Acta Veterinaria-Beograd 2016, 66 (3), 317-335
UDK: 636.2.09:618.19-002
DOI: 10.1515/acve-2016-0028

Research article

IDENTIFICATION OF COWS SUSCEPTIBLE TO MASTITIS BASED ON SELECTED GENOTYPES BY USING DECISION TREES AND A GENERALIZED LINEAR MODEL

ZABORSKI Daniel[1]*, PROSKURA Witold Stanisław[1], WOJDAK-MAKSYMIEC Katarzyna[2], GRZESIAK Wilhelm[1]

[1]Department of Ruminants Science, West Pomeranian University of Technology, Szczecin, Poland;
[2]Department of Genetics and Animal Breeding, West Pomeranian University of Technology, Szczecin, Poland

(Received 01 February; Accepted 13 June 2016)

The aim of the present study was to: 1) check whether it would be possible to detect cows susceptible to mastitis at an early stage of their utilization based on selected genotypes and basic production traits in the first three lactations using ensemble data mining methods (boosted classification tress – BT and random forest – RF), 2) find out whether the inclusion of additional production variables for subsequent lactations will improve detection performance of the models, 3) identify the most significant predictors of susceptibility to mastitis, and 4) compare the results obtained by using BT and RF with those for the more traditional generalized linear model (GLZ). A total of 801 records for Polish Holstein-Friesian Black-and-White cows were analyzed. The maximum sensitivity, specificity and accuracy of the test set were 72.13%, 39.73%, 55.90% (BT), 86.89%, 17.81%, 59.49% (RF) and 90.16%, 8.22%, 58.97% (GLZ), respectively. Inclusion of additional variables did not have a significant effect on the model performance. The most significant predictors of susceptibility to mastitis were: milk yield, days in milk, sire's rank, percentage of Holstein-Friesian genes, whereas calving season and genotypes (lactoferrin, tumor necrosis factor alpha, lysozyme and defensins) were ranked much lower. The applied models (both data mining ones and GLZ) showed low accuracy in detecting cows susceptible to mastitis and therefore some other more discriminating predictors should be used in future research.

Key words: lactoferrin, tumor necrosis factor alpha, lysozyme, defensins, mastitis susceptibility, classification trees.

INTRODUCTION

Mastitis is one of the costliest and most frequent diseases in dairy cattle worldwide [1]. Its incidence in Polish dairy cows is approx. 30 – 50%, whereas the financial losses resulting from its occurrence in various European Union countries and the USA have been estimated at approx. EUR 693 per cow annually or approx. USD 2 billion

*Corresponding author: e-mail: daniel.zaborski@zut.edu.pl

in total per annum [2,3]. Mastitis is mainly caused by environmental factors such as milking machine effects, inappropriate milking and herd management, suboptimal feeding, and poor hygiene, but genetic and physiological factors (e.g. genetic selection for maximum milk yield) also play an important role in its etiology [4-6]. Mastitis leads to many adverse consequences including reduced milk yield, changes in milk quality which render it unsuitable for sale, shortened productive life of the cow and lower immunity to other diseases, premature culling, and increased labor, diagnosis, veterinary and medicine costs [7-11]. Considerable financial savings can be made by preventing mastitis and treating infected animals effectively. The incidence of clinical mastitis can be reduced by selection for resistance to it. However, genetic evaluation of mastitis resistance is difficult due to its low heritability. In such cases, indirect selection can be performed based on traits strongly correlated with mastitis such as somatic cell count [11].

From among the many genes with a potential effect on mastitis occurrence, four are quite frequently mentioned, i.e. genes coding for lysozyme, lactoferrin, tumour necrosis factor alpha and those encoding defensins. Lysozyme, produced by polymorphonuclear leukocytes, is an antimicrobial enzyme present in many different types of body tissues and fluids which protects the epithelium of various organs, including the bovine mammary gland, against bacterial and other microbial infections [12]. It is one of the most crucial components involved in the so-called non-specific humoral response of the immune system, whose concentration in bovine milk strongly depends on the udder health state, i.e. it is significantly increased in the colostrum and mastitis milk compared with normal milk. Moreover, it has been shown that lysozyme has a synergistic effect against *E. coli* and *Microcccus luteus* together with immunoglobulins and lactoferrin [13]. This function of lysozyme is especially important as recent studies on the role of *E. coli* in the pathogenesis of recurrent bovine mastitis indicate that some of its strains are able to produce *curli* fimbriae and cellulose as the components of their extracellular matrix and to form biofilm structures *in vitro*, which may facilitate the adaptation of this bacterial species to the mammary gland environment and cause persistent intramammary infections *in vivo* [14]. On the other hand, the genetic variants of the lactoferrin gene, whose glycoprotein product shows antibacterial, antiviral, antitumor and anti-inflammatory properties, have been successfully associated with cows' susceptibility to mastitis and its main indicator trait, i.e. the level of somatic cell count (SCC) in milk [15]. Being a multifunctional molecule, lactoferrin plays a crucial role in the innate host defense. It can bind iron ions, which is significant for its antibacterial effect (especially against *E. coli*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus*) [16]. According to a recent study by Rajić Savić et al. (2014) [17], *S. aureus* is the predominant species (88% of isolates) among highly virulent coagulase-positive staphylococci, which are resistant to penicillin, produce pigments and haemolysins, and are responsible for chronic infections of the bovine mammary gland. The third of the above-mentioned genes encodes tumor necrosis factor alpha (TNF-α), which belongs to the group of key mediators in the local inflammatory immune response.

TNF-α initiates a cascade of events and increases vascular permeability, which leads to the migration of macrophages and neutrophils to an infection site. As a cytokine, it has a wide spectrum of effects due to the large number of sites where its receptors are located and its ability to activate different signal transduction pathways that affect the expression of many genes [18]. A significant association of the interaction between TNF-α genotype and parity with immunity to mastitis (the number of mastitis cases and infected quarters) has been reported by Wojdak-Maksymiec et al. (2013) [19].

Finally, β-defensins produced by leukocytes are a family of related small cationic peptides present in different tissues of the organism and some of their types can also be found in the mammary gland. Clusters of β-defensin genes are considered candidate regions for SCC in bovine milk [20]. Their role in the immune response consists in providing protection against infectious agents such as bacteria, viruses and fungi through aggregation, pore formation and membrane depolarization. They are also involved in immunomodulation (e.g. by inducing the expression of co-stimulatory molecules on monocytes and myeloid dendritic cells as well as the chemotaxis of immune cells to the infection site) and have potential developmental functions [21].

One of the approaches to the identification of cows with an increased risk of developing mastitis during their life-span is based on the use of statistical methods for classification tasks and, more specifically, data mining techniques, which have become increasingly popular in the recent years in various animal farming applications [22-25]. For example, decision trees are tree-like representations of a learned function used for approximating a target variable (such as health status) with the discrete values from a given dataset and consist of nodes connected with branches, of which terminal nodes (leaves) contain the target values [26]. In decision trees such as classification and regression trees (CART), the entire dataset is first divided into two groups to maximize the homogeneity of cases within the nodes by searching every value of each predictor. Then, a recursive algorithm is applied so that these two groups are split again into two subgroups in order to reduce further node impurity according to the available values of predictors. The splits are continued until the resulting nodes have so few cases that no further division is possible. The final stage of tree model building is its pruning, as too large trees have a tendency to overfit the training data. The basic method to find the optimal tree is V-fold cross validation [27]. Predictions made with decision trees are based on sorting new cases down the tree until the leaf is reached. In some decision tree types, multi-way divisions are also possible whereby splitting can be performed according to all the variants of a categorical variable [28]. To enhance the classification power of single decision trees, ensemble methods have been developed where bagging or boosting is applied. The first solution is used in the so-called random forest (RF), in which a large number of trees are grown in parallel on a subset of records and the predictions made by individual trees are subsequently averaged. The latter approach, on the other hand, is applied in a method known as boosting trees (BT), where trees are grown sequentially and each successive tree tries to reduce the total error by modeling the residuals generated by previous trees [27].

Finally, a generalized linear model (GLZ) can be considered as a reference model in which the values of a dependent variable are predicted from a linear combination of explanatory variables that are connected to the dependent variable via a link function. Different link functions can be applied, such as a logarithmic, power, or logit link function [29].

Taking into account the above-mentioned losses resulting from mastitis and the possibility of using data mining methods to reduce them, the aim of the present study was to check whether it would be possible to detect cows susceptible to mastitis at an early stage of their utilization on the basis of selected genotypes and basic production records by using BT and RF. Another aim was to find out whether the inclusion of additional production variables for subsequent lactations would improve detection performance of the models. Finally, the most significant predictors of susceptibility to mastitis were identified and the results obtained by using BT and RF were compared with those for the more traditional GLZ model.

MATERIALS AND METHODS

A total of 801 records for Polish Holstein-Friesian Black-and-White cows kept on a farm located in the West Pomeranian Province were used in the experiment. The animals were housed in an open barn without access to pasture, and fed a total mixed ration prepared from maize silage. The ration was supplemented with concentrate and diet supplements, whose amount was determined individually for each cow. Water was accessible for the cows *ad libitum* from automatic drinkers. The animals were milked twice a day in a herringbone milking parlour and mastitis symptoms for all udder quarters were examined during each milking by the staff. All alarming symptoms were reported to a veterinarian, who then either confirmed or ruled out clinical mastitis. In addition, the cows were dried-off approx. six weeks prior to the expected calving date and antibiotic protection was applied if mastitis symptoms were visible during the dry-off period. No antibiotic therapy was used for the other cows.

The data was collected between September 2003 and April 2008. An initial set of 990 records was subsequently reduced to 801 records (obtained for the first to third lactation) after editing (removal of erroneous, incomplete or missing data). It should be emphasized that only cows which had completed at least the first three lactations were retained in the final dataset. The number of predictors included in the models varied depending on the lactation number (only the first lactation, the first and second lactation, or all three lactations completed). The following predictors were common for all lactation numbers: X_1 – LTF – lactoferrin genotype, X_2 – TNF - tumour necrosis factor alpha genotype, X_3 – LYZ – lysozyme genotype, X_4 – DEF – combined defensin genotype, X_5 – HF – percentage of HF genes in a cow's genotype (obtained from farm documentation), X_6 – SIRE – cow's sire rank based on the average somatic cell count (SCC) of his daughters, X_7 – CALS – the first calving season. The predictors that differed between lactations were as follows: X_8 – DIM1 – days in milk (days) and X_9 – MILK1 – milk yield (kg) for the first lactation; X_8 – DIM2 – average days in milk (days) and X_9 – MILK2 – average milk yield (kg) for the first and second lactation; X_8 – DIM3 – average days in milk (days) and X_9 – MILK3 – average milk yield (kg) for the first, second and third lactation (kg). The response variable (Y – MAST) was a category of susceptibility to mastitis: susceptible (mastitis) or immune (healthy). The cows from the "healthy" category had never suffered from mastitis during the first three lactations. As previously mentioned mastitis was diagnosed by a veterinarian and classified as acute, chronic or drying-off with antibiotic protection. The means and standard deviations of continuous predictors are presented in Table 1, whereas the distribution of categorical predictors and that of the response variable are given in Table 2. TNF-α, mLYZ and LTF genotypes were determined according to the procedures described by Wojdak-Maksymiec et al. (2013) [19], whereas DEF genotypes were assayed according to Wojdak-Maksymiec et al. (2012) [30].

Table 1. Means and standard deviations of continuous predictors for the training and test sets

Set	Number of lactations	Training (n_L=606)		Test (n_T=195)		Total (n=801)	
Predictor		Mean	SD	Mean	SD	Mean	SD
HF[1] (%)	1, 2, 3	85.79	13.93	88.16	12.31	86.36	13.58
SIRE	1, 2, 3	164.4	76.6	166.7	74.9	165.0	76.1
DIM1 (days)	1	339.2	67.1	348.8	76.7	341.5	69.6
MILK1 (kg)	1	9746.9	2155.6	10003.9	2402.7	9809.5	2219.5
DIM2 (days)	2	344.9	53.1	349.3	59.3	346.0	54.7
MILK2 (kg)	2	10594.7	2054.5	10790.1	2195.9	10642.3	2090.1
DIM3 (days)	3	341.1	45.1	345.0	52.6	342.1	47.0
MILK3 (kg)	3	10833.4	1885.7	11108.7	2071.5	10900.5	1934.9

1 – the names of variables and their variants are described in the Materials and Methods section.

In brief, DNA isolation was performed with ZymoResearch Genomic DNA Kit™ (ZymoResearch, USA) using Fast-Spin column technology. Next, SimpleProbe real-time PCR assays were developed to determine TNF-α, mLYZ and LTF genotypes. PCRs were carried out in a LightCycler 2.0 (Roche Molecular Systems Inc., Pleasanton, USA). Each batch consisted of 31 samples and a negative control (water) in 20μl capillary tubes. The products were analyzed using real-time fluorescence readout. Amplification was made with Qiagen® Multiplex PCR Kit (Qiagen GmbH, Hilden, Germany). The PCR mix (10 μl) for LTF and mLYZ contained: 5 μl 2× Qiagen PCR Master Mix (final concentration of 3 mM $MgCl_2$); 1 μl each primer (0.2 μM); 1 μl SimpleProbe (0.2 μM); 1 μl water. The thermal profile included: initial denaturation – 95°C/15 min; amplification – 45 cycles: denaturation 95°C/20 s, annealing 57°C/30 s, and elongation 72°C/40 s; melting – 95°C, 40°C and 80°C with a ramp rate of 0.1°C/min; cooling – 30 s.

Table 2. Distribution of categorical predictors and the response (output) variable

Set	Training (n_L=606)		Test (n_T=195)		Total (n=801)	
	n	%	n	%	n	%
Variant			**LTF**			
AB	276	45.54	86	44.10	362	45.19
AA	330	54.46	109	55.90	439	54.81
			TNF			
CC	193	31.85	72	36.92	265	33.08
CT	284	46.86	87	44.62	371	46.32
TT	129	21.29	36	18.46	165	20.60
			LYZ			
CC	584	96.37	179	91.79	763	95.26
CT	22	3.63	16	8.21	38	4.74
			DEF			
DEF1	454	74.92	149	76.41	603	75.28
OTH	152	25.08	46	23.59	198	24.72
			CALS			
Autumn	133	21.95	43	22.05	176	21.97
Winter	181	29.87	63	32.31	244	30.46
Spring	186	30.69	50	25.64	236	29.46
Summer	106	17.49	39	20.00	145	18.10
			MAST – response			
Mastitis	355	58.58	122	62.56	477	59.55
Healthy	251	41.42	73	37.44	324	40.45

LTF – lactoferrin genotype, TNF – tumour necrosis factor alpha genotype, LYZ – lysozyme genotype, DEF – combined defensin genotype, DEF1 – genotype A1A2/B1B2/C1C2, OTH – other 21 combined defensin genotypes (A1A2/B2/C1C2, A2/B1B2/C2, A1A2/B1B2/C1, A2/B1B2/C1C2, A1A2/B1/C1C2, A1A2/B1B2/C2, A1/B1B2/C2, A1/B1B2/C1, A1/B1B2/C1C2, A1A2/B1B2/C2C2, A1A2/B2/C1, A2/B2/C1C2, A1/B1/C1C2, A1A2/B1/C1, A1/B1/C1, A1A2/B1B1/C1, A2/B2/C1, A2/B1B2/C1, A1A2/B1/C2, A1/B2/C1C2, A1A2/B2/C2), CALS – the season of the first calving, MAST – susceptibility to mastitis

For TNF-α, asymmetric real-time PCR was used. The PCR mix contained: 5 μl 2× Qiagen PCR Master Mix (final concentration of 3 mM $MgCl_2$); 0.5 μl forward primer (0.1 μM); 1.5 μl reverse primer (0.3 μM); 0.5 μl SimpleProbe (0.1 μM); 1.5 μl water. *Q-solution* was also used. The following temperature profile was applied: initial denaturation – 95°C/15 min; amplification – 45 cycles: denaturation 95°C/30 s, annealing 57°C/30 s, and elongation 72°C/60 s; melting – 95°C, 40°C and 80°C with a ramp rate of 0.1°C/min; cooling – 30 s.

To identify β-defensin genotypes, PCR was performed in thermal cyclers manufactured by Whatman Biometra GmbH (Gottingen, Germany) with the following set of primers (Proligo France SAS): F: 5'-GCCAGCATGAGGCTCCAT-3' and R: 5'-AACAGGTGCCAATCTGT-3'. The PCR mix included: 50 ng DNA, 200 μM dNTP, 20 pmol primer, 0.75 mM $MgCl_2$, 2 μl 10× *Taq*1 Buffer, 1 U *Taq* polymerase and water to a volume of 20 μl (Fermentas International INC, Burlington, Canada). The PCR was carried out in 35 cycles with the following thermal profile: initial denaturation 94°C/300 s, denaturation 94°C/60 s, annealing 63.5°C/60 s, extension 72°C/90 s, and final extension 72°C/60 s. The amplified fragments were then digested with *Taq*I at 65°C/16 h, electrophoresed on a 2% agarose gel with ethidium bromide, and visualized under UV light (Vilber Lourmat Deutschland GmbH, Eberhardzell, Germany).

Because the A1A2/B1B2/C1C2 genotype was predominant (approx. 75% of all cases) and all other genotypes had a low frequency of occurrence, the DEF variable was dichotomized into two categories (DEF1 including the A1A2/B1B2/C1C2 genotype and OTHER including all the rest of genotypes).

The whole dataset of 801 records was randomly divided into two subsets: a training set (L) of 606 records (75.7%) and a test set (T) of 195 records (24.3%). The first one was used to train machine-learning models and to estimate GLZ parameters, while the latter one served as a basis for verifying their prognostic abilities on new data. For RF and BT, a subset of the training set was also randomly created (the so-called validation set) in order to monitor the course of training and prevent overtraining, i.e. too close fit of the model to the training data. The proportion of mastitic to non-mastitic records in the whole dataset was approx. 1.5:1. The quality and prediction abilities of RF and BT were compared with a more traditional classification method in the form of a GLZ.

In the case of BT, the Gini index was used as a measure of node impurity, the prior probabilities were estimated from the training sample, the value of the learning rate was 0.1 and the proportion of randomly selected cases in a subsample was 0.5. Moreover, a validation set consisting of 30% of the total number of training cases was used. In the case of RF, the number of randomly selected predictors for each tree in the forest was four, the minimum number of cases in a node was 20, and the minimum number of cases in a child node was five. The other parameters were the same as those for BT. The last model used in the present study was a GLZ with a binomial distribution of a dependent variable and a logit link function, which was determined according to the following formula [31]:

$$\log itP(Y) = a_0 + \sum_{i=1}^{N} a_i X_i$$

where: logitP(Y) is the logit of the probability of a cow being susceptible to mastitis, a_0 is an intercept, a_i is the model coefficient estimated during its development, and X_i is the i^{th} predictor.

After the model was constructed, we verified its assumptions, i.e. the normal distribution of residuals (with the Shapiro-Wilk test) and the lack of predictor collinearity (based on the variance inflation factor).

To compare the quality of data mining methods and the GLZ, three basic measures were used for the L set, i.e. sensitivity (Se – the proportion of correctly classified cows susceptible to mastitis), specificity (Sp – the proportion of correctly classified immune cows), and accuracy (Acc – the proportion of correctly indicated animals from both categories). Also, a test for proportions was applied to check significant differences among the probabilities generated by individual models. Statistical significance was set at $P \leq 0.05$.

The next step following the model training stage was to identify the variables that were most significant for the determination of a cow's susceptibility to mastitis. In the case of BT, we used an importance measure based on prediction statistics computed for each predictor during each split in each successive tree. In the process of tree construction, the predictor yielding the best possible split in a tree node was always selected and the mean value of the prediction statistic was determined for all the variables and all successive trees in a boosting sequence. Finally, the importance value for the predictor was normalized by assuming the maximum value for the most significant input as unity and reducing the values for the remaining predictors accordingly [32]. A similar approach based on an importance measure was used to determine the significance of predictors in the RF model, whereas the Wald statistic (and its corresponding degrees of freedom) was used for the GLZ models.

In the third stage of the present study, the models' performance was verified on an independent T set consisting of new cases which had not been previously used during the model construction stage. To evaluate the prediction performance of the models, Se, Sp and Acc were computed in the same way as for the L set. In addition, the posterior probability of true positive responses (PSTP) and true negative responses (PSTN) was calculated to show the proportion of cases (cows) assigned by the model to one of the two classes that really belonged to that class. Thus, the posterior probabilities indicated the reliability of predictions. To test the significance of the differences among probabilities, the test for proportions was applied again with $P \leq 0.05$ as a significance level. Finally, the area under the receiver operating characteristic (ROC) curves (AUC) was additionally calculated to assess more easily the discrimination power of the models constructed. AUC is a measure of a classifier's performance and can be used for comparing the results of different classification schemes. Its maximum value is one, which corresponds to the ROC curve crossing the [0.1] point on the plot (100% Se and 0% false alarm rate), whereas the value of 0.5 indicates very poor model quality [33]. All the computations were performed using Statistica® 12 (StatSoft, Inc., Tulsa, OK, USA).

RESULTS

The final number of basic trees in the BT model was seven, eight and one for the first, second and third lactation, respectively, whereas the number of trees in the RF model was 240, 225 and 300, respectively. The coefficients of the GLZ for the first, second and third lactation are given in Tables 3 – 5, respectively. As can be seen from these tables, none of the predictors included in the GLZ models had a statistically significant effect on the category of a dependent variable, i.e. mastitis susceptibility. Moreover, although there was no collinearity between predictors (variance inflation factor for all explanatory variables in all the three models below four), the assumption about the normal distribution of residuals was violated in all the cases.

Table 3. Estimated GLZ parameters for the first lactation

	Variant	Estimate	Standard error	Wald statistic	p	OR
Intercept		-0.8097	0.6975	1.3479	0.2457	0.4450
HF[1]		0.0069	0.0061	1.2833	0.2573	1.0070
SIRE		0.0011	0.0011	1.0002	0.3173	1.0011
DIM1		-0.0012	0.0018	0.4362	0.5089	0.9988
MILK1		0.0001	0.0001	1.3272	0.2493	1.0001
LTF	AB	-0.1208	0.0844	2.0469	0.1525	0.8862
TNF	CC	0.1044	0.1236	0.7137	0.3982	1.1101
	CT	0.0713	0.1131	0.3979	0.5282	1.0739
LYZ	CC	0.1270	0.2218	0.3277	0.5670	1.1354
DEF	DEF1	-0.0033	0.0968	0.0011	0.9730	0.9967
CALS	Autumn	0.0691	0.1529	0.2040	0.6515	1.0715
	Winter	-0.1500	0.1380	1.1808	0.2772	0.8607
	Spring	0.1316	0.1390	0.8959	0.3439	1.1406

1 – the names of variables and their variants are described in the Materials and Methods section, OR – odds ratio

In the quality evaluation of the models, their Se, Sp and Acc based on the L set were compared. The respective values for BT, RF and GLZ are given in Table 6. The highest Se (i.e. the ability of the model to indicate correctly susceptible cows) was exhibited by RF or GLZ (0.87 – 0.94 depending on the lactation number) and this Se differed significantly (P≤0.05) from that for BT in all lactations. On the other hand, BT and RF were characterized by the highest Sp (i.e. the ability of the model to indicate correctly resistant cows), which amounted to 0.27 – 0.49 and differed significantly (P≤0.05) from that for the GLZ. There was also a significant difference (P≤0.05) in Sp between BT, RF and GLZ in the third lactation. Finally, the greatest Acc (the ability to indicate properly cows from both categories) was found for RF (0.66 – 0.67), and it was significantly different (P≤0.05) from Acc for the other classifiers in all lactations.

Table 4. Estimated GLZ parameters for the second lactation

	Variant	Estimate	Standard error	Wald statistic	p	OR
Intercept		-0.5713	0.7740	0.5448	0.4605	0.5648
HF[1]		0.0076	0.0061	1.5534	0.2126	1.0076
SIRE		0.0011	0.0011	1.0782	0.2991	1.0011
DIM2		-0.0016	0.0022	0.4865	0.4855	0.9984
MILK2		0.0000	0.0001	0.5521	0.4574	1.0000
LTF	AB	-0.1249	0.0843	2.1925	0.1387	0.8826
TNF	CC	0.1060	0.1233	0.7393	0.3899	1.1119
	CT	0.0698	0.1131	0.3810	0.5371	1.0723
LYZ	CC	0.1264	0.2218	0.3245	0.5689	1.1347
DEF	DEF1	0.0030	0.0966	0.0010	0.9753	1.0030
CALS	Autumn	0.0639	0.1528	0.1750	0.6757	1.0660
	Winter	-0.1479	0.1380	1.1487	0.2838	0.8625
	Spring	0.1379	0.1388	0.9860	0.3207	1.1478

1 – the names of variables and their variants are described in the Materials and Methods section, OR – odds ratio

Table 5. Estimated GLZ parameters for the third lactation

	Variant	Estimate	Standard error	Wald statistic	p	OR
Intercept		-1.1068	0.8593	1.6591	0.1977	0.3306
HF[1]		0.0071	0.0061	1.3586	0.2438	1.0071
SIRE		0.0011	0.0011	0.9879	0.3202	1.0011
DIM3		-0.0007	0.0026	0.0824	0.7741	0.9993
MILK3		0.0001	0.0001	1.2267	0.2681	1.0001
LTF	AB	-0.1197	0.0849	1.9863	0.1587	0.8872
TNF	CC	0.1010	0.1231	0.6733	0.4119	1.1063
	CT	0.0743	0.1129	0.4329	0.5106	1.0771
LYZ	CC	0.1446	0.2233	0.4195	0.5172	1.1556
DEF	DEF1	0.0017	0.0965	0.0003	0.9860	1.0017
CALS	Autumn	0.0657	0.1529	0.1846	0.6674	1.0679
	Winter	-0.1578	0.1383	1.3025	0.2538	0.8540
	Spring	0.1321	0.1389	0.9041	0.3417	1.1412

[1] – the names of variables and their variants are described in the Materials and Methods section, OR – odds ratio

Some differences in Se were also observed between lactations. Se for BT in the first and second lactation (0.71) was significantly higher ($P \leq 0.05$) than that in the third lactation (0.62), whereas the opposite trend was observed for Sp. In the case of RF, significantly higher Se (P≤0.05) was recorded in the first (0.94) and third (0.93) lactation compared with the second one (0.87), while Sp was significantly higher (P≤0.05) in the second lactation (0.37) in comparison with the third one (0.27). No significant differences in any probability between lactations were revealed for the GLZ.

Table 6. Model quality on the training set

	Lactation 1			Lactation 2			Lactation 3		
	Se	**Sp**	**Acc**	**Se**	**Sp**	**Acc**	**Se**	**Sp**	**Acc**
BT	0.7099^{Aa}	0.3665^{Aa}	0.5677^{a}	0.7099^{Aa}	0.3904^{Aa}	0.5776^{a}	0.6225^{Ba}	0.4861^{Ba}	0.5660^{a}
RF	0.9380^{Ab}	0.2988^{a}	0.6733^{b}	0.8676^{Bb}	0.3665^{Aa}	0.6601^{b}	0.9324^{Ab}	0.2709^{Bb}	0.6584^{b}
GLZ	0.9155^{b}	0.1474^{b}	0.5974^{a}	0.9155^{b}	0.1394^{b}	0.5941^{a}	0.9070^{b}	0.1434^{c}	0.5908^{a}

Se – sensitivity, Sp – specificity, Acc – accuracy, BT – boosted classification trees, RF – random forest, GLZ – generalized linear model, a, b, c – different small letters within columns denote statistical significance at P≤0.05, A, B – different capital letters within rows denote statistical significance at P≤0.05

The next stage of the present study was to identify the most influential predictors of susceptibility to mastitis for the tree models and the GLZ. The results for BT and RF are shown in Figs. 1 and 2, respectively. In both cases, MILK and DIM were the most significant factors affecting predisposition to mastitis, followed by SIRE and HF, although the detailed sequence of importance varied slightly between lactations. CALS and, more importantly, the genotypes were ranked much lower. The sequence of predictors for the GLZ is shown in Fig. 3, but none of them had a statistically significant effect according to the Wald test.

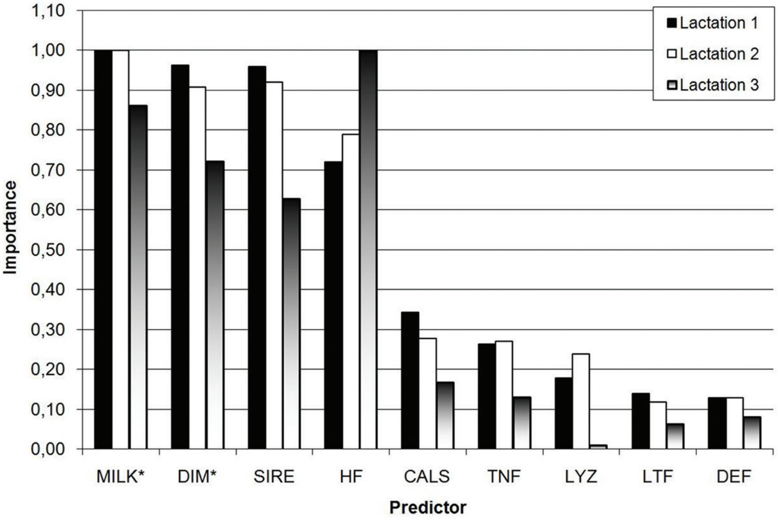

Figure 1. Predictor importance for boosted classification trees
* - predictors differed according to the lactation number (total milk yield and days in milk for the first lactation, average milk yield and days in milk for the first and second lactation, or average milk yield and days in milk for the first, second and third lactation)

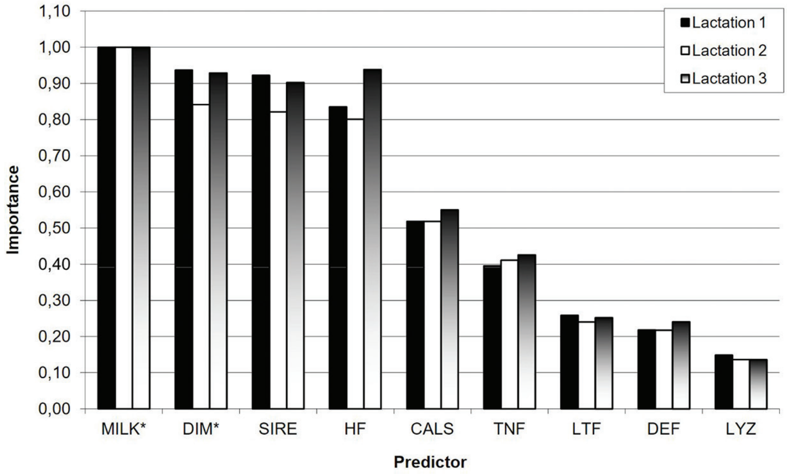

Figure 2. Predictor importance for random forest
* - predictors differed according to the lactation number (total milk yield and days in milk for the first lactation, average milk yield and days in milk for the first and second lactation, or average milk yield and days in milk for the first, second and third lactation)

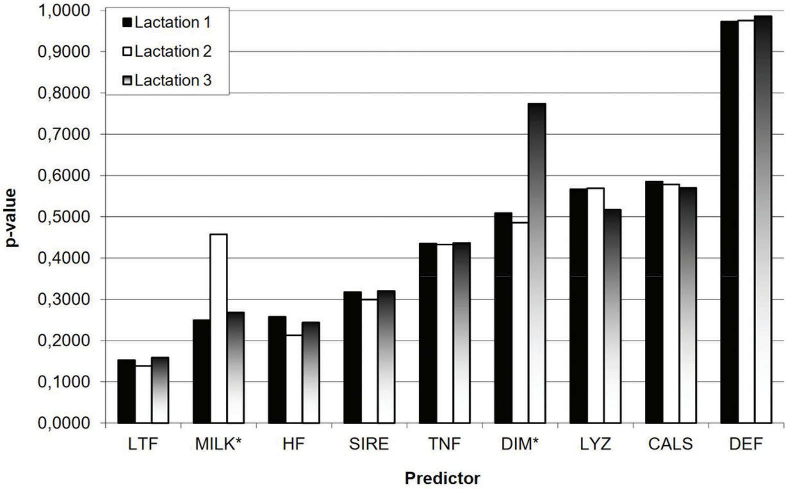

Figure 3. Predictor importance for the generalized linear model
* - predictors differed according to the lactation number (total milk yield and days in milk for the first lactation, average milk yield and days in milk for the first and second lactation, or average milk yield and days in milk for the first, second and third lactation)

After assessing predictor importance on the L set, the models' detection performance was evaluated on the independent T test, whose cases had not been used previously for model construction. Such a set made it possible to verify objectively the models' capacity to indicate cows susceptible to mastitis. The results of this analysis are shown in Table 7. The only significant differences (P≤0.05) between the models were found in Se and Sp. The highest Se (0.84 – 0.89) was observed for RF and GLZ in the first and third lactation, whereas a significant difference in Se in the second lactation was revealed between BT and GLZ. On the other hand, BT was characterized by the highest Sp (0.21 – 0.40), which was particularly evident in the first and third lactation. No other significant differences in the analyzed probabilities among various model types were observed on the T set. However, some significantly different Se and Sp values were noted for the same model types in different lactations: Se for BT was significantly lower (P≤0.05) in the third lactation (0.62) compared with the second one (0.72), while the opposite difference in Sp for the same model type existed between the first (0.40) and the second (0.21) lactation (P≤0.05). No other differences depending on lactation number were found.

Table 7. Model detection performance on the test set

	Se	Sp	PPSTP	PPSTN	Acc	AUC
			lactation 1			
BT	0.6557^{a}	0.3973^{Aa}	0.6452	0.4085	0.5590	0.5583
RF	0.8689^{b}	0.1370^{b}	0.6272	0.3846	0.5949	0.4805
GLZ	0.8934^{b}	0.0822^{b}	0.6193	0.3158	0.5897	0.4426
			lactation 2			
BT	0.7213^{Aa}	0.2055^{Ba}	0.6027	0.3061	0.5282	0.5090
RF	0.7951	0.1781	0.6178	0.3421	0.5641	0.4268
GLZ	0.9016^{b}	0.0685^{b}	0.6180	0.2941	0.5897	0.4373
			lactation 3			
BT	0.6230^{Ba}	0.3151^{a}	0.6032	0.3333	0.5077	0.4695
RF	0.8443^{b}	0.1096^{b}	0.6131	0.2963	0.5692	0.4798
GLZ	0.8689^{b}	0.0685^{b}	0.6092	0.2381	0.5692	0.4175

Se – sensitivity, Sp – specificity, PPSTP – posterior probability of true positive response, PPSTN – posterior probability of true negative response, Acc – accuracy, AUC – area under the receiver operating characteristic curve, BT – boosted classification trees, RF – random forest, GLZ – generalized linear model, a, b – different small letters within columns denote statistical significance between models at P≤0.05, A, B – different capital letters within columns denote statistical significance between lactations at P≤0.05

The last stage of the present study was the calculation of AUC for the three model types evaluated. The highest AUC values were observed for BT in the first two lactations (0.51 – 0.56), whereas those for all the other models in all three lactations were below 0.5, which confirms rather poor detection performance of the constructed classifiers.

DISCUSSION

In general, it should be noted that the initial hypothesis on the possibility of using the four genotypes (lactoferrin, lysozyme, tumour necrosis factor alpha and defensins) together with a few additional predictors available at an early stage of animal development (the proportion of HF genes, a sire's tendency towards transmitting high SCC to his daughters) supplemented with basic production records was not confirmed. The results obtained for the detection performance of the three models investigated in the present study were relatively poor, taking into account the Acc on the T set and, especially, its corresponding AUC. The AUC value below 0.5 in most cases shows that the model did not discriminate any better than a random guess. Although Se was relatively high for all the models on the T set (0.62 – 0.90), it was accompanied by low Sp (0.07 – 0.40), which means that if such decision-support tools were used in practice, many cows would be classified incorrectly as susceptible to mastitis (so-called false alarms) despite their actual resistance to mastitis (at least as evaluated for the first three lactations). The inclusion of additional information on the average milk yield and lactation length for the two subsequent lactations did not lead to any substantial improvement in the model performance either. However, it was not possible to utilize this information in the form of separate predictors (e.g. milk yield for the first lactation and milk yield for the second lactation) due to the limited sample size. It is difficult to compare directly the results of the present work with those of other authors because most research in this field [1,34-39] has been aimed at detecting single mastitis cases on the farm rather than diagnosing cows that are susceptible to this disease. Nevertheless, some comparisons can be made. In the study by Chagunda et al. (2006) [40] on the application of a dynamic deterministic biological model to mastitis detection using lactate dehydrogenase as a disease indicator, Se and Sp were 0.82 and 0.99, respectively, at a threshold mastitis risk of 0.7. On the other hand, in the work of Krieter et al. (2007) [35] on mastitis detection by artificial neural networks based on milk electrical conductivity and flow rate, the Sp on the T set was 0.51 – 0.75 at the assumed minimal Se of 0.80, whereas Cavero et al. (2008) [34], using the same method and similar inputs, obtained Se in the range of 0.63 to 0.93 and Sp ranging from 0.38 to 0.87, depending on the mastitis criterion (SCC above 100,000 cells×ml^{-1} or above 400,000 cells×ml^{-1}), so these values were, in general, much higher than those in our study. Still higher detection performance was characteristic of the neural networks constructed by Sun et al. (2010) [41], for which Se, Sp and Acc were 0.79 – 0.87, 0.91 – 0.92 and 0.87 – 0.91, respectively, depending on the input variables used and the network structure. Another detection and monitoring model based on the on-line recording of SCC [42] yielded an Se of 0.28 to 0.43 when reporting new mastitis cases, and Se between 0.55 to 0.89 when indicating on-going intramammary infections. The lowest proportion of false alarms observed in this study was 0.07. Finally, Se and Sp of decision tree algorithms evaluated through a 10-fold cross-validation or on a separate test set were 0.05 – 0.57 and 0.93 – 1.00, depending on the tree structure and misclassification costs [43] as well as 0.40 – 0.67 and 0.99, depending on the time window in which mastitis cases were observed [44].

As for the AUC, which was very small in the present study, Yang et al. (1999) [39] reported AUC values ranging between 0.77 and 0.87 for different proportions of mastitic to non-mastitic records in the training set, while Kamphuis et al. (2010) [43] obtained transformed partial AUC values (with Sp set at above 0.97) ranging between 0.56 and 0.65, depending on the model structure and misclassification costs.

On the other hand, such a poor result obtained in the present study is not a complete surprise since resistance/susceptibility to mastitis is a complex polygenic trait and it should not be expected that the polymorphisms of four selected genes will suffice to distinguish accurately between these two health categories. The small effect of genetic predictors on the class of response variable was also confirmed by the sequence of explanatory variables in the tree models where milk yield, lactation length, sire and proportion of HF genes were ranked high, with the first calving season and genotypes at the last positions. Moreover, none of the analyzed explanatory variables turned out to have a significant effect on the predisposition to mastitis in the GLZ model based on a more traditional parametric approach to classification tasks, which served as the reference classifier in the present study. Another limitation which has already been mentioned was the relatively small sample size (only 801 records from one herd), which also made it more difficult to observe any significant relationships among the investigated variables.

The models applied in the present study (boosted classification trees, random forest, and generalized linear model) showed low accuracy in detecting cows susceptible to mastitis. Moreover, adding more information to the models (the average milk yield and days in milk for the subsequent lactations) did not improve their performance significantly. Also, the effect of the selected genotypes (lactoferrin, tumour necrosis factor alpha, lysozyme, and combined defensin genotypes) on susceptibility to mastitis was relatively small, and therefore more discriminating predictors (including more genotypes) and a larger sample size will have to be used in future research in order to detect problematic animals more accurately.

Acknowledgements

This work was supported by the Polish Ministry of Science and Higher Education under Grant number 517-01-028-3962/17.

Authors' contributions

ZD carried out statistical analyses, interpreted the results and drafted the manuscript. PSW participated in the acquisition and collation of data and manuscript drafting. WMK conceived the study, participated in its design and coordination and critically revised the successive versions of manuscript. GW participated in the design of the study, helped in results interpretation and critically revised the manuscript. All authors read and approved the final manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy of integrity of any part of the work are appropriately investigated and resolved.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

1. Heald CW, Kim T, Sischo WM, Cooper JB, Wolfgang DR: A computerized mastitis decision aid using farm-based records: an artificial neural network approach. J Dairy Sci 2000, 83:711-720.
2. Siebert LJ: Genome wide association study identifies loci sssociated with mastitis antibiotic therapy phenotypes following experimental challenge with *Streptococcus Uberis*. Plant and Animal Genome XXIII. San Diego, CA. 2015.
3. Viguier C, Arora S, Gilmartin N, Welbeck K, O'kennedy R: Mastitis detection: current trends and future perspectives. Trends Biotechnol 2009, 27:486-493.
4. Mein G, Reinemann D, Schuring N, I O: R-MM-1: milking machines and mastitis risk – a storm in a teatcup. 43rd National Mastitis Council Annual Meeting. Charlotte, NC. 2004.
5. Sender G, Korwin-Kossakowska A, Hameid KGA, Prusak B: Ocena wpływu polimorfizmu wybranych genów na występowanie mastitis u krów. Med Weter 2006, 62:563-565.
6. Sordillo LM: Factors affecting mammary gland immunity and mastitis susceptibility. Livestock Production Science 2005, 98:89-99.
7. Bruckmaier RM, Ontsouka CE, Blum JW: Fractionized milk composition in dairy cows with subclinical mastitis. Vet Med 2004, 49:283-290.
8. Cavero D, Tölle KH, Rave G, Buxadé C, Krieter J: Analysing serial data for mastitis detection by means of local regression. Livest Sci 2007, 110:101-110.
9. Halasa T, Huijps K, Østerås O, Hogeveen H: Economic effects of bovine mastitis and mastitis management: A review. Vet Quart 2007, 29:18-31.
10. Schabauer L, Wenning M, Huber I, Ehling-Schulz M: Novel physico-chemical diagnostic tools for high throughput identification of bovine mastitis associated gram-positive, catalase-negative cocci. BMC Vet Res 2014, 10:1-11.
11. Makovickỳ P, Makovickỳ P, Nagy M, Rimárová K, & Diabelková J: Genetic parameters for somatic cell count, logscc and somatic cell score of breeds: Improved Valachian, Tsigai, Lacaune and their Crosses. Acta Vet-Beograd 2014, 64:386–396.
12. Liu X, Wang Y, Tian Y, Yu Y, Gao M, Hu G, Su F, Pan S, Luo Y, Guo Z, Quan F, Zhang Y: Generation of mastitis resistance in cows by targeting human lysozyme gene to β-casein locus using zinc-finger nucleases. Proc R Soc B 2014, 281:20133368.
13. Barłowska J, Litwińczuk Z, Brodziak A, Król J: Somatic cell count as the factor conditioning productivity of various breeds of cows and technological suitability of milk. In: Dairy Cows: Reproduction, Nutritional Management and Diseases. New York: Nova Science Publishers; 2013, 91-126.
14. Milanov D, Prunić B, Velhner M, Todorović D, Polaček V: Investigation of biofilm formation and phylogenetic typing of Escherichia coli strains isolated from milk of cows with mastitis. Acta Vet-Beograd 2015, 65:202–216.
15. Bukhari S, Das AK, Kumar N, Raghuwanshi P, Taggar RK, Chakraborty D, Kumar D, Vohra V, Gupta P: Genetic polymorphism of promoter region of lactoferrin gene and

its association with mastitis resistance in Jersey crossbred cattle. Indian J Anim Res 2015, 49:165-167.

16. Singh AP, Ramesha KP, Isloor S, Divya P, Rao A, Basavaraju M, Das DN, Munde U: Single nucleotide polymorphisms in lactoferrin gene are associated with lactoferrin content in milk and somatic cell count in Deoni (Bos indicus) cows. Pak Vet J 2015, 35:303-308.
17. Rajić Savić N, Katić V, Velebit B: Characteristics of coagulase positive staphylococci isolated from milk in cases of subclinical mastitis. Acta Vet-Beograd 2014, 64:115–123.
18. Ranjan S, Bhushan B, Panigrahi M, Kumar A, Deb R, Kumar P, Sharma D: Association and expression analysis of single nucleotide polymorphisms of partial tumor necrosis factor alpha gene with mastitis in crossbred cattle. Anim Biot 2015, 26:98-104.
19. Wojdak-Maksymiec K, Szyda J, Strabel T: Parity-dependent association between TNF-alpha and LTF gene polymorphisms and clinical mastitis in dairy cattle. BMC Vet Res 2013, 9:114.
20. Kościuczuk EM, Lisowski P, Jarczak J, Krzyżewski J, Zwierzchowski L, Bagnicka E: Expression patterns of β-defensin and cathelicidin genes in parenchyma of bovine mammary gland infected with coagulase-positive or coagulase-negative Staphylococci. BMC Vet Res 2014, 10:246.
21. Meade KG, Cormican P, Narciandi F, Lloyd A, O'farrelly C: Bovine β-defensin gene family: opportunities to improve animal health? Physiol Genomics 2014, 46:17-28.
22. Piwczyński D: Using classification trees in statistical analysis of discrete sheep reproduction traits. J Cent Eur Agric 2009, 10:303-309.
23. Piwczyński D, Nogalski Z, Sitkowska B: Statistical modeling of calving ease and stillbirths in dairy cattle using the classification tree technique. Livest Sci 2013, 154:19-27.
24. Piwczyński D, Sitkowska B: Statistical modelling of somatic cell counts using the classification tree technique. Arch Tierz 2012, 55:332-345.
25. Piwczyński D, Sitkowska B, Wiśniewska E: Application of classification trees and logistic regression to determine factors responsible for lamb mortality. Small Ruminant Res 2012, 103:225-231.
26. Zia H, Harris N, Merrett G, Rivers M: Predicting discharge using a low complexity machine learning model. Comput Electron Agric 2015, 118:350-360.
27. Brillante L, Gaiotti F, Lovat L, Vincenzi S, Giacosa S, Torchio F, Segade SR, Rolle L, Tomasi D: Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical-mechanical characteristics in wine grapes. Comput Electron Agric 2015, 117:186-193.
28. Hill MG, Connolly PG, Reutemann P, Fletcher D: The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. Comput Electron Agric 2014, 108:250-257.
29. Pulido-Calvo I, Gutiérrez-Estrada JC, Díaz-Rubio E, De La Rosa I: Assisted management of water exchange in traditional semi-intensive aquaculture ponds. Comput Electron Agric 2014, 101:128-134.
30. Wojdak-Maksymiec K, Strabel T, Szyda J, Mikolajczyk K: Clinical mastitis and combined defensin polymorphism in dairy cattle. J Anim Vet Adv 2012, 11:2230-2237.
31. Cornou C, Lundbye-Christensen S: Modeling of sows diurnal activity pattern and detection of parturition using acceleration measurements. Comput Electron Agric 2012, 80:97-104.
32. Statsoft, Inc: Electronic Statistics Textbook. Tulsa, OK: StatSoft; 2013.

33. Ariana DP, Lu R, Guyer DE: Near-infrared hyperspectral reflectance imaging for detection of bruises on pickling cucumbers. Comput Electron Agric 2006, 53:60-70.
34. Cavero D, Tölle KH, Henze C, Buxadé C, Krieter J: Mastitis detection in dairy cows by application of neural networks. Livest Sci 2008, 114:280-286.
35. Krieter J, Cavero D, Henze C: Mastitis detection in dairy cows using neural networks. Agrarinformatik im Spannungsfeld zwischen Regionalisierung und globalen Wertschöpfungsketten. Stuttgart, Germany. 2007.
36. Montgomery ME, White ME, Martin SW: A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows. Can J Vet Res 1987, 51:495.
37. Nielen M, Schukken YH, Brand A, Haring S, Ferwerda-Van Zonneveld RT: Comparison of analysis techniques for on-line detection of clinical mastitis. J Dairy Sci 1995, 78:1050-1061.
38. Nielen M, Spigt MH, Schukken YH, Deluyker HA, Maatje K, Brand A: Application of a neural network to analyse on-line milking parlour data for the detection of clinical mastitis in dairy cows. Prev Vet Med 1995, 22:15-28.
39. Yang XZ, Lacroix R, Wade KM: Neural detection of mastitis from dairy herd improvement records. T ASAE 1999, 42:1063-1071.
40. Chagunda MGG, Friggens NC, Rasmussen MD, Larsen T: A model for detection of individual cow mastitis based on an indicator measured in milk. J Dairy Sci 2006, 89:2980-2998.
41. Sun Z, Samarasinghe S, Jago J: Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks. J. Dairy Res 2010, 77:168-175.
42. Sørensen LP, Bjerring M, Løvendahl P: Monitoring individual cow udder health in automated milking systems using online somatic cell counts. J Dairy Sci 2016, 99:608-620.
43. Kamphuis C, Mollenhorst H, Feelders A, Pietersma D, Hogeveen H: Decision-tree induction to detect clinical mastitis with automatic milking. Comput Electron Agric 2010, 70:60-68.
44. Kamphuis C, Mollenhorst H, Heesterbeek JaP, Hogeveen H: Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision-tree induction. J Dairy Sci 2010, 93:3616-3627.

IDENTIFIKACIJA OSETLJIVOSTI KRAVA NA MASTITIS ZASNOVANA NA ODABRANIM GENOTIPOVIMA KORIŠĆENJEM SISTEMA ODLUKE I GENERALIZOVANOG LINEARNOG MODELA

ZABORSKI Daniel, PROSKURA Witold Stanisław, WOJDAK-MAKSYMIEC Katarzyna, GRZESIAK Wilhelm

Cilj ispitivanja je bio da se: 1) proveriti mogućnost detektovanja, u ranoj fazi eksploatacije, krava prijemčivih na mastitis na osnovu odabranih genotipova i proizvodnih rezultata u prve tri laktacije, i uz upotrebu metoda objedinjavanja podataka (*boosted* *classification trees* – BT and *random forest* – RF), 2) proveri da li uključivanje nekih drugih proizvodnih promenljivih veličina tokom kasnijih laktacija može da poboljša detekciju performansi modela, 3) identifikuju najznačajnije osobine koje mogu da predvide prijemčivost na mastitis, 4) uporede rezultati dobijeni upotrebom BT i RF, sa onima koji su dobijeni tradicionalnom linearnom metodom (GLZ). Ukupno je analiziran 801 podatak koji se odnosio na poljsko frizijsko crno-belo goveče. Maksimalna osetljivost, specifičnost i tačnot testa bila je za BT: 72.13%, 39.73%, 55.90%, za RF 86.89%, 17.81% i 59.49% i za GLZ 90.16%, 8.22% i 58.97%. Obuhvatanje drugih promenljivih veličina nije značajno povećalo značajnost efekta modela. Najznačajniji elementi predviđanja prijemčivosti na mastitis su bili: količina mleka, broj dana u laktaciji, ocena bika i procenat holštajn-frizijskih gena. Sa druge strane, sezona telenja kao i genotipovi (laktoferin, α-faktor nekroze tumora, lizocim i defenzini) su daleko manje bili značajni. Primenjeni modeli, su pokazali nizak nivo tačnosti u detekciji krava koje su bile prijemčive na mastitis. Može se zaključiti da je potrebno da se obuhvate neki drugi kriterijumi i elementi na osnovu kojih bi se uspešnije mogla predvideti povećana osetljivost na mastitis kod krava.