



# Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models



Mansour Ebrahimi<sup>a,e</sup>, Manijeh Mohammadi-Dehcheshmeh<sup>a</sup>, Esmaeil Ebrahimie<sup>a,b,c,d,f,\*</sup>, Kiro R. Petrovski<sup>a,f,g,\*\*</sup>

<sup>a</sup> School of Animal and Veterinary Sciences, The University of Adelaide, Adelaide, 5371, Australia

<sup>b</sup> Genomics Research Platform, School of Life Sciences, Melbourne, La Trobe University, Victoria, 3086, Australia

<sup>c</sup> School of Information Technology and Mathematical Sciences, Division of Information Technology Engineering & Environment, University of South Australia, South Australia, 5095, Australia

<sup>d</sup> School of Computer Science, Faculty of Engineering, Computer and Mathematical Sciences, The University of Adelaide, South Australia, 5005, Australia

<sup>e</sup> School of Basic Sciences, University of Qom, Qom, Iran

<sup>f</sup> Australian Centre for Antimicrobial Resistance Ecology, School of Animal and Veterinary Sciences, The University of Adelaide, South Australia, Australia

<sup>g</sup> Davies Research Centre, School of Animal and Veterinary Sciences, The University of Adelaide, South Australia, Australia

## ARTICLE INFO

**Keywords:**  
Machine learning  
Mastitis

## ABSTRACT

Sub-clinical bovine mastitis decreases milk quality and production. Moreover, sub-clinical mastitis leads to the use of antibiotics with consequent increased risk of the emergence of antibiotic-resistant bacteria. Therefore, early detection of infected cows is of great importance. The Somatic Cell Count (SCC) day-test used for mastitis surveillance, gives data that fluctuate widely between days, creating questions about its reliability and early prediction power. The recent identification of risk parameters of sub-clinical mastitis based on milking parameters by machine learning models is emerging as a promising new tool to enhance early prediction of mastitis occurrence. To develop the optimal approach for early sub-clinical mastitis prediction, we implemented 2 steps: (1) Finding the best statistical models to accurately link patterns of risk factors to sub-clinical mastitis, and (2) Extending this application from the farms tested to new farms (method generalization). Herein, we applied various machine learning-based prediction systems on a big milking dataset to uncover the best predictive models of sub-clinical mastitis. Data from 364,249 milking instances were collected by an electronic automated in-line monitoring system where milk volume, lactose concentration, electrical conductivity (EC), protein concentration, peak flow and milking time for each sample were measured. To provide a platform for the application of the models developed to other farms, the Z transformation approach was employed. Following this, various prediction systems [Deep Learning (DL), Naïve Bayes, Generalized Linear Model, Logistic Regression, Decision Tree, Gradient-Boosted Tree (GBT) and Random Forest] were applied to the non-transformed milking dataset and to a Z-standardized dataset. ROC (Receiver Operating Characteristics Curve), AUC (Area Under The Curve), and high accuracy demonstrated the high sensitivity of GBT and DL in detecting sub-clinical mastitis. GBT was the most accurate model (accuracy of 84.9%) in prediction of sub-clinical bovine mastitis. These data demonstrate how these models could be applied for prediction of sub-clinical mastitis in multiple bovine herds regardless of the size and sampling techniques.

## 1. Introduction

Mastitis of the mammary gland and upper tissues in cows is a costly disease in the dairy industry worldwide, contributing to the economic

loss of approximately 100 Euros per cow [1,2]. Mastitis affects the general health of a herd and is associated with increased use of antimicrobials and associated resistance. As an example, the association of methicillin-resistant *Staphylococcus aureus* (MRSA) ST398 with clinical

\* Corresponding author. School of Animal and Veterinary Sciences, The University of Adelaide, Adelaide, 5371, Australia.

\*\* Corresponding author.

E-mail addresses: [esmaeil.ebrahimie@adelaide.edu.au](mailto:esmaeil.ebrahimie@adelaide.edu.au) (E. Ebrahimie), [kiro.petrovski@adelaide.edu.au](mailto:kiro.petrovski@adelaide.edu.au) (K.R. Petrovski).

and subclinical mastitis is well documented [3]. Therefore, developing useful new models to predict the incidence of the disease and monitor mastitis prevalence is a priority. Clinical mastitis is forty times less common than sub-clinical mastitis which is harder to detect and the lack of clinical symptoms [4].

The detection of sub-clinical mastitis is commonly performed by enumeration of somatic cells in the harvested milk (Somatic Cell Count, SCC). SCC is an indicator of the inflammatory status of mammary glands at the time of milking [5]. Other detection systems are based on testing the Electrical Conductivity (EC) of milk. EC is an indicator of the leakage of blood components into the milk that occurs during mastitis. For good prediction any test should provide data that are sufficiently stable between test days and they should reflect the profile of pathogenic prevalence. High prevalence of minor pathogens may hamper the sensitivity and specificity of SCC and EC tests and their daily variability has led to significant concerns about their reliability. Although importance of utilizing tests has been highlighted in previous studies [6], the aforementioned limitations these tests demonstrate, the urgency for developing new and more reliable tools for the prediction of sub-clinical mastitis.

Automated monitoring devices have simplified the necessity of milk sampling and analysis over successive days and have provided vital production parameters such as fat content, lactose levels and protein amount, volume collected, time and date of milking and milk peak flows. The amount of SCC can also be measured for each sample and added to the data gathered [7]. Access to large datasets detailing the quality and the number of milking samples allowed us to train machine learning algorithms in order to predict the possible occurrence of sub-clinical mastitis based on previously collected data and occurrence of mastitis.

Thus far, few studies have used machine learning to investigate mastitis data generated by automated milking systems, but instead have mainly applied multivariate analysis. Recently, machine learning tools have been employed and found useful in the prediction of many events based on already available data. They have also been used extensively in biological studies, since their pattern recognition abilities can facilitate difficulties associated with analyzing biological data bases [8]. Machine learning methods aim to discern meaningful relationships between regular and target features or variables, in order to distinguish possible cryptic patterns among them. Different prediction models use different learning or training strategies and, when they become trained, their performance or accuracy in unknown instances may be calculated. When the prediction accuracy or performance is sufficiently high (usually more than 85%), the model is said to be well-trained and suitable to be employed to predict future instances [8].

Recently, we mined a large dataset of milking parameters of a commercial herd of Friesian cows (2011–2013) from New Zealand. We used feature selection models, including Weighting by the information gain ratio, Weighting by rule, Weighting by deviation, Weighting by the chi-squared statistic, Weighting by the Gini Index, Weighting by uncertainty, Weighting by Relief, Weighting by Support Vector Machine (SVM) and Weighting by Principal Component Analysis (PCA) [9]. Outcome from this study revealed the importance of low lactose concentration and high EC of milk as an predictive pattern for early sub-clinical mastitis detection [9]. The Decision Tree Random Forest model found the hierarchical pattern of milking and could accurately predict of sub-clinical mastitis. In the pattern discovered, lactose concentration was placed at the top of the tree as the key predictor and the addition of EC and milk to the prediction tree increased the accuracy of prediction [10].

Herein, different prediction systems, Deep Learning (DL), Naïve Bayes (NB), Generalized Linear Model (GLM), Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) were applied to the cleaned automated milking dataset in order to develop the best sub-clinical mastitis prediction method based on milk composition and SCC factors. Ten-fold cross-validation was used to compute the accuracy of each prediction model.

## 2. Material and methods

### 2.1. Data collection

Data collection was performed at a commercial New Zealand dairy farm in Ongaonga, Hawkes Bay, from July 2011 to June 2013. Milk data samples were collected twice a day from 2,400 mixed-age Holstein Friesian cows and their crosses. A range of features, including milk volume (milk Vol), weight, fat, protein (Prot), lactose (Lac), EC, peak flow (Peak-flow), and milking time (Milking-time) was measured by an electronic automated monitoring system. An inline detector (Cell-Sense®) was used to measure SCC 1 min after milking as described previously [9].

### 2.2. Dataset preparation

The original dataset had 346,248 records and eight measurements (features). The prevalence of subclinical mastitis was 19.6% (68005 records out of 346248 records). The eight features (milk Vol, fat, Prot, EC, milking time, peak flow, and SCC) were measured for each milk sample. Based on the Australian definition of mastitis (SCC  $\geq 250,000$  cells/mL), the samples were classified into two groups: non-mastitis and mastitis.

Outlier detection and removal were performed by computing Euclidian distance-based outlier detection. We imported the final cleaned dataset into Rapid Miner software (RapidMiner 9.0.1, Rapid-I GmbH, Stochumer Str. 475, 44,227 Dortmund, Germany), and SCC variable was set as output (label/target) variable with the rest of variables as dependent features.

### 2.3. Z-standardization dataset

As the dataset used for prediction of sub-clinical mastitis was collected from only a single farm, to provide a reliable base for generalization of the findings of this study, the dataset was transformed by Z-Standardization (for each feature, subtracting the mean and dividing by the standard deviation). The resulting dataset was also imported into the RapidMiner and prediction models were applied to this transformed dataset, similarly to the original dataset.

### 2.4. Prediction models

The following prediction models were developed and employed to train the relationship between target or label feature (SCC with two class of Yes or No) and other dependent variables (Lac, Vol, EC, Milking-time, Fat, Prot, and Peak-flow).

#### 2.4.1. Deep Learning (DL)

This employed model was based on a multi-layer feed-forward neural network and was trained with stochastic gradient descent (SGD) using back-propagation with 50 hidden layers (including drop-out layers) with rectifier and max-out activation functions for 10 epochs and auto-tuning samples training per iteration. DL discovers structure in a big dataset by application of the backpropagation model to discover how a machine needs to alter its internal parameters to enhance the representation in each layer from representation in the previous layer [11].

#### 2.4.2. Naïve Bayes (NB)

NB is a high-bias, low-variance classifier [12]. NB is able to build a robust model even in a small dataset. It is simple to use and computationally inexpensive.

#### 2.4.3. Generalized linear model (GLM)

The GLM algorithm fits generalized linear models to the data by maximising the log-likelihood [13]. The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates.

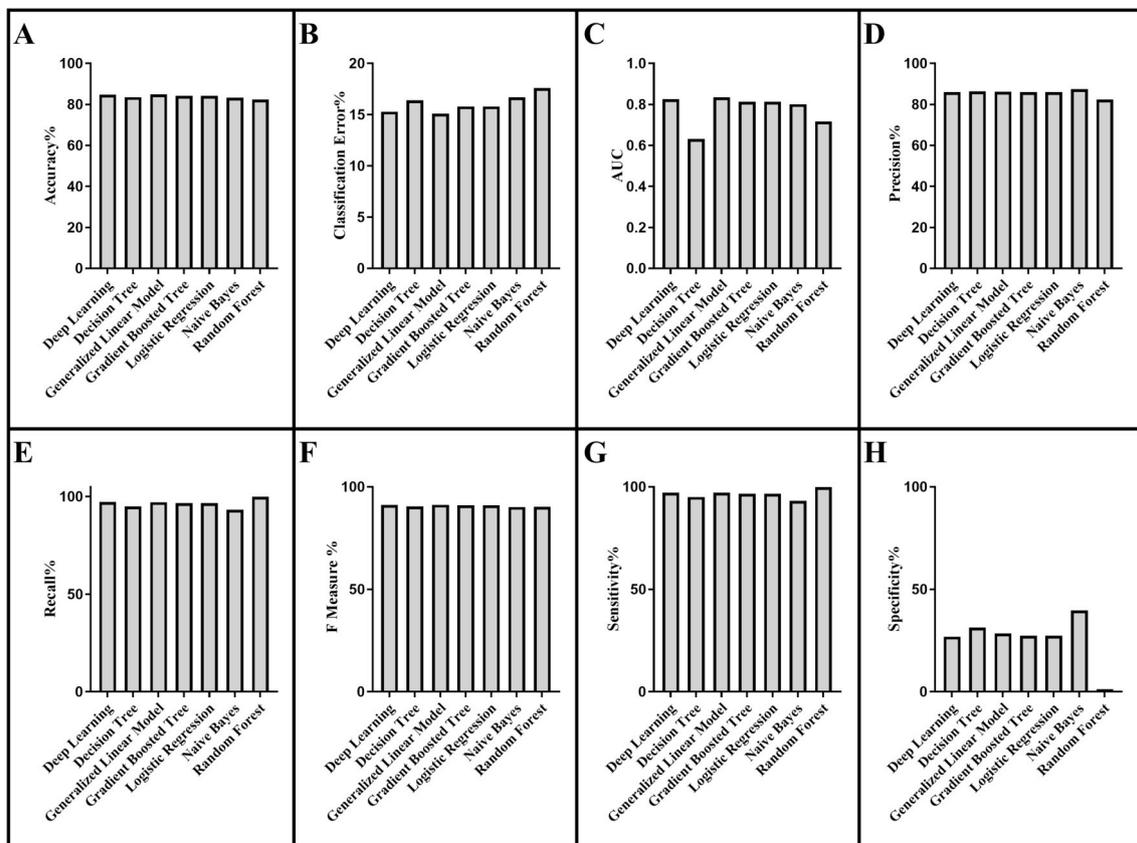


Fig. 1. Comparison of the performance of seven machine learning models in prediction of sub-clinical mastitis, run on the original dataset, in prediction of sub-clinical mastitis from in-line collected milking parameters.

The model-fitting computation is parallel, extremely fast, and scales extremely well for models with a limited number of predictors with non-zero coefficients.

#### 2.4.4. Logistic Regression (LR)

LR is a simplified version of the GLM operator. To perform LR, the Family parameter is set automatically to binomial, and the link parameter to logit. Logistic Regression implementation can handle training data with the binominal label and both nominal and numerical feature attributes [14].

#### 2.4.5. Decision tree (DT)

DT performs recursive partitioning, repeatedly splitting the values of attributes used to extract the possible patterns between the target and regular variables [15,16].

#### 2.4.6. Random Forest (RF)

RF is based on the voting model of all possible tree induction. This operator generates a specified number of random trees [17].

#### 2.4.7. Gradient Boosted Tree (GBT)

The GBT model is a combination of regression and classification tree models, such as Decision tree models. GBT improves prediction power results through progressively improving estimations. Additionally, GBT employs nonlinear regression procedure that helps in improving the accuracy of trees. A series of decision trees is created that produce an ensemble of weak prediction models. While boosting trees increases the accuracy, it also decreases speed and human interpretability. The gradient boosting method generalizes tree boosting to minimize these issues [18,19].

### 2.5. Validation and comparison of predictive algorithms

The models were trained and tested with 10-fold cross-validation, as previously described [20]. To perform ten-fold cross-validation, after shuffling, the dataset was divided into ten equally sized sub-samples. Prediction models were trained on nine sub-sample sets and the last remaining sub-sample was used as unseen data to evaluate the performance of the model for predicting the right class (here SCC class). The procedure was repeated ten times (10-folds), and the average accuracy, Classification Error, Area Under Curve (AUC), Precision, Receiver Operating Characteristic Curve (ROC), Recall, F Measure, Sensitivity, and Specificity of ten runs was calculated by dividing the percentage of correct predictions over the total number of examples.

The schematic overview of the employed computational pipeline in this study is presented in Fig. Supplementary 1. Parameters set for algorithms employed in this study are presented in Supplementary 2.

## 3. Results

### 3.1. Data cleaning

The original dataset had 346,248 milking samples, each with eight milking features. After cleaning low-quality samples and removal of outliers, 297,004 remained in the dataset. Seven different prediction algorithms described (DL, NB, GLM, LR, DT, GBT and RF) were run on the cleaned dataset and various criteria for each prediction model were calculated based on 10-fold cross-validation. A similar procedure was also applied to Z-Standardized dataset (see Materials and Methods).

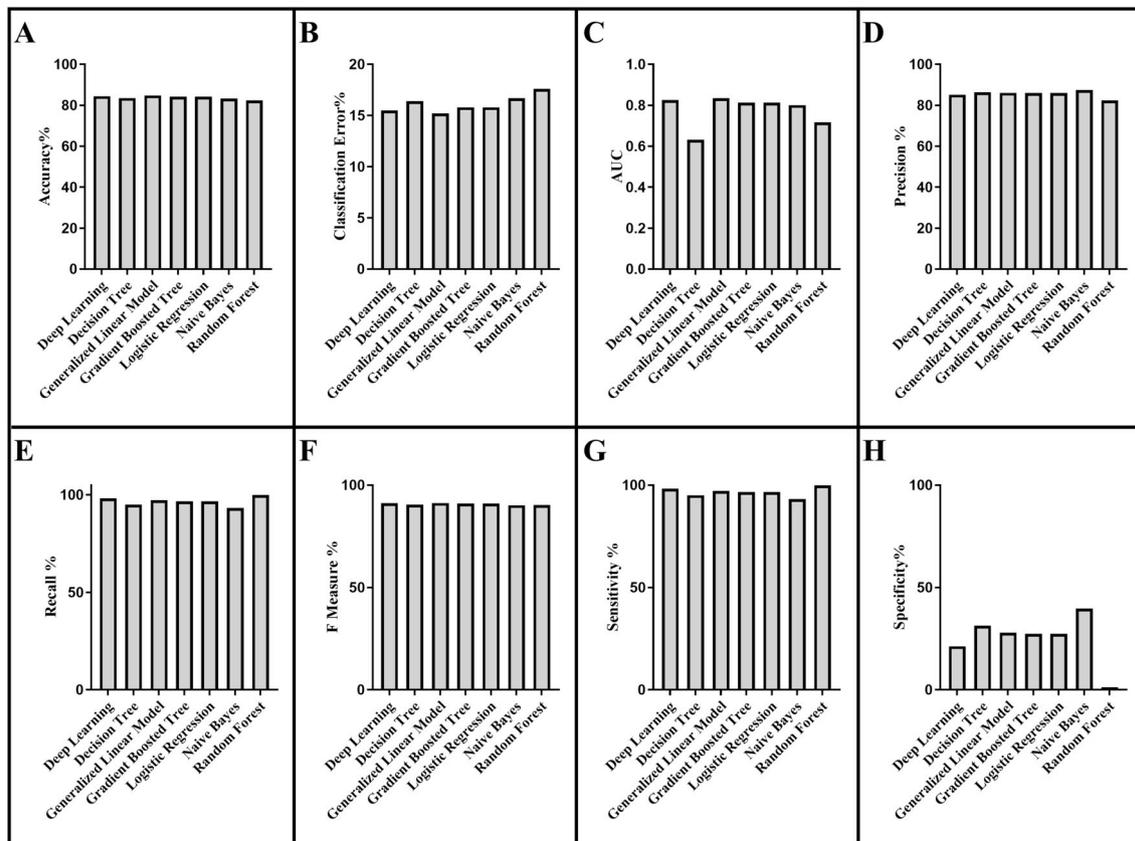


Fig. 2. Comparison of the performance of seven predictive models in prediction of sub-clinical mastitis, run on the Z-standardized dataset, in prediction of sub-clinical mastitis from in-line collected milking parameters.

### 3.2. Performance comparison between machine learning models in prediction of sub-clinical mastitis

#### 3.2.1. Accuracy

The best percentage of accuracy in predicting SCC class (mastitis or non-mastitis) was 84.9% belonging to the GBT model, followed by LR, GLM, DT, and NB. The lowest accuracy was 82.3% seen for the RF model (Fig. 1).

When the models were run on the Z-Standardized data, again GBT demonstrated the highest percentage of accuracy (84.4%) and RF showed the lowest accuracy in modelling SCC class (82.3%) (Fig. 2).

#### 3.2.2. classification error

As this criterion makes calculation based on the model accuracy, the error rate for RF was higher than that of the other models (17.6%), whereas the lowest error computed was for the DL model (Fig. 1). Similarly, the classification error was higher for RF when the data were Z-Standardized. The lowest percentage of Classification Error was seen for GBT (Fig. 2).

#### 3.2.3. AUC

In the non-transformed dataset, DL showed the highest AUC value (0.826), followed by GBT and LR (0.813) (Fig. 1). DT and RF showed remarkably lower AUC with values of 0.632 and 0.716 respectively (Fig. 1).

When the dataset was Z-Standardized, the order of GLM, DL, GBT/LR, NB, RF, and DT was observed for decreasing AUC (Fig. 2).

#### 3.2.4. Precision

For SCC prediction, the best precision was gained by using NB model (87.4%). The precision percentages of GLM, LR, DL, DT, and GBT were close to NB (in a range between 86% and 86.4%). The lowest precision

was 82.3% observed in the RF model (Fig. 1).

In the Z-Standardized dataset, the highest precision of 87.4% remained for NB, and the lowest precision of 82.3% was still observed for the RF model (Fig. 2).

#### 3.2.5. Recall

In the prediction of SCC using non-transformed data, the percentage of recall for all prediction models were higher than 93%. The highest value of 99.9% was obtained for RF (Fig. 1). The same percentages and order were computed for recall in the Z-Standardized data (Fig. 2).

#### 3.2.6. F measure

The F measure indices were higher than 90% for all models, in both non-transformed and Z-Standardized datasets (Figs. 1 and 2).

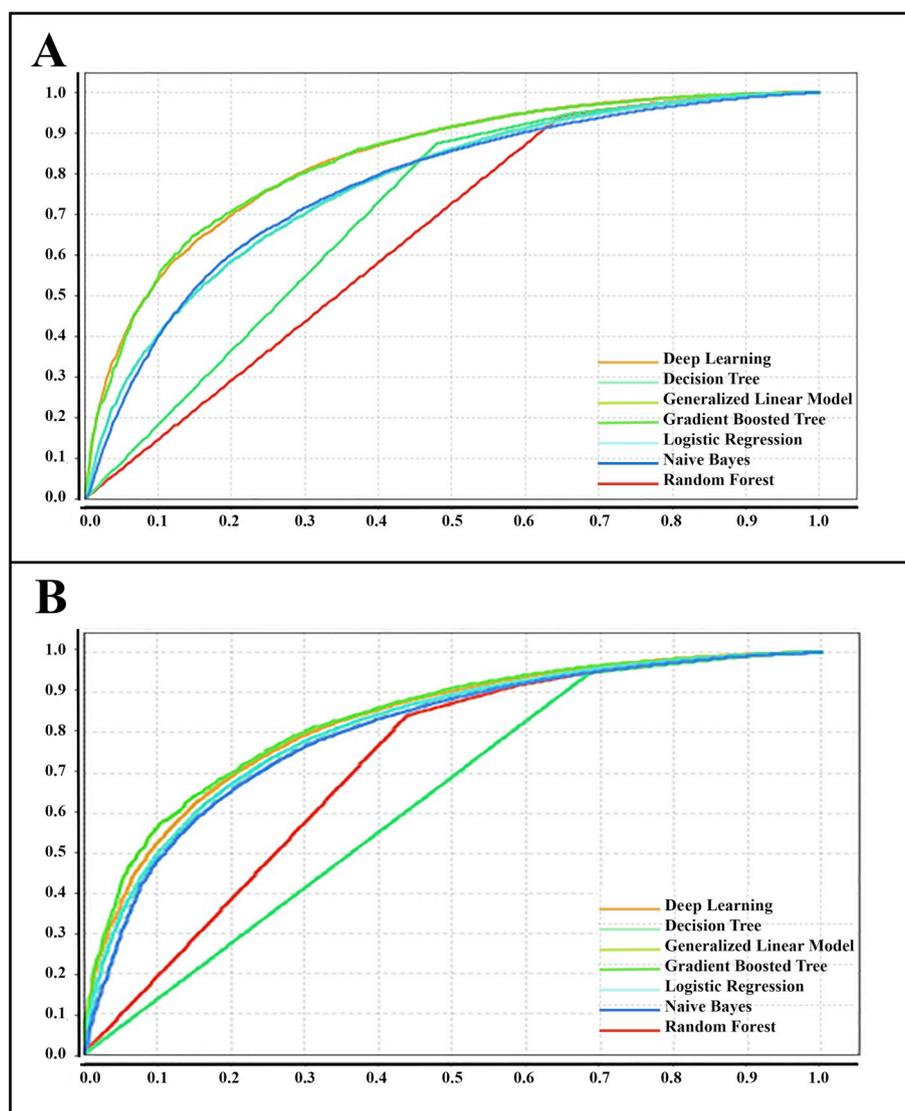
#### 3.2.7. Sensitivity

Noticeable high sensitivity (>93%) was observed for all models that document high distinguishing power of machine learning models in the correct identification of SCC. The RF model performed the best with the highest sensitivity of 99.9% (Fig. 1). The same pattern was observed when the models were run on the Z-Standardized dataset (Fig. 2).

#### 3.2.8. Specificity

As presented in Figs. 1 and 2, the models were diverse in defining specificity. Compared with the other performance criteria, machine learning models received lower power in the identification of healthy samples. In the non-transformed dataset, NB showed the highest specificity of 39.7% and RF was the lowest with 1.2% (Fig. 1).

In the Z-Standardized dataset, the best and worst percentage of specificity were the same computed to the non-transformed dataset (39.7% specificity for NB and 1.2% specificity for RF).



**Fig. 3.** Comparing Receiver Operating Characteristic (ROC) curves of seven machine learning models in prediction of sub-clinical mastitis, run on non-transformed (A) and Z-Standardized (B) datasets of in-line collected milking parameters.

### 3.2.9. ROC curve

Fig. 3 shows ROC comparisons of seven prediction models. As can be inferred in Fig. 3, the best ROC curves belong to GBT and DL while the worst one belonged to RF.

Lift Charts comparison between machine learning models in prediction of sub-clinical mastitis.

Lift charts of DL, DT, GLM, GBT, NB, and RF prediction models on non-transformed and Z-Standardized milking datasets are shown in Fig. 4 and Fig. 5. The x-axis of each chart represents the percentage of the test data, and the y-axis represents the percentage of predicted values. The Lift charts demonstrate that DL, GBT, and GLM generated higher rates of accurate prediction, using the smaller size of the dataset.

### 3.3. Overall weights of milking features to predict sub-clinical mastitis

Overall weights of milking parameters in the prediction of sub-clinical mastitis based on DL, DT, GLM, GBT, NB, LR, and RF are presented in Fig. 6. For each milking variable, a weight between zero to one was computed based on overall weight assumption by predicting models (using the non-transformed milking dataset). Values closer to 1 show higher importance of that particular milking parameter.

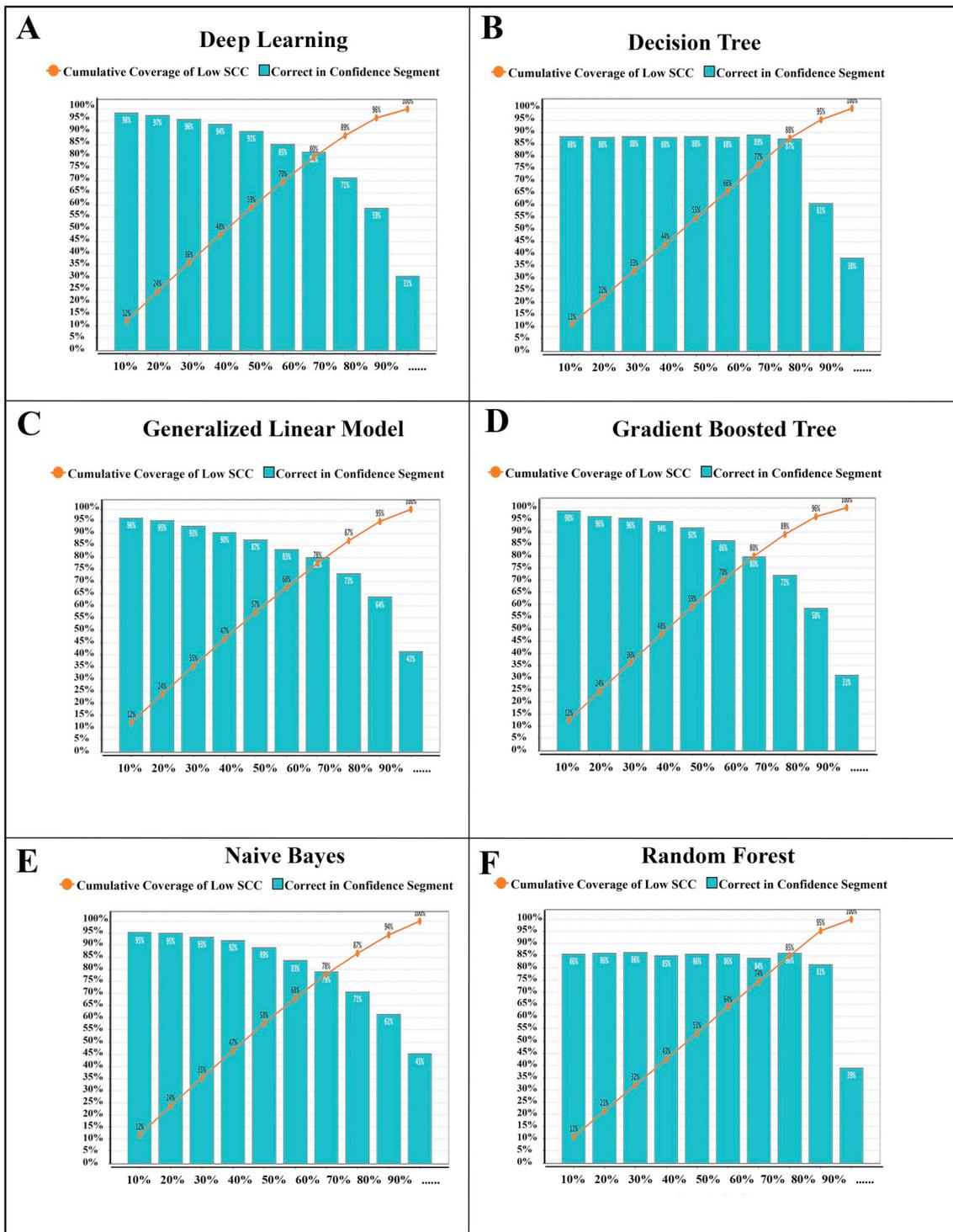
The highest weight was assigned to EC (equal to 1), followed by

lactose (0.59). The lowest weights belonged to milking time and protein content (Fig. 6). Milk volume was described as more important than milking time. Similar results obtained using Z-Standardized dataset (data not shown).

## 4. Discussion

Advanced technologies have provided an opportunity to collect large quantities of data on a daily basis using automated milking sensors and facilities. Generation of large dataset necessitates new and efficient data mining techniques to efficiently analyze data and extract valuable knowledge [21]. In addition to measuring general milk composition features, the numbers of SCC, as an index of mastitis, are also being determined for each milk sample. Having access to these valuable datasets provides a new opportunity to examine various data mining models for the development of predictive tools [22].

Recently, we showed that mastitis prevalence could be predicted by hierarchical pattern recognition in milking variables via RF models, to distinguish healthy cows from those with mastitis [10]. In our previous study, RF models, as well as attribute weighting models, showed high importance of Lactose content, EC and milk volume in the prediction of sub-clinical mastitis [9,10].



**Fig. 4.** Lift charts of Deep Learning, Decision Tree, Generalized Linear Model, Gradient Boosted Tree, Naive Bayes and Random Forest models in prediction of sub-clinical mastitis based using the non-transformed milking dataset.

Here, we performed a comprehensive study evaluating a range of different machine learning models for their ability to predict sub-clinical mastitis. We used milking features of a large dataset (346,248 milking records). Seven most well-recognised prediction algorithms (Deep Learning, Naïve Bayes, Generalized Liner Model, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosted Tree) were applied to the original dataset and to a Z-standardized dataset. Following this, the performance of each prediction model was then evaluated using 10-fold cross-validation.

Individual source (farm) dependency is a significant concern in predictive modelling. This highlights the importance of developing accurate models to run on data collected from different farms. To check whether the trained predictive algorithms in this study can be generalized and applied to the other farms, we prepared a new dataset by normalizing the data (Z-standardization). This method has the potential to eliminate batch effects or farm specificity in prediction performances. All the models developed were trained and tested on both regular and Z-standardized datasets. Interestingly, no significant difference was found

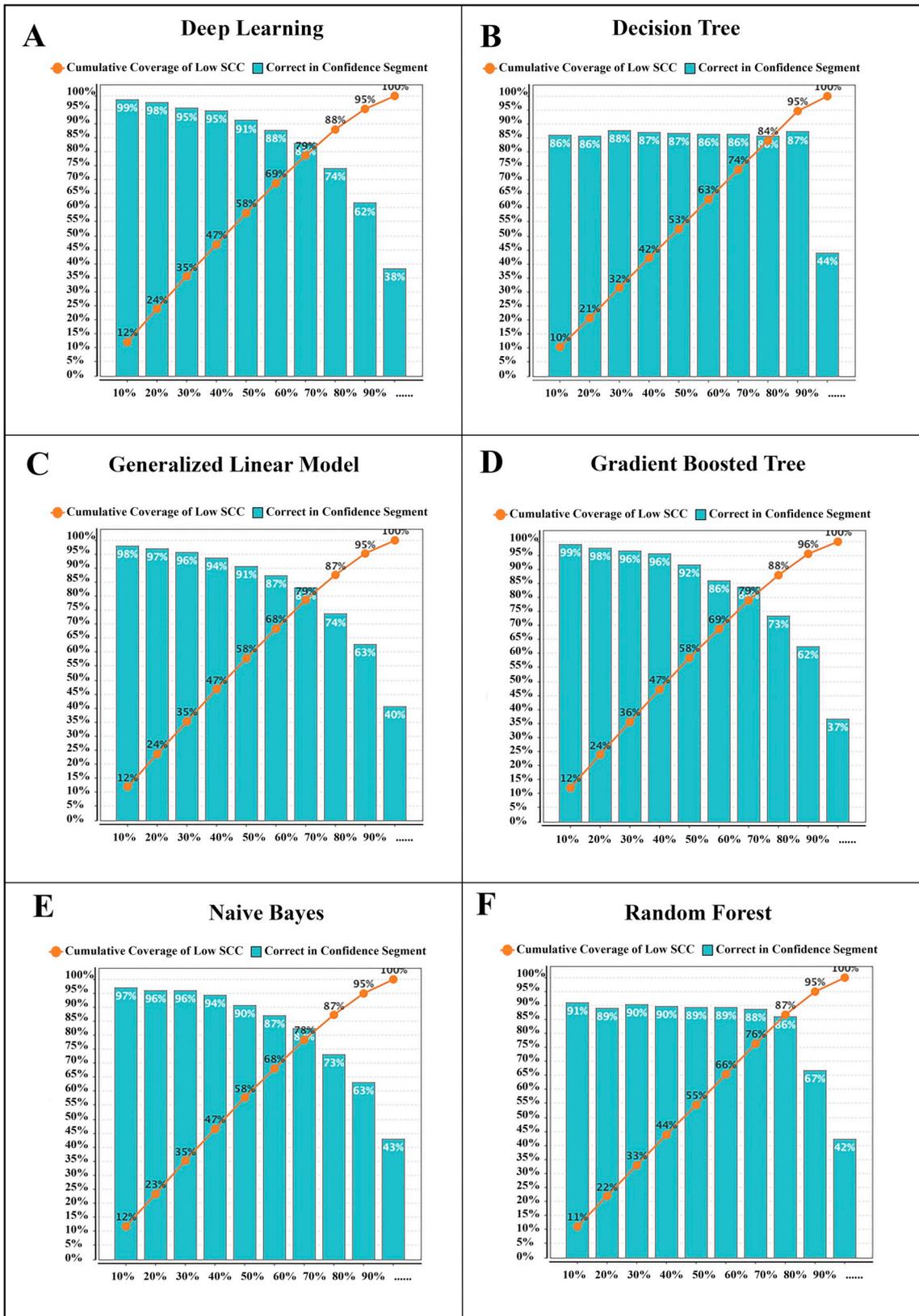
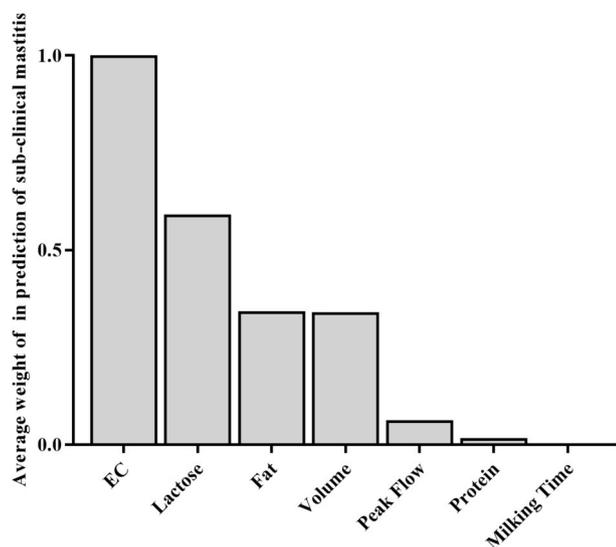


Fig. 5. Lift charts of Deep Learning, Decision Tree, Generalized Linear Model, Gradient Boosted Tree, Naïve Bayes and Random Forest models in prediction of sub-clinical mastitis, run using the Z-Standardized milking dataset.



**Fig. 6.** Overall weights of milking parameters in prediction of sub-clinical mastitis based on Deep Learning, Decision Tree, Generalized Linear Model, Gradient Boosted Tree, Naïve Bayes, Logistic Regression and Random Forest models. For each milking variable, a weight between zero and one was computed based on overall weight assumption by predicting models (Normalised weights vary between 0 and 1). Values closer to 1 show higher importance of a milking parameter.

between the outcomes of the non-transformed dataset and Z-standardized dataset. Therefore, it can be expected that the proposed models in this study are generally applicable to datasets from different farms that have appropriate collection facilities and are capable of producing similar milking data. We hypothesize that the difference between non-transformed and Z-standardized datasets were most likely non-significant due to the availability of data at various times of the year, accounting for calving, drying off and environmental stress, all of which can impact on mastitis incidence.

For assessing the quality of predictive models, accuracy is often the first notable criterion, since it measures the ratio of correct predictions to the total number of cases evaluated. GBT and DL were the most accurate models for predicting sub-clinical mastitis based on milking parameters, having a high accuracy of about 85% in both non-transformed and Z-standardized datasets. Adaptive learning rate, rate annealing, momentum training, dropout and layer regularisation enable high predictive accuracy of DL. The lowest accuracy was obtained for RF. Lift chart analysis also showed that GBT and DL maintain the high level of prediction accuracy even with the smaller sample size of the dataset. Consequently, GBT and DL would be methods of choice for extending these results to other farms with smaller sized records.

The ROC curve shows the relationship between sensitivity (False Positive Rate) and specificity (True Positive Rate) and is one of the best graphical ways to compare the performances of prediction models. GBT along with DL showed the best area under the curve in predicting the true positive rate against false positive rate, supporting GBT and DL as the most robust and efficient models in predicting sub-clinical mastitis.

In line with our previous findings [9], we herein observed that EC and Lactose concentration have the highest weights in the prediction of sub-clinical mastitis. Bacterial infection alters the ionic composition of milk both in dairy cows and goats. EC-based automation system is proposed as a cheap tool for fast diagnosis of mastitis [23,24] as well as excellent tool for herd health monitoring in both dairy cows and goats [24–26]. EC peak signals have been shown to have a high association with the health status of the herd [26]. EC changes in harvested milk are caused by leakiness of the mastitic mammary glands. The decrease in lactose levels of milk from infected cows can be linked to subclinical mastitis, as confirmed by previous study [9]. Lactose is produced

exclusively by mammary glands and its concentrations are relatively stable.

#### 4.1. Sensitivity

The observed high sensitivity (>93%) of all employed machine learning models in this study demonstrates the high distinguishing power of these models in the reliable identification of sub-clinical mastitis. However, general low specificity of the models tested in this study demonstrates that machine learning models harbor lower power to identify healthy samples. This limitation needs to be addressed in future studies by integrative models such as meta-analysis from different farms and machine learning models [8,27].

## 5. Conclusion

For the first time, this study shows that the two prediction models (GBT and DL) are able to accurately forecast sub-clinical mastitis based on multiple milking variables, confirming the power of machine-based prediction in solving one of the oldest problems in dairy cattle industry. As automated data collection on dairy farms is becoming routine, developing, training and testing such predictive models is applicable to identify infected cows, thereby minimizing economic losses. Similar results with the Z-standardization dataset show the opportunity of extending the outcomes to other farms.

#### Conflicts of interest

The authors declare that there is no conflict of interest.

#### Acknowledgment

This research was supported by use of the Nectar Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS). Also, this study was supported by School of Animal and Veterinary Sciences of The University of Adelaide, based on the research grant awarded to A/Professor Kiro Petrovski.

Authors would like to thank farm owners and staff for their cooperation and provision of data. We would like to greatly thank Emeritus professor Jeremy Timmis (School of Biological Sciences, The University of Adelaide) and Dr. Jean Winter (Adelaide Medical School, The University of Adelaide) for editing of the manuscript.

#### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.combiomed.2019.103456>.

#### References

- [1] K.R. Petrovski, M. Trajcev, G. Buneski, A review of the factors affecting the costs of bovine mastitis, *J. S. Afr. Vet. Assoc.* 77 (2006) 52–60.
- [2] H. Seegers, C. Fourichon, F. Beaudeau, Production effects related to mastitis and mastitis economics in dairy cattle herds, *Vet. Res.* 34 (2003) 475–491.
- [3] W. Vanderhaeghen, T. Cerpentier, C. Adriaensen, J. Vicca, K. Hermans, P. Butaye, Methicillin-resistant *Staphylococcus aureus* (MRSA) ST398 associated with clinical and subclinical mastitis in Belgian cows, *Vet. Microbiol.* 144 (2010) 166–171.
- [4] D. Watson, M. McColl, H. Davies, Field trial of a staphylococcal mastitis vaccine in dairy herds: clinical, subclinical and microbiological assessments, *Aust. Vet. J.* 74 (1996) 447–450.
- [5] A. Henderson, C. Hudson, A.J. Bradley, V. Sherwin, M.J. Green, Prediction of intramammary infection status across the dry period from lifetime cow records, *J. Dairy Sci.* 99 (2016) 5586–5595.
- [6] S.C. Archer, A.J. Bradley, S. Cooper, P.L. Davies, M.J. Green, Prediction of *Streptococcus uberis* clinical mastitis risk using Matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) in dairy herds, *Prev. Vet. Med.* 144 (2017) 1–6.

- [7] D.B. Jensen, H. Hogeveen, A. De Vries, Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis, *J. Dairy Sci.* 99 (2016) 7344–7361.
- [8] S. Sharifi, A. Pakdel, M. Ebrahimi, J.M. Reecy, S.F. Farsani, E. Ebrahimie, Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle, *PLoS One* 13 (2018), e0191227.
- [9] E. Ebrahimie, F. Ebrahimi, M. Ebrahimi, S. Tomlinson, K.R. Petrovski, 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 (2018) 193–200.
- [10] E. Ebrahimie, F. Ebrahimi, M. Ebrahimi, S. Tomlinson, K.R. Petrovski, Hierarchical pattern recognition in milking parameters predicts mastitis prevalence, *Comput. Electron. Agric.* 147 (2018) 6–11.
- [11] Y. LeCun, Y. Bengio, G.J.N. Hinton, *Deep learning* 521 (2015) 436.
- [12] D.D. Lewis, Naive (Bayes) at Forty: the Independence Assumption in Information Retrieval, *European Conference on Machine Learning*, Springer, 1998, pp. 4–15.
- [13] J.A. Nelder, R.W.M. Wedderburn, Generalized linear models, *J. R. Stat. Soc. Ser. A* 135 (1972) 370–384.
- [14] D.G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, *Logistic Regression*, Springer, 2002.
- [15] A.A. Jamali, R. Ferdousi, S. Razzaghi, J. Li, R. Safdari, E.J.D.d.t. Ebrahimie, DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins 21 (2016) 718–724.
- [16] S.R. Safavian, D. Landgrebe, cybernetics man, A Survey of Decision Tree Classifier Methodology, *IEEE Trans. Syst. Man Cybern. B Cybern.* 21 (1991) 660–674.
- [17] L.J.M.I. Breiman, *Random Forests* vol. 45, 2001, pp. 5–32.
- [18] L. Lombardo, M. Cama, C. Conoscenti, M. Märker, E.J.N.H. Rotigliano, Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy) 79 (2015) 1621–1648.
- [19] J. Ye, J.-H. Chow, J. Chen, Z. Zheng, Stochastic gradient boosted distributed decision trees. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ACM, 2009, pp. 2061–2064.
- [20] M.R. Bakhtiarzadeh, M. Moradi-Shahrbabak, M. Ebrahimi, E. Ebrahimie, Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology, *J. Theor. Biol.* 356 (2014) 213–222.
- [21] K.P. Gaddis, J. Cole, J. Clay, C. Maltecca, Benchmarking dairy herd health status using routinely recorded herd summary data, *J. Dairy Sci.* 99 (2016) 1298–1314.
- [22] I. Santman-Berends, J. Swinkels, T. Lam, J. Keurentjes, G. van Schaik, Evaluation of udder health parameters and risk factors for clinical mastitis in Dutch dairy herds in the context of a restricted antimicrobial usage policy, *J. Dairy Sci.* 99 (2016) 2930–2939.
- [23] J. Díaz, G. Romero, R. Muelas, M. Alejandro, C. Peris, Effect of intramammary infection on milk electrical conductivity in Murciano-Granadina goats, *J. Dairy Sci.* 95 (2012) 718–726.
- [24] A. Gáspárdy, G. Ismach, Á. Bajcsy, G. Veress, S. Márkus, I. Komlósi, Evaluation of the on-line electrical conductivity of milk in mastitic dairy cows, *Acta Vet. Hung.* 60 (2012) 145–155.
- [25] C. Kamphuis, D. Pietersma, R. Van der Tol, M. Wiedemann, H. Hogeveen, Using sensor data patterns from an automatic milking system to develop predictive variables for classifying clinical mastitis and abnormal milk, *Comput. Electron. Agric.* 62 (2008) 169–181.
- [26] M. Zaninelli, L. Rossi, A. Costa, F.M. Tangorra, A. Agazzi, G. Savoini, Use of electrical conductivity sensors to monitor health status and quality of milk in dairy goats, *Int. J. Health. Anim. Sci. Food Saf.* 2 (2015).
- [27] M. Farhadian, S.A. Rafat, K. Hasanpur, M. Ebrahimi, E. Ebrahimie, Cross-species meta-analysis of transcriptomic data in combination with supervised machine learning models identifies the common gene signature of lactation process, *Front. Genet.* 9 (2018).