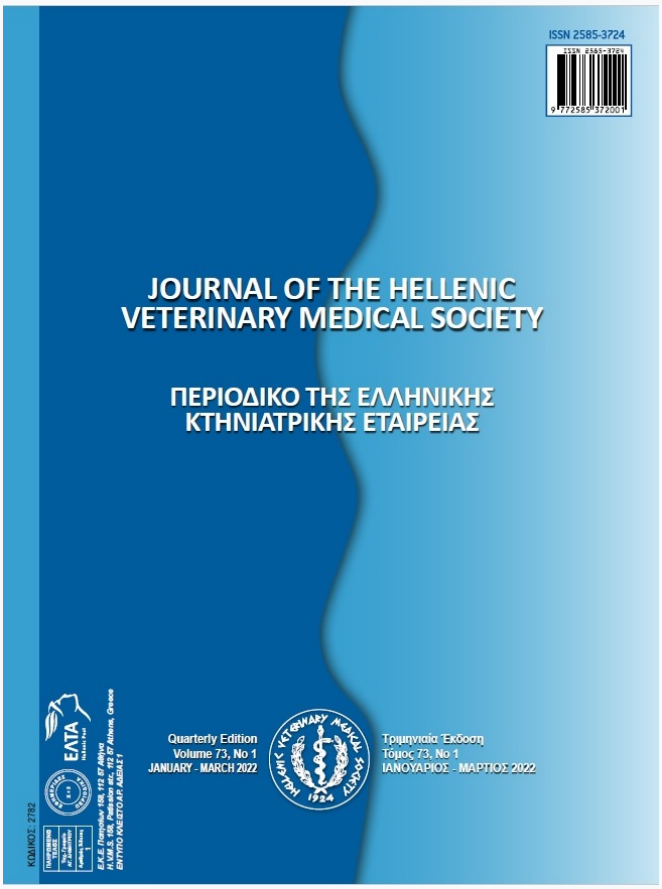
Journal of the Hellenic Veterinary Medical Society

Vol 73, No 1 (2022)



Use of Multivariate Adaptive Regression Splines, Classification Tree and Roc Curve in Diagnosis of Subclinical Mastitis in Dairy Cattle

Yasin ALTAY, İbrahim AYTEKİN, Ecevit EYDURAN

doi: 10.12681/jhvms.25864

To cite this article:

ALTAY, Y., AYTEKİN, İbrahim, & EYDURAN, E. (2022). Use of Multivariate Adaptive Regression Splines, Classification Tree and Roc Curve in Diagnosis of Subclinical Mastitis in Dairy Cattle. *Journal of the Hellenic Veterinary Medical Society*, *73*(1), 3817–3826. https://doi.org/10.12681/jhvms.25864

J HELLENIC VET MED SOC 2022, 73 (1): 3817-3826
ΠΕΚΕ 2022, 73 (1): 3817-3826

Research article
Ερευνητικό άρθρο

Use of Multivariate Adaptive Regression Splines, Classification Trees and ROC Curve in Diagnosis of Subclinical Mastitis in Dairy Cattle

Y. Altay[1], İ. Aytekin[2], E. Eyduran[3]

[1] *Eskisehir Osmangazi University, Department of Animal Science, Biometry and Genetics Unit, Eskisehir, Turkey*

[2] *Selcuk University, Agricultural Faculty, Animal Science Department, Konya, Turkey*

[3] *Iğdır University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Quantitative Methods, Iğdır, Turkey*

ABSTRACT: Subclinical mastitis is one of the most significant diseases that cause economic losses in dairy cattle farming. This research was conducted on 112 heads of Holstein Friesian dairy cattle to reveal the relationship between subclinical mastitis and milk composition and milk quality.In the study, CMT (California Mastitis Test) and CSCC (Classified Somatic Cell Count) used in the diagnosis of subclinical mastitis were used as a binary response variable i.e. healthy and unhealthy. Potential predictors included here were lactation number, days in milk (DIM), darkness-lightness ranges between 0=black and 100=white (L*), green-red ranges between - a*=-60 and a*=+60 (a*), blue-yellow ranges between -b*=-60 and b*=+60 (b*), redness-yellowness (Hue°), vividness-dullness (Chroma), milk fat, milk protein, lactose, milk freezing point, solid non-fat SNF, density, solids, pH, and electrical conductivity. Classification and Regression Tree (CART), Chi-Squared Automatic Interaction Detection (CHAID), Exhaustive Chi-Squared Automatic Interaction Detection (Ex-CHAID), Quick, Unbiased, Efficient, Statistical Tree (QUEST), and multivariate adaptive regression splines (MARS) were used as data mining algorithms that help to make an accurate decision about detecting influential factors increasing the risk of subclinical mastitis.

In conclusion, better classification performances of CART and MARS data mining algorithms were determined compared with those of the remaining algorithms to correctly discriminate between healthy and unhealthy cows.

Keywords: CMT, CSCC, Classification trees, MARS algorithm, Subclinical Mastitis, Milk quality.

Corresponding Author:
Yasin Altay, Faculty of Agriculture, Department of Animal Science, Biometry and Genetics Unit, University of Eskisehir Osmangazi, Eskişehir, Turkey
E-mail address: yaltay@ogu.edu.tr

Date of initial submission: 21-01-2021
Date of revised submission: 18-04-2021
Date of acceptance: 23-04-2021

INTRODUCTION

A major part of total milk production in developed countries is produced by dairy cattle that are more suitable to intensive production compared to small ruminants i.e. sheep and goat.Dairy milk is a significant nutrient-rich food for growing healthy human generations and a vital animal product for progressing the country's economy. The quality of milk that is indispensable for food safety is closely associated with the udder health of cows. Mastitis is one of the important diseases for dairy cattle in the world, and it occurs as a result of inflammation of the udder tissues due to infection.It is an animal welfare problem that negatively affects milk quality and causes economic losses in milk yield per cow (Green et al., 2002; Sharma et al., 2011; Walsh et al., 2011). Variability of mastitis disease was ascribed to genetic (breed, herd) and environmental factors i.e. lactation number, lactation stage, parity, calving month, age, calving year, feeding, and managerial conditions (Aytekin et al., 2018; Sinha et al., 2021). Akdag et al. (2017) reported that udder-related traits were necessary to be considered in dairy cattle breeding strategies. Ural (2013) evaluated the relationship between subclinical mastitis and udder traits in Holstein-Friesian dairy cattle reared in the Aydın province of Turkey and emphasized that it was necessary to ascertain udder traits decreasing SCC amount of the cow's milk for selection strategies. Subclinical mastitis is a significant determinant that adversely affects profitability in a dairy farm, depending on the health status of cows that can change based on managerial conditions.

Detection of subclinical mastitis in dairy cattle is made by California Mastitis Test (CMT), White Side Test (WST), Surf Field Mastitis Test, Sodium Lauryl Sulphate Test (SLST), Somatic Cell Count (SCC), Electrical Conductivity (EC), milk colour-related sensor device, biochemical analyses, and the occurrence of pathogens in milk (Eyduran et al., 2005; Islam et al., 2011; Gáspárdy et al., 2012; Aytekin and Boztepe, 2014; Hoque et al., 2015; Mpatswenumugabo et al., 2017). El-Sayed et al., (2015) studied the relationship of bacteria counts with SCC, EC, and chemical composition of the cow's milk as part of subclinical mastitis diagnosis. Da Costa Ribeiro et al., (2016) mentioned that chlorine and lactose amounts of milk can be used in the diagnosis of subclinical mastitis.

EC, CMT, and SCC diagnosis techniques were appraised relatively as part of the detection of subclinical mastitis (Špakauskas et al., 2006; Kandiwa et al., 2017). It was reported that high SCC amounts in the cow milk decreased casein, lactose, and fat amounts of milk, which shortened milk shelf-life and obstructed conversion of the milk to other milk products (Eyduran et al., 2005). Ribeiro et al, . (2016) reported that changes in lactose, protein, and fat could be with enhancing SCC amount. El-Sayed et al., (2015) reported that EC in the milk of a healthy animal had the range of 4.4 and 5.5 (mS/cm).

Relationships between udder characteristics, milk composition, lactation milk yield and milk quality and composition etc. of subclinical mastitis (SCC or CMT) in dairy cattle were investigated (Sharma et al., 2011; Kaşıkçı et al., 2012; Akdag et al., 2017; De Oliveira Mouraet al., 2017).From the point of improving new breeding strategies, it is imperative to determine the relationship between subclinical mastitis and the amount, quality, composition of the cow's milk together with environmental factors mentioned above. However, there are few previous studies in dairy cattle. Moura et al., (2017) evaluated the correlation between subclinical mastitis (EC and SCC) and the Physico-chemical composition of the milk in Zebu cows. Tiwari et al., (2017) utilized ROC (Receiver Operator Characteristics) analysis to estimate threshold values regarding EC, pH, and SCC that help to diagnose subclinical mastitis of crossbreed cows raised under subtropical conditions. Aytekin et al., (2018) comprehensively investigated the relationship of subclinical mastitis (CMT) with calving month, electrical conductivity, colour, composition, and quality traits of milk in cattle breeds i.e. Brown Swiss and Holstein-Friesian with the methodology of tree-structured CART) data mining algorithm. To make a better decision about the reliable diagnosis of the subclinical mastitis with EC, SCC, and CMT in dairy cattle, application of powerful statistical techniques i.e. Logistic Regression (Altay et al., 2019; Kılıç and Keskin, 2019), CART (Aytekin et al., 2018), support vector machines (SVM, Mammadova and Keskin 2015), artificial neural network (ANN, Mikail and Keskin, 2015, Fuzzy Logic (Coskun and Zulkadir, 2018) and adaptive neuro-fuzzy inference system (ANFIS, Mikail and Keskin, 2015) may be a noteworthy opportunity for animal breeders. However, the applicability of the MARS data mining algorithm, which is a non-parametric regression method that discloses the high dimensional relationship between sets of dependent and explanatory variables without necessitating distributional and functional assumptions of the variables, has not yet been recognized in subclinical

mastitis diagnosis with the aid of EC, SCC and CMT together with quality, composition, and colour traits of the cow's milk in dairy cattle. Therefore, the main aims of this investigation were to find milk quality, composition, and colour traits affecting subclinical mastitis based on CMT and CSCC as binary variables i.e. healthy and unhealthy through CART, CHAID, Exhaustive CHAID, QUEST, and MARS classification algorithms, and to obtain evidence to select the best subclinical mastitis diagnosis method between CMT and CSCC, depending upon classification performances of the data mining algorithms.

MATERIALS AND METHODS

Materials

In this study, 112 head of Holstein Friesian cattle used were obtained from a private farm in Konya, Turkey. Milk samples were collected in the morning milkingduringthe summer of 2019.Milk sampling was performed with primiparous dairy cows averaging 152.30±14.95 (Mean±SE) days in milk (DIM) and was taken from cows without a clinical history of mastitis in herd. All cows fed ad libitum with a mixture of concentrated feed and forage such as straw, alfalfa, fescue grass, and corn silage.

Milking, Milk samples, and Milk analysis

Holstein Friesian cows were housed in a free-stall barn and fed ad libitum with a mixture of concentrated feed and forage as total mixed ration (TMR). Dairy cows milked three times daily in a 2 x 15 parallel milking parlor ENGS, EcoHerd, Version 1.01). Since there is no classification tree-based power analysis, a power analysis based on logistic regression, which is the simplest classification, has been performed. The probability of mastitis = 0.3235, odds ratio = 4.3731, and the number of animals with 90% power calculated by reference to the study (Altay et al., 2019) is 97.In order to have the power of the test in the study over 90% and to provide a homogeneous data structure, the number of animals was determined as 112.The data of 112 cows with first lactation number were obtained from the herd management system. Milk samples were obtained from 112 dairy cattle that had been milked with two milkers by using sampling equipment during milking time to represent homogeneous of all milk. After the morning milking, analyses were immediately conducted. Fat (%), protein (%), lactose (%), freezing point (°C), SNF (%), density (kg/m^3), total solids (%), pH, and conductivity (uS/cm) traits, which were examined as milk components, were immediately analyzed two times with the help of an ultrasonic milk analyser (LACTOSCAN MMC30, Milkotronic Ltd, Bulgaria). Somatic Cell Count (SCC) in milk was analyzed by an electronic counter (Nucleocounter SCC-100, Chemometec, Denmark). California mastitis test (CMT) scores of all samples were determined by using a same solution, equipment and expert. Milk samples were homogeneously taken from each cow at the milking by using milk sampler. Then, milk samples were placed in a plastic test paddle, divided into 4 separate wells, in order to determine mastitis status. CMT solution was added on the milk samples taken and after mixing same direction in an oval shape for about 20 seconds, it was diagnosed by the expert (Shitandi and Kihumbu, 2004). All milk samples were screened for subclinical mastitis by the CMT to determine the healthy and unhealthy status of cows. Also, colour characteristics of milk samples were measured for CIELAB system measuring parameters by the Minolta Chroma Meter CR-400 (Konica Minolta, Inc., Osaka, Japan) (CIELAB, 1976). By using The L*, a*, and b* colour values, Hue°and Chroma values were calculated using the formula Hue° =Tan^{-1} x (b*/a*) and Chroma =$\sqrt{}$ (a^ (*2) +) b^ (*2).

Statistical analysis

CSCC and CMT as a subclinical mastitis diagnosis test were binary dependent variables i.e. healthy and unhealthy. Animals whose CSCC amount is less than 200 000 in 1 cc milk were accepted as healthy; otherwise, they were considered unhealthy. Classification performances of CART (Breiman et al., 1984), CHAID (Kass, 1980), Exhaustive CHAID (Biggs et al., 1991), QUEST (Loh and Shih, 1997), and MARS (Friedman, 1991) algorithms were evaluated comparatively based on accuracy, sensitivity, specificity and area under ROC curve. CART, CHAID, Exhaustive CHAID, and QUEST produces a tree structure to yield the highest accuracy rates as soon as possible. CART (Kovalchuk et al., 2017; Kovalchuk et al., 2018) and QUEST work according to binary node splitting rule, but CHAID and Exhaustive CHAID algorithms run based on multiway node splitting rule (Akin et al., 2018). As a modified form of the CART algorithm, MARS is used to find predictors with hinges function for a better solution of binary logistic regression. Maximum tree depth was used for CART (5), QUEST (5), and both CHAID algorithms (3) by default. In the 5-fold cross-validation, the whole data set (112 records) was randomly separated into 10 approx. equal parts of 21 or 22 records, from which nine were

used to train a given type of prediction model and one served as an independent test set. This process was repeated 5 times. The minimum number of parent and child nodes was 10 and 5 for decision trees i.e. CART, CHAID, Exhaustive CHAID, and QUEST. Optimal trees of the decision tree algorithms were produced after their resubstition costs were very close to corresponding cross-validation costs (Tyasi et al., 2021). Accuracy is an algorithm's proportion of correctly classifying healthy and unhealthy animals. Sensitivity is the algorithm's proportion of correctly classifying unhealthy animals. Specificity is the algorithm's proportion of correctly classifying healthy animals (Grzesiak and Zaborski, 2012). The confusionmatrix for the classifier algorithms is given in Table 1.

Table 1. Confusion table for the classifier algorithms

		Predicted as	
		Unhealthy	**Healthy**
Observed	**Unhealthy**	A	B
	Healthy	C	D

The expressions A, D, B, and C gave in the following equation represent the numbers of true positive, true negative, false positive, and false negative, respectively. The formula developed by (Hanley and McNei, 1982) was used to determine AUC (AUCse).

Accuracy = (A+D) / (A+B+C+D)

Sensitivity = A / (A+B)

Specificity = D / (C+D)

Error proportion = 1 - Accuracy

$$^{se}AUC = \sqrt{\frac{AUC(1-AUC)+(n_A-1)(q1-AUC^2)+(n_B-1)(q2-AUC^2)}{n_A n_B}}$$

n_A=A+C and n_B=B+D

$$q1 = \frac{AUC}{2-AUC} \quad \text{and} \quad q2 = \frac{2AUC^2}{1+AUC}$$

Pairs of algorithms in the area under ROC curve were compared based onthe z test.

Statistical analyses associated with CART, CHAID, Exhaustive CHAID, and QUEST were IBM SPSS 23 (IBM Corp. Released, 2015). MARS analysis was performed using earth (v5.1.2; Milborrow, 2019) and caret (v6.0.86; Kuhn, 2020) packages of R software (R Core Team, 2020; Kuhn and Johnson, 2013; Eyduran et al., 2019; Akin et al., 2020). The trial version 19.5.1 of the MedCalc software was used to calculate the area under ROC curve and comparison (AUC) and to compare pairs of algorithms in the area. Also, logistic regression-based power analysis used to determine the sample size in the study was performed in G*Power package program version 3.1.7 (Faul et al., 2013).

RESULTS AND DISCUSSION

Table 2 presents descriptive statistics of milk quality, composition, and colour traits for each diagnostic test. Although the method averages are close to each other in terms of the traits considered in general (except for SSC), it was observed that there were differences between the average traits of healthy and unhealthy animals regardless of the method.

Table 2. Descriptive statistics of parameters of milk quality for each diagnosis test

Variables	Methods	Diagnosis	N	Minimum	Maximum	Mean±SE	StDev	CoefVar
DIM (day)	**CMT**	**Healthy**	63	6.00	499.00	123.20±13.20	105.00	85.16
		Unhealthy	49	11.00	424.00	181.40±16.70	116.80	64.39
	CSCC	**Healthy**	77	6.00	499.00	128.50±12.50	110.10	85.72
		Unhealthy	35	22.00	424.00	193.20±18.50	109.70	56.76
Morning Milk (kg)	**CMT**	**Healthy**	63	6.90	21.60	13.14±0.49	3.87	29.48
		Unhealthy	49	1.90	19.10	11.01±0.59	4.12	37.43
	CSCC	**Healthy**	77	5.60	21.60	12.98±0.43	3.80	29.27
		Unhealthy	35	1.90	18.80	10.52±0.73	4.30	40.86
L	**CMT**	**Healthy**	63	81.42	89.75	86.90±0.21	1.64	1.89
		Unhealthy	49	82.98	89.35	86.25±0.20	1.38	1.59
	CSCC	**Healthy**	77	81.42	89.75	86.86±0.18	1.61	1.86
		Unhealthy	35	82.98	88.56	86.07±0.22	1.28	1.49
a	**CMT**	**Healthy**	63	-3.55	4.11	-2.34±0.14	1.09	-46.33
		Unhealthy	49	-3.56	-1.71	-2.70±0.07	0.49	-17.98
	CSCC	**Healthy**	77	-3.55	4.11	-2.33±0.11	0.99	-42.39
		Unhealthy	35	-3.56	-1.95	-2.87±0.08	0.45	-15.55

b	**CMT**	**Healthy**	63	-0.45	6.95	2.65±0.19	1.50	56.56
		Unhealthy	49	-0.08	8.05	3.63±0.30	2.12	58.29
	CSCC	**Healthy**	77	-0.45	6.95	2.58±0.17	1.46	56.38
		Unhealthy	35	0.16	8.05	4.16±0.37	2.17	52.05
H	**CMT**	**Healthy**	63	-66.73	59.40	-38.65±3.05	24.21	-62.63
		Unhealthy	49	-66.14	1.83	-48.11±2.50	17.51	-36.40
	CSCC	**Healthy**	77	-66.73	59.40	-39.19±2.67	23.42	-59.75
		Unhealthy	35	-66.14	-3.86	-50.70±2.70	15.95	-31.47
C	**CMT**	**Healthy**	63	1.91	8.07	3.80±0.15	1.20	31.55
		Unhealthy	49	2.37	8.80	4.68±0.26	1.82	38.84
	CSCC	**Healthy**	77	1.91	8.07	3.73±0.13	1.14	30.57
		Unhealthy	35	2.38	8.80	5.18±0.32	1.87	36.05
Fat (%)	**CMT**	**Healthy**	63	1.97	4.98	3.53±0.09	0.70	19.68
		Unhealthy	49	2.22	4.96	3.85±0.09	0.60	15.70
	CSCC	**Healthy**	77	1.97	4.98	3.56±0.07	0.66	18.49
		Unhealthy	35	2.22	4.96	3.93±0.11	0.64	16.37
Protein (%)	**CMT**	**Healthy**	63	2.71	3.53	3.24±0.02	0.14	4.33
		Unhealthy	49	2.62	3.52	3.24±0.02	0.15	4.67
	CSCC	**Healthy**	77	2.62	3.53	3.22±0.02	0.16	4.83
		Unhealthy	35	3.06	3.52	3.27±0.02	0.11	3.40
Lactose (%)	**CMT**	**Healthy**	63	4.06	5.28	4.85±0.03	0.21	4.31
		Unhealthy	49	3.92	5.27	4.84±0.03	0.23	4.73
	CSCC	**Healthy**	77	3.92	5.28	4.83±0.03	0.23	4.82
		Unhealthy	35	4.57	5.27	4.89±0.03	0.17	3.54
Freezing Point (°C)	**CMT**	**Healthy**	63	-0.62	-0.47	-0.56±0.01	0.03	-4.90
		Unhealthy	49	-0.63	-0.45	-0.57±0.01	0.03	-5.38
	CSCC	**Healthy**	77	-0.62	-0.45	-0.56±0.01	0.03	-5.39
		Unhealthy	35	-0.63	-0.53	-0.57±0.01	0.02	-4.21
SNF (%)	**CMT**	**Healthy**	63	7.39	9.62	8.83±0.05	0.38	4.30
		Unhealthy	49	7.13	9.59	8.81±0.06	0.42	4.73
	CSCC	**Healthy**	77	7.13	9.62	8.79±0.05	0.42	4.83
		Unhealthy	35	8.32	9.59	8.90±0.05	0.31	3.53
Density (kg/m^3)	**CMT**	**Healthy**	63	1023.97	1033.29	1030.59±0.19	1.52	4.98
		Unhealthy	49	1023.74	1032.43	1030.24±0.22	1.55	5.12
	CSCC	**Healthy**	77	1023.74	1033.29	1030.41±0.19	1.70	5.58
		Unhealthy	35	1027.49	1032.43	1030.49±0.19	1.12	3.68
Solids (%)	**CMT**	**Healthy**	63	0.61	0.79	0.73±0.01	0.03	4.27
		Unhealthy	49	0.59	0.79	0.72±0.01	0.03	4.69
	CSCC	**Healthy**	77	0.59	0.79	0.72±0.01	0.03	4.78
		Unhealthy	35	0.68	0.79	0.73±0.01	0.02	3.51
pH	**CMT**	**Healthy**	63	6.30	7.00	6.53±0.03	0.22	3.34
		Unhealthy	49	5.64	7.08	6.64±0.04	0.31	4.61
	CSCC	**Healthy**	77	6.26	7.00	6.52±0.02	0.2	3.24
		Unhealthy	35	5.64	7.08	6.71±0.05	0.32	4.77
Conductivity (µS/cm)	**CMT**	**Healthy**	63	3.35	6.22	5.05±0.09	0.74	14.58
		Unhealthy	49	3.36	7.31	4.91±0.13	0.87	17.76
	CSCC	**Healthy**	77	3.35	7.31	5.14±0.09	0.75	14.56
		Unhealthy	35	3.36	6.05	4.67±0.14	0.82	17.56

Table 3 presents the classification performances of the tested data mining algorithms for each subclinical mastitis diagnosis. The areas under ROC (AUC) were found to be significant in all algorithms used in the diagnosis of subclinical mastitis ($P < 0.05$).

The best classification performance for the CMT diagnosis test was recorded for MARS data mining algorithms with sensitivity (0.857), specificity (0.809), and accuracy (0.830) rates above 0.80. MARS algorithm correctly classified 85.7 (%) of unhealthy cows, 80.9 (%) of healthy cows, and 83.0 (%) of all the cows. Based on the CMT diagnosis test, MARS had the biggest area under ROC curve of 0.869 as the best agreement between sensitivity and specificity in the CMT diagnosis test. The best classifier algorithm for detecting subclinical mastitis using CMT was found as MARS, followed by CART. The present CART findings were in near agreement with those recorded by Aytekin et al. (2018) who found that 77.2 (%) of the Brown Swiss and Holstein-Friesian cows were unhealthy (sensitivity), 95.7 (%) of them were healthy in CMT diagnostic test.

Among classifier algorithms whose performances were tested for CSCC subclinical mastitis diagnosis test in the current work, CART was determined to be a promising algorithm that had a sensitivity (0.917), specificity (0.725), accuracy (0.848) rates, and the largest area of 0.890 under ROC curve. CART algorithm correctly classified 91.7 (%) of the cows detected as unhealthy by CSCC, correctly classified 84.8 (%) of healthy and unhealthy cows. Based on the CSCC subclinical diagnosis test, MARS had the biggest area under ROC curve of 0.890 as the best agreement between sensitivity and specificity in CSCC.

As a result of the CSCC diagnosis test, the best agreement between the sensitivity and specificity rates was recorded in the CART classifier according to the largest area under ROC curve constructed for the classifier and the highest sensitivity rate of 91.7 (%). Among the combinations of the algorithm and the diagnosis test, the best one was understood to be the combination of the CART algorithm and CSCC diagnosis test. It could be suggested that CSCC that correctly captured 91.7 (%) of unhealthy cows was the best subclinical diagnosis test in the CART classifier. However, the MARS algorithm correctly classified 93.5 (%) of healthy cows in the CSCC diagnosis test as a result of the highest specificity rate estimated for the algorithm. Higher accuracy estimates of the algorithms evaluated for the CSCC diagnosis test here were obtained (0.786 to 0.848) compared to those recorded by Mammadova and Keskin (2015) with an accuracy rate of 50 (%) for support vector machines and Mikail and Keskin (2015) with the accuracy rates of 36 and 65 (%) for ANN and ANFIS algorithms in SCC diagnosis test. Aytekin et al, . (2018) emphasized easy interpretation of the visual results obtained from CART in the CMT diagnosis test.

The flexible MARS model providing very high classification quality for CMT diagnosis test was produced by two milk colour traits i.e. “a” and “C”.

$GLM_{unhealthy}$ = -0.9504965 - 1.043345 * max (0, a + 2.97) + 1.154928 * max (0, C - 4.46). In this case, the probability of being unhealthy for any cow based on the CMT diagnosis test can be estimated with the help of $P_{UNHEALTHY} = \exp^{GLMunhealthy} / (1 + \exp^{GLMunhealthy})$ where exp the base of natural logarithm whose value is 2.718. For example, the probability of being unhealthy for a cow with a= -2.20 and C=3.29 based on the CMT diagnosis test can be calculated as follows:

$GLM_{unhealthy}$ = -0.9504965 - 1.043345 * max (0, -2.20 + 2.97) + 1.154928 * max (0, 3.29 - 4.46).

$GLM_{unhealthy}$ = -0.9504965 - 1.043345 * max (0, -2.20 + 2.97) + 1.154928 * max (0, 3.29 - 4.46).

Where, max (0, -2.20 + 2.97) = max (0, 0.77) = 0.77 max (0, 3.29 - 4.46) = max (0, -1.17) =0

$GLM_{unhealthy}$ = -0.9504965 - 1.043345 * 0.77 + 1.154928 *0

$GLM_{unhealthy}$ = -0.9504965 - 1.043345 * 0.77 = -1.75387215

$P_{UNHEALTHY} = \exp^{GLMunhealthy} / (1 + \exp^{GLMunhealthy})$

$P_{UNHEALTHY} = \exp^{(-1.75387215)} / (1 + \exp^{(-1.75387215)}) = 0.1731024 / (1+0.1731024)$

$P_{UNHEALTHY}$ = 0.1475595 is expressed as the probability of being unhealthy for a cow with a= -2.20 and C=3.29).

Table 3. Classification performances of the algorithms for each diagnosis test

Methods	Algorithm	Sensitivity	Specificity	AUC	Accuracy	Associated Criterion	P-value
	CART	0.716	0.839	0.742	0.750	0.5788	0.000
	CHAID	0.644	0.773	0.634	0.670	0.5631	0.015
CMT	Exhaustive CHAID	0.716	0.667	0.740	0.696	0.3717	0.000
	QUEST	0.644	0.720	0.628	0.661	0.5382	0.020
	MARS	0.857	0.809	0.869	0.830	0.3894	0.000
	CART	0.917	0.725	0.890	0.848	0.3394	0.000
	CHAID	0.800	0.773	0.710	0.795	0.4864	0.000
CSCC	Exhaustive CHAID	0.800	0.773	0.710	0.795	0.4864	0.000
	QUEST	0.791	0.762	0.696	0.786	0.4853	0.010
	MARS	0.514	0.935	0.776	0.804	0.4843	0.004

Comparison of the algorithms in the area under the ROC curve for both subclinical mastitis diagnostic tests is given in Table 4. The classification performance order for CMT diagnostic test was MARS > CART = Exhaustive CHAID > CHAID = QUEST in area under ROC curve. The classification performance order recorded for CSCC diagnostic test was CART > MARS > CHAID=Exhaustive CHAID=QUEST in terms of the area under ROC curve. CSCC was found superior to CMT in the area under ROC curve of the CART classifier. But, no significant difference was found between CSCC and CMT tests in the area under ROC curve for CHAID, QUEST, and MARS classifiers. A significant point is that an area of 0.869 under ROC curve was estimated as the second-largest value numerically (Table 4). The present ROC curve results indicated that most of the algorithms showed similar trends for CMT and CSCC in comparing subclinical mastitis tests within CHAID, Exhaustive CHAID, QUEST, and MARS algorithms, which was in agreement with those reported by some authors who reported the suitability of EC with CMT and SCC (Kaşıkçı et al., 2012; Aytekin et al., 2018). Mammadova and Keskin (2015) obtained 50% accuracy, 92% specificity, and 89% sensitivity using support vector machines algorithm as well as 43% accuracy, 79% specificity, and 75% sensitivity using binary logistic regression analysis based on SCC subclinical mastitis diagnosis test. Both accuracy rates reported by Mammadova and Keskin (2015) were lower than those of the tested algorithms employed for CSCC and CMT diagnosis tests here.

Table 4. Comparison of the algorithms in the area under ROC curve

Methods	CART	CHAID	Exhause-CHAID	QUEST	MARS
CMT	0.742 ± 0.050^{Bb}	0.634 ± 0.054^{Ca}	0.740 ± 0.047^{Ba}	0.628 ± 0.054^{Ca}	0.869 ± 0.036^{Aa}
CSCC	0.890 ± 0.035^{Aa}	0.710 ± 0.052^{Ba}	0.710 ± 0.058^{Ba}	0.696 ± 0.059^{Ba}	0.776 ± 0.058^{Ba}

$^{A, B}$The difference between the algorithms with the capital letter in CMT or CSCC row is significant (comparison of the algorithms)
$^{a, b}$The difference between diagnostic tests with the letter in any algorithm column is significant (comparison of the subclinical mastitis diagnostic tests)

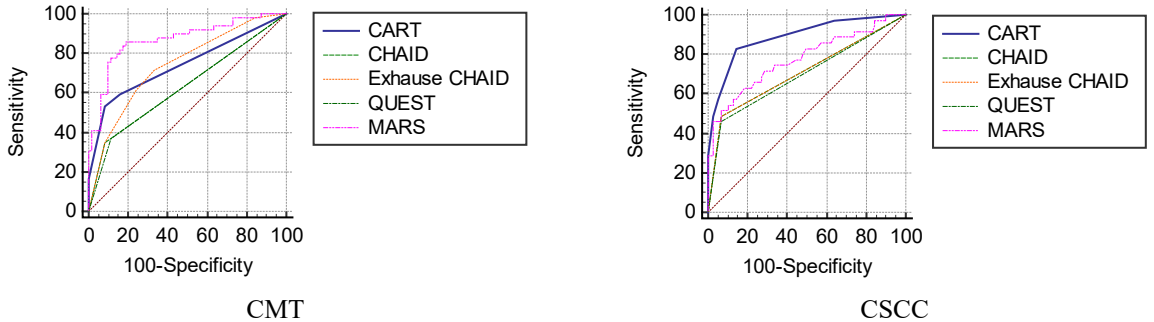


Figure 1. ROC curves of classifier algorithms for each diagnosis test

In the CSCC method used in the diagnosis of subclinical mastitis, Aytekin et al, . (2018) reported that milk color features a *, C *, fat and DIM compatible markers, while in CMT method, EC, “L” and “a”, fat, freezing point, and calving month properties. It has been identified as an important indicator.

As part of CSCC as a diagnostic test for subclinical mastitis, CART was selected as the best classifier (Table 1). 68.8 (%) of the 112 cows in the CSCC diagnosis test were characterized as healthy, and 31.2 (%) of them were classified as unhealthy (Node 0).

The CART classification tree of the CSCC diagnosis test is presented in Figure 2. At the top of the CART classification tree structure, Node 0 was divided into two smaller subgroups i.e. Node 1 (a subgroup of the cows with C < 5.195) and Node 2 (a subgroup of the cows with C > 5.195) according to C milk colour trait at the first tree depth. 81.4 (%) of the cows with C < 5.195 were determined as healthy (Node 1) ; however, 73.1 (%) of the cows with C > 5.195 were characterized as unhealthy (Node 2).

At the second tree depth, Node 1 was branched into two smaller subgroups i.e. Node 3 and Node 4 according to the DIM trait. 96.6 (%) of the cows with C < 5.195 and DIM < 53.5 days were found as healthy (Node 3). 73.7 (%) of the cows with C < 5.195 and DIM > 53.5 days were described as healthy.

At the second tree depth, Node 2 was split into two subgroups i.e. Node 5 and Node 6 according to “a” milk colour trait.

All of the cows with C > 5.195 and a < - 3.195 were determined as unhealthy (Node 5), whereas 56.2 (%) of the cows with C > 5.195 and a > - 3.195 were identified as unhealthy (Node 6). Cut-off values of 5.195 C and - 3.195 a could provide some hints for breeders to detect subclinical mastitis.

At the third tree depth, Node 4 was partitioned into two smaller subgroups i.e. Node 7 and Node 8 according to the DIM trait. In the CART tree diagram, 77.8 (%) of the cows with C < 5.195 and 54.5 < DIM < 77.5 days were recorded as unhealthy (Node 7), but 83.3 (%) of the cows with C < 5.195 and DIM > 77.5 days were characterized as healthy (Node 8).

At the fourth tree depth of the CART algorithm, Node 8 was divided into two smaller subgroups i.e. Node 9 and Node 10 according to milk fat trait. 88.4 (%) of the cows with C < 5.195, DIM > 77.5 days, and milk fat < 4.250 (%) were ascertained as healthy. 60 (%) of the cows with C < 5.195, DIM > 77.5 days and milk fat > 4.250 (Node 10) were found as unhealthy.

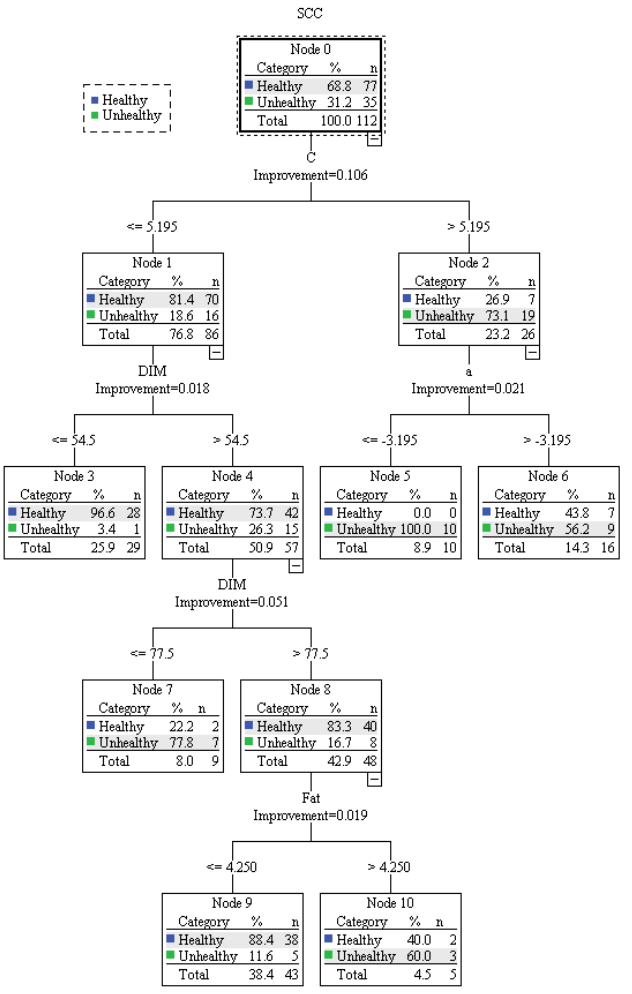


Figure2. CART classification tree of the CSCC diagnosis test.

Aytekin et al. (2018) reported that the CART classification tree created for the CMT subclinical mastitis diagnosis test consists of calving month, EC, milk fat, milk freezing point and milk color (L* and a*). Whereas, the MARS data mining algorithm gave the best classification performance for the CMT and produced an easier interpretable equation for more correctly classifying healthy and unhealthy cows when compared with the results of CART, ANN, ANFIS, SVM algorithms published elsewhere (Mammadova and Keskin 2015; Mikail and Keskin 2015). In this regard, the use of the MARS algorithm that helps to find cut-off values of significant milk traits discriminating unhealthy and healthy cows is advisable together with subclinical mastitis diagnosis tests.

CART analysis produced by Aytekin et al., (2018) reflected that subclinical mastitis risk of cows with

EC > 5.695 and fat > 3.160 (%) was 3.62 times more in comparison with that of those with EC > 5.695 and fat < 3.160. In the study, the risk of cows (60%) with milk fat > 4.250 (Node 10) was found to be 5.17 times more than the risk of cows with fat < 4. 250 (11.6%) (Node 9) when C < 5.195 and DIM > 77.5 days were considered. CSCC and CMT diagnosis tests produced almost similar tendenciesbased on QUEST, MARS, and both CHAID algorithms (Table 3). This finding was nearly in agreement with the previous statements of some authors (Kaşıkçı et al., 2012; Aytekin et al., 2018).

Wide variation in mastitis disease diagnosis was attributed to genetic (breed, herd) and environmental factors i.e. lactation number, lactation stage, lactation period, parity, udder traits, calving month, age, calving year, feeding and managerial conditions, especially statistical analysis techniques.

With the scope of subclinical mastitis detection, the usability of CHAID and especially MARS algorithms were scarce in the literature to reveal the relationship of the mastitis risk with milk quality, composition, colour traits. In this respect, further studies are still required for subclinical mastitis diagnosis with the support of data mining algorithms.

CONCLUSION

In herd management, it is desirable to identify unhealthy animals as soon as possible to obtain healthy products for animal health, and also human health. Also, the failure of farms to distinguish between healthy and unhealthy animals will cause economic losses because of management. With the development of data mining methods in recent years, it will benefit herd management by closing the gaps in diagnosing mastitis. In diagnosis, increasing the detection of mastitis in unhealthy animals and decreasing the margin of error will make a great contribution to herd management.

In the current work, the relationship of subclinical mastitis with milk quality, colour, and composition traits was investigated based on data mining algorithms i.e. CART, CHAID, Exhaustive CHAID, QUEST, and MARS. Among these algorithms, MARS gave the best performance by correctly classifying unhealthy and healthy animals in CMT diagnosis, while the CART algorithm was found to be the best classifier in CSCC ($P < 0.05$). According to the results of CART analysis obtained for SCC, all cows with a < -3.195 and C > 5.195 in their milk could be said to be at risk of subclinical mastitis. The subclinical mastitis risk of cows with C > 5.195 was observed to be four times more than that of those with C < 5.195 in their milk. No significant differences of CSCC and CMT diagnosis tests in the area under ROC curve for other algorithms except for CART were obtained. For a generalization of the current results, the examination of much larger populations in different cattle breeds was recommendable.

In conclusion, the application of CART and MARS algorithms may a good choice for cattle breeders to find cut-off values of influential milk traits correctly discriminating healthy and unhealthy cows.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Akin M, Eyduran SP, Eyduran E, Reed BM (2020) Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. Plant Cell, Tissue and Organ Culture (PCTOC), 140 (3), 661-670.

Akin M, Hand C, Eyduran E, Reed BM (2018) Predicting minor nutrient requirements of hazelnut shoot cultures using regression trees, Plant Cell, Tissue and Organ Culture (PCTOC), 132 (3), 545-559.

Altay Y, Kılıç B, Aytekin I, Keskin I (2019) Determination of factors affecting mastitis in Holstein Friesian and Brown Swiss by using logistic regression analysis, Selcuk Journal of Agriculture and Food Sciences, 33 (3), 194-197.

Akdag F, Gürler H, Teke B, Ugurlu M, Koçak O. (2017) The effect of the difference in the evaluation of cmt scores and scores in jersey cows on milk yield, milk components and subclinical mastitis diagnosis, Journal of the Faculty of Veterinary Medicine, 43 (1), 44-52.

Aytekin I, Eyduran E, Keskin I (2018) Detecting the relationship of california mastitis test (CMT) with electrical conductivity, composition and quality of the milk in Holstein-Friesian and Brown Swiss cattle breeds using cart analysis, Fresenius Environmental Bulletin, 27 (6), 4559-4565.

Aytekin İ, Boztepe S (2014) Somatic cell count, importance and effect factors in dairy cattle, Turkish Journal of Agriculture - Food Science and Technology, 2, 112-121.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). Classification and regression trees. Chapman & Hall/CRC

Biggs D, De Ville B, Suen E (1991) A method of choosing multiway partitions for classifi cation and decision trees. Journal of Applied Statistics, 18 (1), 49-62.

Coskun FS, Zulkadir U (2018) The use of fuzzy logic approach in evaluation of subclinic mastitis, Selcuk Journal of Agriculture and Food Sciences, 32 (3), 436-439.

Da Costa Ribeiro AB, Sifuentes J, Zanol D, Lombarde LL (2016) Evaluation of an electrical conductivity portable device as an alternative for subclinical mastitis detection, Revista de Salud Animal, 38, 131-135.

De Oliveira Moura E, Do Nascimento Rangel AH, Borba LHF, Júnior JGBG, Da Costa Lima GF, De Lima Júnior DM, De Aguiar EM (2017) Electrical conductivity and somatic cell count in zebu cow's milk, Semina: Ciências Agrárias, 38 (5), 3231-3240.

El-Sayed SM, Awad IE, Shalapy SM (2015) A study on the bacteria causing subclinical mastitis in dairy cows and its effect on somatic cell count and milk chemical composition parameters, Zagazig Veterinary Journal, 43 (1), 26-35.

Eyduran E, Ozdemir T, Yazgan K, Keskin S (2005) The effects of lactation rank and period on somatic cell count (scc) in milks of holstein cows, Van Veterinary Journal. 16, 61-65.

Eyduran E, Akin M, Eyduran SP (2019) Application of multivariate adaptive regression splines in agricultural sciences through R software. Ankara: Nobel Academic Publishing.

Faul F, Erdfelder E, Buchner A, Lang AG (2013) G* Power 3.1. 7 [computer software]. Uiversitt Kiel, Germany. Retried from http:// www. psycho. uniduesseldorf. de/abteilungen/aap/gpower3/download-and-register.

Friedman JH (1991) Multivariate adaptive regression splines, Annals of Statistics. 19, 1-141.

IBM Corp. Released, (2015). IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

Islam MA, Islam MZ, Rahman MS, Islam MT (2011) Prevalence of subclinical mastitis in dairy cows in selected areas of Bangladesh, Bangladesh Journal of Veterinary Medicine, 9 (1), 73-78.

Gáspárdy A, Ismach G, Bajcsy Á, Veress G, Márkus S, Komlósi I (2012) Evaluation of the on-line electrical conductivity of milk in mastitic dairy cows, Acta Veterinaria Hungarica. 60, 145-55.

Green LE, Hedges VJ, Schukken YH, Blowey RW, Packington AJ (2002). The impact of clinical lameness on the milk yield of dairy cows. Journal of Dairy Science, 85 (9), 2250-2256.

Grzesiak W, Zaborski D (2012) Examples of the use of data mining methods in animal breeding. In: Data mining applications in engineering and medicine (ed. A Karahoca). InTech, Rijeka, Croatia, pp. 303-324. https://doi.org/10.5772/50893

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143, 29-36.

Hoque MN, Das ZC, Talukder AK, Alam MS, Rahman ANMA. (2015). Different screening tests and milk somatic cell count for the prevalence of subclinical bovine mastitis in Bangladesh. Tropical Animal Health and Production, 47 (1), 79-86.

Kandiwa E, Iraguha B, Mushonga B, Hamudikuwanda H, Mpatswenumugabo JP. (2017). Comparison of cow-side diagnostic tests for subclinical mastitis of dairy cows in Musanze district, Rwanda. Journal of the South African Veterinary Association, 88 (1), 1-6.

Kass GV (1980) An exploratory technique for investigating large quantities of categorical data, Journal of the Royal Statistical Society: Series C (Applied Statistics), 29 (2), 119-127.

Kaşıkçı G, Cetin O, Bingol EB, Gunduz MC (2012) Relations between electrical conductivity, somatic cell count, California mastitis test and some quality parameters in the diagnosis of subclinical mastitis in dairy cows, Turkish Journal of Veterinary and Animal Sciences. 36, 49-55.

Kılıç B, Keskin I (2019) Determination of factors effective in diagnosis of mastitis in holstein cattle by logistic regression analysis, Journal of Bahri Dagdas Animal Research, 8 (2), 46-55.

Kovalchuk IY, Mukhitdinova Z, Turdiyev T, Madiyeva G, Akin M, Eyduran E, Reed BM (2017) Modeling some mineral nutrient requirements for micropropagated wild apricot shoot cultures, Plant Cell, Tissue and Organ Culture (PCTOC), 129 (2), 325-335.

Kovalchuk IY, Mukhitdinova Z, Turdiyev T, Madiyeva G, Akin M, Eyduran E, Reed BM (2018) Nitrogen ions and nitrogen ion proportions impact the growth of apricot (prunus armeniaca) shoot cultures, Plant Cell, Tissue and Organ Culture (PCTOC), 133 (2), 263-273.

Kuhn M (2020) Caret: Classification and regression training. Retrieved from https://CRAN.R-project.org/package=caret

Kuhn M, Johnson K (2013) Applied predictive modeling. New York: Springer.

Loh WY, Shih YS (1997) Split selection methods for classification trees, Statistica Sinica, 7, 815-840.

Mammadova NM, Keskin I (2015) Application of neural network and adaptive neuro-fuzzy inference system to predict subclinical mastitis in dairy cattle, Indian Journal of Animal Research. 49, 671-679.

Mikail N, Keskin I (2015) Subclinical mastitis prediction in dairy cattle by application of fuzzy logic, Pakistan Journal of Agricultural Sciences. 52, 1101-1107.

Milborrow S (2019) Earth: Multivariate adaptive regression splines. Retrieved from https://CRAN.R-project.org/package=earth

Mpatswenumugabo JP, Bebora LC, Gitao GC, Mobegi VA, Iraguha B, Kamana O, Shumbusho B (2017) Prevalence of subclinical mastitis and distribution of pathogens in dairy farms of Rubavu and Nyabihu districts, Rwanda. Journal of Veterinary Medicine, 1-8, https://doi.org/10.1155/2017/8456713.

R Core Team (2020) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sharma N, Singh NK, Bhadwal MS (2011) Relationship of somatic cell count and mastitis: An overview, Asian-Australasian Journal of Animal Sciences, 24 (3), 429-438.

Shitandi A, Kihumbu G (2004) Assessment of the California mastitis test usage in smallholder dairy herds and risk of violative antimicrobial residues. Journal of Veterinary Science, 5 (1), 5-10.

Sinha R, Sinha B, Kumari R, VineethMR, Verma A, Gupta ID. (2021). Effect of season, stage of lactation, parity and level of milk production on incidence of clinical mastitis in Karan Fries and Sahiwal cows. Biological Rhythm Research, 52 (4), 593-602.

Špakauskas V, Klimien I, Matuseviius A (2006) A comparison of indirect methods for diagnosis of subclinical mastitis in lactating dairy cows, Veterinarski arhiv. 76, 101-109.

Ural AD (2013) The relationships among some udder traits and somatic cell count in holstein-friesian cows, Kafkas Universitesi Veteriner Fakultesi Dergisi, 19 (4), 601-613.

Tiwari S, Mohanty TK, Patbandha TK, Kumaresan A, Bhakat M, Kumar N, Baithalu RK (2017) Critical thresholds of milk scc, ec and ph for detection of sub-clinical mastitis in crossbred cows reared under subtropical agroclimatic condition, International Journal of Livestock Research, 8 (6), 152-159.

Tyasi TL, Eyduran E, Celik S (2021) Comparison of tree-based regression tree methods for predicting live body weight from morphological traits in Hy-line silver brown commercial layer and indigenous Potchefstroom Koekoek breeds raised in South Africa, Trop Anim Health Prod, https://doi.org/10.1007/s11250-020-02443-y

Walsh SW, Williams EJ, Evans ACO. (2011). A review of the causes of poor fertility in high milk producing dairy cows. Animal Reproduction Science, 123 (3-4), 127-138.