computing the distances between the new test instance and all training instances, sorting the distances to determine the K-nearest neighbors, and then allocating the test instance to the class that owns the majority of K-nearest neighbors (Ali *et al*., 2019).

The most common distance function is Euclidean distance as

$$\boldsymbol{d_{Euclidean}(x,y) = \sqrt{\sum_{i}^{m}(x_i - y_i)^2}},$$

,

where x = x_1, x_2,..., x_m, and y = y_1, y_2,..., y_m represent the m attribute values of two records (Task, 2014).

For tuning this model, 10-folds of CV for each of 40 candidate values of K (number of neighbors), the distance metric, and the weights (totaling 400 fits) were run and the optimal parameters were k=1, distance metric was the Euclidean distance, and the weights are uniform.

The CART-DT

A DT is a flowchart-like tree in which each internal node refers to a choice between several alternatives, and each leaf node represents an output class (decision). A DT starts with a root node and then continues to split till reaches the final classification or decision at the leaf node (Sonia Singh, 2014).

CART-DT was applied in our study, it constructs binary trees as each internal node has only two splitting edges. The selection of the best attributes depends on the Gini index criteria and the final tree is pre-pruned by the cost–complexity Pruning. It is also characterized by its ability to handle both categorical and numerical features and outliers (Charbuty and Abdulazeez, 2021). The DT optimal hyperparameters were obtained after fitting 10-fold of CV for each of 120 candidate values of grid search (totaling 1200 fits). The splitting criteria were Gini index, the maximum depth of DT was 3, Min_samples_leaf which indicates the minimum samples the leaf node must possess was 1, Min_samples_split which indicates the minimum sample number an internal node must possess before splitting was 2.

The RF

The RF is a grouping of a large number of ensemble DTs in which each tree depends on a random vector value sampled independently and with the same distribution for all trees in the forest (Kullarni and Sinha, 2013). Substantial gains in classification and regression accuracy can be achieved by using ensembles of trees, where each tree in the ensemble is grown by a random parameter called bootstrap aggregating or simply bagging. Final predictions are obtained by either majority voting or averaging, based on results from all decision trees in the forest (Klusowski,

2018).

10-fold CV for each iteration of each 486 grid search value in a total of 4860 fits were run to tune the model and stand on the optimal parameters for the RF. The Gini index splitting criteria was the best, the n_estimator that represents the number of RF trees was 50, the max_depth was 3, Min_samples_leaf was 1, and Min_samples_split was 2.

Validation and performance metrics

The first metric is called a confusion matrix is shown in Table 3. It visualizes the performance of a Supervised ML algorithm for binary data classification. Each column of the matrix represents the number of predictions in each class, while each row represents the instances in the real class. It enables us to calculate the total number of false negatives (Incorrectly identified negative class (TFN), false positives (Incorrectly identified positive class (TFP), true negatives (Correctly identified negative class (TTN), and the total true positives (Correctly identified positive class (TTP) for each class (Mathkunti and Rangaswamy, 2020).

The confusion matrix is then used for calculating the 5 used performance metrics for comparing the ML models:

Accuracy = ((TTP+TTN))/((TTP+TTN+TFP+TFN))

Recall= TFP/((TFP+TTN))

Precision = TTP/((TTP+TFP))

F1-score = 2 × (Percision × Recall)/(Percision + Recall)

AUC represents the degree or measure of separability that is how much the model is capable of differentiating between the classes.

AUC= 1/2 .(Sensitivity+Specificity) (Itoo *et al*., 2021).

In this article, we used the online Google Colab framework, an excellent, easy, and fast environment for Python 3 coders to implement ML algorithms.

Table 3. Confusion Matrix.

Actual classes	Predicted classes	
	Negative	Positive
Negative	TTN	TFP
Positive	TFN	TTP

TTN: Total true negatives; TFP: Total false positives; TFN: Total false negatives; TTP: Total true positives.

Results

Table 4 presents the performance metrics comparison for the six established ML models in the testing phase. The SVM algorithm gained the highest accuracy (74%), followed by LR and DT, which both displayed a 73% accuracy rate, RF (71% accuracy), and Gaussian NB (70% accuracy). However, out of all the models, the KNN model had the lowest accuracy (62%).

Table 4. The models' performance metrics of the test dataset.

Algorithms	Accuracy	Precision	Recall	F1 score	AUC
SVM	0.74	0.7	0.54	0.61	0.70
LR	0.73	0.69	0.55	0.61	0.70
Gaussian NB	0.7	0.58	0.76	0.66	0.71
KNN	0.62	0.5	0.53	0.51	0.6
DT	0.73	0.64	0.64	0.64	0.71
RF	0.71	0.68	0.47	0.56	0.67

AUC: Area under the ROC curve; SVM: Support vector machine; LR: Logistic regression, Gaussian NB: Gaussian Naïve Bayes; KNN: K-Nearest Neighbor; DT: Decision tree; RF: Random forest.

According to the precision results, SVM and LR showed relatively the same highest rates (70% and 69% respectively) followed by RF and DT (68% and 64% respectively), while NB and KNN came last with rates of 58% and 50%, respectively. Among all metrics, the recall showed the worst results (ranging from 47% to 55%) in the majority of algorithms. Except in NB, it scored 76% and 64% in DT. The results of the F1-score performance criterion fluctuated within a small range from 66% to 51% in all models. Finally, Gaussian NB and DT showed the highest AUC = 71%. From all the above, only the DT algorithm showed balanced results of all the performance metrics indicating that the results of DT were more reliable and accurate so it has been voted as the best ML algorithm for CM prediction. Figure 6 presents the AUC reported in each ROC curve for the six algorithms.

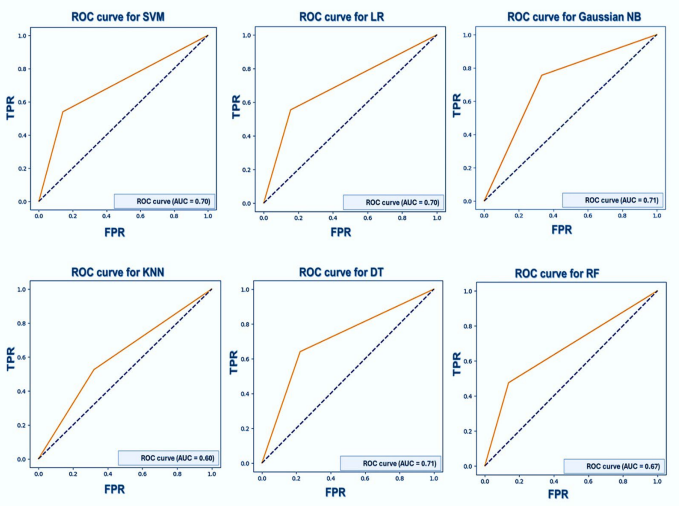

Fig. 6. Comparing Receiver Operating Characteristic (ROC) curves of six ML methods [Support Vector Machine (SVM), Logistic Regression (LR), Gaussian Naïve Bayes (NB), K-Nearest Neighbors (kNN), Decision Tree (DT), and Random Forest (RF)] run on the test set, in the prediction of CM of Holstein Friesian dairy cows. In each plot, the area under the curve (AUC) was reported.

Hence DT was the best model, we plotted the feature importance shown in Figure 7, which is determined by how much each feature contributes to reducing the uncertainty in the target variable. This is typically measured by the amount of reduction in the Gini impurity that is achieved by splitting on a particular feature. It appeared that the DIM was the most effective feature at reducing uncertainty in the target variable by about 47%, followed by the Age at last season sharing by 30%, 17% by the lactation order, and 3% for both 305 ME and DMY features so that they are considered the most important features by the DT model.

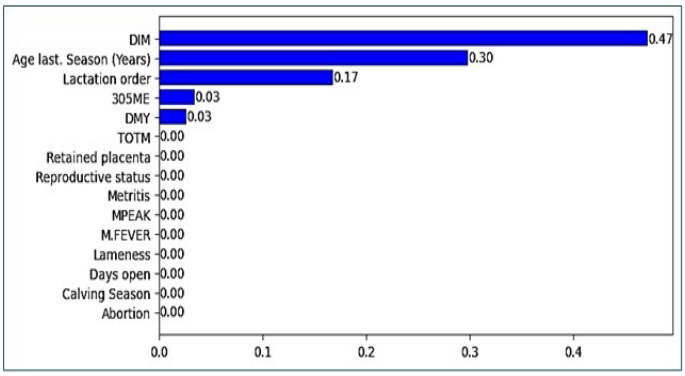

Fig. 7. Features importance plot of the Decision tree algorithm.

Figure 8 presents the DT graph which in general consists of 7 internal nodes (the first starting root node and 6 other child nodes) and 8 leaf nodes. Each box in the tree represents a node and it consists of the value of the split feature, the value of gini impurity before the split, the number of samples before the split, the values of classes after samples split, and the majority class. For example, at the root node if DIM is equal to or less than a standardized value of 0.393 this means that out of 3276 samples, 2029 samples would fall in the mastitis category and the other samples would continue to split regarding other features, and so on till reaching one of the final decisions (Mastitis or Healthy) at the leaf nodes.

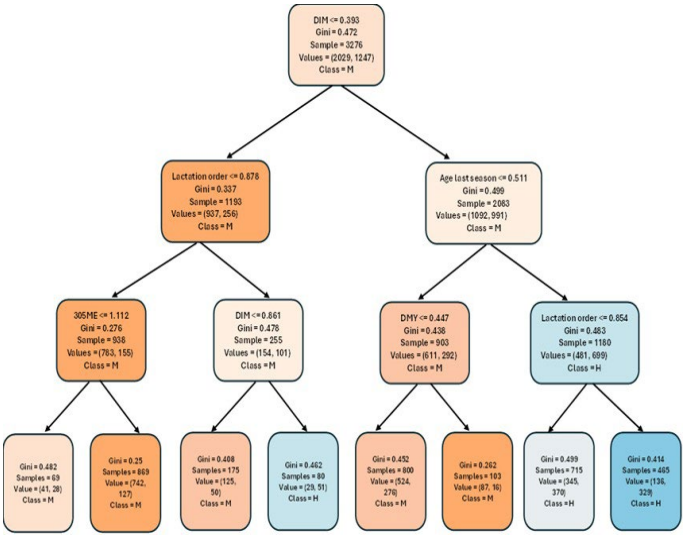

Fig. 8. The Decision tree graph.

Discussion

This article, up to our knowledge, is the first study in Egypt, that was established for CM prediction using a newly introduced domain of ML and depending on the AMS software data. The main aim is to maximize the benefit from the every day recorded dairy data in the decision-making process, especially regarding animal health and welfare. We used 15 features of milk production parameters and health records that have been mentioned before. But none of them was SCC, EC, or pH as the participating Egyptian farm follows the routine of once a month screening the SCC for the entire herd, which makes SCC and the other 2 parameters not available for us for CM prediction. We tried to optimize the performance of the studied ML models. All data pre-processing (outliers detection and transformation, data standardization, and Autoencoder feature selection) and model tunning (10-fold CV and grid search) procedures were practiced efficiently and many train/test datasets splits were tried starting from 70%/30% train/test split up to 90%/10% train/test split.

The effectiveness of six of the ML models was compared. The accuracy of all models ranged from 62% to 74% and the AUC results ranged from 60% to 71%. We computed 5 different performance metrics for each algorithm but the final optimal algorithm was selected depending on AUC, as AUC is preferred over accuracy as it takes into account both the TPR and FPR of the model across different cut-off thresholds of the ROC curve (Ling *et al*., 2003). The greater the AUC, the better the accuracy, and 0.7 – 0.9 AUC indicated moderate accuracy (Akobeng, 2007). The DT and Gaussian NB showed the highest AUC of 71%, but regarding the other 4 metrics the DT was the best sustainable model so it is considered the best optimal CM predictive ML model of our study.

Regarding the results of the DT variable importance, DIM, age at the last season, lactation order, 305ME, and DMY were found to be the most important and indicative features for CM prediction. Consistent with these findings, Fadul-Pacheco *et al.* (2021) found that DIM, lactation, milk yield (MY), EC, and age of cow at 1st lactation were the features of importance for both 72% and 68% AUC respectively RF and Extreme Gradient Boosting algorithms built for CM detection on the long-term while, the disease events of (retained placenta, abortion, metritis, ketosis, and previous CM) were of no importance and Luo *et al.* (2023) also got a 98% accurate DT CM predictive model with important indicative features of standard deviation and mean MY, lactation days, EC, and lying time.

Regarding DIM which showed the highest importance score, Faris *et al.* (2021) revealed that the lactation stage was a significant risk factor for CM and the prevalence of CM during the early DIM (1-90 days) was higher than in the mid-stage (91-180 days) and the late stage (> 180 days). Also, Koeck *et al.* (2012) stated that the majority of CM cases were during the first month of lactation by a ratio of 32.7%. Such results might be due to the stress of peak milk production during the early stage of lactation (Chegini *et al*., 2016) or due to the diminished antioxidant defense mechanism and higher oxidative stress as a result of increased lipid peroxidase wastes due to the high demand during the early DIM (Sharma *et al*., 2011). Another explanation, this might be due to the delayed diapedesis of the neutrophils into the udder cells making them more sensitive to microbiological agents (Boujenane *et al*., 2015).

The age of the animal at the last season, when we obtained data, was the second important variable in the DT model. As the cow ages and increases the lactation times, the teat canal becomes dilated and partially opened permanently making it highly susceptible to catching infection from the external environment (Shittu *et al*., 2012).

The third indicative feature was the lactation order which in other words intended as parity. The CM cases showed a higher mean of lactation order than the healthy cases this was similar to the finding of Nakov *et al.* (2014). They reported that as the number of parties increases, the risk of CM increases.

The final features of importance were the DMY and 305 ME which showed the same score of importance and both revealed lower mean values (26.6 kg and 10673.5kg respectively) in Mastitic cases than in non-mastitic ones (31.5 kg and 11184.1 kg respectively). The mixed linear regression model revealed a negative correlation ($P<0.001$) between the total score of the four udder quarters inflammation and total milk production (Wahyu Harjanti and Sambodho, 2020). Also, Adriaens *et al.* (2021) emphasized the fact that CM can significantly cause a reduction in milk production by more than 100 kg.

Conclusion

DT algorithm is the best model for a moderate performance of 71% for the AUC under the ROC curve in CM prediction under our data conditions. This study emphasized that it is important to integrate ML data analysis in dairy farms to maximize the benefits of AMS and sensor data. This may have a positive impact on early CM detection and prevention improving animal health and welfare and maintaining the farm profitability.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Adriaens, I., Van Den Brulle, I., Geerinckx, K., D'Anvers, L., De Vliegher, S., Aernouts, B., 2021. Milk losses linked to mastitis treatments at dairy farms with automatic milking systems. Preventive Veterinary Medicine 194, 105420.

Akobeng, A.K., 2007. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatrica, International Journal of Paediatrics 96, 644–647.

Ali, N., Neagu, D., Trundle, P., 2019. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Applied Sciences 1, 1–15.

Ardelean, E.R., Coporîie, A., Ichim, A.M., Dînşoreanu, M., Mureşan, R.C., 2023. A study of autoencoders as a feature extraction technique for spike sorting. PLoS ONE 18, 1–29.

Bausewein, M., Mansfeld, R., Doherr, M.G., Harms, J., Sorge, U.S., 2022. Sensitivity and Specificity for the Detection of Clinical Mastitis by Automatic Milking Systems in Bavarian Dairy Herds. Animals 12, 1–18.

Bender, R., 2009. Introduction to the use of regression models in epidemiology. Mukesh Verma (ed.), Methods in Molecular Biology, Cancer Epidemiology, Vol. 471© 2009 Humana Press, a part of Springer Science + Business Media, Totowa, NJ Cancer Epidemiology, pp. 179–195.

Bi, Q., Goodman, K.E., Kaminsky, J., Lessler, J., 2019. What is Machine Learning? A Primer for the Epidemiologist. American Journal of Epidemiology 188, 2222-2239.

Boujenane, I., El Aimani, J., By, K., 2015. Incidence and occurrence time of clinical mastitis in Holstein cows. Turkish Journal of Veterinary and Animal Sciences 39, 42–49.

Charbuty, B., Abdulazeez, A., 2021. Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends 2, 20–28.

Chegini, A., Hossein-Zadeh, N.G., Hosseini-Moghadam, H., Shadparvar, A.A., 2016. Estimation of genetic and environmental relationships between milk yield and different measures of mastitis and hyperkeratosis in Holstein cows. Acta Scientiarum - Animal Sciences 38, 191–196.

Cheng, W.N., Han, S.G., 2020. Bovine mastitis: risk factors, therapeutic strategies, and alternative treatments — A review. Asian-Australasian Journal of Animal Sciences 33, 1699–1713.

Cockburn, M., 2020. Review: Application and prospective discussion of machine learning for the management of dairy farms. Animals 10, 1–22.

Dhoble, A.S., Ryan, K.T., Lahiri, P., Chen, M., Pang, X., Cardoso, F.C., Bhalerao, K.D., 2019. Cytometric fingerprinting and machine learning (CFML): A novel label-free, objective method for routine mastitis screening. Computers and Electronics in Agriculture 162, 505–513.

Dong, B., Wang, X., Cao, Q., 2022. Performance Prediction of Listed Companies in Smart Healthcare Industry: Based on Machine Learning Algorithms. Journal of Healthcare Engineering 2022, 1–7.

Fadul-Pacheco, L., Delgado, H., Cabrera, V.E., 2021. Exploring machine learning algorithms for early prediction of clinical mastitis. International Dairy Journal 119, 105051.

Faris, D., El-Bayoumi, K., El-Taranany, M., Abdel-Hamed, A., Kamel, E., 2021. Prevalence and Risk Factors of Clinical Mastitis in Holstein Cows under Subtropical Egyptian Conditions. Benha Veterinary Medical Journal 41, 19–23.

Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 9, 1–16.

Ghafoor, N.A., Sitkowska, B., 2021. MasPA: A Machine Learning Application to Predict Risk of Mastitis in Cattle from AMS Sensor Data. AgriEngineering 3, 575–583.

Gokul Krishna, R., Periyasamy, S.V., Roshan Khan, S.B., Mohan Raj, T., 2023. Exploring the Potential of Machine Learning for Early Cattle Disease Diagnosis. 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 853–857.

Han, J., Kamber, M., Pei, J., 2012. Data Preprocessing. In: Data Mining. pp. 83–124.

Hogeveen, H., Steeneveld, W., Wolf, C.A., 2019. Production Diseases Reduce the Efficiency of Dairy Production: A Review of the Results, Methods, and Approaches Regarding the Economics of Mastitis. Annual Review of Resource Economics 11, 289–312.

Hossain, M.E., Kabir, M.A., Zheng, L., Swain, D.L., McGrath, S., Medway, J., 2022. A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. Artificial Intelligence in Agriculture 6, 138–155.

Iliou, T., Anagnostopoulos, C.N., Nerantzaki, M., Anastassopoulos, G., 2015. A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance. Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS) 2015, 1–5.

Itoo, F., Meenakshi, Singh, S., 2021. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology (Singapore) 13,1503–1511.

Keceli, A.S., Catal, C., Kaya, A., Tekinerdogan, B., 2020. Development of a recurrent neural networks-based calving prediction model using activity and behavioral data. Computers and Electronics in Agriculture 170, 105285.

King, M.T.M., LeBlanc, S.J., Pajor, E.A., Wright, T.C., DeVries, T.J., 2018. Behavior and productivity of cows milked in automated systems before diagnosis of health disorders in early lactation.

Journal of Dairy Science 101, 4343–4356.
Klusowski, J.M., 2018. Complete Analysis of a Random Forest Model. Journal of Machine Learning Research,13, 1063–1095.
Koeck, A., Miglior, F., Kelton, D.F., Schenkel, F.S., 2012. Alternative somatic cell count traits to improve mastitis resistance in Canadian Holsteins. Journal of Dairy Science 95, 432–439.
Kullarni, V.Y., Sinha, P.K., 2013. Random Forest Classifier: A Survey and Future Research Directions. International Journal of Advanced Computing 36, 1144–1156.
Lee, E.P.F., Lozeille, J., SoldÃ¡n, P., Daire, S.E., Dyke, J.M., Wright, T.G., 2001. An empirical study of the naive Bayes classifier. Physical Chemistry Chemical Physics 3,
Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., Zhang, C., 2021. CleanML: A study for evaluating the impact of data cleaning on ml classification tasks. Proceedings - International Conference on Data Engineering 2021, 13–24.
Ling, C.X., Huang, J., Zhang, H., 2003. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics 6852, 329–341.
Luo, W., Dong, Q., Feng, Y., 2023. Risk prediction model of clinical mastitis in lactating dairy cows based on machine learning algorithms. Preventive Veterinary Medicine 221, 106059.
Mathkunti, N.M., Rangaswamy, S., 2020. Machine Learning Techniques to Identify Dementia. SN Computer Science 1, 1–6.
Nakov, D., Hristov, S., Andonov, S., Trajchev, M., 2014. Udder-related risk factors for clinical mastitis in dairy cows. Veterinarski Arhiv 84, 111–127.
Nusinovici, S., Tham, Y.C., Chak Yan, M.Y., Wei Ting, D.S., Li, J., Sabanayagam, C., Wong, T. Y., Cheng, C.Y., 2020. Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of Clinical Epidemiology 122, 56–69.
Ruegg, P.L., 2017. A 100-Year Review: Mastitis detection, management, and prevention. Journal of Dairy Science 100, 10381–10397.
Schölkopf, B., 2003. An Introduction to Support Vector Machines. Recent Advances and Trends in Nonparametric Statistics 2003, 3–17.
Sharma, N., Singh, N.K., Singh, O.P., Pandey, V., Verma, P.K., 2011. Oxidative stress and antioxidant status during transition period in dairy cows. Asian-Australasian Journal of Animal Sciences 24, 479–484.
Shittu, A., Abdullahi, J., Jibril, A., Mohammed, A.A., Fasina, F.O., 2012. Sub-clinical mastitis and associated risk factors on lactating cows in the Savannah Region of Nigeria. BMC Veterinary Research, 8, 134.
Sonia Singh, P.G., 2014. Comparative Study ID3,CART AND C4.5 Decision Tree Algorithm. International Journal of Advanced Information Science and Technology 27, 98.
Task, C., 2014. Chapter 7; k-Nearest Neighbor Algorithm. pp. 149–164.
Vik, J., Stræte, E.P., Hansen, B.G., Nærland, T., 2019. The political robot – The structural consequences of automated milking systems (AMS) in Norway. NJAS - Wageningen Journal of Life Sciences 90–91, 1-9.
Volkmann, N., Kulig, B., Hoppe, S., Stracke, J., Hensel, O., Kemper, N., 2021. On-farm detection of claw lesions in dairy cows based on acoustic analyses and machine learning. Journal of Dairy Science 104, 5921–5931.
Wahyu Harjanti, D., Sambodho, P., 2020. Effects of mastitis on milk production and composition in dairy cows. IOP Conference Series: Earth and Environmental Science 518, 012032.
Zhou, X., Xu, C., Wang, H., Xu, W., Zhao, Z., Chen, M., Jia, B., Huang, B., 2022. The Early Prediction of Common Disorders in Dairy Cows Monitored by Automatic Systems with Machine Learning Algorithms. Animals 12, 1251.