Annales Mathematicae et Informaticae
Accepted manuscript
DOI: 10.33039/ami.2024.02.004
URL: https://ami.uni-eszterhazy.hu

Predicting somatic cell count in milk samples using machine learning*

Bence Tarr[a], István Szabó[a], János Tőzsér[b]

[a]Institute of Technical Sciences,
Hungarian University of Agriculture and Life Sciences,
Gödöllő, Hungary
tarr.bence.gyula@uni-mate.hu
szabo.istvan.prof@uni-mate.hu

[b]Department of Animal Science, Albert Kázmér Faculty,
Széchenyi István University,
Mosonmagyaróvár, Hungary
tozser.janos@ga.sze.hu

Abstract. Milk quality is an important factor both for the farmers to be able to sell their products and for the milk industry to be able to plan its production based on quantity and quality. Milk quality has a direct link with cow health, more specifically with utter health. One of the most common utter diseases is mastitis. It always captures a lot of interest based on its frequency and cost as a dairy disease which eventually leads to an involuntary and premature culling of milking cows and decreased milk yield. The genetic evaluation of mastitis is very difficult as it is a low heritable trait and categorical in nature [2]. That is why it is necessary to find markers that could predict the occurrence of mastitis. One of the widely used such markers is the somatic cell count (SCC) [9] which is considered to be the most suitable indicator trait for mastitis resistance given its medium to high genetic correlation with mastitis and its greater heritability than mastitis. The SCC is also easy to record in the practice. The selection for lower SCC in milk has a positive effect on the incidence of mastitis. The selection against high SCC also does not deteriorate the immune system of cattle and decreases the risk of infection at the same time. The genetic evaluation [1] of this trait is mostly based on somatic cell score (SCS), a logarithmic transformation of SCC to achieve normality of distribution. In our study, we used the milk database of Holstein cows from 3 different farms. From each farm, we had

*Special thanks to Laszló Dégen and Attila Monostori at ÁT Kft. for supporting this research.

Submitted: July 27, 2023
Accepted: February 8, 2024
Published online: February 9, 2024

altogether 8000 samples tested. The samples were analyzed using chemical methods every month for a year. 11 different types of data were recorded from each sample. Our aim was to find the best mixture of recorded data that would predict the value of linearized somatic cell count. After the logarithmic linearization the SCC results were divided into 3 main groups (based on the probability of mastitis). Thus our prediction problem turned into a classification problem. We used machine learning to train our algorithm. We experimented with different types of classification methods and found good results for the prediction of SCC in milk samples. We changed the input variables as not all the 9 measured input variables will be necessary for good prediction results. Our preliminary results show that using machine learning it is possible to build a model that can be used to predict mastitis in dairy cows based on variables generally analyzed during milk quality checking tests.

Keywords: somatic cell count, machine learning, prediction using AI, milk quality

AMS Subject Classification: 92-02

1. Introduction

Precision agriculture is no more a question of the future but the reality of today. The ever-growing demand for more food, better animal husbandry and less environmental impact puts a high stress on agriculture experts to find better ways to optimize production. The importance of the quality of cattle breeding is growing as the demand for better food and better milk is emerging. Monitoring the quality of milk is important for several reasons: overall animal health, milk quantity, and manufacturing of dairy products. So monitoring milk quality and finding new tools to help farmers plan their production is of greater importance than ever. [12] There are already existing methods for milk quality monitoring. However the most accurate solution to this is laboratory testing which is done usually monthly. This method can not be than on farms and takes time. Since laboratory testing had been a standard monitoring method for many years, there are a lot of historical data available to work with. Milk consists of several important food constituents like fats, proteins, carbohydrates, several minerals, and vitamins. The quantity of these constituents will determine the quality of the milk as well as provide information about the animal's health. We collected the milk data from three different farms. There already exists mathematical methods for how laboratories analyze their data. The focus of milk data analysis is to find milk samples which reflect a possible health problem in the animal. When a possible infected milk sample is found, then the cow where the sample came from must be checked by an expert for further possible medication. Somatic cell count is the most frequently used indicator of subclinical mastitis in dairy cattle. The most important cause of increased SCC is a bacterial infection of the mammary gland [7]. Other factors influence SCC like age, stage of lactation, season, stress, food and many others. These factors are considered less important than bacterial factors. Milk samples

for SCC analysis as part of DHI programs are routinely collected at milking time. There are several factors which influence somatic cell counts and the high value not necessarily mean an animal with a real clinical problem. Usually, we consider infected milk to have an SCC value of 300,000 and 250,000 cells/mL [11]. Apart from an infection, several other factors to consider to evaluate high SCC values are lactation days, number of calving and also on-site weather circumstances. An easy-to-use, on-site prediction method for SCC value in milk yields great economic importance. Mastitis if realized in time can be cured faster, and cheaper with using much less medication.

2. Method

In this study our aim was to prove that with a carefully selected machine learning model, it is possible to build an algorithm that can predict SCC based on other milk constituent values that are easier to monitor. Input variables were treated before creating the training and testing databases. A few input parameters were transformed into categorical values for biological reasons and to help better model creation. Lactation days (0:n<100, 1:100<n<200, 2:n>200) were categorized into three categories. The number of calving was also divided into three categories: 1,2,3+. As a first step general data cleaning steps were made. We deleted all outlier and zero values from our database. We developed a software environment where all steps of data cleaning, transformation, modification and the selection teaching and testing of the model can be done automatically in a single workflow. Our solution was written in Python using Pandas and scikit-learn modules. Since SCC values have a huge variance we will certanly need to convert them to a better usable categorical output variable, so we will be able to use tree-based models. Our dependent variable (linear SCC score) was transformed to create 3 categories, thus we needed tree based algorithms that are suitable for multi-class classification. Using the logarithmic linearization method we created three SCC groups. We denoted the groups with 0, 1 and 2 values each of them corresponding to healthy, possible infected and infected categories. The distribution of the SCC categories can be seen in Table 1.

Table 1. Instances of our output variable.

Category	**No. of instances**
0	17 500
1	5 200
2	2 600

As we expected we were facing a very unbalanced output dataset. This is due to the fact that in the selected milk samples the majority of the animals were healthy with a relatively low number of possible infected milk samples. We had to find workarounds to balance categories to be able to build a better model. We

were considering two solutions to balance our dataset before teaching our model: The first one is Up-sampling of the dataset. In this case when we duplicated the underrepresented categories of the dependent variable and at the same time we altered the input variables. We used a random multiplication factor for each of the input variables (a factor of 0.98-1.02). To create a fully balanced dataset we also applied another solution to the same dataset. Additional farm data was used to make our dataset bigger, but in this case only the under-represented values were inserted into the final database. After all these treatments to the experimental dataset, we reached a very balanced dataset which was suitable for model creation using multi-class machine learning algorithms. To find the best prediction model we had to decide which ML algorithms would be used for the experiment. Based on our practice with other datasets 4 ML methods were selected for deeper testing: Random Forest, Support Vector Method, Decision Tree Regressor and Extra Trees Classifier [10]. We created two subsets from our data: one dataset for training and one dataset for testing. Our dataset was divided into two parts on an 80% (training) and 20% (testing) basis, the entries were selected randomly from the original dataset. The distribution of the output variable in both datasets was the same as in our original balanced dataset. To find the suitable input variables which will give us the best result for prediction we used the correlation matrix. Several trial runs were made to select the most suitable features.

3. Experiment

3.1. Dataset

The initial dataset covers 3 years (2019–2021) of laboratory data from 3 different farms. Farmers generally use monthly laboratory checking to monitor their milk quality so all together we used 36 months of data from each herd. The size of the dataset was 25000 measurements and each measurement resulted in 11 different parameters for each milk sample. Considering biological, chemical and statistical methods we focused on the following parameters: casein, lactoferrin, somatic cell count (SCC), proteins and milk fat. In this study our goal was to create a prediction model for somatic cell count, we have to examine the SCC variable at the first step. The statistical description of the SCC variable in our dataset can be seen in Table 2.

3.2. Somatic Cell count

Somatic cells (SCC) found in cow milk are a mixture of milk-producing cells and immune cells. SCC value can be used for estimating mammary health and milk quality [3] as cells are secreted into the milk generally. SC is related to mastitis as its main role is to fight infection and repair damaged tissues. Milk somatic cell counts (SCCs) are widely used as a marker to monitor the milk quality and animal health in dairy herds. The SCC is categorized based on the number of cells per ml of milk.

Table 2. Statistical evaluation of SCC values in our dataset.

No. samples	26 6686
Mean	266686
Std.	1245000
Min	2000
25%	50000
50%	150000
75%	401000
Max	900000

- If SCC $<$ 100,000 then it is considered an 'uninfected' cow
- If SCC $>$ 100,000 but $<$ 300,000 the cow will need special attention
- If SCC $>$ 300,000 it means Cows are highly likely to be infected

In real life the variance of SCC is large we need to convert these values into another scale for better prediction results. For this purpose the linearized somatic cell count number is used [6]. For analysis of the SCC variable on the test day, the transformation we used the following formula [5]:

$$\mathrm{SCC}t = [\log 2(\mathrm{SCC}/100{,}000)] + 3$$

The result of SCC conversion into categorical numbers can be seen in Table 3.

Table 3. Linear scores after SCC logaritmization.

Category Number	**SCC**	**Category Number**	**SCC**
1	25 000	6	800 000
2	50 000	7	1 600 000
3	100 000	8	3 200 000
4	200 000	9	6 400 000
5	400 000	–	–

Linear SCC scores were divided into 3 groups to create the final categorical variables to be predicted (not infected = 0, possible infection = 1 and infected = 2). These will be the values of the dependent variable for the model to predict.

3.3. Lactoferrin

Careful selection of input variables is the key to finding the best prediction model. Lactoferrin (Lfe), is an iron-binding glycoprotein. It plays a key role in the defence

mechanisms of the mammary gland, contributing to the prevention of microbiological infectious diseases. Lactoferrin also may limit the oxidative degeneration of cellular components during inflammation and involution of the mammary gland. Lfe concentration in milk was significantly associated with somatic cell count (SCC) in laboratory experiments [4]. However, in real life the correlation between Lfe and SCC was not so strong, animals with high Lactoferrin rarely had mastitis. So the problem to be solved was to find other parameters that together with Lfe can be used to predict SCC.

4. Verification

To benchmark our model's real-life accuracy we calculated the accuracy of the model separately for all categories and then calculated an average accuracy to validate our results. The accuracy score function we evaluated for our results computes the accuracy as the count of correct predictions. In the case of multi-label classification, this function returns the subset accuracy. In case the predicted categories for a sample match with the original measured set of categories, then the subset accuracy is 1; otherwise it is 0. For calculating the accuracy in multi label classification the following outcomes are defined:

- False positives (FP): This is when a classifier predicts a label that does not match the input data.
- False negatives (FN): This is when a classifier misses a label that exists in the input data.
- True positives (TP): This is when a classifier correctly predicts the existence of a label.
- True negatives (TN): This is when a classifier correctly predicts the inexistence of a label.

Accuracy is the proportion of examples that were correctly classified. It is the sum of the number of true positives and true negatives, divided by the number of examples in the dataset.

$$\texttt{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{FP} + \mathrm{FN} + \mathrm{TP} + \mathrm{TN}}.$$

We checked the accuracy of the model independently for all three categories, we assumed that the average accuracy is calculated as the average of the three sub-results. A good model that would be suitable for real-life usage should give prediction success of over 80% and an over 90% success rate our model would be accurate enough to use instead of laboratory measurements.

5. Results

We run 10 different combinations of input variables on all 4 ML algorithms. To find the best combination of input variables, we used biological and statistical tools. Based on biological and chemical observations Lactoferrin (LF) and Protein were included in our input mix. For other candidates, we calculated the correlation between the variables and used these results as a hint for creating the input mixture. We have created a correlation heatmap of our variables. The correlation heatmap displays the correlation between multiple variables as a color-coded matrix. It's like a color chart that shows us how closely related different variables are. In the correlation heatmap, each variable is represented by a row and a column, and the cells show the correlation between them. The color of each cell represents the strength and direction of the correlation, with darker colors indicating stronger correlations. Our correlation heatmap in terms of our output variable (SCC) can be seen in Figure 1 .

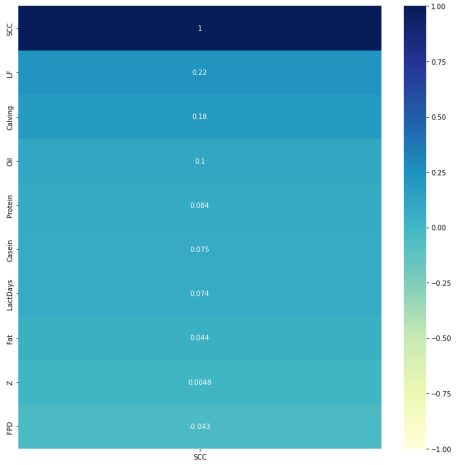

Figure 1. Heatmap of all possible variables. Showing the correlations.

The best model result used the following input variables: LNPC, Z, Urea, SUGAR, LF, FAT, PROTEIN. The calculations for average accuracy and the accuracy for each separate category (positive TRUE values) are shown in Table 4.

As we can see all selected algorithms gave us a true value of 80% or above. The best algorithm seems to be the Extra Trees Classifier. We used mainly tree-based models primarily due to their robust nature and because they can be used on any type of data (categorical/continuous), can be used on data that is not normally distributed, and require little data transformations. ExtraTrees is an ensemble machine learning model that trains several decision trees and aggregates the results from the group of decision trees to output a prediction. It is also

Table 4. Accuracy results of different models in this study.

ML algorithm	LSCC=0	LSCC=1	LSCC=2	Average
Random Forest	0.88	0.86	0.85	0.86
SVM	0.86	0.85	0.80	0.84
Decision Tree Regressor	0.83	0.80	0.82	0.82
Extra Trees Classifier	0.89	0.88	0.86	0.88

called an extremely randomized tree since to ensure sufficient differences between individual decision trees, it randomly selects the values where it splits a feature and creates child nodes [8]. It would be very interesting to see which input variable (which milk constituent) plays a greater role in the best algorithm to predict the SCC value. As from biological considerations, it is not yet clear which parameters will predict the future change in SCC. So we further investigated the algorithm in the best model. To find the importance of each input feature in our model we used scikit-learn's feature importance investigation tools. Our feature importance technique assigns a score to input features based on how important they are in the model that predicts the target variable. Investigating our best model, we calculated the feature score for each input variable. The result can be seen in Figure 2.

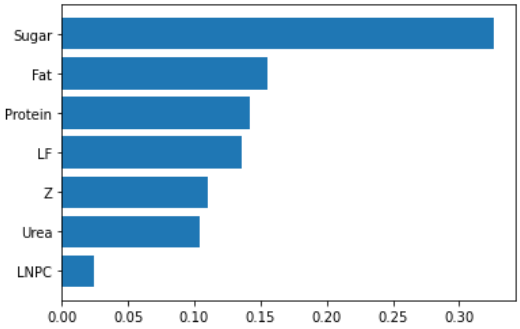

Figure 2. The result of the feature importance scoring for the best model's input parameters.

From the feature importance scoring results we can conclude that in our prediction algorithm sugar, fat, protein and lactoferrin variables play the greatest role in the prediction of SCC.

The accuracy checking for each output category separately was important to prove that our algorithm is not biased. As our model gave similarly good prediction values for less represented cases in our validation database it can be used in real life as well. This experiment shows that with enough data a good model can be built which can be used safely to predict the SCC category of a milk sample. Using this method the chemical analyses of milk samples can be cheaper or faster.

6. Conclusion

As a conclusion, we can see that regularly collected milk data can be used to predict somatic cell count in milk samples. Lactoferrin which seemed biologically the most significant impact factor on SCC values gives a much better result when used with other input variables. However, in the best-performing model lactoferrin was not the most significant input variable. The result of the feature importance scoring can serve as a basis to build another prediction model SCC scores well ahead of time, to give more time to farmers to watch the affected cows for possible mastitis. For this, we will need further data with more datestamp which can be obtained from farms using milking robots. We also proved that the logarithmic SCC value classified into 3 categories is well-suitable for prediction purposes and it also satisfies the milk health needs of the market.

Since most of the collected milk samples are from healthy cows the balancing of the input dataset must be done prior to teaching the algorithm. Regular balancing methods proved to be successful in balancing the dataset and for machine learning classification. Our results proved that multiclass machine learning can be used to predict high levels of somatic cell count in milk samples. This algorithm can serve as a basis for future prediction of somatic cell count it can help in predicting early mammary health of milking cows. If the farmer can start medication right before critical somatic cell count values are reached in milk samples which can lead to more effective milking of the farm.

References

[1] M. Alam, C. Cho, T. Choi, B. Park, J. Choi, Y. Choy, S. Lee, K. Cho: *Estimation of Genetic Parameters for Somatic Cell Scores of Holsteins Using Multi-trait Lactation Models in Korea*, Asian-Australas J Anim Sci 28.3 (2015), pp. 303–310, doi: 10.5713/ajas.13.0627.

[2] A. A. E. Amin: *Estimates of Heritability for Somatic Cell Count, Test-Day Milk Yield and Some Udder-Teat Characteristics in Saudi Dairy Goats using Random Regression Animal Model*, Adv. Anim. Vet. Sci 6.3 (2018), pp. 128–134, doi: 10.17582/journal.aavs/2018/6.3.128.134.

[3] C. Burvenich, et al.: *Physiological and Genetic Factors That Influence the Cows Resistance to Mastitis, Especially during Early Lactation*, Proceedings of the 5th IDF Mastitis Congress, Symposium on Immunology of Ruminant Mammary Gland (2000).

[4] J. B. Cheng, J. q. Wang, D. P. Bu, G. L. Liu, C. G. Zhang, H. Y. Wei, L. Y. Zhou, J. Z. Wang: *Factors Affecting the Lactoferrin Concentration in Bovine Milk*, Journal of Dairy Science 91 (2007), pp. 970–976, doi: 10.3168/jds.2007-0689.

[5] S. Dabdoub, G. Shook: *Phenotypic relations among milk yield, somatic cell count and clinical mastitis*, Journal of Dairy Science 67.1 (1984), pp. 163–164.

[6] L. Dégen, A. Monostori: *Az új tőgyegészségügyi rendszer bemutatása*, Állattenyésztési Teljesítményvizsgáló Kft.

[7] I. Dohoo, K. E. Leslie: *Evaluation of changes in somatic cell counts as indicators of new intramammary infections*, Preventive Veterinary Medicine 10.3 (1991), pp. 225–237, doi: 10.1016/0167-5877(91)90006-N.

[8] P. Geurts, D. Ernst, L. Wehenkel: *Extremely randomized trees*, Machine Learning 63 (2006), pp. 3–42, doi: 10.1007/s10994-006-6226-1.

[9] K. C. Halasa Tariq: *Differential Somatic Cell Count: Value for Udder Health Management*, Frontiers in Veterinary Science 7 (2020), doi: 10.3389/fvets.2020.609055.

[10] W. A. Muhammad, J. Reynolds, Y. Rezgui: *Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees*, Journal of Cleaner Production 203 (2018), pp. 810–821, doi: 10.1016/j.jclepro.2018.08.207.

[11] N. Sharma, N. K. Singh, M. S. Bhadwal: *Relationship of Somatic Cell Count and Mastitis: An Overview*, Asian-Australasian Journal of Animal Sciences 24.3 (2011), pp. 429–438, doi: 10.5713/ajas.2011.10233.

[12] I. Tímár: *Versenyképesség a magyar tejágazatban*, PhD thesis, Budapesti Corvinus Egyetem, 2004.