



## Exploiting machine learning methods with monthly routine milk recording data and climatic information to predict subclinical mastitis in Italian Mediterranean buffaloes

T. Bobbo,<sup>1,2</sup> R. Matera,<sup>3</sup> G. Pedota,<sup>4</sup> A. Manunza,<sup>1</sup> A. Cotticelli,<sup>3</sup> G. Neglia,<sup>3\*</sup> and S. Biffani<sup>1</sup>

<sup>1</sup>Consiglio Nazionale delle Ricerche (CNR), Istituto di Biologia e Biotecnologia Agraria (IBBA), 20133 Milan, Italy

<sup>2</sup>Department of Agricultural and Environmental Sciences, University of Milan, 20133 Milan, Italy

<sup>3</sup>Department of Veterinary Medicine and Animal Production, Federico II University, 80137 Naples, Italy

<sup>4</sup>Associazione Regionale Allevatori della Basilicata, 85100 Potenza, Italy

### ABSTRACT

Mastitis has detrimental effects on the world's dairy industry, reducing animal health, milk production and quality, as well as income for farmers. In addition, consumers' growing interest in food safety and rational usage of antibiotics highlights the need to develop novel strategies to improve mastitis detection, prevention, and management. In the present study we applied machine learning (ML) analyses to predict presence or absence of subclinical mastitis in Italian Mediterranean buffaloes, exploiting information collected the previous month during routine milk recording procedures, as well as climatic data. The data set included 3,891 records of 1,038 buffaloes from 6 herds located in Basilicata Region (South Italy). Prediction models were developed using 4 different ML algorithms (Generalized Linear Model, Support Vector Machines, Random Forest, and Neural Network) and 2 data set splitting approaches for the creation of the training and test sets (by record or by animal ID number, always with 80% of the data used for model training and the remaining 20% for model testing). Support Vector Machine was the best method to predict high or low somatic cell count at the subsequent test-day record in the validation set, and therefore it was used to estimate the contribution of each feature to the best model. Independently from the data set splitting approach, the most important features were somatic cell score, differential somatic cell count, electrical conductivity, and milk production. Among climatic data, the most informative were temperature and relative humidity. When the data were split by animal ID, an improvement in models' predictive performance on the test set was observed, suggesting this as the most appropriate data splitting approach in data sets with repeated measures to avoid data leakage. Ac-

cording to different metrics, Neural Network was the best method for making predictions on the test set. Our findings confirmed the promising role of ML methods to improve prevention and surveillance of subclinical mastitis, exploiting the large amount of data currently available to identify animals that would possibly have high somatic cell count the subsequent month.

**Key words:** machine learning, mastitis, climatic data, data leakage, Italian Mediterranean buffaloes

### INTRODUCTION

Mastitis, an inflammatory condition of the udder, has become a critical issue in the world's dairy industry, affecting animal health, milk production and quality, and income for farmers (Halasa et al., 2007). Mediterranean buffaloes (*Bubalus bubalis*) have been generally considered less susceptible to udder infections compared with dairy cows, thanks to morphological characteristics of the teat canal and sphincter that reduce the possible invasion of mastitis-causing pathogens (Fagiolo and Lai, 2007). Nevertheless, mastitis has a detrimental effect also on the buffalo dairy sector, which suffers from poor scientific knowledge about this disease in comparison to the bovine dairy sector (Puggioni et al., 2020). Recently, efforts have been made to improve mastitis detection, management and selection in dairy buffaloes. Indeed, novel indicators of mammary gland inflammation derived from traditional SCC, previously developed for improving selection for mastitis resistance in Italian Holstein cattle (Bobbo et al., 2018), were investigated in dairy buffaloes (Costa et al., 2021). Moreover, the dynamics of the different cell types (e.g., macrophages and neutrophils) that compose total SCC have been explored (Alterisio et al., 2021). Differential somatic cell count (**DSCC**), a novel parameter that represents the proportion of lymphocytes and neutrophils on the total SCC, has been recently introduced in the routine milk recording scheme of dairy buffaloes. The combi-

---

Received May 11, 2022.

Accepted September 27, 2022.

\*Corresponding author: neglia@unina.it

nation of SCC and DSCC has been demonstrated to better define the udder health status of dairy cattle and enhance a rational use of antibiotics (Bobbo et al., 2020). In addition, a novel cathelicidin ELISA has been developed for detecting buffalo mastitis (Puggioni et al., 2020). However, there is still a need for filling the gap in knowledge, possibly by using information that is currently available and not fully exploited. For instance, great advantage could be taken of the large amount of data provided by automatic milking recording systems, as well as by monthly test-day (**TD**) milk recording procedures. Such information, easily accessible, could be used to train machine learning (**ML**) algorithms for the prediction of specific traits of interest, such as phenotypes that are difficult to measure, or the possible occurrence of a disease. Machine learning offers a new approach for data analysis and has already been applied in several areas of dairy research (e.g., feeding, behavior, reproduction, and health) for supporting management of farms (Cockburn, 2020). Early detection and prevention of mastitis would represent a valuable asset from both the economic and health point of view. Previous studies reported in the literature have attempted to predict mastitis in dairy cattle, defined by the presence of high milk SCC (Ebrahimi et al., 2019; Anglart et al., 2020; Bobbo et al., 2021) or of mastitis-causing pathogens (Sharifi et al., 2018; Hyde et al., 2020), by applying different ML algorithms. Nevertheless, in livestock research, where data sets with repeated measures are often used for ML data analysis, there has been little discussion on the issue of data leakage related to data splitting and model overfitting (Satola and Bauer, 2021; Ji et al., 2022). Data leakage occurs when the training set used to create the model contains information about the target to be predicted.

Following the approach reported by Bobbo et al. (2021), in the present study we exploited information already collected in the frame of the monthly routine milk recording procedure of Italian Mediterranean buffaloes, as well as climatic data (features at time  $t - 1$ ), to predict which animals will present high or low milk SCC level at the subsequent TD (outcome at time  $t$ ). In addition, we compared results obtained using 2 different data splitting approaches to evaluate the possible effects of data leakage.

## MATERIALS AND METHODS

### Ethics Statement

Animal welfare and use committee approval was not needed for this study because data sets were obtained

from pre-existing databases based on routine animal recording procedures.

### Data Collection and Editing

Buffaloes involved in the current study were reared on commercial farms and were not subjected to any invasive procedure. Test-day data, recorded during monthly routine milk recording procedures, were provided by the Italian Breeders Association (Rome, Italy). Data included information about herd, animals (ID number, date of calving, stage of lactation, and parity order), date of sampling, daily milk production (kg/d), milk composition [fat (%), protein (%), casein (%), lactose (%), pH, and urea (mg/100 mL)], SCC (cells/mL), DSCC (%), BHB (mmol/L), electrical conductivity (**EC**, mS), and milk coagulation properties [rennet coagulation time (min) and curd firmness 30 min after rennet addition (mm)]. The original data set, which included records collected from August 2019 to February 2021, was edited to select animals with at least 2 TD records within lactation, and with less than 360 DIM. In addition, only consecutive TD records separated by a time interval lower than 6 wk were selected. This approach, also applied by Bobbo et al. (2021), was adopted to reduce data fragmentation over time. Among milk traits, outliers beyond 4 standard deviations, possibly resulting from errors in sampling or recording procedures, were considered as missing values, and only full records were selected for subsequent analysis. Average daily milk production and SCC of contemporary groups—that is, animals sampled in the same herd and day (herd-test-date, **HTD**)—were also determined (**milk-HTD** and **SCC-HTD**, respectively). Finally, the 2 SCC-related traits (SCC and **SCC-HTD**) were log-transformed to SCS and **SCS-HTD** to achieve normality, whereas no transformation was required for DSCC. The outcome to be predicted—that is, presence or absence of subclinical mastitis at the subsequent monthly TD—was coded as a binary trait and was based on SCC: animals were classified as healthy ( $\text{SCC} \leq 200,000 \text{ cells/mL}$ ) or mastitic ( $\text{SCC} > 200,000 \text{ cells/mL}$ ). The threshold of 200,000 cells/mL was selected based on the published literature (Moroni et al., 2006; Costa et al., 2020, 2021). The prevalence of subclinical mastitis ( $\text{SCC} > 200,000 \text{ cells/mL}$ ) was 40.3%. After editing, the data set included 3,891 records of 1,038 buffaloes in 6 herds. Each record included information of 2 subsequent monthly TD: animal and milk data collected at the previous TD and outcome (healthy vs. mastitic) at the subsequent TD.

In addition, climatic information of the sampling location and date were retrieved from the NASA Pre-

dition of Worldwide Energy Resource (POWER) Data Access Viewer (Sparks, 2018), which allowed access to daily averaged data by providing latitude and longitude values of the 6 herds and the desired date range. In particular, parameters of interest were as follows: All Sky Surface Shortwave Downward Irradiance ( $\text{MJ/m}^2$  per day), All Sky Surface UV Index (dimensionless), Temperature at 2 Meters ( $^{\circ}\text{C}$ ), Relative Humidity at 2 Meters (%), Precipitation Corrected (mm/day), Surface Pressure (kPa), Wind Speed at 2 Meters (m/s), and Wind Direction at 2 Meters (Degrees). For a detailed description of climatic variables see Supplemental Table S1 (<https://data.mendeley.com/datasets/pdmy7czpz4/1>; Bobbo et al., 2022).

Finally, a total of 27 features were considered: parity (from 1 to  $\geq 6$ ), stage of lactation (DIM: 10 classes, 9 of 30 d each and the last one including DIM  $>300$  d), year and month of calving (18 levels), year and month of sampling (10 levels), milk production, fat, protein, casein, lactose, pH, urea, SCS, DSCC, BHB, EC, milk-HTD, SCS-HTD, the 2 milk coagulation properties, and the 8 climatic parameters.

### Data Processing, Recursive Feature Elimination, and Model Building

Four different ML methods were adopted to develop subclinical mastitis prediction models: Generalized Linear Models (**GLM**; Nelder and Wedderburn, 1972), Support Vector Machine (**SVM**; Cortes and Vapnik, 1995), Random Forest (**RF**; Breiman, 2001), and Neural Network (**NN**; McCulloch and Pitts, 1943). Two approaches were used for splitting the data, to evaluate whether results could be biased by possible overfitting due to data leakage in time series data sets:

- (1) Splitting by record. The data set was randomly split into 2 subsets: 80% of the data was used to train and evaluate the models, and the remaining 20% was excluded from model building and held out as a test set. Random sampling was performed within each outcome class, thus preserving the original outcome rate in training, validation, and test sets. Splitting by record, the same animals, but with different TD records, can be found in all created subsets.
- (2) Splitting by animal ID. The data set was randomly split so that 80% of the animals (and all their relative TD records) were included in the training subset used for model building and evaluation, and the remaining 20% were included in the test set. Original class distribution of the outcome was preserved. Splitting by animal ID, buffaloes

in the test set were not included in the training subset.

Recursive feature elimination using a 10-fold cross-validation repeated 100 times with the RF method (Svetnik et al., 2004) was applied to eventually reduce the number of features, automatically selecting the most predictive ones to identify the most parsimonious model with best performance—that is, with highest accuracy of prediction. Then, a stratified 10-fold cross-validation repeated 100 times was employed to train and evaluate the models. In particular, the training data set was divided into 10 subsets of equal size. Splitting the data by record, partitions of the 10-fold cross-validation were randomly selected; splitting by animal ID, data were split into the 10 subsets based on groups (ID). At each of the 10 iterations, prediction models were trained on 9 subsets and evaluated on the last one, changing the validation subset every time. This entire process was repeated 100 times, for a total of 1,000 iterations. Therefore, 100 mean accuracy and kappa values of each 10-fold cross-validation were then averaged to obtain the final metrics of each method reported in the tables. Data standardization was performed within cross-validation. Tuning details of each model are reported in the supplemental information file (<https://data.mendeley.com/datasets/pdmy7czpz4/1>; Bobbo et al., 2022). Data analysis was performed using Caret v. 6.0-86 (Kuhn, 2021) and Tidyverse v. 1.3.1 (Wickham et al., 2019) packages of R software v. 4.1.2 (R Core Team, 2021).

### Comparison of Methods Predicting Performance on Validation and Test Sets

Comparison of methods predicting performance on the validation set was first performed by means of accuracy and Cohen's kappa values. Feature importance (i.e., the estimation of the contribution of each variable to the best model) was then computed. Importance values were then scaled to 0 (least important) and 100 (most important). Predictive ability of all models on the test set was then assessed, and method comparisons were based on different metrics: sensitivity, specificity, accuracy, positive predictive value, negative predictive value, Cohen's kappa value, and F1 score. False positive, false negative, and total error rates of each method were also calculated. Receiver operating characteristic curve analysis was performed using pROC package v. 1.17.0.1 (Robin et al., 2011), and area under the receiver operating characteristic curve (**AUC**) was measured. Finally, Matthew's correlation coefficient (**MCC**) was calculated according to the following formula:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}},$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

## RESULTS

### **Data Processing, Recursive Feature Elimination, and Model Building**

Four ML methods (GLM, SVM, RF, and NN) were applied to develop subclinical mastitis prediction models, using animals and milk information collected during monthly routine milk recording procedures and climatic data. Training and test sets were obtained using 2 approaches: dividing the original data set by records or by animal ID, so that the same animals could or could not be present in the 2 sets of data; that is, they could be totally unknown or not when testing the model.

Before model building and training, a recursive feature elimination was applied to eventually reduce the number of features and remove uninformative ones. Splitting the data set both by record and by animal ID, all 27 features were retained in the most parsimonious yet accurate model (Figures 1a and 1b). Nevertheless, a sort of plateau can be reached with the first 7 most important features (SCS, SCS-HTD, milk-HTD, EC, milk, parity, and DSCC).

### **Comparison of Methods Predicting Performance on Validation Set**

Evaluation and comparison of the predicting performance of the 4 ML algorithms on the validation set was based on accuracy and kappa value. Splitting the data set by record, accuracy ranged between 75.4% (NN) and 76.1% (SVM), and kappa between 0.476 (NN) and 0.489 (SVM) (Figure 2a). Splitting the data by animal ID, slightly lower values were reported, with accuracy ranging from 74.8% (RF) to 75.3% (SVM), and kappa from 0.446 (RF) to 0.457 (GLM; Figure 2b). In both cases, SVM was the best method to predict presence or absence of subclinical mastitis in the validation set, and therefore it was used for estimating the contribution of each variable to the best model. Results of the feature importance using SVM on the validation set suggested that, independently from the data set splitting approach, SCS at the previous TD was the most important feature, followed by SCS-HTD and DSCC (Figures 3a and 3b). Two other

important variables were milk-HTD and EC. Among climatic data, the most informative were temperature and relative humidity.

### **Comparison of Methods Predicting Performance on Test Set**

Comparison of the prediction performance of the 4 ML algorithms on test set, obtained by splitting the original data set with 2 different approaches, was based on several metrics, summarized in Table 1. Splitting the data set by record, accuracy of prediction ranged from 73.9% (SVM) to 75.4% (NN), whereas kappa values were ranged between 0.447 (SVM) and 0.480 (NN). The NN method also showed the highest F1 score (0.676) and MCC (0.482), followed by GLM (0.670 and 0.476, respectively). Similar findings but with slightly greater scores were obtained by splitting the data set by animal ID. Indeed, NN proved to be the best-performing method, with prediction accuracy of 76.2%, kappa value of 0.518, F1 score of 0.726, and MCC of 0.522. The SVM method, which was the most accurate in predicting subclinical mastitis on the validation set, was instead the worst-performing on the test set.

Considering all 4 methods, the greatest AUC values were observed by splitting the data set by animal ID rather than by record: 84.1% versus 81.2% for GLM, 83.3% versus 80.2% for SVM, 84.1% versus 79.0% for RF, and 84.0% versus 81.4% for NN (Figures 4a and 4b).

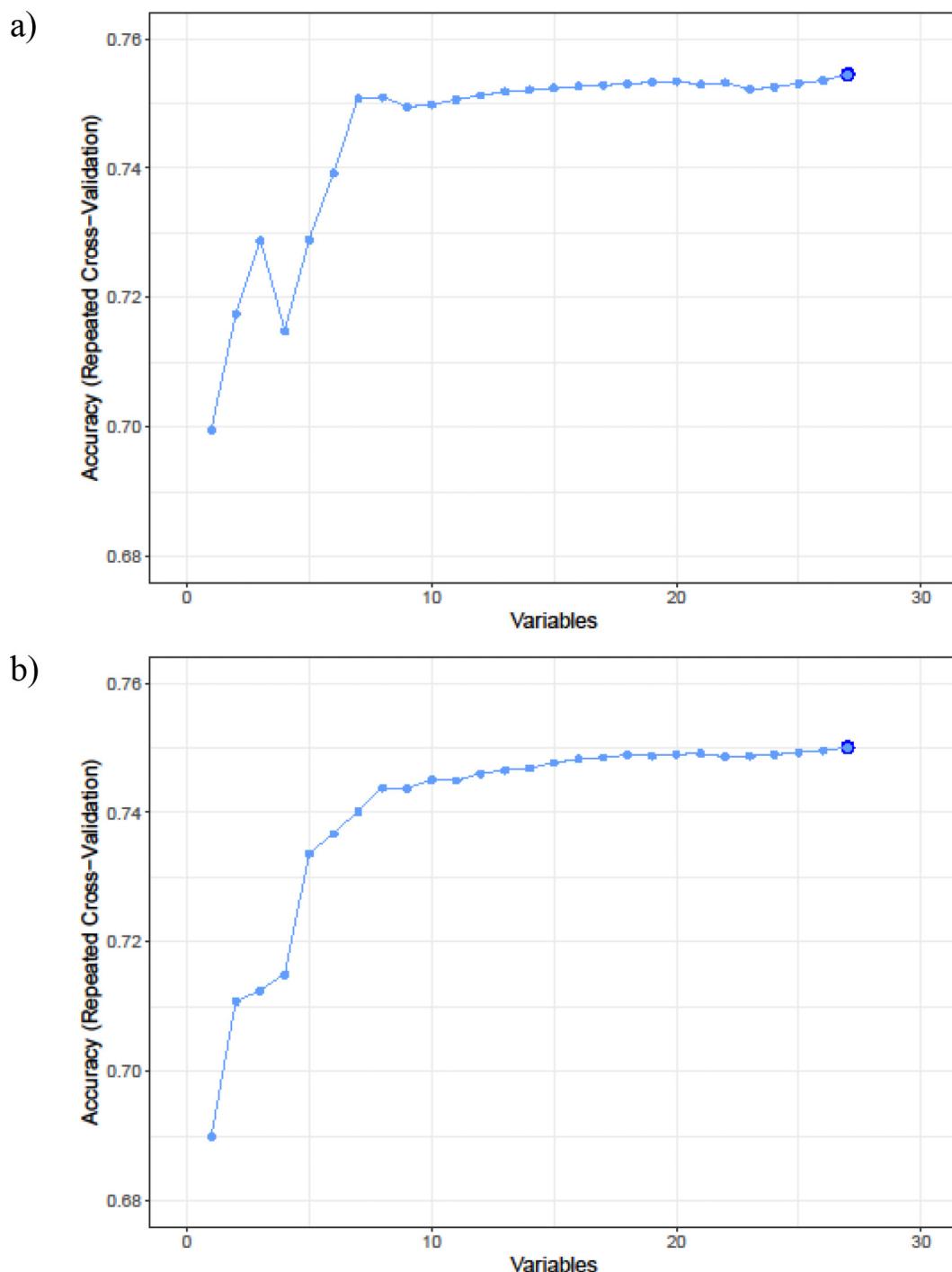
## DISCUSSION

In the current study, we predicted whether Italian Mediterranean buffaloes will present high or low SCC in the milk collected at the subsequent TD, applying ML analyses on easily accessible and already available information (i.e., milk data collected the previous month during monthly routine milk recording, as well as climatic data related to the sampling location). Although Mediterranean buffaloes seem to be more robust and resistant to diseases than dairy cows, their health and production are also affected by mastitis (Puggioni et al., 2020). Therefore, strategies for early detection and prevention of subclinical mastitis are of paramount importance for both economic and health aspects. From this perspective, our study highlighted the pivotal role of ML analysis for exploiting the large amounts of data that are available nowadays, with the aim of improving disease surveillance and, consequently, farm management strategies.

Subclinical mastitis prediction models were developed using 4 different ML methods, one linear (GLM), one

with a distance-based approach (SVM), an algorithm based on decision trees (RF), and one that works like the human brain trying to perform pattern recognition (NN). We decided to compare results obtained using 2 different data set splitting approaches. Indeed, train-

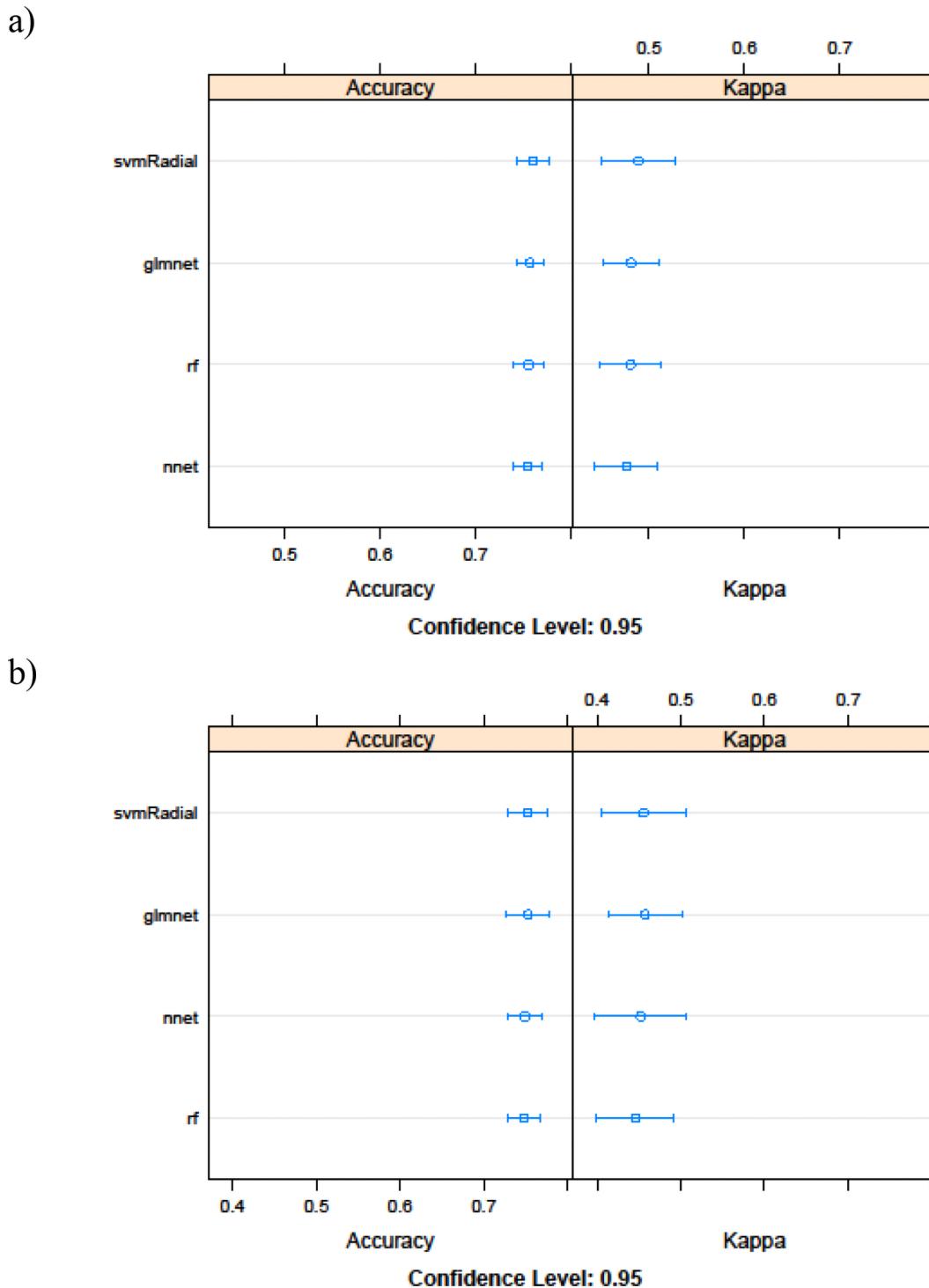
ing and test sets were created by dividing the original data set by records (i.e., the same animals, but with different TD records, can be found in both sets of data) or by animal ID (i.e., animals in the test set were not included in model building and were totally unknown).



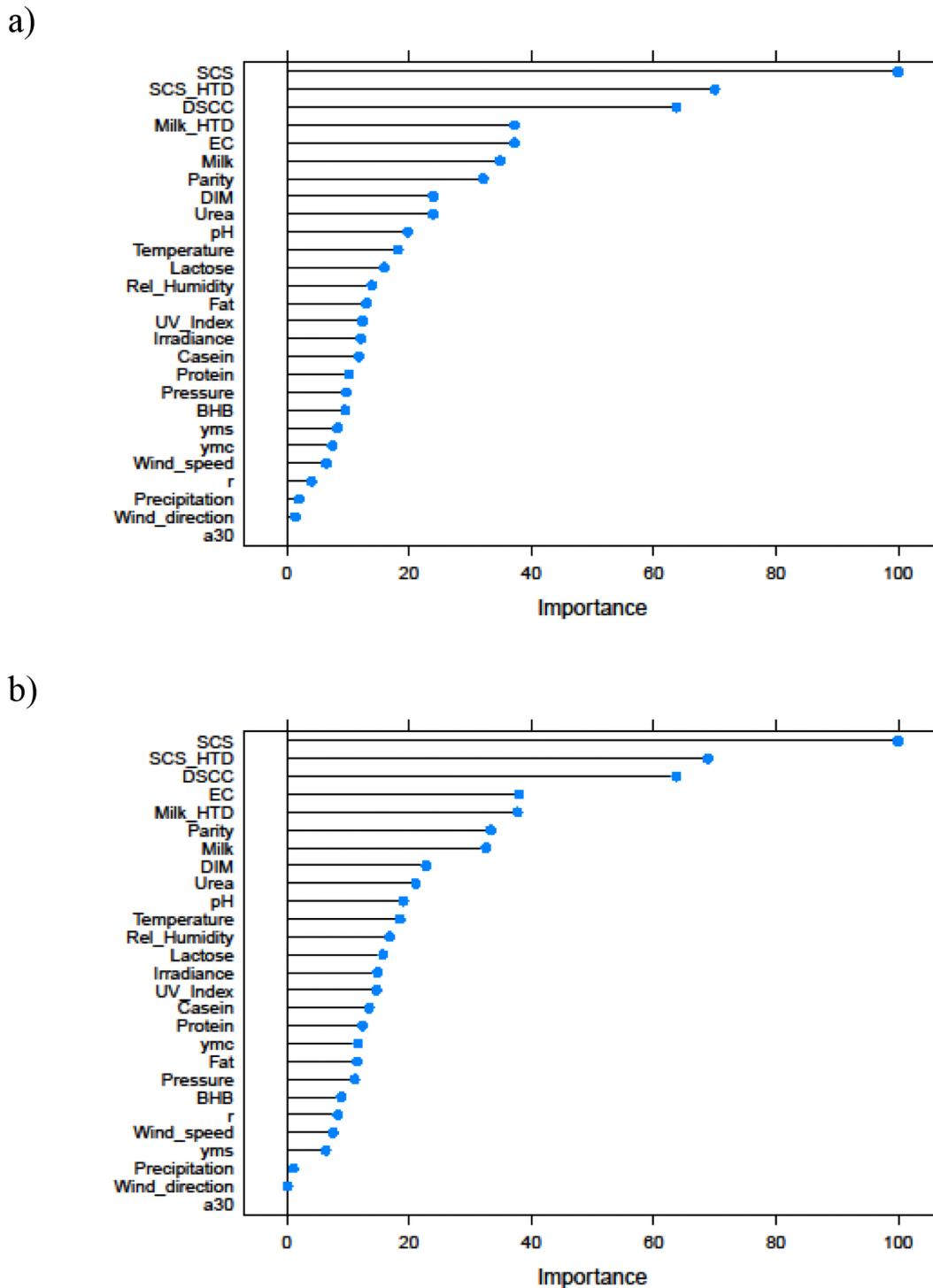
**Figure 1.** Results of the recursive feature elimination, a function that implements backward feature selection, incorporating 27 to 1 features in the model, using the training set obtained by splitting the original data set (a) by record and (b) by animal ID. The number of features is reported on the x-axis, and the model accuracy from the 10-fold cross-validation repeated 100 times on the y-axis.

A common approach during model building is to randomly divide the data set into multiple subsets, so that training and fine-tuning of the model are performed

using a k-fold cross-validation as resampling procedure (Ebrahimi et al., 2019; Anglart et al., 2020; Bobbo et al., 2021). In addition, data sets can also be split to



**Figure 2.** Metrics (accuracy and Cohen's kappa value) for the comparison of methods predicting performance on the validation set, obtained by splitting the original data set (a) by record and (b) by animal ID. Prediction models were developed using 4 machine learning methods: Generalized Linear Model (glmnet), Support Vector Machines (svmRadial), Random Forest (rf), and Neural Network (nnet). Error bars represent 95% CI.



**Figure 3.** Plot of the feature importance, scaled from 0 (least important) to 100 (most important), showing the ranking for the prediction of presence or absence of subclinical mastitis in the validation set, obtained by splitting the original data set (a) by record and (b) by animal ID. Evaluated features, using Support Vector Machine as the predictive method, are as follows: individual SCS and SCS of contemporary group (scs and scs\_htd), differential SCC (DSCC), electrical conductivity (EC), individual milk production and milk production of contemporary group (milk and milk\_HTD), parity, stage of lactation (DIM), milk composition traits (urea, pH, lactose, fat, casein, protein), BHB, year and month of sampling (yms), year and month of calving (ymc), rennet coagulation time (r), curd firmness 30 min after rennet addition (a30), and climatic data (temperature, relative humidity, UV index, irradiance, pressure, precipitation, wind speed, and wind direction).

**Table 1.** Metrics for the comparison of methods predicting performance on test set, obtained by splitting the original data set by record and by animal ID: accuracy, 95% CI, sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), Cohen's kappa value, F1 score, and Matthew's correlation coefficient (MCC)<sup>1</sup>

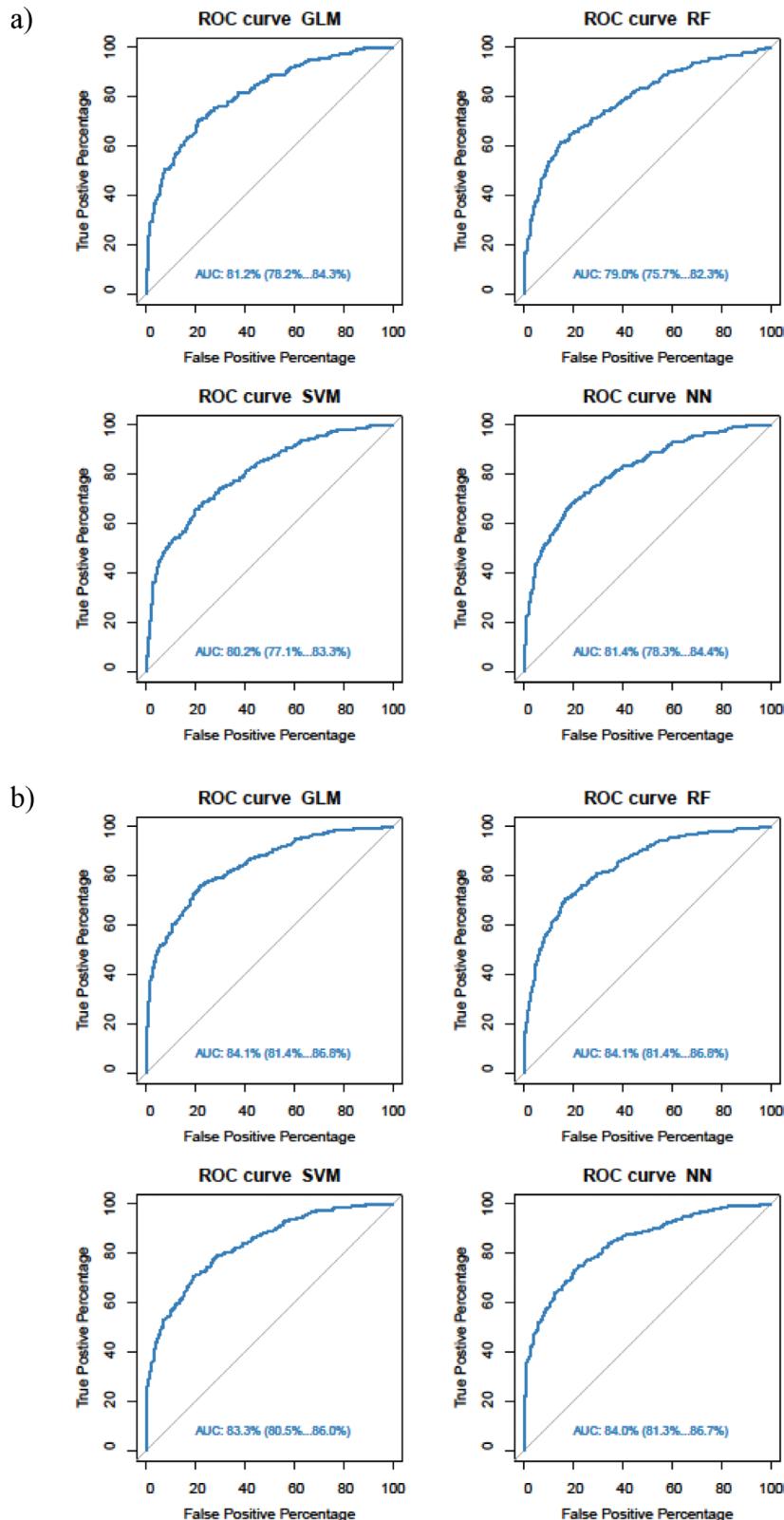
Method	Accuracy	95% CI	Se	Sp	PPV	NPV	Kappa	F1 score	MCC
Splitting by record									
GLM	0.752	0.720–0.782	0.624	0.838	0.723	0.767	0.473	0.670	0.476
SVM	0.739	0.707–0.770	0.618	0.821	0.700	0.760	0.447	0.656	0.450
RF	0.740	0.708–0.771	0.627	0.817	0.699	0.764	0.452	0.661	0.453
NN	0.754	0.723–0.784	0.634	0.836	0.724	0.771	0.480	0.676	0.482
Splitting by animal ID									
GLM	0.760	0.729–0.789	0.661	0.845	0.787	0.742	0.512	0.719	0.518
SVM	0.749	0.717–0.778	0.651	0.834	0.772	0–0.734	0.489	0.706	0.495
RF	0.759	0.728–0.788	0.645	0.857	0.796	0.736	0.509	0.713	0.517
NN	0.762	0.731–0.791	0.677	0.836	0.781	0.749	0.518	0.726	0.522

<sup>1</sup>Prediction models were developed using 4 machine learning methods: Generalized Linear Model (GLM), Support Vector Machines (SVM), Random Forest (RF), and Neural Network (NN).

use part of the data for training with cross-validation and to hold out a portion of the data as external test set (e.g., 80/20%, 90/10%, 50/50%). The test set is important in order to obtain non-inflated estimates due to possible overfitting; indeed, model predictive performance on test sets is generally lower. In such cases, data are typically divided by randomly selecting a certain proportion of records (Anglart et al., 2020; Bobbo et al., 2021) or of farms (Hyde et al., 2020), or numbers of milkings (Ankinakatte et al., 2013). Nevertheless, records in time series data sets or in data sets with repeated measures of the same individual (e.g., animals with several TD) might be highly correlated; therefore special attention should be paid to choosing the most appropriate data splitting approach. In such cases, data should be split based on ID rather than by records, to avoid possible overfitting due to data leakage. Indeed, the aim of predictive modeling is to develop a model that makes accurate predictions on novel unseen data. Splitting by record data sets with repeated measures, data leakage might occur; that is, the data you are using for model training might contain the information you are trying to predict. In our study, when splitting by record, we observed slightly better predictive performances on the validation set and lower performance on the test set. This can be the result of overfitting, although, in our study, to minimize data leakage, recursive feature elimination as well as data standardization were performed within cross-validation. When the data were split by animal ID (both in the creation of the training and test sets and during cross-validation), an improvement in models' predictive performance on the test set was observed, suggesting this as the most appropriate data splitting approach according to our data structure.

Comparisons of the predicting performance of the 4 ML algorithms on both validation and test sets were based on several metrics, including F1score, AUC, and

MCC, which are independent from the outcome rate. Results of the feature importance based on the most accurate method (SVM) on the validation set revealed that, independently from how the data set was split, SCS recorded at the previous TD was, as expected, the most important feature for predicting the presence or absence of subclinical mastitis at the subsequent TD, followed by the other 2 SCC-related traits (SCS\_HTD and DSCC). In addition to individual SCS, average SCS of contemporary groups was included to represent herd hygiene conditions. Our results confirmed the important information provided by DSCC, a novel indicator of udder health status, to be used in combination with SCC to better screen for udder health status, as previously observed for dairy cattle (Bobbo et al., 2020). Indeed, DSCC and SCS are different traits, as their phenotypic and genetic correlations differ from unity (i.e., 0.66), as reported by (Bobbo et al., 2019). Other important variables were milk\_HTD, a proxy for herd management level, individual milk production, and EC. The negative correlation between buffaloes' milk production and SCS has already been reported in the literature (Tripaldi et al., 2010; Costa et al., 2020). In addition, previous ML studies on dairy cows (Ebrahimie et al., 2018; Ebrahimie et al., 2019) have found EC to be one of the most important features in the prediction of subclinical mastitis based on automatic milking parameters. Indeed, udder infection alters the volume of milk produced, as well as its ionic composition due to leakage of components through the blood-milk barrier. Parity order and stage of lactation also showed relevant contributions to the best model; indeed, they are well known factors affecting SCC variation (Cerón-Muñoz et al., 2002). Among climatic data, the most informative were temperature and relative humidity. In livestock, heat stress is known to negatively affect both milk production and animal health (Bernabucci et al., 2010, 2014). The temperature-humidity index,



**Figure 4.** Receiver operating characteristic (ROC) curves of 4 machine learning methods [Generalized Linear Model (GLM), Support Vector Machines (SVM), Random Forest (RF), and Neural Network (NN)] run for predicting the presence or absence of subclinical mastitis on the test set, obtained splitting the original data set (a) by record and (b) by animal ID. In each plot, area under the curve (AUC) and 95% CI are reported.

which represents the combined effect of air temperature and humidity, is a parameter commonly used to evaluate the degree and the consequences of heat stress (Bernabucci et al., 2014; Matera et al., 2022). A recent study conducted on Italian Mediterranean buffaloes (Matera et al., 2022) has confirmed the negative effect of temperature-humidity index variation on udder health, defined by SCC. In the present study, traits related to solar radiation (UV\_index and irradiance) also showed moderate relevance. Climate variables such as temperature, relative humidity, and solar radiation have previously been found to slightly affect milk production and composition (Sharma et al., 1988). In addition, the inclusion of meteorological parameters (e.g., precipitation, sunshine hours, and soil temperature) in milk production forecast models resulted in a slight improvement in the prediction accuracy, with sunshine hours having the largest effect (Zhang et al., 2020).

## CONCLUSIONS

The findings of our study confirmed ML methods to be a promising tool to improve prevention and surveillance of subclinical mastitis, exploiting the large amount of data currently available. Given consumers' growing concerns about food safety, quality, and antibiotic usage, further studies are needed to advance mastitis detection, management, and selection. Indeed, given the high economic value of Protected Designation of Origin (PDO) Mozzarella di Bufala cheese, special attention should be paid to the health and well-being of Italian Mediterranean buffaloes and their milk quality. We are confident that our research will serve as a basis for practical implementation of these methodologies in dairy management systems, as well as in the application of complex phenotypes in genetic and genomic evaluations.

## ACKNOWLEDGMENTS

This research was funded by Italian Ministry of Agriculture (MIPAAF – DISR 07) – Programma di Sviluppo Rurale Nazionale 2014/2020 (Rome, Italy). Caratterizzazione delle risorse genetiche animali di interesse zootecnico e salvaguardia della biodiversità. Sottomisura: 10.2 – Sostegno per la conservazione, l'uso e lo sviluppo sostenibili delle risorse genetiche in agricoltura. Project: “Bufala Mediterranea Italiana – Tecnologie innovative per il miglioramento Genetico – BIG” Prot. N. 0215513 11/05/2021. CUP ANASB: J29J21003720005; CUP UNINA: J69J21003020005. Climatic data were obtained from the NASA Langley Research Center POWER Project funded through the

NASA Earth Science Directorate Applied Science Program. The authors thank the Associazione Nazionale Allevatori Specie Bufalina (ANASB; Caserta, Italy) for providing the data. The authors have not stated any conflicts of interest.

## REFERENCES

- Alterisio, M. C., P. Ciaramella, and J. Guccione. 2021. Dynamics of macrophages and polymorphonuclear leukocytes milk-secreted by buffaloes with udders characterized by different clinical status. *Vet. Sci.* 8:204. <https://doi.org/10.3390/vetsci8100204>.
- Anglart, D., C. Hallén-Sandgren, U. Emmanuelson, and L. Rönnegård. 2020. Comparison of methods for predicting cow composite somatic cell counts. *J. Dairy Sci.* 103:8433–8442. <https://doi.org/10.3168/jds.2020-18320>.
- Ankinakatte, S., E. Norberg, P. Løvendahl, D. Edwards, and S. Højsgaard. 2013. Predicting mastitis in dairy cows using neural networks and generalized additive models: A comparison. *Comput. Electron. Agric.* 99:1–6. <https://doi.org/10.1016/j.compag.2013.08.024>.
- Bernabucci, U., S. Biffani, L. Buggiotti, A. Vitali, N. Lacetera, and A. Nardone. 2014. The effects of heat stress in Italian Holstein dairy cattle. *J. Dairy Sci.* 97:471–486. <https://doi.org/10.3168/jds.2013-6611>.
- Bernabucci, U., N. Lacetera, L. H. Baumgard, R. P. Rhoads, B. Ronchi, and A. Nardone. 2010. Metabolic and hormonal acclimation to heat stress in domesticated ruminants. *Animal* 4:1167–1183. <https://doi.org/10.1017/S175173111000090X>.
- Bobbo, T., S. Biffani, C. Taccioli, M. Penasa, and M. Cassandro. 2021. Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. *Sci. Rep.* 11:13642. <https://doi.org/10.1038/s41598-021-93056-4>.
- Bobbo, T., R. Matera, G. Pedota, A. Manunza, A. Cotticelli, G. Neglia, and S. Biffani. 2022. Supplementary\_Information\_file\_JDS.2022-22292. Mendeley Data, V1. <https://doi.org/10.17632/pdmy7czpz4.1>.
- Bobbo, T., M. Penasa, and M. Cassandro. 2019. Short communication: Genetic aspects of milk differential somatic cell count in Holstein cows: A preliminary analysis. *J. Dairy Sci.* 102:4275–4279. <https://doi.org/10.3168/jds.2018-16092>.
- Bobbo, T., M. Penasa, and M. Cassandro. 2020. Combining total and differential somatic cell count to better assess the association of udder health status with milk yield, composition and coagulation properties in cattle. *Ital. J. Anim. Sci.* 19:697–703. <https://doi.org/10.1080/1828051X.2020.1784804>.
- Bobbo, T., M. Penasa, R. Finocchiaro, G. Visentin, and M. Cassandro. 2018. Alternative somatic cell count traits exploitable in genetic selection for mastitis resistance in Italian Holsteins. *J. Dairy Sci.* 101:10001–10010. <https://doi.org/10.3168/jds.2018-14827>.
- Breiman, L. 2001. Random forests. *Mach. Learn.* 45:5–32. <https://doi.org/10.1023/A:1010933404324>.
- Cerón-Muñoz, M., H. Tonhati, J. Duarte, J. Oliveira, M. Muñoz-Berrocal, and H. Jurado-Gámez. 2002. Factors affecting somatic cell counts and their relations with milk and milk constituent yield in buffaloes. *J. Dairy Sci.* 85:2885–2889. [https://doi.org/10.3168/jds.S0022-0302\(02\)74376-2](https://doi.org/10.3168/jds.S0022-0302(02)74376-2).
- Cockburn, M. 2020. Review: Application and prospective discussion of machine learning for the management of dairy farms. *Animals (Basel)* 10:1690. <https://doi.org/10.3390/ani10091690>.
- Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20:273–297. <https://doi.org/10.1007/BF00994018>.
- Costa, A., M. De Marchi, G. Neglia, G. Campanile, and M. Penasa. 2021. Milk somatic cell count-derived traits as new indicators to monitor udder health in dairy buffaloes. *Ital. J. Anim. Sci.* 20:548–558. <https://doi.org/10.1080/1828051X.2021.1899856>.
- Costa, A., G. Neglia, G. Campanile, and M. De Marchi. 2020. Milk somatic cell count and its relationship with milk yield and qual-

- ity traits in Italian water buffaloes. *J. Dairy Sci.* 103:5485–5494. <https://doi.org/10.3168/jds.2019-18009>.
- Ebrahimi, M., M. Mohammadi-Dehcheshmeh, E. Ebrahimi, and K. R. Petrovski. 2019. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep learning and gradient-boosted trees outperform other models. *Comput. Biol. Med.* 114:103456. <https://doi.org/10.1016/j.combiomed.2019.103456>.
- Ebrahimi, E., F. Ebrahimi, M. Ebrahimi, S. Tomlinson, and K. R. Petrovski. 2018. A large-scale study of indicators of sub-clinical mastitis in dairy cattle by attribute weighting analysis of milk composition features: Highlighting the predictive power of lactose and electrical conductivity. *J. Dairy Res.* 85:193–200. <https://doi.org/10.1017/S0022029918000249>.
- Fagioli, A., and O. Lai. 2007. Mastitis in buffalo. *Ital. J. Anim. Sci.* 6(Suppl. 2):200–206. <https://doi.org/10.4081/ijas.2007.s2.200>.
- Halasa, T., K. Huijps, O. Østerås, and H. Hogewege. 2007. Economic effects of bovine mastitis and mastitis management: A review. *Vet. Q.* 29:18–31. <https://doi.org/10.1080/01652176.2007.9695224>.
- Hyde, R. M., P. M. Down, A. J. Bradley, J. E. Breen, C. Hudson, K. A. Leach, and M. J. Green. 2020. Automated prediction of mastitis infection patterns in dairy herds using machine learning. *Sci. Rep.* 10:4289. <https://doi.org/10.1038/s41598-020-61126-8>.
- Ji, B., T. Banhazi, C. J. C. Phillips, C. Wang, and B. Li. 2022. A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. *Biosyst. Eng.* 216:186–197. <https://doi.org/10.1016/j.biosystemseng.2022.02.013>.
- Kuhn, M. 2021. caret: Classification and regression training. R package version 6.0-86. <https://CRAN.R-project.org/package=caret>.
- Matera, R., A. Cotticelli, M. Gómez Carpio, S. Biffani, F. Iannacone, A. Salzano, and G. Neglia. 2022. Relationship among production traits, somatic cell score and temperature-humidity index in the Italian Mediterranean buffalo. *Ital. J. Anim. Sci.* 21:551–561. <https://doi.org/10.1080/1828051X.2022.2042407>.
- McCulloch, W. S., and W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5:115–133. <https://doi.org/10.1007/BF02478259>.
- Moroni, P., C. Sgoifo Rossi, G. Pisoni, V. Bronzo, B. Castiglioni, and P. J. Boettcher. 2006. Relationships between somatic cell count and intramammary infection in buffaloes. *J. Dairy Sci.* 89:998–1003. [https://doi.org/10.3168/jds.S0022-0302\(06\)72165-8](https://doi.org/10.3168/jds.S0022-0302(06)72165-8).
- Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. *J. R. Stat. Soc. [Ser A]* 135:370–384. <https://doi.org/10.2307/2344614>.
- Puggioni, G. M. G., V. Tedde, S. Uzzau, J. Guccione, P. Ciaramella, C. Pollera, P. Moroni, V. Bronzo, and M. F. Addis. 2020. Evaluation of a bovine cathelicidin ELISA for detecting mastitis in the dairy buffalo: Comparison with milk somatic cell count and bacteriological culture. *Res. Vet. Sci.* 128:129–134. <https://doi.org/10.1016/j.rvsc.2019.11.009>.
- R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. <https://doi.org/10.1186/1471-2105-12-77>.
- Satola, A., and E. A. Bauer. 2021. Predicting subclinical ketosis in dairy cows using machine learning techniques. *Animals (Basel)* 11:2131. <https://doi.org/10.3390/ani11072131>.
- Sharifi, S., A. Pakdel, M. Ebrahimi, J. M. Reecy, S. Fazeli Farsani, and E. Ebrahimi. 2018. Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. *PLoS One* 13:e0191227. <https://doi.org/10.1371/journal.pone.0191227>.
- Sharma, A. K., L. A. Rodriguez, C. J. Wilcox, R. J. Collier, K. C. Bachman, and F. G. Martin. 1988. Interactions of climatic factors affecting milk yield and composition. *J. Dairy Sci.* 71:819–825. [https://doi.org/10.3168/jds.S0022-0302\(88\)79622-8](https://doi.org/10.3168/jds.S0022-0302(88)79622-8).
- Sparks, A. H. 2018. nasapower: A NASA POWER global meteorology, surface solar energy and climatology data client for R. *J. Open Source Softw.* 3:1035. <https://doi.org/10.21105/joss.01035>.
- Svetnik, V., A. Liaw, C. Tong, and T. Wang. 2004. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. Pages 334–343 in *Multiple Classifier Systems*. F. Roli, J. Kittler, and T. Windeatt, ed. Springer.
- Tripaldi, C., G. Palocci, M. Miarelli, M. Catta, S. Orlandini, S. Amatiste, R. Di Bernardini, and G. Catillo. 2010. Effects of mastitis on buffalo milk quality. *Asian-Australas. J. Anim. Sci.* 23:1319–1324. <https://doi.org/10.5713/ajas.2010.90618>.
- Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. 2019. Welcome to the Tidyverse. *J. Open Source Softw.* 4:1686. <https://doi.org/10.21105/joss.01686>.
- Zhang, F., J. Upton, L. Shalloo, P. Shine, and M. D. Murphy. 2020. Effect of introducing weather parameters on the accuracy of milk production forecast models. *Inf. Process. Agric.* 7:120–138. <https://doi.org/10.1016/j.inpa.2019.04.004>.

## ORCIDS

- T. Bobbo  <https://orcid.org/0000-0003-0328-8903>
- R. Matera  <https://orcid.org/0000-0003-2204-0022>
- A. Cotticelli  <https://orcid.org/0000-0002-5279-9577>
- G. Neglia  <https://orcid.org/0000-0002-0989-6072>