Institut für Tierwissenschaften

Challenges related to statistical methods and sensor systems for the daily prediction of health disorders in individual dairy cows

Dissertation

zur Erlangung des Grades
Doktor der Agrarwissenschaften (Dr. agr.)

der Landwirtschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

von

Christian Post

aus Ostercappeln

Bonn 2023

Referentin:	Prof. Dr. Dr. Helga Sauerwein
Korreferent:	Prof. Dr. Wolfgang Büscher
Tag der mündlichen Prüfung:	16.12.2022

Angefertigt mit Genehmigung der Landwirtschaftlichen Fakultät der Universität Bonn

I. Table of Contents

II. List of Tables

III. List of Figures

IV. List of Abbreviations

ADA	AdaBoost
AUC	Area Under the Curve
ALT	Activity, Lying, Temperature
AMS	Automatic Milking System
CI	Confidence Interval
CMS	Conventional Milking System
CSV	Comma-Separated Values
CUSUM	Cumulative Sum
DIM	Days In Milk
DT	Decision Tree
EC	Electrical Conductivity
EWMA	Exponentially Weighted Moving Average
ET	Extra Trees classifier
FN	False Negative
FP	False Positive
GNB	Gaussian Naïve Bayes
HIV	Human Immunodeficiency Virus
ID	Identifier
KNN	K-Nearest Neighbors
LDH	Lactate Dehydrogenase
LOESS	Locally Estimated Scatterplot Smoothing
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
MR	Milk Recording
n	Number
NY	New York

NPV	Negative Predictive Value
PPV	Positive Predictive Value
r	Pearson correlation coefficient
r^2	Coefficient of determination
RF	Random Forest
RF-I	Random Forest feature Importance
RG	Risk Group
RGB	Red, Green, Blue
RM	Rolling Mean
ROC	Receiver Operator Characteristic
RT	Risk Time
SCC	Somatic Cell Count
SD	Standard Deviation
SFS	Sequential Forward Selection
SMOTE	Synthetic Minority Over-sampling Technique
SPC	Statistical Process Control
SQL	Structured Query Language
SVM	Support Vector Machine
THI	Temperature Humidity Index
TierSchNutztV	Tierschutz-Nutztierhaltungsverordnung (German animal welfare regulation)
TMR	Total Mixed Ration
TN	True Negative
TP	True Positive
USA	United States of America

1 Abstract

1.1 English Abstract

The use of digital support systems has become a standard in dairy farming, and significant effort has been made to detect individual animals that are in need for a treatment by using sensor data. However, developing systems that are of actual use for practical farming remains a challenge. A variety of combinations of sensor systems and statistical algorithms for the detection of problems related to mastitis and lameness has been investigated to this date. These studies are mostly limited to certain conditions, models, and dependent variables that describe the disease to be detected, which makes them difficult to compare with one another and adapt to new datasets. An inherent challenge is the low prevalence of disease or treatment events, respectively, when looking at each individual cow on a day-by-day basis. As a result, even models with seemingly suitable combinations of sensitivity and specificity produce a large portion of false positive alarms.

Within the framework of this dissertation, two studies addressing these challenges were carried out and published. The aim of the first study was to use available (sensor) data from the Frankenforst experimental farm of the University of Bonn to develop classification models with which cows in need of treatments for lameness and mastitis could be detected. A wide range of statistical and machine learning algorithms were tested, as well as different ways of re-sampling the training data to achieve class balance between days with and without treatment. The ExtraTrees Classifier model achieved the best results with a mean area under curve (AUC) of 0.79 for mastitis and 0.71 for lameness treatments, while the sampling methods had no significant influence on the results.

In a second subsequent study, the four best models from the first study were re-trained with an expanded data base to include data from two additional experimental farms in order to test the transferability of the developed models to previously unknown data from two additional farms. For this purpose, the models were trained on data from one farm and then tested on the remaining farms' data. In addition, special attention was paid to the possibility of reducing false alarms by forming risk groups and risk times in the data and testing their potential to increase the relative frequency of occurrence of treatment days and thus a higher positive predictive value. It was found that the models showed poorer detection performances for data from an unknown farm, especially for lameness treatments, so that in conclusion training on data from the same farm was recommended. Regarding the subgroups

with a higher risk for a treatment, especially the cows with a previous treatment of the same category (mastitis or lameness) in the current lactation showed a significantly increased risk compared to the test data without grouping. However, the resulting increased positive predictive values (up to 20%) are still not sufficient for satisfactory use in practice. Here it was shown that the problem of the high number of false alarms is predominantly based in stochastics, which challenges the use of such models for the daily detection of cows, or livestock in general, in need of a treatment.

Overall, it is essential for the discussion to shift towards how these sensor systems can meet the requirements and expectations for practical decision support systems. Future research needs to focus on the transferability of the individual experimental conditions, the specificity of the features derived from sensor data, the choice of statistical models and especially the daily prevalence of the condition to be detected.

1.2 Deutsche Kurzfassung

Digitale Assistenzsysteme sind mittlerweile zum Standard in der Milchviehhaltung geworden. Das vorrangige Ziel dieser Systeme ist es, mithilfe von Sensordaten einzelne Tiere zu erkennen, die einer Behandlung bedürfen. Es bleibt jedoch eine Herausforderung, solche Systeme auch den Anforderungen der praktischen Nutztierhaltung anzupassen. Bisher wurde eine Vielzahl an Kombinationen von Sensorsystemen und statistischen Modellen, größtenteils zur Erkennung von Problemen rund um Mastitis und Lahmheit, untersucht. Diese Studien beschränken sich oft auf eine bestimmte Stallumgebung, ein Modell und eine bestimmte Zielvariable, welche die zu erkennende Erkrankung beschreibt; dies erschwert deren Vergleich und Anwendung auf neue Datensätze. Eine besondere Herausforderung ist außerdem die niedrige Auftretenshäufigkeit von Erkrankungs- bzw. Behandlungsereignissen, sobald die Daten auf Tagesbasis betrachtet werden. Als Resultat produzieren selbst Modelle mit scheinbar geeigneten Kombinationen von Sensitivität und Spezifität eine hohe Anzahl falsch-positiver Alarme.

Im Rahmen dieser Dissertation wurden zwei Studien durchgeführt und publiziert, welche diese Herausforderungen adressiert haben. Das Ziel der ersten Studie war es, bereits auf dem Versuchsbetrieb Frankenforst der Universität Bonn verfügbare (Sensor-)daten zu nutzen, um Klassifikationsmodelle zur Erkennung von behandlungsbedürftigen Kühen (Mastitis oder Lahmheit) zu entwickeln. Eine breite Auswahl an Algorithmen aus der Statistik und des maschinellen Lernens wurde getestet, sowie verschiedene Arten eines Ausgleichs der

Häufigkeit von Fällen mit und ohne Behandlung in den Daten (Re-sampling). Bei der Wahl des Modelles zeigte das ExtraTrees-Klassifikationsmodell mit einer mittleren area under the curve (ROC-AUC) von 0,79 für Mastitis und 0,71 für Lahmheit die besten Ergebnisse, während das Re-sampling keinen Einfluss hatte.

In einer zweiten Studie wurden dann die vier besten Modelle der ersten Studie mithilfe einer erweiterten Datenbasis neu trainiert. In dieser waren zusätzlich Daten zweier weiterer Versuchsbetriebe enthalten, anhand derer die Übertragbarkeit der entwickelten Modelle auf unbekannte Betriebe getestet werden sollte. Dazu wurden die Modelle mit den Daten eines Betriebes trainiert und dann mit den Daten der zwei verbleibenden Betriebe getestet. Zusätzlich wurde besonderes Augenmerk auf eine mögliche Reduzierung falscher Alarme durch die Bildung von Risikogruppen und -zeiten und gelegt mit dem Ziel, die relative Häufigkeit von Behandlungstagen und damit den positiven Vorhersagewert zu erhöhen. Es zeigte sich, dass Modelle bei der Anwendung auf Daten eines unbekannten Betriebes schlechtere diagnostische Kennzahlen produzierten, besonders bei der Klassifikation von Lahmheitsbehandlungen, sodass empfohlen wird, Modelle wenn möglich mit Daten des Betriebes zu trainieren, auf dem sie eingesetzt werden sollen. In den Risikogruppen, besonders bei Kühen mit einer vorherigen Behandlung derselben Kategorie (Mastitis oder Lahmheit) in der laufenden Laktation, zeigte sich eine signifikant erhöhte Wahrscheinlichkeit für eine Behandlung, verglichen mit den Daten ohne Gruppierung. Auch wenn die Klassifikation in diesen Gruppen zu höheren positiven Vorhersagewerten (bis zu 20%) führte, sind diese noch nicht ausreichend für einen Einsatz in praktischen Betrieben. Diese Studie zeigte, dass die hohe Anzahl falscher Alarme im Wesentlichen stochastische Gründe hat, und weiterhin eine große Herausforderung für den Einsatz solcher Modelle für die tägliche Erkennung eines Behandlungsbedarfs von Kühen oder anderen Tieren darstellt.

Insgesamt zeigte sich, dass sich die Diskussion um Sensorsysteme in der Milchviehhaltung mehr auf die Erfüllung der Anforderungen und Erwartungen praktischer Betriebe konzentrieren sollte. Die zukünftige Forschung muss sich auf die Übertragbarkeit der unter Versuchsbedingungen entwickelten Modelle, die Spezifität der aus den Sensordaten abgeleiteten Variablen, die Auswahl des geeigneten statistischen Modells, und besonders die Auftretenshäufigkeit des zu erkennenden Ereignisses fokussieren.

2 Introduction and Literature Review

The structures in dairy farming are in constant flux. Innovations in the field of husbandry, feeding and milking technology have enabled a steady increase in the number of cows per farm. For comparison, the average number of cows on German dairy farms was still around 27 in 1995, whereas it was around 68 in 2020 (Statistisches Bundesamt 2020). This development means that a single person working on the farm has to supervise more cows today than ever before. The tasks of animal supervision include the control of feeding, estrus observation, but above all the health and well-being of the animal. Farmers are obliged to visually inspect an animal directly at least once per day (§ 4 (1) TierSchNutztV). Disturbances in animal health must be detected as early as possible so that appropriate measures can be taken in time, and pain and suffering of the cows can be avoided.

To fulfil these tasks, the use of digital support systems has become standard. Dairy farming was a pioneer of the precision livestock farming sector, as transponders for animal identification and the first pedometers for heat detection have been developed since the 1970s (Bridle 1976; Kiddy 1977). This detection of the individual animal is the basis for the collection of continuous, individual animal-related data, which now include not only feeding, but also data from the milking parlor (milk quantity, milk flow, electrical conductivity or conductance of the milk, and milking duration), movement data (pedometer or accelerometer, position data), as well as other physiological data (body weight, rumen pH value, body temperature). All these data have in common that they are automatically and continuously collected at fixed or flexible intervals and sent to a central database for further processing: Milking parlor data at each milking time (two to three times per day) or each milking of the individual animal in the milking robot (animal-specific), aggregated movement data at different intervals from one day (De Mol et al. 2013), one or two hours (Maatje et al. 1997; Garcia et al. 2014), to 15 minutes (Alsaaod et al. 2012), and data on (concentrate) feed and water intake at each visit to the feed or water trough. This collection and storage of data represents the first of a total of four levels of a sensor system (Rutten et al. 2013). The bare measurement and representation of the recorded (raw) data alone (level I) does not tell anything about the condition of an animal, e.g., estrus or disease. Rather, it requires interpretation and integration with additional information about the individual cow, which represents the level II. In small, manageable herds, this interpretation can be done by the farmers themselves. But larger herds require computer-aided interpretation in the form of threshold values or predictive models. Level III describes the integration of additional information, e.g., economics, and in Level IV either the user or the system itself makes a decision. Existing systems differ in the extent to which levels III and IV are already automated. A feeding computer measures the amount of concentrate intake and then integrates this data with additional information about the individual cow, such as milk yield, body weight and days in milk, and finally autonomously allocates a recalculated amount of concentrate to that cow. In other cases, such as an estrus or mastitis alert, the farmer still makes the decision about insemination or medicinal treatment of the cow. These tasks also need to be carried out by a human person and are not likely to be fully automated in the near future. However, there are already possibilities for automated selection of animals that the system considers conspicuous. According to Rutten et al. (2013) there were still no publications in which levels III and IV were integrated into the respective sensor system in a fully automated way at the time of their meta-study. These additional linkages and the decision making were still left to the users.

In the following literature review, previous studies on sensor data, their relationship with health disorders, and their suitability for the development of detection models will be presented in more detail.

2.1 Sensor data and their relationship with the detection of health disorders

2.1.1 Review of sensor systems used in previous studies

For a sensor to be suitable for use in models that are intended to detect health disorders or predict diseases, it is necessary that there is a correlation with the disorder to be detected. In dairy cows, these are mostly mastitis, lameness, and metabolic problems, e.g. hypocalcemia (Brügesch et al. 2013). The greater this correlation, i.e., the higher the correlation between sensor data (or the information extracted from it) and the trait to be detected, the more meaningful and thus useful the sensor system is overall. A wide variety of sensors have been tested in the literature in connection with the detection of health disorders in dairy cows. An overview is given in Table 2-1 which lists the sensor or sensor system used, the variable measured (or derived using internal algorithms) and the health disturbance associated with it, along with the corresponding studies.

This overview shows that with regard to mastitis and lameness, there are preferences as to which sensors are used for the detection of the respective disorder and where these data are collected. Studies on mastitis detection rely mainly on data from the milking process, i.e., milk yield, milk components, and conductivity (Kramer et al. 2009; Lukas et al. 2009; Miekley et al. 2013b, among others) as well as the less frequently used variables milk color and milk temperature (Kamphuis et al. 2008a; Nielen et al. 1995), while studies on lameness detection mostly rely on sensors measuring movement and behavior, i.e. pedometer, accelerometer, weighing platforms, feed visits (De Mol et al. 2013; Alsaaod et al. 2012; Pastell et al. 2010, among others). However, there is also an overlap of sensors used for multiple disorders. In a study by Garcia et al. (2014), data from an automatic milking system (AMS) on the number of milkings, milking interval and milk flow were integrated into a lameness detection model. Kamphuis et al. (2013) studied cows that were milked in a milking carousel; the variables milk quantity, milking duration and milking order were used to distinguish between lame and non-lame cows, in addition to activity and weight data. Activity, measured as impulses from a pedometer or accelerometer, as a variable for lameness detection was used in the studies of Stangaferro et al. (2016) and Miekley et al. (2013b), among others. Stangaferro et al. (2016) found significantly less activity (measured by an accelerometer tag on the cows' collar) in animals with mastitis both before and after the day of diagnosis.

Table 2-1. Overview of sensors used in previous studies and their measured or derived variables and associated health disorders.

Sensor	Raw or derived data	Detected health disorder			Studies
		Mastitis	Lameness	Metabolic	
Real-time milk analyzer	Milk Yield	X	X		(Kramer et al. 2009; Lukas et al. 2009; Miekley et al. 2013b; Nielen et al. 1995; De Mol et al. 2013; Kamphuis et al. 2013; Huybrechts et al. 2014)
	Electrical conductivity (EC)	X			(Cavero et al. 2007; Kamphuis et al. 2008b; Lukas et al. 2009; Miekley et al. 2013b; Nielen et al. 1995)
	Milk components (fat, protein, lactose, blood, SCC[1])	X			(Jensen et al. 2016)
	Milk color (RGB values)	X			(Kamphuis et al. 2008a; Steeneveld et al. 2010)
	Milk temperature	X			(Nielen et al. 1995)
	Milk flow, number of milkings (AMS[2])		X		(Garcia et al. 2014)
	Milking duration, milking order		X		(Kamphuis et al. 2013)
Weighing Troughs	Feed and water intake	X	X		(Kramer et al. 2009; Miekley et al. 2013b; Garcia et al. 2014; Lukas et al. 2008; Palmer et al. 2012)
	Number and duration of feeding and drinking visits	X	X		(Kramer et al. 2009; Miekley et al. 2013b; Palmer et al. 2012)
				X	(Goldhawk et al. 2009)

[1]somatic cell count; [2]automatic milking system

Table 2-1 (continued)**.**

Sensor	Raw or derived data	Detected health disorder			Studies
		Mastitis	Lameness	Metabolic	
Pedometer	Activity[3]	X	X		(Miekley et al. 2013b; Alsaaod et al. 2012; Kamphuis et al. 2013)
				X	(Edwards and Tozer 2004)
Pedometer (ALT[3])	Lying time, lying position, ambient temperature		X		(Alsaaod et al. 2012)
Accelerometer	Lying time, lying and standing bouts, activity		X		(De Mol et al. 2013; Nechanitzky et al. 2016; Beer et al. 2016)
	Number, distance and duration of strides		X		(Beer et al. 2016)
	Magnitude and direction of acceleration		X		(Kofler et al. 2012; Pastell et al. 2009)
Tag on the ear or collar	Rumination and activity	X	X		(Stangaferro et al. 2016; van Hertem et al. 2013)
noseband sensor	Feed intake, rumination		X		(Beer et al. 2016)
Automatic weighing scale	Live weight		X		(Kamphuis et al. 2013)
Four-unit-weighing platform	Live weight, weight distribution between legs		X		(Pastell et al. 2010; Pastell und Kujala 2007)
Video camera	Hoof position		X		(Song et al. 2008)
Weather station	Ambient temperature, relative humidity	X[4]	X[4]	X[4]	(Lukas et al. 2008)

[3]activity, lying, temperature; [4]measured as number of impulses

The milk variables measured in the milking parlor or in the AMS refer to the quantity of milk (total milk yield, as well as quantity related to time, i.e., the milk flow) and its quality. Variables reflecting the milk quality include the ingredients naturally contained in the milk such as fat, protein, lactose and, from a certain amount undesirable, substances such as somatic cells (expressed as the somatic cell count, SCC) and blood. In addition, the conductivity of the milk is often measured, which increases when the permeability of the milk ducts is disturbed and is used as an indicator for mastitis detection (Hogeveen et al. 2010). Here, a distinction has to be made whether it is the physical conductivity, expressed in mS/cm (Cavero et al. 2007; Kamphuis et al. 2008b), or a reference variable derived from the sensor system (Lukas et al. 2009; Miekley et al. 2013b; Steeneveld et al. 2010). In addition, there are studies that have shown the suitability of milk color as a sensor for mastitis detection (Steeneveld et al. 2010; Kamphuis et al. 2008a). This variable is already being used in a majority of commercial AMS to detect abnormal milk (Steeneveld and Hogeveen 2015).

The sensor systems used to describe the activity or behavior vary between the studies mentioned. The difference between a pedometer and an accelerometer is that the registered strokes of the pedometer are a binary variable (impulse or no impulse), whereas the accelerometer measures acceleration as a continuous value (Kofler et al. 2012; Garcia et al. 2014). However, it is not always the case that this continuous variable has been directly used as a feature in a lameness detection model, i.e., that the trajectories of accelerations on the three spatial axes have been directly compared between cows (Kofler et al. 2012; Pastell et al. 2009). More often, sensor system-internal algorithms aggregate the measured data as number of impulses or steps, or as lying position per time unit, which are then used as variables in the model (Nechanitzky et al. 2016; De Mol et al. 2013; Garcia et al. 2014). Variables describing behavior, such as number and duration of lying, standing and walking, steps per unit time, and feeding behavior alone are not reliable indicators to describe lameness, while sensors and variables describing gait (number of steps, duration between steps, and symmetry between legs) provided higher correlations to lameness (O'Leary et al. 2020).

Models for the detection of health disorders are usually not only based on the raw data generated by the sensors, but calculate additional variables from these. This step is also called feature design or feature extraction (Kelleher et al. 2015; Kuhn and Johnson 2013). The aim of generating additional features, i.e., variables that are used as inputs to the statistical models (Matloff 2017), is to obtain information beyond what the absolute raw data values provide, and make it available to the model. Often used methods of feature design are aggregations (e.g., number, sum, and mean), ratios of two features, or binning/mapping, i.e., converting a continuously measured feature into a discrete, categorical feature (e.g., low, medium or high) (Kelleher et al. 2015). Examples of features created in this way in studies published to date include: Change in daily milk yield of two consecutive days, difference and ratio of average milk yield of the last two weeks, slope of a linear regression of average milk yield of the last seven days, day/night ratio of activity (van Hertem et al. 2013), difference of measured values for body weight, activity and milk yield of two consecutive days (Kamphuis et al. 2013), sums of steps impulses and lying time per day, and number, mean, maximum and minimum duration of lying periods (Alsaaod et al. 2012).

2.1.2 Recording of the target variable: The gold standard

Most recognition models use what is known as supervised learning, in which the model considers the relationship between the features, i.e. the recorded sensor data, and a target variable (Kelleher et al. 2015). This target is based on recordings, measurements, or annotations, in this case cow health or treatment data. The target variable is also referred to as the "gold standard" and aims to represent the characteristic to be recognized later by the model itself as realistically as possible (Rutten et al. 2013). In the studies mentioned in Table 2-1, different approaches can be found for the definition of the target variable: recording of treatments carried out by farmers or veterinarians (Miekley et al. 2013b; van Hertem et al. 2013; Huybrechts et al. 2014, among others), automated recording of the target variable, in this case SCC (Cavero et al. 2008; Mollenhorst et al. 2012), or routine visual inspection of cow health, in this case gait assessment that resulted in lameness scores (Alsaaod et al. 2012; De Mol et al. 2013; Kamphuis et al. 2013, among others). The selection of a recording method for the target variable has implications for the experimental design, as well as the subsequent performance of the model. The choice of treatments as the target means that the model will make similar decisions to those of farmers, but will also make the same mistakes as them, and may not detect subclinically ill animals (Rutten et al. 2013). Regular recording of the health status of cows by veterinarians or trained personnel, on the other hand, has the potential to identify otherwise inconspicuous cows as sick. However, this method of recording the target variable is more time-consuming, requires additional labor and, in the case of necessary examinations (e.g., ultrasonography, somatic cell count, bacteriological examination), increased material costs. The time interval of these examinations also plays a role in the subsequent model quality. If the individual measurements are too far apart in time, cases of disease may be detected too late after their actual occurrence or even overlooked (Rutten et al. 2013). Finally, a scoring system, such as the frequently used five-point scoring system according to (Sprecher et al. 1997) or a modified form according to (Flower and Weary 2006), also harbors the potential for errors if the definition of the individual classes leaves room for interpretation and cows are thus incorrectly classified by the observer (van Nuffel et al. 2015a).

The translation of the ordinal scaled five-point lameness scoring system into a binary variable (lame or not lame) requires a cut-off at a certain score, which may have an impact on subsequent performance. In most of the mentioned studies, cows with a score of ≥3 were classified as "lame" (Alsaaod et al. 2012; De Mol et al. 2013; Kamphuis et al. 2013; Pastell and Kujala 2007), although in the study of (De Mol et al. 2013) cows with a score of 2 ("mildly lame") were excluded from the study and therefore only cows with a score of 1 were considered "not lame", and in other studies (Kamphuis et al. 2013; Pastell et al. 2010) both lower and higher thresholds for the classification as "lame" were tested in parallel.

There is also leeway in defining mastitis status as a binary variable. For the definition of the gold standard for "mastitis" or "healthy", the SCC was used in some of the mentioned studies (Cavero et al. 2008; Cavero et al. 2007; Nielen et al. 1995). Here, the thresholds for defining a mastitis were a SCC of 100,000 or 400,000 (Cavero et al. 2008; Cavero et al. 2007) and 500,000 cells/mL of milk (Nielen et al. 1995), respectively, although in the latter study cows were only considered healthy with an SCC of < 200,000 cells/mL, and periods with measurements between these two thresholds were excluded from the analysis as undefined cases. A universally accepted definition of mastitis in terms of detection models does not yet exist (Dominiak and Kristensen 2017), and any threshold for SCC means a trade-off between detecting subclinical mastitis cases also, or keeping false-positive alarms to a minimum. As an alternative to the binary classification of cow mastitis status, it is possible to have the model estimate this status as a value for mastitis risk between 0 and 1 instead (Friggens et al. 2007). In this study, the features milk yield, SCC, and lactate dehydrogenase (LDH) activity were used in a mixed model to calculate a likelihood value that represented the degree of mastitis. This approach considers the fact that mastitis disease itself is not a binary variable, but differs both between cows and between time points of the same cow. Presenting a probability rather than a binary alarm system can provide valuable additional information for users. However, a weakness of this particular study was that a large proportion of cases that were unclear in terms of health status (482 out of 611) were excluded from the analysis, and the transferability to practice can thus be questioned (Hogeveen et al. 2010).

2.2 Methods for the evaluation of sensor systems

After a prediction model has been developed and tested, the predictions obtained must be validated. This section will further expand on means of evaluation. The common method after obtaining results from model testing is to combine the predictions with the actual states (gold standard) in a contingency table, also called confusion matrix. The procedure for a binary classification is shown in Table 2-2, modified after Fawcett (2006).

Table 2-2. Confusion matrix and performance metrics calculated from the combinations of true and predicted conditions.

		Prediction / classification		
		positive	negative	
True condition / gold standard	positive	True positive, TP	False negative, FN	Sensitivity = TP / (TP + FN)
	negative	False positive, FP	True negative, TN	Specificity = TN / (TN + FP)
		Positive predictive value = TP / (TP + FP)	Negative predictive value = TN / (TN + FN)	

The four states result from the combinations of positive or negative prediction and positive or negative gold standard: True positive (TP), false positive (FP), true negative (TN), or false negative (FN). Two kinds of misclassification are possible: In the context of automated detection of health disorders in cows, a FP classification means that a cow that is actually healthy is wrongly identified as "sick", i.e., having a health disorder or being in need for a treatment. In a predictive sensor system that informs farmers about cows in need of treatment with alarms, this would therefore represent a false alarm. Since farmers do not know the true health status of the cow (according to the gold standard), unless additional information can be added to classify these alarms, all cows with alarms must be checked for health disorders. Conversely, a FN classification, i.e., the absence of an alarm, implies that the cow in question is supposed to be "healthy", when in fact a health disorder is present. This leads to a lack of further assessment and potential omission of a necessary treatment of the cow.

From the ratios of the columns and rows of the confusion matrix, four key figures can be calculated with which the quality of the classification can be assessed. The sensitivity (also called *precision*) indicates how many cows with health disorders are detected by the model, while the specificity implies the number of actually "healthy" cows not falsely identified as "sick" by the model. The positive predictive value (PPV, also called *recall*) shows the proportion of false alarms. While a low PPV means unnecessary work and is directly noticeable to the farmer, a low negative predictive value (NPV) means that necessary treatments are not carried out and only become visible later through worsening of disease conditions and consequential health issues.

The levels of sensitivity and specificity can be compared between individual studies only to a limited extent. This is due to the fact that a certain threshold value for classification (for details see Chapter 2.3) has already been set. A low threshold induces needless examinations of cows that can lead to frustration of farmers (Horseman et al. 2014). A high threshold, on the other hand, leads to a very specific model, but one that leaves many cows in need of treatment undetected. The sensitivity is thus in an (approximately) inverse-proportional relationship to the specificity (Brenner and Gefeller 1997). The determination of the corresponding threshold value is subject to the model developers, or in some cases the users of the model, and therefore prevents a direct comparison between different models in different studies. Therefore, it is necessary to use metrics that are independent of this threshold. One of these metrics is the area under curve (AUC) of the receiver operator characteristic (ROC) curve, sometimes also referred to as *precision-recall curve* (Fawcett 2006). The ROC curve describes the ratio of sensitivity to false positive rate, i.e., $1 - \text{specificity}$, at all possible threshold values (Fawcett 2006; Pedregosa et al. 2011). The false positive rate is plotted on the x-axis and the sensitivity on the y-axis; thus, it allows a depiction of all possible combinations of sensitivity and specificity (see Figure 2-1).

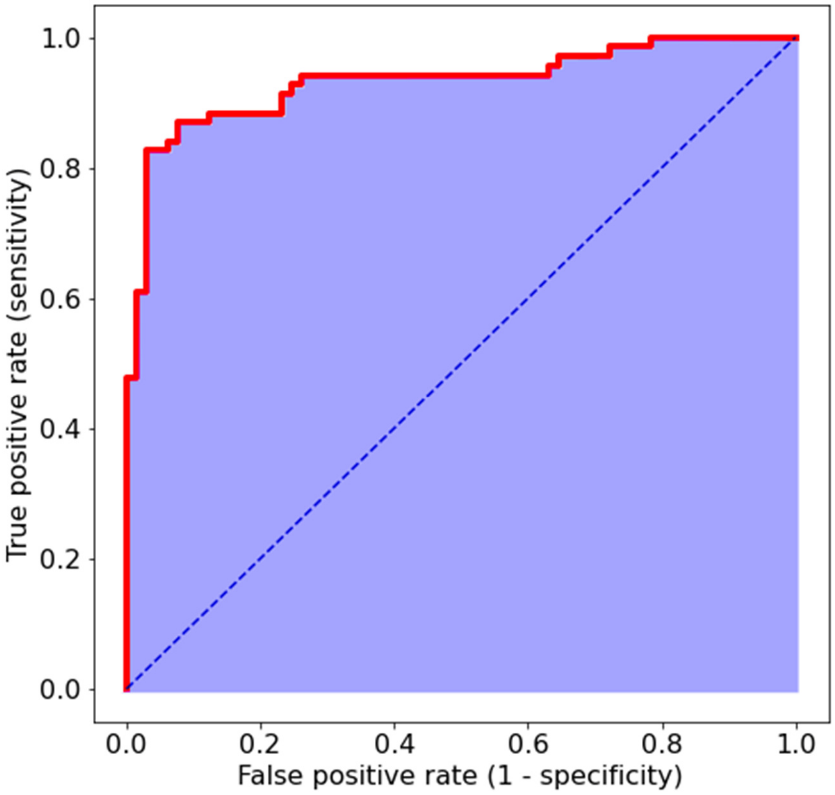

Figure 2-1. Example of a receiver-operator-characteristic (ROC) curve (red) with the area under the curve (AUC) in bright blue and a ROC of a random prediction (blue dashed line).

To convert the two-dimensional ROC curve into a single, one-dimensional value, the area under the ROC curve, the AUC, is calculated. The value of the AUC always lies in the range from 1 to 0 due to the characteristics of the chosen coordinate system. A perfect classification would result in an AUC of 1, while an absolutely random classification (e.g., a coin toss) would result in an AUC of 0.5, which is represented by the diagonal dashed line in Figure 2-1. Values below 0.5 are possible but extremely rare and indicate faulty correlations between features and labels or errors in the chosen training or testing data. In addition, there are finer subdivisions that assess the quality of the level of AUC; these are portrayed in the discussion sections of the included publications (Chapters 3 and 4).

While the level of sensitivity and specificity depends predominantly on the selected cut-off point between positive and negative classification, the predictive values PPV and NPV are significantly influenced by the frequency with which the feature to be detected (in this case, the health disorder) occurs in the data (Brenner and Gefeller 1997). This relationship is of mathematical nature and results from the conditional probabilities for the presence of the disorder (prevalence, frequency of occurrence) and the detection of animals as either "sick" (sensitivity) or "healthy" (specificity). Although the literature reports prevalence values for lameness of over 50% (Barker et al. 2010; Von Keyserlingk et al. 2012) and for subclinical mastitis of approximately 30% (Plozza et al. 2011), the proportion of days on which clinical signs are actually treated is very low in relation to the total number of days: cow-day prevalence rates for mastitis of 0.6% (Kamphuis et al. 2008b), 0.1–0.9% (Kramer et al. 2009) or even as low as 0.04% (Steeneveld et al. 2010) have been reported. The data in the study by Cavero et al. (2006) showed a daily prevalence of mastitis treatments of 0.5% and 0.3% in the training and test data, respectively, but when a threshold of SCC of 100,000 cells/mL, i.e., the subclinically sick animals, was also added to the gold standard, these values increased to 27.7% and 15.2%, respectively. For lameness treatments, the prevalence was 0.2% in the study by De Mol et al. (2013), 0.5% in the study by Kamphuis et al. (2013), and 0.8%–1.7% in the study by Miekley et al. (2013a), when additional three to seven days before treatments were included in the data. With such a low prior probability, this means that shifting the ratio between sensitivity and sensitivity (or false positive rate) along the ROC curve in favor of sensitivity increases the number of cases classified as FP by a multiple for each new TP case thus obtained. Therefore, in the publication presented in Chapter 4, special attention will be paid to the PPV.

2.3 Differences in statistical approaches to detect cows with health disorders

In addition to a high variability regarding the sensor combinations used in the studies mentioned in Chapter 2.1.1, there are also great differences between the statistical models used for the detection of health disorders. In order to assess the methodology used in the two own studies (Chapters 3 and 4), an overview of the range of possible algorithms which have been tested with dairy cattle data is presented here. These can be roughly divided into two statistical approaches: Regression (in this particular case: time series models) and classification.

The simplest way to detect a health disorder is to compare raw or preprocessed sensor data related to the condition to be detected with a threshold value (Hogeveen et al. 2010). However, for many sensor data, absolute values vary widely between individual animals and these values must therefore be placed in relation to past values of the same animal in order to detect variation (Hogeveen et al. 2010; O'Leary et al. 2020). Time series models attempt to capture the development of one or more variables of a single cow over time. Animals with abnormal health are identified by changes that lead to a threshold being exceeded. An example of this approach can be found in the study by Cavero et al. (2007). Here, to detect deviations in electrical conductivity as an indicator of mastitis, three methods were tested: a simple moving average of the last ten observations, i.e., milkings, an exponentially weighted moving average (EWMA), and a locally weighted polynomial regression (LOESS). In EWMA, all previous milkings were considered, with exponentially less weight given to events further back in time, to emphasize more recent sensor measurements. In LOESS, a linear function using least squares was superimposed on the last n observations of EC, again giving higher weight to more recent observations. The predictions generated this way were compared with a range of threshold values, and deviations were taken as a sign of mastitis. The best models were found for deviations in the range of 5%–9% and achieved a sensitivity in the range of about 0.85 with a specificity of 0.67–0.81, but also error rates of up to 0.87 (Cavero et al. 2007). A similar time series model, called TSMx, was used in the study by De Mol and Ouweltjes (2001). Again, using up to 30 past measurements of milk yield and conductivity at the AMS, regression equations were created, the residuals of the values for these two variables from the most recent milking were compared with three different confidence intervals (95%, 99%, and 99.9%) and alarms were generated in case of the interval being exceeded. This approach resulted in a high sensitivity (0.88–0.92), but also produced a very high number of false positive alarms, between 520 and 3278 compared to the 48 true mastitis cases in the data set (De Mol and Ouweltjes 2001).

A similar type of prediction is offered by the Statistical Process Control (SPC), specifically the CUSUM (cumulative sum) charts used in the studies of Huybrechts et al. (2014), Miekley et al. (2013a), and Dittrich et al. (2021). In the CUSUM chart, the deviations of the signal from a target value, usually the moving average, are summed up for upwards and downwards deviations separately. If one of these CUSUMs exceeds a control limit that has been defined in advance, an alarm is generated (Huybrechts et al. 2014). The signal used can be either from a single feature (Huybrechts et al. 2014) or multidimensional (Miekley et al. 2013a; Dittrich et al. 2021). Dittrich et al. (2021) extended the CUSUM method to include a prior feature transformation using either Principal Component Analysis or Partial Least Squares to ensure the feature independence required by the CUSUM chart. Variables from leg and neck accelerometers (activity, standing and lying behavior, feeding, and rumination) and milking variables (milk yield, milk flow, and quarter milk EC) were used, and the outcome variable consisted of the combined treatments for lameness, mastitis and metabolic disorders. Despite a high AUC of 0.85–0.90 (depending on the combination of variables used in the model), there was also a high number of false positives in the test data: Out of

19 animals, up to five animals on average produced a false alarm each day (Dittrich et al. 2021).

Similar to time series models, classification models estimate a continuous response variable that is compared to a predetermined cut-off value to make a statement about the condition to be detected. However, in classification, this variable is usually the probability of the sample belonging to one of the two classes (in this case, whether a health disorder is present or not, Kuhn and Johnson 2013). The standard cut-off value for this probability is often assumed to be 0.5, but in principle any value between 0 and 1 can be set, which influences the relationship between sensitivity and specificity presented in Chapter 2.2. There are some algorithms that may also have an output between 0 and 1, though mathematically they do not generate a probability. This applies to neural networks and partial least squares, as their outputs are brought into a range between 0 and 1 by means of a so-called *softmax transformation* (Kuhn and Johnson 2013). However, this distinction is irrelevant for the practical application of such models, and thus is not further discussed. The classification algorithms used in the studies mentioned up to this point include logistic regression (Goldhawk et al. 2009; Kamphuis et al. 2013; van Hertem et al. 2013; Nielen et al. 1995), artificial neural networks (Pastell and Kujala 2007; Cavero et al. 2008), Partial Least Squares Discriminant Analysis (Garcia et al. 2014), Naive Bayes (Jensen et al. 2016; Steeneveld et al. 2010) or Support Vector Machines (Alsaaod et al. 2012). The comparison of these and other classification algorithms is part of Chapter 3 and will accordingly be further elaborated there.

The fuzzy logic models represent a special category of decision-making concepts. In the application of fuzzy logic, described by De Mol and Woldt (2001), features are first converted into linguistic terms, e.g., "low", "moderate" and "high", or "increased" and "not increased" (fuzzification). In a second step, rule sets are then established with these features, which describe the relationships of the individual features and their quality with the target variable "health status" with simple if-then relationships (fuzzy inference). In the third and last step, the conclusions drawn from this whole set of relationships are combined and converted into a prediction (defuzzification). In the field of health disorder detection in dairy cows, this methodology has already been tested in several studies (Kramer et al. 2009; De Mol and Woldt 2001; Kamphuis et al. 2008b). While in the study by Kramer et al. (2009) this approach did not result in improved predictions compared to other algorithms, De Mol and Woldt (2001) applied fuzzy logic to mastitis alarms generated by an AMS, and Kamphuis et al. (2008b) used fuzzy logic for a combination of two sensor systems for mastitis detection (in-line SCC and EC). This led to an increase of the specificity from 0.95 to 0.99 (De Mol and Woldt 2001) and of the PPV from 0.07 and 0.13 to 0.22 (Kamphuis et al. 2008b). Despite this potential, fuzzy logic models have since been used only sporadically for mastitis detection in dairy cows (Zülkadir and Çoşkun 2018) and in goats (Zaninelli et al. 2016).

The studies presented and the variety of sensor combinations and methodological approaches found in them illustrate the need for additional research, be it in testing new or already widely available sensors, the application of models tested in other disciplines, or newly developed models. However, the question of applicability in practice should never be overlooked. A critical point here is the number of cows identified as false positives, which in some of the work presented prevented further implementation in practice.

In the following two chapters, a range of already tested classification models was first validated and compared using data from farms in practice. Then, in the second manuscript, special attention was paid to positive predictive value, and the effects of prior selection of animals at higher risk for a treatment were tested.

Introduction and Literature Review References

Alsaaod, M.; Römer, C.; Kleinmanns, J.; Hendriksen, K.; Rose-Meierhöfer, S.; Plümer, L.; Büscher, W. (**2012**): Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. *Applied Animal Behaviour Science* 142 (3), 134–141.

Barker, Z.; Leach, K.; Whay, H.; Bell, N.; Main, D. (**2010**): Assessment of lameness prevalence and associated risk factors in dairy herds in England and Wales. *Journal of Dairy Science* 93 (3), 932–941.

Beer, G.; Alsaaod, M.; Starke, A.; Schuepbach-Regula, G.; Müller, H.; Kohler, P.; Steiner, A. (**2016**): Use of Extended Characteristics of Locomotion and Feeding Behavior for Automated Identification of Lame Dairy Cows. *PLOS one* 11 (5), e0155796.

Brenner, H.; Gefeller, O. (**1997**): Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statist. Med.* 16 (9), 981–991.

Bridle, J. (**1976**): Automatic dairy cow identification. *Journal of Agricultural Engineering Research* 21 (1), 41–48.

Brügesch, F.; Spindler, B.; Fels, M.; Schallenberger, E.; Kemper, N. (**2013**): BMTW - Häufigkeitsverteilung von Diagnosen in Rinderbeständen im Mittelweserraum auf Basis der Auswertung von tierärztlichen Arzneimittel-Anwendungs- und Abgabe-Nachweisen. *Berliner und Münchener tierärztliche Wochenschrift* 6 (3-4), 169–174.

Cavero, D.; Tölle, K.-H.; Buxadé, C.; Krieter, J. (**2006**): Mastitis detection in dairy cows by application of fuzzy logic. *Livestock Science* 105 (1-3), 207–213.

Cavero, D.; Tölle, K.-H.; Henze, C.; Buxadé, C.; Krieter, J. (**2008**): Mastitis detection in dairy cows by application of neural networks. *Livestock Science* 114 (2-3), 280–286.

Cavero, D.; Tölle, K.-H.; Rave, G.; Buxadé, C.; Krieter, J. (**2007**): Analysing serial data for mastitis detection by means of local regression. *Livestock Science* 110 (1-2), 101–110.

De Mol, R.; André, G.; Bleumer, E.; De Haas, Y.; van der Werf, J.; van Reenen, C. (**2013**): Applicability of day-to-day variation in behavior for the automated detection of lameness in dairy cows. *Journal of Dairy Science* 96 (6), 3703–3712.

De Mol, R.; Ouweltjes, W. (**2001**): Detection model for mastitis in cows milked in an automatic milking system. *Preventive Veterinary Medicine* 49 (1-2), 71–82.

De Mol, R.; Woldt, W. (**2001**): Application of Fuzzy Logic in Automated Cow Status Monitoring. *Journal of Dairy Science* 84 (2), 400–410.

Dittrich, I.; Gertz, M.; Maassen-Francke, B.; Krudewig, K.-H.; Junge, W.; Krieter, J. (**2021**): Combining multivariate cumulative sum control charts with principal component analysis and partial least squares model to detect sickness behaviour in dairy cattle. *Computers and Electronics in Agriculture* 186, 106209.

Dominiak, K.; Kristensen, A. (**2017**): Prioritizing alarms from sensor-based detection models in livestock production - A review on model performance and alarm reducing methods. *Computers and Electronics in Agriculture* 133, 46–67.

Edwards, J.; Tozer, P. (**2004**): Using Activity and Milk Yield as Predictors of Fresh Cow Disorders. *Journal of Dairy Science* 87 (2), 524–531.

Fawcett, T. (**2006**): An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8), 861–874.

Flower, F.; Weary, D. (**2006**): Effect of hoof pathologies on subjective assessments of dairy cow gait. *Journal of Dairy Science* 89 (1), 139–146.

Friggens, N.; Chagunda, M.; Bjerring, M.; Ridder, C.; Hojsgaard, S.; Larsen, T. (**2007**): Estimating Degree of Mastitis from Time-Series Measurements in Milk: A Test of a Model Based on Lactate Dehydrogenase Measurements. *Journal of Dairy Science* 90 (12), 5415–5427.

Garcia, E.; Klaas, I.; Amigo, J.; Bro, R.; Enevoldsen, C. (**2014**): Lameness detection challenges in automated milking systems addressed with partial least squares discriminant analysis. *Journal of Dairy Science* 97 (12), 7476–7486.

Goldhawk, C.; Chapinal, N.; Veira, D.; Weary, D.; Von Keyserlingk, M. (**2009**): Prepartum feeding behavior is an early indicator of subclinical ketosis. *Journal of Dairy Science* 92 (10), 4971–4977.

Hogeveen, H.; Kamphuis, C.; Steeneveld, W.; Mollenhorst, H. (**2010**): Sensors and Clinical Mastitis—The Quest for the Perfect Alert. *Sensors* 10 (9), 7991–8009.

Horseman, S.; Roe, E.; Huxley, J.; Bell, N.; Mason, C.; Whay, H. (**2014**): The use of in-depth interviews to understand the process of treating lame dairy cows from the farmers' perspective. *Animal Welfare* 23 (2), 157–165.

Huybrechts, T.; Mertens, K.; De Baerdemaeker, J.; De Ketelaere, B.; Saeys, W. (**2014**): Early warnings from automatic milk yield monitoring with online synergistic control. *Journal of Dairy Science* 97 (6), 3371–3381.

Jensen, D.; Hogeveen, H.; De Vries, A. (**2016**): Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis. *Journal of Dairy Science* 99 (9), 7344–7361.

Kamphuis, C.; Frank, E.; Burke, J.; Verkerk, G.; Jago, J. (**2013**): Applying additive logistic regression to data derived from sensors monitoring behavioral and physiological characteristics of dairy cows to detect lameness. *Journal of Dairy Science* 96 (11), 7043–7053.

Kamphuis, C.; Pietersma, D.; van der Tol, R.; Wiedemann, M.; Hogeveen, H. (**2008a**): Using sensor data patterns from an automatic milking system to develop predictive variables for classifying clinical mastitis and abnormal milk. *Computers and Electronics in Agriculture* 62 (2), 169–181.

Kamphuis, C.; Sherlock, R.; Jago, J.; Mein, G.; Hogeveen, H. (**2008b**): Automatic detection of clinical mastitis is improved by in-line monitoring of somatic cell count. *Journal of Dairy Science* 91 (12), 4560–4570.

Kelleher, J.; MacNamee, B.; D'Arcy, A. (**2015**): *Fundamentals of machine learning for predictive data analytics*. Cambridge, Massachusetts, USA: The MIT Press.

Kiddy, C. (**1977**): Variation in Physical Activity as an Indication of Estrus in Dairy Cows. *Journal of Dairy Science* 60 (2), 235–243.

Kofler, J.; Mangweth, G.; Altenhofer, C.; Weber, A.; Gasser, C.; Schramel, J.; Tichy, A.; Peham, C. (**2012**): Messung der Bewegung lahmheitsfreier Kühe mittels Accelerometer im Schritt und Vergleich der Beschleunigungswerte nach Kleben eines Klotzes. *Wiener Tierarztliche Monatsschrift* 99 (7), 179.

Kramer, E.; Cavero, D.; Stamer, E.; Krieter, J. (**2009**): Mastitis and lameness detection in dairy cows by application of fuzzy logic. *Livestock science* 125 (1), 92–96.

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

Lukas, J.; Reneau, J.; Linn, J. (**2008**): Water Intake and Dry Matter Intake Changes as a Feeding Management Tool and Indicator of Health and Estrus Status in Dairy Cows. *Journal of Dairy Science* 91 (9), 3385–3394.

Lukas, J.; Reneau, J.; Wallace, R.; Hawkins, D.; Munoz-Zanzi, C. (**2009**): A novel method of analyzing daily milk production and electrical conductivity to predict disease onset. *Journal of Dairy Science* 92 (12), 5964–5976.

Maatje, K.; Loeffler, S.; Engel, B. (**1997**): Predicting Optimal Time of Insemination in Cows that Show Visual Signs of Estrus by Estimating Onset of Estrus with Pedometers. *Journal of Dairy Science* 80 (6), 1098–1105.

Matloff, N. (**2017**): *Statistical regression and classification. From linear models to machine learning.* Boca Raton, London, New York: Chapman & Hall/CRC.

Miekley, B.; Stamer, E.; Traulsen, I.; Krieter, J. (**2013a**): Implementation of multivariate cumulative sum control charts in mastitis and lameness monitoring. *Journal of Dairy Science* 96 (9), 5723–5733.

Miekley, B.; Traulsen, I.; Krieter, J. (**2013b**): Principal component analysis for the early detection of mastitis and lameness in dairy cows. *Journal of Dairy Research* 80 (3), 335–343.

Mollenhorst, H.; Rijkaart, L.; Hogeveen, H. (**2012**): Mastitis alert preferences of farmers milking with automatic milking systems. *Journal of Dairy Science* 95 (5), 2523–2530.

Nechanitzky, K.; Starke, A.; Vidondo, B.; Müller, H.; Reckardt, M.; Friedli, K.; Steiner, A. (**2016**): Analysis of behavioral changes in dairy cows associated with claw horn lesions. *Journal of Dairy Science* 99 (4), 2904–2914.

Nielen, M.; Schukken, Y.; Brand, A.; Deluyker, H.; Maatje, K. (**1995**): Detection of Subclinical Mastitis from On-Line Milking Parlor Data. *Journal of Dairy Science* 78 (5), 1039–1049.

O'Leary, N.; Byrne, D.; O'Connor, A.; Shalloo, L. (**2020**): Invited review: Cattle lameness detection with accelerometers. *Journal of Dairy Science* 103 (5), 3895–3911.

Palmer, M.; Law, R.; O'Connell, N. (**2012**): Relationships between lameness and feeding behaviour in cubicle-housed Holstein–Friesian dairy cows. *Applied Animal Behaviour Science* 140 (3-4), 121–127.

Pastell, M.; Hänninen, L.; De Passillé, A.; Rushen, J. (**2010**): Measures of weight distribution of dairy cows to detect lameness and the presence of hoof lesions. *Journal of Dairy Science* 93 (3), 954–960.

Pastell, M.; Tiusanen, J.; Hakojärvi, M.; Hänninen, L. (**2009**): A wireless accelerometer system with wavelet analysis for assessing lameness in cattle. *Biosystems Engineering* 104 (4), 545–551.

Pastell, M.; Kujala, M. (**2007**): A probabilistic neural network model for lameness detection. *Journal of Dairy Science* 90 (5), 2283–2292.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. (**2011**): Scikit-learn. Machine learning in Python. *Journal of machine learning research* 12 (Oct), 2825–2830.

Plozza, K.; Lievaart, J.; Potts, G.; Barkema, H. (**2011**): Subclinical mastitis and associated risk factors on dairy farms in New South Wales. *Australian Veterinary Journal* 89 (1-2), 41–46.

Rutten, C.; Velthuis, A.; Steeneveld, W.; Hogeveen, H. (**2013**): Invited review. Sensors to support health management on dairy farms. *Journal of Dairy Science* 96 (4), 1928–1952.

Song, X.; Leroy, T.; Vranken, E.; Maertens, W.; Sonck, B.; Berckmans, D. (**2008**): Automatic detection of lameness in dairy cattle—Vision-based trackway analysis in cow's locomotion. *Computers and Electronics in Agriculture* 64 (1), 39–44.

Sprecher, D.; Hostetler, D.; Kaneene, J. (**1997**): A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. *Theriogenology* 47 (6), 1179–1187.

Stangaferro, M.; Wijma, R.; Caixeta, L.; Al-Abri, M.; Giordano, J. (**2016**): Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part II. Mastitis. *Journal of Dairy Science* 99 (9), 7411–7421.

Statistisches Bundesamt (**2020**): *Anzahl der Milchkühe je Betrieb in Deutschland in den Jahren 1995 bis 2020*. Statista. Hg. v. Statista GmbH. Available online: https://de.statista.com/statistik/daten/studie/28755/umfrage/anzahl-der-milchkuehe-je-halter-in-deutschland-seit-1990/ (accessed on 08 April 2021).

Steeneveld, W.; Hogeveen, H. (**2015**): Characterization of Dutch dairy farms using sensor systems for cow management. *Journal of Dairy Science* 98 (1), 709–717.

Steeneveld, W.; Van der Gaag, L.; Ouweltjes, W.; Mollenhorst, H.; Hogeveen, H. (**2010**): Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *Journal of Dairy Science* 93 (6), 2559–2568.

TierSchNutztV (**2006**): *Verordnung zum Schutz landwirtschaftlicher Nutztiere und anderer zur Erzeugung tierischer Produkte gehaltener Tiere bei ihrer Haltung (Tierschutz-Nutztierhaltungsverordnung - TierSchNutztV)*. Available online: https://www.gesetze-im-internet.de/tierschnutztv/BJNR275800001.html (accessed on 08 April 2021).

Van Hertem, T.; Maltz, E.; Antler, A.; Romanini, C.; Viazzi, S.; Bahr, C.; Schlageter-Tello, A.; Lokhorst, C.; Berckmans, D.; Halachmi, I. (**2013**): Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *Journal of Dairy Science* 96 (7), 4286–4298.

Van Nuffel, A.; Zwertvaegher, I.; Pluym, L.; van Weyenberg, S.; Thorup, V. M.; Pastell, M.; Sonck, B.; Saeys, W. (**2015**): Lameness detection in dairy cows: Part 1. How to distinguish between non-lame and lame cows based on differences in locomotion or behavior. *Animals* 5 (3), 838–860.

Von Keyserlingk, M.; Barrientos, A.; Ito, K.; Galo, E.; Weary, D. (**2012**): Benchmarking cow comfort on North American freestall dairies: lameness, leg injuries, lying time, facility design, and management for high-producing Holstein dairy cows. *Journal of Dairy Science* 95 (12), 7399–7408.

Zaninelli, M.; Tangorra, F.; Costa, A.; Rossi, L.; Dell'Orto, V.; Savoini, G. (**2016**): Improved Fuzzy Logic System to Evaluate Milk Electrical Conductivity Signals from On-Line Sensors to Monitor Dairy Goat Mastitis. *Sensors* 16 (7), 1079.

Zülkadir, U.; Çoşkun, F. (**2018**): The Use of Fuzzy Logic Approach in Evaluation of Subclinic Mastitis. *Selcuk Journal of Agriculture and Food Sciences* 32 (3), 436–439.

3 Using Sensor Data to Detect Lameness and Mastitis Treatment Events in Dairy Cows: A Comparison of Classification Models

Christian Post [1], Christian Rietz [2], Wolfgang Büscher [3] and Ute Müller [1]

[1] Institute of Animal Science, Physiology and Hygiene Unit, University of Bonn, 53115 Bonn, Germany; ute-mueller@uni-bonn.de

[2] Department of Educational Science, Faculty of Educational and Social Sciences, University of Education Heidelberg, 69120 Heidelberg, Germany; christian.rietz@ph-heidelberg.de

[3] Institute for Agricultural Engineering, Livestock Technology Section, University of Bonn, D-53115 Bonn, Germany; buescher@uni-bonn.de

Received: 30 April 2020; Accepted: 9 July 2020

Published in: *Sensors* 2020, 20 (14), 3863; https://doi.org/10.3390/s20143863

Abstract

The aim of this study was to develop classification models for mastitis and lameness treatments in Holstein dairy cows as the target variables based on continuous data from herd management software with modern machine learning methods. Data was collected over a period of 40 months from a total of 167 different cows with daily individual sensor information containing milking parameters, pedometer activity, feed and water intake, and body weight (in the form of differently aggregated data) as well as the entered treatment data. To identify the most important predictors for mastitis and lameness treatments, respectively, Random Forest feature importance, Pearson's correlation and sequential forward feature selection were applied. With the selected predictors, various machine learning models such as Logistic Regression (LR), Support Vector Machine (SVM), K-nearest neighbors (KNN), Gaussian Naïve Bayes (GNB), Extra Trees Classifier (ET) and different ensemble methods such as Random Forest (RF) were trained. Their performance was compared using the receiver operator characteristic (ROC) area-under-curve (AUC), as well as sensitivity, block sensitivity and specificity. In addition, sampling methods were compared: Over- and undersampling as compensation for the expected unbalanced training data had a high impact on the ratio of sensitivity and specificity in the classification of the test data, but with regard to AUC, random oversampling and SMOTE (Synthetic Minority Over-sampling) even showed significantly lower values than with non-sampled data. The best model, ET, obtained a mean AUC of 0.79 for mastitis and 0.71 for lameness, respectively, based on testing data from practical conditions and is recommended by us for this type of data, but GNB, LR and RF were only marginally worse, and random oversampling and SMOTE even showed significantly lower values than without sampling. We recommend the use of these models as a benchmark for similar self-learning classification tasks. The classification models presented here retain their interpretability with the ability to present feature importances to the farmer in contrast to the "black box" models of Deep Learning methods.

Keywords: classification; sensor data; lameness; mastitis; machine learning

1. Introduction

Supporting herd managers to identify animals with health problems is an important task of precision livestock farming. The automation of dairy farms as well as the size of dairy herds is continuously increasing [1] which makes it necessary to support farmers with digital decision support systems, and ensure the welfare of the cows not only for economic reasons, but also against the background of animal protection laws and ethics. A large number of studies already exist that have developed and evaluated models for classifying cows in need of treatment for mastitis [2,3,4,5] and lameness [6,7,8,9] with different machine learning methods, such as logistic regression [8,10], support vector machines [6], Bayesian classifiers [3,4] and neural networks [2,11]. These studies are usually limited to testing a single model with different conditions, so they can only be compared to a limited extent. These studies use different independent variables as features, and the machine learning models used differ in their ability to represent the importance of the features in relation to the target [12].

Furthermore, the reference method for defining the positive case varies between systematic veterinary examinations of all animals [13], assessment of the gait of cows [6,10,14] and milk [14], respectively, and records of mastitis or lameness treatments [3,7]. A lack of a standard for the target characteristic makes it impossible to compare different publications, especially in the case of lameness detection, as the visual assessment of lameness is to some extent subjective, and there are more than 20 different scoring scales [15]. In addition, cows and days were sometimes selectively included in the test data [13] or unexplained cases were removed from the test data [8,16]. These circumstances also change the frequency of occurrence of the target variable, i.e., the probability of whether an animal has undergone treatment or not, in the data. An artificially high frequency allows for higher combinations of sensitivity and specificity than what would be the case in practice [15].

The aim of the present study was to apply a variety of machine learning models, e.g., logistic regression, support vector machines and decision tree-based models, and different methods of sampling (random under- and oversampling, SMOTE) to a practical data set in order to identify the most important features and make daily classifications of cows for mastitis and lameness treatments. The results are compared using the receiver operator characteristic (ROC) as well as the combination of sensitivity and specificity, and will finally be used to make recommendations for further experiments of this kind.

2. Materials and Methods

2.1. Data Source and Preprocessing

Raw data was collected from the dairy herd at Frankenforst research farm of the University of Bonn. The herd on average consists of 65 German Holstein dairy cows with 305-day milk yield of 9605 kg. All data from the farm was collected, processed via an SQL database system and presented as CSV files. The raw data comprised a period from June 2015 to October 2018 and contained in total 167 different animals. The data set consisted of individual animal information (animal ID, parity, days in milk) with one record per cow per day (n = 80,307). Milking data (milk yield, duration, milk flow, and conductivity) was recorded twice a day. Feeding (roughage and concentrate from automated feeders) and drinking data contained the amount per visit, number and time of visits as well as visits without intake. Activity was measured as the sum of impulses from a pedometer at a 2 h resolution, and climate data (temperature, humidity) from a nearby weather station was present as daily aggregations. The data also contained monthly milk recordings (fat, protein and lactose content, and somatic cell count in 1000 cells/mL), which was copied to each day for each cow until the next successive recording, so that each day contained the information about the data from the last milk recording. Lastly, all recordings of veterinary treatments, hoof trimmings and other routine measures were added to the data, each with a category and diagnosis. Cows in need of a treatment were identified by farm staff during the work routine, and treatments were conducted by a trained veterinarian.

Before further processing, the data was checked for plausibility: for feed intake, water intake, and visit duration all values that were more than 3 standard deviations above or below the herd mean were removed, as well as absolute values below 5 kg and above 100 kg for feed and 200 kg for water intake. Values above 15 kg for daily concentrate intake were also discarded. Values deleted in this process, as well as missing values in the raw data, were linear interpolated up to 7 d. Recordings with remaining missing values were discarded after this process.

2.1.1. Additional Aggregation

The Pedometer activity and the feeding visit data were further aggregated with calculations presented in [8] (see Table 2-1 for variable descriptions). In addition, for each variable except parity, days in milk and weeks in milk, several aggregations over multiple consecutive days were calculated to capture their development over time for each cow, which are also described in Table 3-1.

From the data, all days with recorded lameness and mastitis treatments were extracted as the dependent variables (targets). For each lactation with at least one treatment, only the first treatment was considered and the remaining days were discarded. Because of a high probability that the sensor data on the treatment day was influenced by the treatment itself, the day of treatment was then shifted back one day. In addition, all 3 days prior to a treatment were also considered as a positive target to calculate block sensitivity (for a definition see Section 2.5), as done in [7,17].

2.1.2. Data Splitting

For each classification of lameness and mastitis cases the data set was split 10 times into training and testing data for model evaluation. Sixty six percent of individual cows were randomly sampled and all data points from those cows formed the training set. Consequentially, the test set contained the remaining data. In both data sets, days in lactations that contained a lameness or mastitis treatment were discarded after the first treatment occurred. For lactations without a treatment, a random day was chosen for cut-off instead. The data in the training set was further reduced to four weeks per lactation. Finally, the data in both sets was scaled by subtracting each feature’s mean and dividing by its standard deviation (Z-score normalization), because some models, like KNN and SVM, assume that all data is within a similar range.

2.2. Feature Selection

To lower the training time and overfitting of some algorithms (see discussion section), the number of independent variables (features) in the data was reduced. First an estimator (in this case Random Forest) was fit to the training data to obtain feature importances (RF-I) and biserial correlation (r), i.e., the average decrease in impurity for a feature over all trees [18]. The features were then sorted by their RF-I and the best 100 were kept. Further reduction was done by using Sequential Forward Selection (SFS) [19]. Here, the estimator is initialized with an empty feature subset, and in each iteration, one additional feature is added to the subset, performance of the estimator is measured and the feature associated with the best performance is removed from the feature set and added to the subset, until it reached a size of 20 features.

Table 3-1. Per variable description of additional aggregation of sensor data, daily (feed and water intake data, and pedometer activity) and over multiple consecutive days (all variables except parity, days in milk and weeks in milk).

Aggregation	**Description**
Daily	
Mean	Arithmetic mean
SD	Standard deviation
Median	Median
Sum	Sum of values
Max	Highest single value
Min	Lowest single value
Range	Max-Min
3 highest (Sum)	Sum of the 3 highest values
6 highest (Sum)	Sum of the 6 highest values
3 lowest (Sum)	Sum of the 3 lowest values
6 lowest (Sum)	Sum of the 6 lowest values
Sum Day	Sum of values from 04:01 to 20:00
Sum Night	Sum of values from 20:01 to 04:00
Day/Night ratio	Sum Day / Sum Night
Multiple days	
d-1	Value of previous day
d-2	Value 2 days before
d-3	Value 3 days before
RM	Rolling Mean of previous 7 days
RMdiff	Difference of current day’s value to RM
RMprev	Rolling mean of previous week (d-8 to d-14)
slope	Slope of a linear regression from the recent 7 values

The measure of performance was Matthew's correlation coefficient since it is a performance metric suitable for imbalanced data [20]. This analysis was performed for both target variables (lameness and mastitis treatments).

2.3. Sampling Methods

The known low occurrence of treatments in the data resulted in a high imbalance between days with and without treatment. A common method of addressing this is to equalize the distributions of both classes by sampling. Three commonly used methods from the Python module imblearn [21] were applied to the training data: (1) Random Oversampling populates the data set with copies of randomly selected data points of the minority class (days with a treatment). (2) In Random Undersampling data points from the majority class (days without treatment) are removed at random. (3) SMOTE (Synthetic Minority Over-sampling Technique) is a variant of oversampling, where instead of copying existing data points, new (synthetic) data points of the minority class are created by creating a random vector between a data point and its k neighbors (k = 3 in this case) in the multidimensional feature space [22]. Each of these three methods resulted in a data set with an equal number of days with and without a treatment. The sampled training data sets as well as the non-sampled data were then used to train the classification models.

2.4. Classification Models

Data processing, model building and presentation of results was done with the Python language (Python Software Foundation, Wilmington, DE, USA). The following classification models that were used were all part of the module Scikit-learn [23]: Logistic Regression (LR), Support Vector Machine (SVM), K-nearest neighbors (KNN), Gaussian Naïve Bayes (GNB), Decision Tree Classifier (DT), Random Forest (RF), Extremely randomized trees, or ExtraTrees (ET), and AdaBoost (ADA). These represent a wide range of commonly used machine learning techniques.

LR calculates the probability of a data point to belong to one of two classes, in this case a day with a treatment. This is done by estimating the model parameters via maximum likelihood estimation. Given the probability, a threshold is then used to classify the data point [18]. To compensate for possible multicollinearity in the feature set and improve coefficient estimates, the l2 penalty (ridge regression) was added to the model. An SVM constructs a linear decision surface, also known as hyperplane, in a multi-dimensional feature space that has the widest possible margin to separate data points of both classes. To achieve non-linear separation, the input vectors are mapped to a higher dimensional space with a kernel function [24]. In KNN classification, a data point is assigned to one class by a majority vote of its k neighbors. The metric used to identify the neighbors is the Euclidean distance in the feature space [25]. In this study the default value of 5 was used for the number of neighbors. GNB estimates the probability of a cow having a treatment or not given the corresponding feature vector [18]. Though GNB makes the naïve assumptions that the input features are normally distributed and independent, the obtained binary classifications work reasonable in practice, even with violated assumptions [26]. A DT classifies data points by asking questions about the feature vector (called interior nodes) that eventually lead to one of many labelled end points (called leaf nodes). At each node, the feature to split the data is determined by a measure of purity of the daughter nodes, typically the Gini impurity or the information gain [27].

Ensemble Methods

A model ensemble is a classification model that, instead of constructing a single model to make a prediction, generates a set of different models and combines their predictions into a single estimation. This is done by bagging, where each model of the ensemble is trained independently on a random bootstrapped sample of the training data, and boosting, where a strong classifier is built from a set of multiple weak classifiers [28]. In a RF, decision trees are built with only a limited, randomly drawn selection of features at each node [29]. In contrast to decision trees, the random sampling of each tree makes the random forest less susceptible to over-fitting [18,29]. An important hyperparameter for this algorithm is the number of features selected at each node, which here was set to √F, where F is the number of total features in the data set [18]. Furthermore the number of trees was set to 500 and the maximum depth was not limited. ET is a derivation of the random forest algorithm, where in addition to picking features at random for each split, the threshold that splits the data is also drawn from a randomly generated set of splits. This is done to reduce the variance even more in comparison to the random forest [23]. ADA creates an ensemble of weak classifiers that are restricted in their depth (also called stumps). These learners are used to give predictions on a modified set of the data, where each data point is given a weight. These weights decrease and increase for correctly and incorrectly classified data, respectively, meaning that those data points that are difficult to classify gain influence with each iteration. The classification of test data is done through a majority vote [18].

A Voting Classifier simply is an ensemble of arbitrarily selected classifiers. The decision on labelling a sample can either be done through hard voting, where the assigned class label is the majority vote of all classifiers, or soft voting, which takes into account the uncertainty of each classifier by using the weighted average class probability [23]. In this experiment, two different configurations were used: SoftVoting1 (RF, KNN, and GNB) and SoftVoting2 (LR, ADA and KNN).

2.5. Evaluation

For each classifier, a prediction of the probability of belonging to the class label 1, i.e., in need of a treatment, was given for each data point in the test data. The resulting vector of probabilities compared with the vector of true labels was used to create a Receiver Operator Characteristic (ROC) curve, which plots the rate of true positives over the false positive rate (1 − specificity) for all different thresholds (thresholds that result in redundant combinations are excluded) [23]. This curve allowed to calculate the Area Under Curve (AUC), which is a value between 0 and 1, where 0.5 describes a random classification, and 1 would be a perfect match between classification and the target variable.

Before classification of the testing data, a subset of 33% of the training data was selected at random and used as a validation data set for which a classification was made. From the resulting ROC, the threshold was selected where the sensitivity was at least 0.8, and this threshold was used for classification of the testing data set.

The resulting true positives, false positives, true negatives and false negative were entered into a confusion matrix to calculate the specificity, as well as the positive and negative predictive values. In addition, the block sensitivity was calculated, where a true positive was at least one correct classification of the three days before a treatment, and a false negative if neither of these days was classified as a treatment.

All results are presented as the mean ± the 95% confidence interval of the mean. The measures of performance (AUC, sensitivity, block sensitivity and specificity) for classification models per sampling method (no sampling, Random Over- and Undersamling, SMOTE) were compared using the Welch’s test due to violated homogeneity of variance. These tests were implemented with SPSS version 26.0 (IBM Corp, Armonk, NY, USA) with significant differences at $p \leq 0.05$.

3. Results

After processing and plausibility checks, 53,970 records remained in the data from 112 individual cows and 235 cows' individual lactations, respectively. This corresponds to approx. 67% of the original data. Data from 55 cows was removed due to short lactations <10 d or incomplete sensor data.

3.1. Feature Importance

The process of variable creation described in Section 2.2 resulted in a total of 471 different independent variables (features). The total number of variables by category can be seen in Table 3-2. The most features were generated by sensors where data was available in a higher than daily resolution (feed and water troughs, pedometer activity). The milking variables included daily milkings as well as the last monthly milk recording. From all features, 83 represented data of a single day, while the other 350 described temporal relationships.

Table 3-2. Number of features by category, before feature selection.

Category	Number of Features
Animal dependent variables	
Feed and water intake and visits	189
Activity	127
Milking	77
Concentrate intake	25
Body weight	8
Other[1]	3
Animal independent variables	
Climate	4

[1] parity, days and weeks in milk

Feature importances (RF-I) and correlations (r) with the target variable (mean + 95%-CI) of the 20 most important features for mastitis treatments are shown in Table 3-3.

3.1.1. Mastitis Treatments

The highest RF-I (0.039) as well as the highest correlation with the mastitis treatment events (0.176) was shown by the somatic cell count from the last monthly milk recording. It is remarkable that of the further variables listed, only two others can be associated with the daily milking data: The slope and the difference from the rolling mean of the evening milk conductivity (RF-I = 0.013, r = −0.076 and RF-I = 0.010, r = 0.080, respectively). The other 19 most important features consisted of temporal derived variables from the feeding troughs and the concentrate feeder. The slope of the absolute concentrate intake and the deviation from the allowance were both negatively correlated with the treatment. The feed intake variables showed positive correlations while the correlations of the feeding visits were negative.

3.1.2. Lameness Treatments

RF-I for lameness classification were lower overall than for mastitis. Total time at the feeding trough with intake was most important (RF-I = 0.013) and showed a relatively high negative correlation (r = −0.105). The number of feeding and drinking visits and the time spent at the through also had negative correlation values. The four present activity features showed the standard deviation and range of daily values and had a negative correlation. Two climate features (temperature and THI) had a RF-I of 0.009 and a correlation of −0.061. The proportion of temporal derived features among the most important was lower than with mastitis.

3.2. Classification Results

3.2.1. Results for Training Data

For a part of 33% of the training data (=validation data, sampled and without sampling) classifications were made by each machine learning method to obtain a limit value for the classification of the test data based on the resulting combinations of sensitivity and specificity. From these results, AUC as well as sensitivity and specificity could be calculated, providing an overview of the application of the methods to sampled data. These data are presented in Table 3-4 (mean values ± 95%-CI for each sampling method, results for all machine learning methods combined). From this, it can be seen that for both mastitis and lameness treatments, random oversampling and SMOTE are higher than random under sampling (0.76 and 0.71) and without sampling (0.80 and 0.76) for AUC, with 0.95 and 0.91 respectively.

3.2.2. Results for Testing Data

All following results refer to the classification of the test data. The mean AUC results of the classification models trained on non-sampled data are shown in Figure 3-1. For mastitis, ET, GNB, GridSearchDT, LR, RF and the Soft Voting ensembles showed the highest mean AUC with ET at 0.79 (0.73–0.84). Grid search improved the Decision Tree model, but not AdaBoost. There were no differences between the two Soft Voting ensembles (AUC 0.74 vs. 0.73 for mastitis and 0.69 vs. 0.66 for lameness, respectively). The overall variance in the mastitis classifications was higher than in lameness, as indicated by the larger CI.

Table 3-3. The 20 most important variables ranked by RF-I (mean ± 95%-CI) with respective r-values (mean ± CI, all correlations with $p < 0.001$).

	Mastitis Treatments Classification			Lameness Treatments Classification		
Rank	**Feature**	**RF-I** [1]	**r**	**Feature**	**RF-I**	**r**
1	Last Milk recording SCC [2]	0.039 ± 0.009	+0.176 ± 0.016	Feeding time with intake	0.013 ± 0.005	−0.105 ± 0.014
2	Concentrate intake, slope	0.014 ± 0.005	−0.076 ± 0.016	Feed intake Sum day, RMprev [3]	0.012 ± 0.006	+0.079 ± 0.010
3	Milk conductivity p.m., slope	0.013 ± 0.006	+0.082 ± 0.024	Activity, SD, RMprev	0.012 ± 0.006	−0.072 ± 0.012
4	Feed intake (Median), RMprev	0.011 ± 0.004	+0.067 ± 0.010	Feeding visits with intake	0.011 ± 0.005	−0.080 ± 0.013
5	Feed intake (S.D.), RMprev	0.011 ± 0.004	+0.064 ± 0.008	Activity (Range), RMprev	0.010 ± 0.006	−0.067 ± 0.011
6	Feeding visit duration (mean), RM [4]	0.011 ± 0.004	+0.009 ± 0.006	Activity (Max), RMprev	0.010 ± 0.004	−0.066 ± 0.011
7	Feed intake 6 highest (Sum), RMprev	0.011 ± 0.006	+0.068 ± 0.009	Air temperature	0.010 ± 0.004	−0.061 ± 0.015
8	Feed intake 3 highest (Sum), RMprev	0.010 ± 0.005	+0.068 ± 0.009	THI [5]	0.009 ± 0.003	−0.061 ± 0.015
9	Feeding visit duration (mean), d−3	0.010 ± 0.004	−0.002 ± 0.006	Feed intake (SD), RM	0.009 ± 0.004	+0.098 ± 0.011
10	Conc. intake abs. deviation, RM	0.010 ± 0.004	−0.047 ± 0.014	Feeding time with intake, RM	0.009 ± 0.006	−0.062 ± 0.008
11	Milk conductivity p.m., RMdiff [6]	0.010 ± 0.005	+0.080 ± 0.017	Feeding time with intake, RMdiff	0.009 ± 0.003	−0.095 ± 0.018
12	Feed intake (Max), RMprev	0.009 ± 0.006	+0.065 ± 0.01	Drinking time with intake	0.009 ± 0.005	−0.065 ± 0.009
13	Feed intake (Mean), RMprev	0.009 ± 0.005	+0.067 ± 0.01	Feeding time with intake, slope	0.008 ± 0.003	−0.090 ± 0.019
14	Feeding visits with intake, RMprev	0.009 ± 0.006	−0.060 ± 0.005	Feed intake (Median)	0.007 ± 0.001	+0.107 ± 0.010
15	Feed intake (S.D.), RM	0.008 ± 0.003	+0.052 ± 0.007	Feed intake, 6 highest (Sum), RM	0.006 ± 0.003	+0.101 ± 0.009
16	Conc. intake rel. deviation, RM	0.008 ± 0.004	−0.036 ± 0.011	Feed intake per visit	0.006 ± 0.003	+0.106 ± 0.011
17	Feeding visit duration (Mean), RMprev	0.008 ± 0.005	+0.034 ± 0.008	Activity, 3 highest (Sum), RMprev	0.006 ± 0.003	−0.067 ± 0.011
18	Feed intake 6 highest (Sum), RM	0.008 ± 0.003	+0.062 ± 0.008	Feeding visits with intake, d-1	0.006 ± 0.003	−0.068 ± 0.011
19	Feed intake (Range), RMprev	0.008 ± 0.005	+0.065 ± 0.010	Feed intake, RMprev	0.006 ± 0.004	+0.063 ± 0.015
20	Feeding visits with intake, RM	0.008 ± 0.005	−0.057 ± 0.005	Drinking visits, RM	0.006 ± 0.003	−0.059 ± 0.014

[1] Random Forest-Importance, [2] somatic cell count, [3] rolling mean of previous week, [4] rolling mean, [5] temperature humidity index, [6] Difference of current day's value to RM.

Table 3-4. Mean AUC, Sensitivity and Specificity (± 95%-CI) for validation data (33% of sampled training data), means for all machine learning methods.

Sampling of Training Data	**AUC** [1]	**Sen.** [2]	**Spe.** [3]
Mastitis treatments			
No sampling	0.80 ± 0.02 ^{b}	0.72 ± 0.04 ^{c}	0.72 ± 0.05 ^{b}
Random Undersampling	0.76 ± 0.01 ^{c}	0.81 ± 0.02 ^{b}	0.59 ± 0.04 ^{c}
Random Oversampling	0.95 ± 0.01 ^{a}	0.89 ± 0.02 ^{a}	0.91 ± 0.02 ^{a}
SMOTE [4]	0.95 ± 0.01 ^{a}	0.88 ± 0.01 ^{a}	0.91 ± 0.02 ^{a}
Lameness treatments			
No sampling	0.76 ± 0.02 ^{b}	0.70 ± 0.04 ^{b}	0.68 ± 0.05 ^{b}
Random Undersampling	0.71 ± 0.01 ^{c}	0.80 ± 0.02 ^{b}	0.53 ± 0.03 ^{c}
Random Oversampling	0.91 ± 0.02 ^{a}	0.89 ± 0.02 ^{a}	0.84 ± 0.04 ^{a}
SMOTE	0.91 ± 0.02 ^{a}	0.87 ± 0.01 ^{a}	0.83 ± 0.04 ^{a}

[1] Area Under ROC-Curve; [2] Sensitivity; [3] Specificity; [4] Synthetic Minority Over-sampling Technique; ^{a,b,c} superscript letters indicate significant differences at $p \leq 0.05$ between sampling methods within treatments.

To compare the sampling methods, Figure 3-2 shows the differences for AUC, sensitivity, block sensitivity and specificity between the sampling methods. For mastitis, Random Undersampling resulted in a higher AUC than both oversampling methods. For lameness the difference between Random Undersampling and Oversampling was not significant, but there was still a difference to the models that used SMOTE. Without sampling and with Random Undersampling the ratio of sensitivity to specificity is moving in favor of sensitivity for both mastitis (Figure 3-2a) and lameness (Figure 3-2b) treatments.

In order to test the effects of including the data from feed and water troughs, the complete evaluation was carried out both including these features and without. Table 3-5 shows the results for AUC, sensitivity and specificity, averaged over all sampling methods and machine learning models (mean values ± 95%-CI). It becomes clear that similar AUCs (0.67 and 0.66) were achieved in the classification of mastitis treatments both with and without the feed and water data, while the classification of lameness treatments suffered from the exclusion of these features (AUC of 0.62 vs. 0.55).

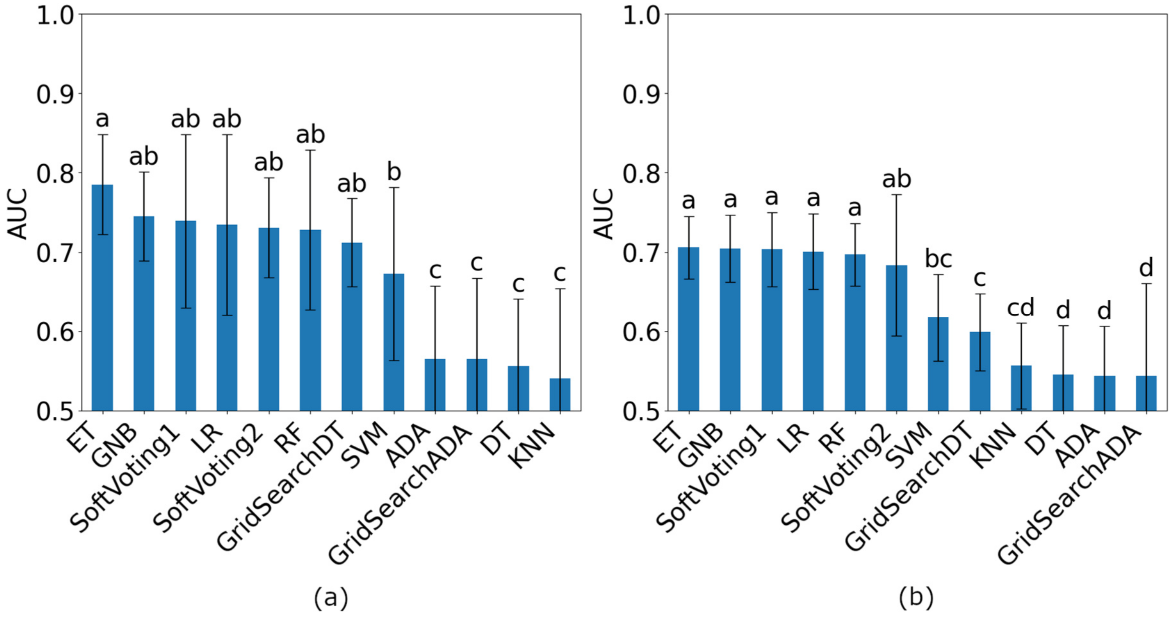

Figure 3-1. Mean test data AUC (Mean ± CI) for models trained on non-sampled data. (a) Mastitis treatments; (b) Lameness treatments. Different letters indicate significant ($p < 0.05$) differences between classification models. AUC: Area Under ROC-Curve; ET: ExtraTrees Classifier; GNB: Gaussian Naïve Bayes; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; ADA: AdaBoost; DT: Decision Tree; KNN: K-Nearest Neighbors.

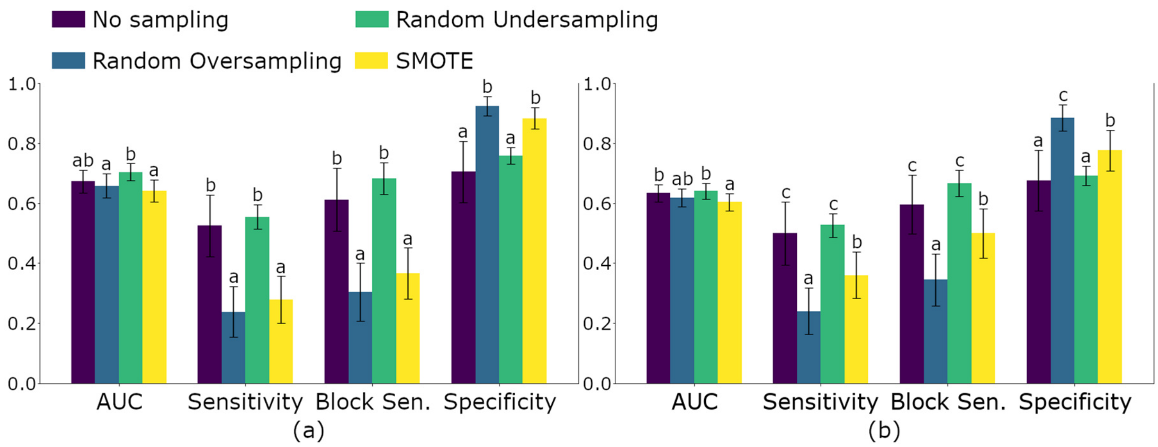

Figure 3-2. Mean test data AUC, Sensitivity, Block Sensitivity and Specificity (± 95%-CI) for each sampling method. (a) Mastitis treatments; (b) Lameness treatments. Different letters indicate significant ($p < 0.05$) differences between sampling methods. AUC: Area Under ROC-Curve; SMOTE: Synthetic Minority Over-sampling Technique.

Table 3-5. Mean AUC, Sensitivity and Specificity (± 95%-CI) for classification of treatments with or without inclusion of data from feed and water troughs, means include all machine learning models and sampling methods.

Feed and Water Data Included	AUC [1]	Sen. [2]	Block Sen.	Spe. [3]
		Mastitis treatments		
Yes	0.67 ± 0.01	0.40 ± 0.02	0.49 ± 0.03	0.82 ± 0.02
No	0.66 ± 0.01	0.39 ± 0.02	0.51 ± 0.02	0.82 ± 0.01
		Lameness treatments		
Yes	0.62 ± 0.01 ^{a}	0.41 ± 0.02	0.53 ± 0.02 ^{a}	0.76 ± 0.02 ^{a}
No	0.55 ± 0.01 ^{b}	0.38 ± 0.02	0.50 ± 0.02 ^{b}	0.69 ± 0.02 ^{b}

[1] Area Under ROC-Curve; [2] Sensitivity; [3] Specificity; [a,b] superscript letters indicate significant differences at $p \leq 0.05$ within treatments.

To finally compare the machine learning models in their ability to correctly classify mastitis and lameness treatments, Table 3-6 ranks them by AUC (mean ± 95%-CI) in descending order. The highest mean AUCs for the classification of mastitis treatments in the test data were obtained from LR, ET, GNB, Soft Voting 1 and 2, RF, and Grid Search DT with values between 0.75 and 0.69. For lameness treatments GNB, Soft Voting 1, ET, LR, RF and Soft Voting 2 yielded the highest AUCs between 0.70 and 0.66. In both treatment categories KNN, Grid Search ADA, ADA and DT resulted in the lowest AUC values between 0.58 and 0.54.

4. Discussion

4.1. Feature Importance

Feature importance obtained from Random Forest (RF-I) is commonly used to reduce the dimensionality of the model input. It is also interpreted to understand connections between the input and the output data, however, there are certain caveats such as that there is a bias against variables with only few categories [30] as well as a bias when variables are highly correlated [31]. This is why Sequential Forward Selection (SFS) was used as an iterative approach to reduce the feature set to its final size for model training and classification. Here, because only the feature that increases the applied measure (here: Matthews Correlation Coefficient, MCC) the most, non-correlated variables are favored.

Table 3-6. Mean testing data AUC, Sensitivity, Block Sensitivity and Specificity (± 95%-CI) for each classification model, averaged over all sampling methods, for mastitis and lameness treatments.

Mastitis Treatments		Lameness Treatments	
Classification Method	**AUC [1]**	**Classification Method**	**AUC [1]**
LR [2]	0.75 ± 0.02 ^{a}	GNB	0.70 ± 0.01 ^{a}
ET [3]	0.75 ± 0.02 ^{a}	Soft Voting 1	0.69 ± 0.01 ^{a}
GNB [4]	0.75 ± 0.02 ^{a}	ET	0.68 ± 0.01 ^{ab}
Soft Voting 1	0.74 ± 0.02 ^{a}	LR	0.68 ± 0.01 ^{ab}
Soft Voting 2	0.73 ± 0.02 ^{a}	RF	0.67 ± 0.02 ^{ab}
RF [5]	0.72 ± 0.02 ^{ab}	Soft Voting 2	0.66 ± 0.02 ^{abc}
Grid Search DT [6]	0.69 ± 0.03 ^{bc}	SVM	0.62 ± 0.03 ^{bc}
SVM [7]	0.65 ± 0.03 ^{c}	Grid Search DT	0.60 ± 0.02 ^{cd}
KNN [8]	0.58 ± 0.02 ^{d}	KNN	0.57 ± 0.02 ^{de}
Grid Search ADA [9]	0.56 ± 0.01 ^{d}	Grid Search ADA	0.54 ± 0.01 ^{e}
ADA	0.56 ± 0.01 ^{d}	ADA	0.54 ± 0.01 ^{e}
DT	0.55 ± 0.02 ^{d}	DT	0.54 ± 0.01 ^{e}

[1] Area Under ROC-Curve; [2] Logistic Regression; [3] Extra Trees Classifier; [4] Gaussian Naïve Bayes; [5] Random Forest; [6] Decision Tree; [7] Support Vector Machine; [8] K-nearest Neighbors; [9] AdaBoost; $^{a,\ b,\ c,\ d,\ e}$ superscript letters indicate significant ($p < 0.05$) differences between classification methods within treatments.

MCC was chosen because it takes into account all parts of the contingency table and thus is suited and recommended for imbalanced classification problems [20].

The advantage of this feature selection method is that it retains interpretability of the models for the farmer, which is not possible when using other methods such as Principal Component Analysis and Deep Learning, where the feature importance is harder to interpret. In practical applications, models that show the correlations between an alarm and the input variables could offer additional help for the farmer. Random Forest feature selection is suitable for reducing the number of input variables to the most important ones while retaining the interpretability of the final models.

For mastitis treatments classification, the SCC of the last monthly milk recording highlighted as the most important predictor with an RF-I of 0.039 and r = 0.176, despite the fact that time gaps between the last measurement and treatments could be as large as a month. An elevated SCC is a direct indicator for both subclinical and clinical mastitis, which also explains the comparably high positive correlation with the treatments. Former studies also showed that including SCC as a feature improved classification: On-line measured SCC improved the positive predictive value for clinical mastitis from 0.11 to 0.32 compared to electrical conductivity alone [32]. The incorporation of cow information that included previous SCC from milk recording also improved AUC of mastitis alerts from 0.62 (only information from AMS) to 0.78 [3]. The knowledge of the last SCC could have influenced the decision to conduct a mastitis treatment, but there were no systematic accumulations of treatments in the week after milk recording results were received. The on-farm protocol is that for a cow to be treated there have to be signs of abnormal milk, detected by visual observation or with the California mastitis test. Compared to the importance of the other features (for mastitis as well as lameness treatments) the RF-I of monthly SCC is by far the highest. Because it can be assumed that SCC is a predictor with a direct relationship to udder health, it is questionable whether the features with RF-I < 0.014 should be viewed as having only an indirect relationship to the target variable. A raise in milk conductivity is also an indicator commonly used to detect mastitis, although the correlation to SCC is low (r = 0.48) [32] and the absolute value is dependent on the animal [33]. This explains the occurrence of two conductivity variables that capture the change over time in the 20 most important features for mastitis treatment classification. The somatic cell count from the monthly milk recordings is by far the most important predictor for the classification of udder treatments, as it is directly related to mastitis.

Other important features for mastitis treatments classification were derived from feeding data (concentrate and roughage intake, and feeding visits). Concentrate intake slope and deviation from allowance seemed to be important with RF-I of 0.014 and 0.010, respectively, but are dependent on the cow's milk yield, since maximum daily allowance is automatically adjusted by milk kg and days in milk. The features for roughage intake included the current and previous rolling means and were positively correlated with treatments (RF-I = 0.008–0.011, r = 0.052–0.068). This might be the result of cows with a high milk yield (and thus a higher feed intake) being at an up to 1.44 times higher risk for mastitis compared to cows with low milk yield [34]. Feeding data are not highly correlated with mastitis treatments. Therefore, they are less suitable as predictors for classifying udder treatments.

Features derived from pedometer activity and behavior are most commonly used as predictors for lameness, e.g., differences in average activity between days [7,10], and accelerometer data like number and duration of lying and standing events [35]. The mean number of step impulses of lame cows is supposed to be lower than for healthy cows [6].

This is expected and also shown in then own study: The activity features from the data all had a negative correlation with the lameness treatments, although this effect was small ($r = -0.072$). Pedometer activity has a significant, albeit small, correlation with lameness treatments.

In data from 118 cows (44 with a lameness-related treatment within a period of 10 months), activity variables (e.g., sum, mean and standard deviation of 12 daily 2 h-activity values, deviation from previous day, difference between previous weeks) showed a correlation of $r = 0.23 \pm 0.06$ on average [8]. Here, neck collars were used that delivered activity indices based on number, intensity and direction of impulses. The resulting sensor data might have a more direct correlation to lameness than impulses measured by a leg pedometer. But the authors of [8] built their database by strictly excluding cows with treatments other than lameness, which also contributed to a higher correlation with lameness treatments because there was less noise in the data from other treatments. In the own study, the aggregated impulse data from leg pedometers was used, because this sensor is available on most practical farms nowadays. As expected, the results show that they cannot be considered a sensor with a direct correlation to lameness treatment, unlike the monthly SCC for mastitis. This can be explained with a high variance between individual lameness events and their impact on sensor variables [9]. Additionally, there are more different underlying conditions for lameness (e.g., sole ulcers, sole hemorrhages, or digital dermatitis), which have a different impact on a cow’s gait [36]. Other sensors like automatic gait score assessment with cameras, or leg weight distribution can potentially improve classification, as seen in [37] where lameness (based on scores) was detected from leg weight distributions on a weighing platform, with an AUC of up to 0.88. This value cannot be compared to the own study though, because the underlying data set consisted only of 7 d and a single classification, based on clinical examinations of all cows. These advanced sensor systems are not widely available in practice. This emphasizes the need for sensors that are more directly related to lameness treatments than just the number of activity impulses.

Data of roughage and water intake and visits is obtained from weighing troughs that are only used in experimental dairy farms, but not usable in practice. Information about a cow’s feeding and drinking visits can be approximated with the use of tracking [38] and accelerometer systems [39], but the amount of roughage and water intake of individual cows remains unknown. When excluding all features derived from the weighing trough data in the own study, a mean AUC of only 0.55 was obtained for lameness classification, confirming the importance of those features. Two features from water troughs, drinking time

(RF-I = 0.009, r = −0.065) and rolling mean of the number of drinking visits (RF-I = 0.006, r = −0.059) were found among the 20 most important features. This indicates a potential for water trough visits as a feature for lameness classification. From a technical point of view, it is conceivable that in the future these characteristics could be recorded with the help of sensors. Data from feeding and water troughs improve the classification of lameness treatments, but are not available as such in practice.

4.2. Sampling Methods

The goal of over- and undersampling is to mitigate effects of severe class imbalance, as seen in the data used in the own study, where samples labeled as positive were less than 1% of the data set. The own results showed no improvement in AUC for testing data when using sampling methods on the training data. Few studies on the classification of treatments or health related events in dairy cows explicitly use over- or undersampling techniques. One study that tried to identify cows with a high locomotion score used SMOTE to balance data with 45 lame and 1613 non-lame days to achieve a sensitivity of 1.00 (95%-CI: 0.19–1.00) and specificity of 0.8 (95%-CI: 0.71–0.87) on validation data (10% of total data), but they did not compare that to a model trained on non-sampled data [40]. When classifications were made for sampled data (validation data, see Section 2.5), SMOTE and Random Oversampling, compared to non-sampled data, both improved the AUC for mastitis (0.95 vs. 0.80) and lameness (0.91 vs. 0.76). In the study of [41] different intensities of random sampling were applied to a credit score classification problem. The portion of negative samples in the data ranged from 1% to 30% and showed that some models (LR, linear SVM and quadratic linear discriminant analysis) suffer from high imbalances and resulted in AUC values of nearly 0.50 for data with 1% negative samples, while decision tree-based models were less affected (highest AUC for the 1% negative set of 0.90). However, in that study the testing data was also sampled. But for a robust estimation of a model's capabilities in a practical setting, sampling cannot be applied to the testing data, because that would require a priori knowledge of the true underlying condition (a cow is needing a treatment), which is not present in an applied use case. To make estimations for sensitivity, specificity, and AUC, the class distributions in the testing set should reflect those of real world data [18]. Over- and undersampling is commonly used to correct the imbalanced ratio between days with and without treatment in the training data, but this is not advisable for the testing data. Therefore, in the testing data this ratio should correspond as closely as possible to data from practical farms.

4.3. Interpretation of the Final Classification Models

Due to differences in the study design and the composition of the data, a direct comparison of the own results of the classification models with those of other studies dealing with the classification of dairy cows in need of treatment is only possible to a certain extent, even if the statistical methods are the same [42].

Studies that have also developed models to classify lameness or mastitis treatments as target variables differ in the independent variables (features) used. Studies that have classified cows for lameness treatment have either focused on feeding related variables such as feed intake, trough time or number of visits [43], or only on ALT pedometer data such as number of impulses and resting time [6] with a resulting accuracy of 0.76, or used multiple data sources (live weight, pedometer activity, milk yield) and additional individual animal information [10] to obtain an AUC of 0.74 compared to models with only a single parameter that yielded AUCs of 0.60–0.66. The own models used sensor data that are widely available in practice (pedometer activity, milking parlor data, concentrate intake, live weight, climate data, and monthly milk recordings) as well as data from feeding and drinking troughs (visits and intake). The classification for mastitis treatments in previous studies was based on activity and rumination time [5] with a sensitivity of 0.55, or change in milk yield and conductivity [44] to obtain a sensitivity of 0.48 when specificity was set to 0.98. [4] included both data from real-time milk analyzers (SCC, fat, protein) and non-sensory information (parity, season, and weeks in milk) in their model, which resulted in an AUC of 0.89. In our own study mainly sensor data and information were used that are already available on many farms, and we recommend this approach when developing classification models for practical application. The feed and water intakes are a special case, because the data acquisition is only possible on experimental farms with weighing troughs. Therefore, the evaluations were carried out also without these features. The results show that for the classification of mastitis treatments, the exclusion all features from feed and water troughs did not affect the AUC, while for the lameness treatment classification the mean AUC dropped from 0.62 to 0.55. This emphasizes the importance of those features and shows the need to develop and improve sensors that measure feeding behavior of individual cows.

The data sets also differ between the studies in terms of the different pre-selection options. This selection takes place before the separation into training and test data and therefore affects the relationship between treated and non-treated cows, which leads to a higher probability of correctly positive classifications. In the study of [6] a total of 549 days out of eleven cows were selected, of which about half were classified as "lame" based on lameness scores. Other studies excluded unclear cases from their data: Exclusion of cows with a scoring system from 1 (non-lame) to 4 (severe lameness) [16] or 1 to 5 [35] and exclusion of all animals with a score of 2 in both cases, or data for cows with more than 50% missing sensor data for activity were excluded [35]. Another possibility is to limit the test data to a certain number of days before treatment, e.g., 3 weeks, resulting in a higher proportion of days with treatment [8]. In the studies mentioned above, it is questionable which values for sensitivity and specificity would have resulted from less selected data. In our own study, only the days after treatment were not considered in the test data, missing values were interpolated. Values for sensitivity and specificity of up to 0.79 for mastitis treatments and 0.71 for lameness treatments could be achieved on the non-sampled test data. Excluding data and thus artificially increasing the number of positive cases in the test data leads to higher values for AUC, sensitivity and specificity and is therefore not recommended. Test data sets should, as far as possible, be similar to the data that will need to be classified later in practical use on farms. Many studies pre-select both training and test data, which leads to higher AUCs, but these are not achieved in practice. In future studies, classification models from sensor data should be tested on practical data.

In addition to the above-mentioned methods of processing the test data set, it is possible to increase the number of days with treatment by also defining a certain number of days before a treatment is carried out as a "day with treatment", i.e., as a positive target variable. In our own study, 3 days were used, as in [7] for mastitis and lameness treatments. This allows the calculation of block sensitivity, where at least one of the three days prior to treatment must have been positively classified. In [7] block sensitivities of 0.77 for mastitis and 0.74 for lameness treatments were achieved, but no sensitivities per day were given for comparison. The highest block sensitivities in our own experiment were 0.80 for mastitis treatments and 0.81 for lameness treatments. In a practical setting on a farm, the exact day before a necessary treatment is not important as long as the need for treatment is visible. A disadvantage is that this can lead to animals in need of treatment being detected too early without clinical signs and then incorrectly registered as healthy, as has been discussed in other literature [9,42]. The calculation of block sensitivity reflects the use of a model in practice better than pure sensitivity for the exact day before treatment and is therefore recommended for the evaluation of classification models for mastitis and lameness treatments.

Finally, the recording, or definition, of the target variable also influences the interpretation of the results. In our own study, mastitis and lameness treatments were carried out by veterinarians and stable personnel, respectively. Also, in other studies on classification models of dairy cow data, treatments of mastitis [3,4,7,43,45] and of lameness [7,8,43] were used as target variables. Although these records are made by qualified professionals and are therefore considered reliable, it was discussed to what extent this type of data collection incorrectly records cows without clinical signs as not requiring treatment [15,46]. Other studies have used lameness scores or clinical assessment to define lameness [10,35,36] and cell count measurements or cytobacteriological tests for mastitis [5,14]. This type of recording potentially reflects the health status of the individual animals better, but is much more personnel and time intensive and is therefore not suitable for evaluations over periods of several years. In addition, the use and storage of treatment data collected under practice conditions allows the underlying classification models to continuously adapt and learn from newly entered data to improve the classification of future treatment events.

So far there are few studies that systematically compare the application of different machine learning models for the classification of farm animal data. A study by [47] compared the methods RF, SVM, KNN and ADA for the classification of three different types of grazing behavior in sheep. There, RF gave the best results (highest accuracy of 0.92), with KNN and ADA worse by 0.05 on average. The authors particularly emphasized RF’s ability to correctly assess non-linearly related data and its robustness against statistical noise. The authors of [48] compared different methods (NB, DT, RF, Bayesian network and bagging) to predict insemination success in Holstein dairy cows based on phenotypic and genotypic traits. Results for AUC ranged from 0.61 to 0.75, with RF showing significantly better results than all other methods tested. Two studies from other disciplines are also mentioned here, credit assessment [41] and Alzheimer diagnostics [12], as they also systematically compared classification models on a binary target variable. In the study by [41] data sets with different proportions of good and bad credit scores were created and classified by means of down-sampling. Decision tree-based models (RF and gradient boosting) provided the highest AUC (up to 0.90), while other models such as LR and SVM showed only random classifications (AUC = 0.5) for the data sets with the lowest proportion of bad scores. The Alzheimer study, which used features of electrical brain activity, compared RF, SVM, LR and neural networks and found only minimal differences in AUC (0.83–0.87). The authors emphasize the advantages of models that output feature importance (RF, LR) compared to (non-linear) SVM and neural networks, whose decision making is much more difficult to interpret [12] and are therefore considered "black boxes". In our own study, models based on decision trees (ET, RF) or containing them (Soft Voting 1), but also LR and GNB resulted in the highest average AUC (0.71–0.79 for mastitis treatments and 0.67–0.71 for lameness treatments, respectively).

Random forest models are more robust compared to single decision trees, because the tendency to overfitting and the overall variance within the model is lower [18]. They also perform well when dealing with class imbalances [41]. Logistic Regression has also been used for classification of treatments in other studies [8,10,49] and has the key advantage of interpretability, where not only the absolute importance of each feature, but also the trend can be derived from its coefficients [18,23]. Based on the results of this study the use of ET, LR, RF and GNB can be recommended for further, similar studies.

An AUC of 0.75 or 0.70, as achieved by the best own models for detecting mastitis and lameness treatments with a practical data set, is within the range of studies that have worked with similar data and sensor combinations. A value of 1 would mean a perfect classification. A value above 0.70 is considered a "strong model", while a value below 0.60 is considered a "weak model" [28]. The more directly the features are related to the target/classification variable, the greater the AUC, as can be seen from the example of somatic cell count in relation to mastitis treatments in the own data. Sensors and features with a direct relationship to the target variable are required to achieve a higher AUC. For the detection of dairy cows with classification models no minimum requirements for AUC values are known. Various authors demand minimum sensitivity values of 0.70 or 0.80 with a specificity of 0.99 [15], which would correspond to an AUC of more than 0.90, and cannot be achieved with practical data sets (i.e., not sampled or similar). Only a few studies critically question the practical applicability of models with lower than the required AUC (or sensitivity and specificity combinations), e.g., in a study on calving prediction, where the authors point out the model's limited benefit resulting from a low frequency of occurrence of the target variable "calving" in the data, and the resulting low positive predictive value despite AUC values of up to 0.81 [50]. The authors' own AUC values also make it clear that the application of classification models from practical sensor systems on realistic data sets must be viewed critically and does not reliably lead to the identification of animals in need of treatment or similar classification events, as generally expected. Thus, limits of the use of sensors (the corresponding machine learning methods and all other associated techniques) for finding individual animals in need become apparent.

5. Conclusions

The following recommendations for future sensor-based classification models for single animal-related events result from the comparison of different evaluation variants and models on a comprehensive test data set: (1) The mastitis classification resulted in an overall mean AUC higher by 0.05 than the lameness classification, due to predictors with a higher correlation to the treatment (especially somatic cell count from monthly milk recordings with r = +0.18). (2) Over-and under-sampling of days with treatments in the training data did not improve the AUC when classifying the testing data, where the class balance should not be artificially increased and always reflect practical data. (3) The use of treatments as the target variable is required when using practical data over long periods of time and enables the use of potential self-learning models, which would not be possible with clinical observation. (4) The classification models presented here retain their interpretability with the ability to present feature importance and their significance to the farmer. This is an advantage over "black box" models such as deep neural networks, and should be considered in future models. (5) The best models, LR, ET, RF, GNB, and Soft Voting 1 and 2, obtained a mean AUC of 0.72–0.79 for mastitis and 0.66–0.71 for lameness, respectively, based on testing data from practical conditions and are recommended by us for this type of data.

This study shows that the classification of treatments using practical sensor data achieved similar results as comparable studies, with more classical methods (logistic regression) performing as well as newer methods (ExtraTrees, Gaussian Naïve Bayes). In a follow-up study, the transferability of the sensitivities and specificities obtained to practical data from dairy farms should be investigated and discussed.

Author Contributions: Investigation, C.P.; Project administration, U.M.; Supervision, C.R. and W.B.; Writing-original draft, C.P.; Writing-review & editing, U.M. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by funds of the German Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support program.

Acknowledgments: The authors gratefully acknowledge the assistance of the staff of the Frankenforst Research Station of the University of Bonn (Königswinter, Germany). We also thank the members of the Center of Integrated Dairy Research (CIDRe, University of Bonn, Germany).

Conflicts of Interest: The authors declare no conflict of interest.

Chapter 3 References

1. Barkema, H.; Von Keyserlingk, M.; Kastelic, J.; Lam, T.; Luby, C.; Roy, J.-P.; Leblanc, S.; Keefe, G.; Kelton, D. (**2015**): Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. *Journal of Dairy Science* 98 (11), 7426–7445.

2. Pintado, D. (**2006**): *Automated Mastitis Detection in Dairy Cows Using Different Statistical Methods.* PhD Thesis, Christian-Albrechts-Universität zu Kiel, Kiel, Germany.

3. Steeneveld, W.; Van Der Gaag, L.; Ouweltjes, W.; Mollenhorst, H.; Hogeveen, H. (**2010**): Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *Journal of Dairy Science* 93 (6), 2559–2568.

4. Jensen, D.; Hogeveen, H.; De Vries, A. (**2016**): Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis. *Journal of Dairy Science* 99 (9), 7344–7361.

5. Stangaferro, M.; Wijma, R.; Caixeta, L.; Al Abri, M.; Giordano, J. (**2016**): Use of rumination and activity monitoring for the identification of dairy cows with health disorders: Part II. Mastitis. *Journal of Dairy Science* 99 (9), 7411–7421.

6. Alsaaod, M.; Römer, C.; Kleinmanns, J.; Hendriksen, K.; Rose-Meierhöfer, S.; Plümer, L.; Buscher, W. (**2012**): Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. *Applied Animal Behaviour Sci*ence 142 (3), 134–141.

7. Miekley, B.; Traulsen, I.; Krieter, J. (**2013**): Principal component analysis for the early detection of mastitis and lameness in dairy cows. *Journal of Dairy Research* 80 (3), 335–343.

8. Van Hertem, T.; Maltz, E.; Antler, A.; Romanini, C.; Viazzi, S.; Bähr, C.; Schlageter-Tello, A.; Lokhorst, C.; Berckmans, D.; Halachmi, I. (**2013):** Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *Journal of Dairy Science* 96 (7), 4286–4298.

9. Van Nuffel, A.; Zwertvaegher, I.; Pluym, L.; Van Weyenberg, S.; Thorup, V.; Pastell, M.; Sonck, B.; Saeys, W. (**2015**): Lameness Detection in Dairy Cows: Part 1. How to Distinguish between Non-Lame and Lame Cows Based on Differences in Locomotion or Behavior. *Animals* 5 (3), 838–860.

10. Kamphuis, C.; Frank, E.; Burke, J.; Verkerk, G.; Jago, J. (**2013**): Applying additive logistic regression to data derived from sensors monitoring behavioral and physiological characteristics of dairy cows to detect lameness. *Journal of Dairy Science* 96 (11), 7043–7053.

11. Pastell, M.; Kujala, M. (**2007**): A Probabilistic Neural Network Model for Lameness Detection. *Journal of Dairy Science* 90 (5), 2283–2292.

12. Lehmann, C.; Koenig, T.; Jelic, V.; Prichep, L.; John, R.; Wahlund, L.-O.; Dodge, Y.; Dierks, T. (**2007**): Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *Journal of Neuroscience Methods* 161 (2), 342–350.

13. Nechanitzky, K.; Starke, A.; Vidondo, B.; Müller, H.; Reckardt, M.; Friedli, K.; Steiner, A. (**2016**): Analysis of behavioral changes in dairy cows associated with claw horn lesions. *Journal of Dairy Science* 99 (4), 2904–2914.

14. Mollenhorst, H.; Van Der Tol, P.; Hogeveen, H. (**2010**): Somatic cell count assessment at the quarter or cow milking level. *Journal of Dairy Science* 93 (7), 3358–3364.

15. Dominiak, K.; Kristensen, A. (**2017**): Prioritizing alarms from sensor-based detection models in livestock production—A review on model performance and alarm reducing methods. *Computers and Electronics in Agriculture* 133, 46–67.

16. Garcia, E.; Klaas, I.; Amigó, J.; Bro, R.; Enevoldsen, C. (**2014**): Lameness detection challenges in automated milking systems addressed with partial least squares discriminant analysis. *Journal of Dairy Science* 97 (12), 7476–7486.

17. Cavero, D.; Tölle, K.-H.; Rave, G.; Buxadé, C.; Krieter, J. (**2007**): Analysing serial data for mastitis detection by means of local regression. *Livestock Science* 110 (1-2), 101–110.

18. Kuhn, M.; Johnson, K. (**2013**): *Applied Predictive Modeling*; Springer: New York, USA.

19. Schenk, J.; Kaiser, M.; Rigoll, G. (**2009**): Selecting Features in On-Line Handwritten Whiteboard Note Recognition: SFS or SFFS? In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, Barcelona, Spain, 26–29 July, 1251–1254.

20. Boughorbel, S.; Jarray, F.; El-Anbari, M. (**2017**): Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS one* 12 (6), e0177678.

21. Lemaître, G.; Nogueira, F.; Aridas, C. (**2017**): Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* 18 (1), 559–563.

22. Chawla, N. (**2003**): C4.5 and imbalanced data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In: *Proceedings of the ICML*; CIBC: Toronto, ON, Canada, 66–73.

23. Pedregosa, F.; Varoquaux, G.; Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (**2011**): Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

24. Cortes, C.; Vapnik, V. (**1995**): Support-vector networks. *Machine Learning* 20 (3), 273–297.

25. Cover, T.; Hart, P. (**1967**): Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13 (1), 21–27.

26. Domingos, P.; Pazzani, M. (**1997**): On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* 29 (2-3), 103–130.

27. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. (**2017**): *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA.

28. Kelleher, J.; MacNamee, B.; D’Arcy, A. (**2015**): *Fundamentals of Machine Learning for Predictive Data Analytics*; MIT Press: Cambridge, MA, USA.

29. Breiman, L. (**2001**): Random forests. *Machine Learning* 45 (1), 5–32.

30. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. (**2007**): Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8 (1), 25.

31. Genuer, R.; Poggi, J.-M.; Malot, C. (**2010**): Variable selection using random forests. *Pattern Recognition Letters* 31 (14), 2225–2236.

32. Kamphuis, C.; Sherlock, R.; Jago, J.; Mein, G.; Hogeveen, H. (**2008**): Automatic Detection of Clinical Mastitis Is Improved by In-Line Monitoring of Somatic Cell Count. *Journal of Dairy Science* 91 (12), 4560–4570.

33. Fernando, R.; Rindsig, R.; Spahr, S. (**1982**): Electrical Conductivity of Milk for Detection of Mastitis. *Journal of Dairy Science* 65 (4), 659–664.

34. Oltenacu, P.; Ekesbo, I. (**1994**): Epidemiological study of clinical mastitis in dairy cattle. *Veterinary Research* 25 (2-3), 208–212.

35. De Mol, R.; Andre, G.; Bleumer, E.; Van Der Werf, J.; De Haas, Y.; Van Reenen, C. (**2013**): Applicability of day-to-day variation in behavior for the automated detection of lameness in dairy cows. *Journal of Dairy Science* 96 (6), 3703–3712.

36. Flower, F.; Weary, D. (**2006**): Effect of Hoof Pathologies on Subjective Assessments of Dairy Cow Gait. *Journal of Dairy Science* 89 (1), 139–146.

37. Pastell, M.; Hänninen, L.; De Passillé, A.; Rushen, J. (**2010**): Measures of weight distribution of dairy cows to detect lameness and the presence of hoof lesions. *Journal of Dairy Science* 93 (3), 954–960.

38. Borchers, M.; Chang, Y.; Tsai, I.; Wadsworth, B.; Bewley, J. (**2016**): A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors. *Journal of Dairy Science* 99 (9), 7458–7466.

39. Wolfger, B.; Timsit, E.; Pajor, E.; Cook, N.; Barkema, H.W.; Orsel, K. (**2015**): Technical note: Accuracy of an ear tag-attached accelerometer to monitor rumination and feeding behavior in feedlot cattle1. *Journal of Animal Science* 93 (6), 3164–3168.

40. Schindhelm, K.; Haidn, B.; Trembalay, M.; Döpfer, D. (**2017**): Automatisch erfasste Leistungs- und Verhaltensparameter als Risikofaktoren in einem Vorhersagemodell für Lahmheit bei Milchkühen der Rasse Fleckvieh. In: *Proceedings of the 13. Tagung: Bau, Technik und Umwelt*, Stuttgart-Hohenheim, Germany, 18 September 2017, 228–233.

41. Brown, I.; Mues, C. (**2012**): An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications* 39 (3), 3446–3453.

42. Hogeveen, H.; Kamphuis, C.; Steeneveld, W.; Mollenhorst, H. (**2010**): Sensors and Clinical Mastitis—The Quest for the Perfect Alert. *Sensors* 10 (9), 7991–8009.

43. González, L.; Tolkamp, B.; Coffey, M.; Ferret, A.; Kyriazakis, I. (**2008**): Changes in Feeding Behavior as Possible Indicators for the Automatic Monitoring of Health Disorders in Dairy Cows. *Journal of Dairy Science* 91 (3), 1017–1028.

44. Lukas, J.; Reneau, J.; Wallace, R.; Hawkins, D.; Munoz-Zanzi, C. (**2009**): A novel method of analyzing daily milk production and electrical conductivity to predict disease onset. *Journal of Dairy Science* 92 (12), 5964–5976.

45. Cavero, D.; Tölle, K.-H.; Henze, C.; Buxadé, C.; Krieter, J. (**2008**): Mastitis detection in dairy cows by application of neural networks. *Livestock Science* 114 (2-3), 280–286.

46. Rutten, C.; Velthuis, A.; Steeneveld, W.; Hogeveen, H. (**2013**): Invited review. *Journal of Dairy Science* 96 (4), 1928–1952.

47. Mansbridge, N.; Mitsch, J.; Bollard, N.; Ellis, K.; Miguel-Pacheco, G.; Dottorini, T.; Kaler, J. (**2018**): Feature Selection and Comparison of Machine Learning Algorithms in Classification of Grazing and Rumination Behaviour in Sheep. *Sensors* 18 (10), 3532.

48. Shahinfar, S.; Page, D.; Guenther, J.; Cabrera, V.; Fricke, P.; Weigel, K. (**2014**): Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *Journal of Dairy Science* 97 (2), 731–742.

49. Huzzey, J.; Veira, D.; Weary, D.; Von Keyserlingk, M. (**2007**): Prepartum Behavior and Dry Matter Intake Identify Dairy Cows at Risk for Metritis. *Journal of Dairy Science* 90 (7), 3220–3233.

50. Zehner, N.; Niederhauser, J.; Schick, M.; Umstätter, C. (**2019**): Development and validation of a predictive model for calving time based on sensor measurements of ingestive behavior in dairy cows. *Computers and Electronics in Agriculture* 161, 62–71.

4 The Importance of Low Daily Risk for the Prediction of Treatment Events of Individual Dairy Cows with Sensor Systems

Christian Post [1], Christian Rietz [2], Wolfgang Büscher [3] and Ute Müller [1]

1 Institute of Animal Science, Physiology Unit, University of Bonn, 53115 Bonn, Germany; ute-mueller@uni-bonn.de

2 Department of Educational Science, Faculty of Educational and Social Sciences, University of Education Heidelberg, 69120 Heidelberg, Germany; christian.rietz@ph-heidelberg.de

3 Institute for Agricultural Engineering, Livestock Technology Section, University of Bonn, 53115 Bonn, Germany; buescher@uni-bonn.de

Received: 30 November 2020 / Revised: 29 January 2021 / Accepted: 11 February 2021

Published in: Sensors Special Issue "Sensors for Animal Health Monitoring and Precision Livestock Farming", *Sensors* 2021, 21 (4), 1389; https://doi.org/10.3390/s21041389

Abstract

The prediction of health disorders is the goal of many sensor systems in dairy farming. Although mastitis and lameness are the most common health disorders in dairy cows, these diseases or treatments are a rare event related to a single day and cow. A number of studies already developed and evaluated models for classifying cows in need of treatment for mastitis and lameness with machine learning methods, but few have illustrated the effects of the positive predictive value (PPV) on practical application. The objective of this study was to investigate the importance of low-frequency treatments of mastitis or lameness for the applicability of these classification models in practice. Data from three German dairy farms contained animal individual sensor data (milkings, activity, feed intake) and were classified using machine learning models developed in a previous study. Subsequently, different risk criteria (previous treatments, information from milk recording, early lactation) were designed to isolate high-risk groups. Restricting selection to cows with previous mastitis or hoof treatment achieved the highest increase in PPV from 0.07 to 0.20 and 0.15, respectively. However, the known low daily risk of a treatment per cow remains the critical factor that prevents the reduction of daily false-positive alarms to a satisfactory level.

Sensor systems should be seen as additional decision-support aid to the farmers’ expert knowledge.

Keywords: mastitis, lameness, machine learning, animal welfare

1. Introduction

For dairy farms, various sensor systems offer predictions of health data or diseases. These systems offer support for the farmers in their task to identify the animals in need of veterinary control or treatment. For legal, moral, and ethical reasons, these checks must be carried out daily. In relation to all diseases in dairy cows, most treatments are done for mastitis or lameness [1]. A number of studies have examined the classification of cows for mastitis or hoof treatments using single sensors, or combinations of sensors. In many of these studies, the focus is on the first stage of biological validation: the identification of the subjects known to be affected by a given health issue or not, i.e., the Receiver Operator Characteristic (ROC) curve and the indicators sensitivity and specificity [2,3]. Within the framework of these studies, models are often developed using data sets consisting of defined test populations that show a higher proportion of animals with the condition to be predicted (e.g., cows requiring treatment and limited time windows) [4,5].

Fewer studies complete the second stage of validating the developed algorithms on data sets that correspond to a practical situation. In the second stage, the indicators for assessing the predictive quality (termed “diagnostic value” in medical test procedures) are determined and used for evaluation. For most procedures, this involves the use of the positive predictive value (PPV), which describes the proportion of false-positives in all positive tests, with the negative predictive value (NPV) used less frequently. The direct relationship between the PPV and the frequency of occurrence of the event to be predicted or classified is seldom highlighted in the studies; the lower the risk of an event being predicted, the lower the PPV or the higher the false-positive rate. In the studies investigating the possibility of predicting the need for treatment, it is clear that the probability of occurrence of treatments per day per animal is low. Miekley et al. [6] explicitly reported the value related to the practical data was approximately a 0.5% risk for mastitis treatment per animal per day and approximately 7% for lameness treatments, and Steeneveld et al. [7] found a frequency of 0.04% for mastitis treatments in data from automatic milking systems (AMS). The low risk per animal and day for this predictable treatment event leads directly to a low PPV of an alarm list, i.e., a high number of false-positive classifications. In the study of Miekley et al. [6], the values of the

PPV for mastitis and lameness treatments were approximately 0.01 and 0.10, and in the study by Steeneveld et al. [7], the resulting PPV was 0.01.

It can be assumed that, by applying the developed algorithms to subgroups in which the event occurs more frequently (also known as risk groups [8]), the PPV is higher and the false-positive rate lower. In human medicine, testing procedures are therefore carried out in these groups or in people with corresponding symptoms in order to increase the probability that the associated event could occur in this group of people. For example, screening tests for chlamydia infections in humans with very high values for sensitivity and specificity of 0.98 and 0.97 ("very high" compared to possible predictive models from livestock farming) in a group of people with a prevalence of 3% could still only achieve a PPV of 0.50, so that half of the tested persons received an incorrect initial diagnosis [9]. For this reason, screening tests, e.g., for chlamydia or HIV, are primarily performed in groups of people with a higher prevalence of the disease under investigation [8,10].

In the context of studies on predictive models of disease indicators in dairy farming, the application of algorithms in risk groups and its effect on the PPV has not yet been investigated. The following risk groups would be conceivable within a herd: cows with a prior treatment in the previous or current lactation [11,12,13], cows with an increased cell count during milk performance testing [14,15], or cows during certain periods of lactation. Koeck et al. [1] specified a higher incidence of clinical mastitis (35% of all cases) in the first 30 days in milk (DIM). In the same study, 22% of all lameness cases occurred in the first 30 DIM. For both treatments, the remaining cases were evenly distributed over the rest of the lactation.

However, the application of predictive models to a risk group does not automatically lead to a higher PPV. Zehner et al. [16] have developed predictive models for calving, i.e., an event in a defined risk period (from seven days before the expected calving date), using data from rumination sensors. Moreover, in these few days before the expected calving, the exact time of calving within the hourly evaluated data (168 h or 24 h before calving) represents a rare event per time unit (0.6% or 4%), so that a high number of false alarms resulted in positive prediction values of only 0.01–0.03 for 168 h or 0.06–0.18 for 24 h [16]. The authors concluded that, despite satisfactory values for sensitivity and specificity, their model was not suitable for practical use because of the low PPV.

The objective of this study was to investigate the importance of the frequency of occurrence of treatments in risk groups and the resulting variation in the PPV for the applicability of classification models in practice. For this, machine-learning models for the classification of mastitis and lameness treatments (i.e., a form of health data available in databases) developed in a previous study [17] were validated using data from other dairy farms.

2. Materials and Methods

2.1. Data Source

Raw data were collected from three German Holstein dairy farms. The criteria for the selection of these farms were to be able to automatically record milking data (milk yield, milk flow, and conductivity), activity data (pedometer impulse count), as well as feed intake via weighing troughs. This data was transferred in a standardized form to a shared database system, where it was processed and output as CSV files. Per farm, data were used in periods of 3–3.4 years, with average herd sizes of 65–121 cows (see Table 4-1).

Table 4-1. Overview of the three farms and the amount of data used.

Farm	Time Period	Raw Data Size[1]	Mean Herd Size	Mean Daily Milk Yield (kg)	Mean Lactation Number[2]
A	June 1, 2015 – October 20, 2018	80,307	65	31.3	2.4 ± 1.7
B	January 1, .2014 – May 31, 2017	203,421	163	36.6	2.4 ± 1.4
C	January 1, 2017 – December 31, 2019	133,270	121	35.4	2.3 ± 1.6

[1] records per cow and day, [2] ± standard deviation.

The data collection and processing were carried out analogous to the study in Post et al. [17]. The following types of sensor data were available per cow and day: Milking data (milk yield as the sum of two milkings, milking time, milk flow, conductivity for morning and evening milkings, respectively), feed data (total feed intake, number, average duration of trough visits), pedometer activity (sum of impulses from a pedometer at a 2 h resolution), body weight (kg, averaged for two measurements per day), and animal information (lactation number, days in milk (DIM). Cows with clinical signs were identified during the daily routine and treated by a veterinarian. All treatment data were recorded by farm staff and entered into the database. The milk performance data of the milk recording (milk yield, fat, protein, and lactose content, as well as somatic cell count in 1,000 cells/mL) were recorded monthly for farm A and weekly for farms B and C.

2.2. Data Preprocessing

Before further processing, the aggregation, plausibility checks, and adjustment of the data were carried out according to the same scheme as in the previous study [17]. Lactations with completely missing values in at least one feature, as well as lactations with less than 28 days in the data, were removed. For farm C, records of treatments were missing for a period of six months; data from this period were discarded. Furthermore, on farm C, feed weighing troughs were not installed in all areas of the barn and therefore the records of feed intake of a cow were not continuously available throughout the lactation. Hence only periods with existing data were considered. As in Post et al. [17], additional features were calculated for each variable (except for lactation number, DIM, and monthly milk recording data), which reflected their change over time: Rolling mean (i.e., the moving average) of the last seven days, rolling mean of the previous week, change of the current value to the rolling mean, values of the three previous days, the slope of a linear regression through the last seven days. This step resulted in a total of 182 available features.

Data on treatment records contained information not related to the disease of interest, such as antibiotic treatments for dry-off (udder), hoof care without a diagnosis (claws), estrus synchronization, and silent heat (fertility). These treatment events were ignored. For cows with at least one treatment in a given lactation, the first treatment day was identified for each treatment. A period of 14 days after one treatment, or after the last follow-up treatment if treatment administration occurred over multiple days, was removed from the data, as it was not clear from the data when a cow could be considered “healthy” again, similar to the 14-day exclusion period for follow-up treatments used in the studies by Kamphuis et al. [2] and Jensen et al. [3] (see Figure 4-1).

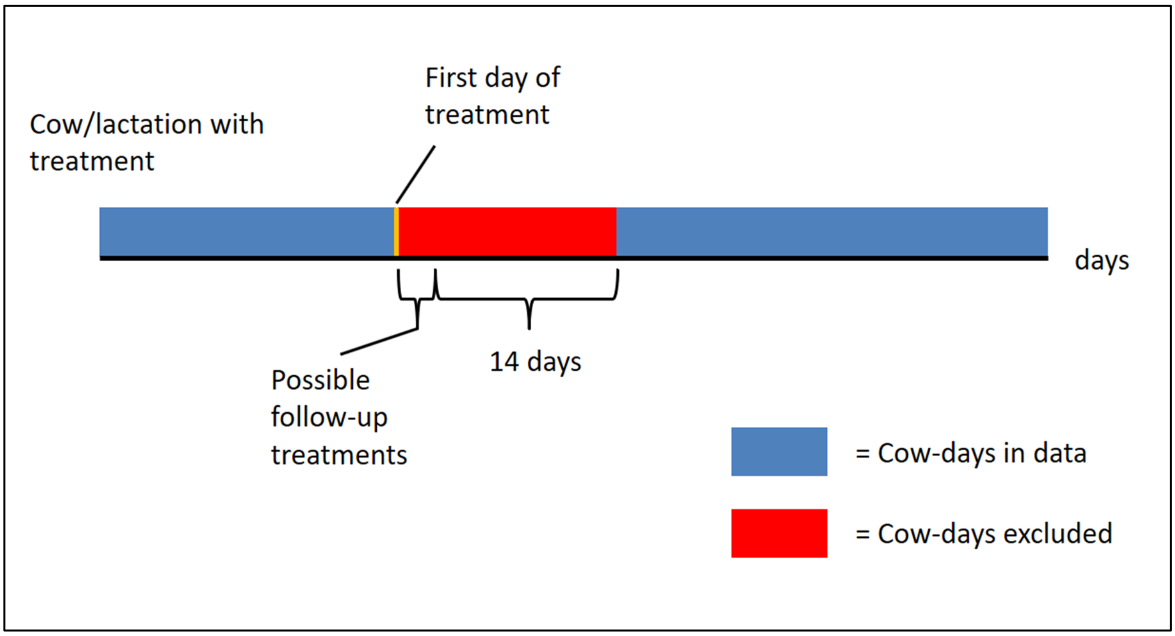

Figure 4-1. Schematic representation of the extraction of data for cows/lactations with and without at least one treatment.

Afterward, the day of treatment was moved forward by one day, firstly to exclude the influence of the treatment on the sensor data, and secondly, because, this way, a classification about the cow's condition tomorrow was already simulated at the end of the previous day

2.3 Training and Testing of the Classification Models

Based on the findings of the previous study [17] four statistical classification models were selected: Random Forest, Logistic Regression, Gaussian Naive Bayes, and ExtraTrees Classifier. These models were part of the Python package Scitkit-Learn [18] and are described in detail in Post et al. [17]. The following steps were performed separately for each of the two categories of mastitis and lameness treatments as the target variable. First, for the training data, all blocks (treatment + previous days) where a treatment other than the target variable was recorded were removed from the data, so that no days were falsely marked as free of treatment. As described in detail in Post et al. [17], the data was normalized (z-score normalization) and feature selection was performed using Sequential Forward Selection (SFS) to identify the most important 20 features per treatment category and facility based on Random Forest mean decrease in impurity.

To evaluate the classification models based on data from the same farm, a 5-fold cross-validation was performed for each farm separately. For this purpose, the cows and lactations were randomly divided into five data sets. For each of these data sets, four were used as training data and the remaining one was used as validation data. The training data set was further reduced to 28 days per lactation, as described in Post et al. [17]. For lactations with a treatment, this was the period before the treatment. For lactations without a treatment, a random period was chosen instead. The models used in this study provided an estimate of the probability that a cow was in need of treatment on a given day. Based on the results from the cross-validation, thresholds for this probability estimate were determined where sensitivity was at least 0.7. These thresholds were stored for later application to the test data. The models were then re-trained on the entire data of one farm and then applied to the data of the two remaining farms. This was done for all cows and days, as well as for only those days on which a cow was classified as a “risk animal” (according to the groups defined in Section 2.4). Alongside this approach, the models were also tested using the combined data from two farms as training data, and then using the remaining farm as testing data, to identify possible differences between these procedures.

2.4. Definition of Risk Groups

To increase the frequency of days with treatments and thus improve the classification for mastitis and lameness treatments, the cow-days (i.e., the data from one cow on a particular day) were filtered based on various criteria, further referred to as risk groups (RG). Three different RG limited the testing data to only cow-days with previous treatments:

- RGtreat-SC/PL: Cows with at least one treatment of the same category in the previous lactation
- RGtreat-SC/SL: Cows with at least one previous treatment of the same category in the same lactation
- RGtreat-OC/SL: Cows with at least one previous treatment of another category in the same lactation

The classification into a risk group applied to a cow from the occurrence of the respective condition until the end of the current lactation. This scheme is shown in Figure 4-2. In RGtreat-SC/PL the cow belongs to this group from the day of calving, (DIM 1 in the data). Membership in RGtreat-SC/SL starts as soon as 14 days after the first treatment, and the days from the treatment until that point have been removed from the data.

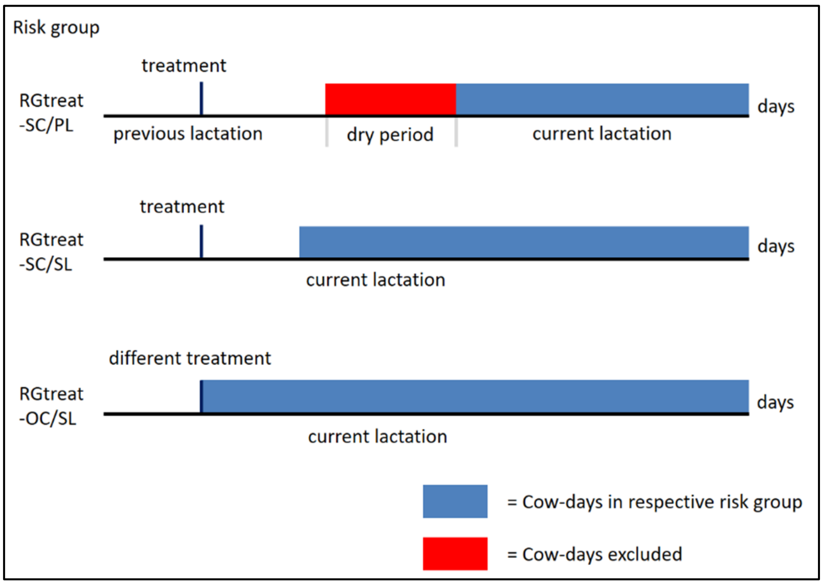

Figure 4-2. Schematic representation of the assignment of the cow-days to one of the three risk groups: RGtreat-SC/PL (at least one treatment of the same category in the previous lactation), RGtreat-SC/SL (at least one previous treatment of the same category in the same lactation) and RGtreat-OC/SL (at least one previous treatment of another category in the same lactation).

RG-SCC contained only cow-days where a cow showed an increased somatic cell count (SCC) during the last milk recording (MR). This includes four criteria, which were derived from the udder health indicators of the German Association for Performance and Quality Testing (“Deutscher Verband für Leistungs- und Qualitätsprüfungen e.V.”, DLQ) [19]:

- Cows with a new infection of the udder (defined by SCC > 100,000 with previous SCC of ≤ 100,000)
- Cows with an infection in the first MR after calving (SCC > 100,000), if last SCC in the previous lactation ≤ 100,000
- Heifers with an infection in the first MR after calving (SCC > 100,000)
- Cows with chronic mastitis (three consecutive MR with SCC > 700,000)

The SCC describes the proportion of somatic cells in 1 mL milk and provides information about the udder health status of a cow. The critical value for udder health is a SCC of > 100,000 / mL, values above this value indicate an infection of the udder [21]. A cow was assigned to RG 2 from the day of MR, if at least one of the above mentioned criteria was detected. This classification was valid either until a SCC of ≤ 100,000/mL was detected in a subsequent MR or, if this did not occur, until the end of lactation. This is illustrated in Figure 4-3.

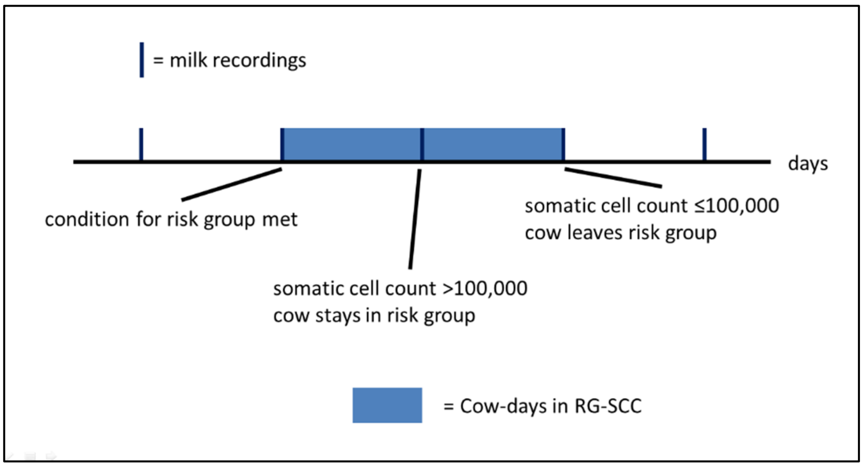

Figure 4-3. Schematic representation of the assignment of cow-days to the risk group RG-SCC (increased somatic cell count after monthly/weekly milk recording).

Lastly, for the formation of the risk group RGtime-100 only cow-days with a value for the DIM of ≤100 were included in the test data. In addition, a test data set was also formed using the same procedure, but with the criterion "DIM ≤60", in order to test the effect of even greater limitation.

2.5 Evaluation

To assess the classification value of trained models on test data sets in which the event to be classified is available as a reference, the frequency of occurrence of the event to be classified in this test data set is determined. This is done by dividing the number of days with treatments by the total number of all days and then displaying it as a percentage. Secondly, the PPV is calculated to interpret the prediction quality. The level of the PPV depends on the risk that the event, in this case, the treatment, will occur.

For each cow and day, the models yielded a probability of belonging to class label “1”, i.e., in need of treatment. These probabilities in combination with the vector of true labels were used to obtain the area under curve (AUC) from the Receiver Operator Characteristic (ROC), as described in detail in Post et al. [17]. In addition to the actual day before a treatment, another two days before treatment were also marked as “treated” in the data, i.e., an alarm was considered true positive within three days before treatment (see Figure 4-4). Subsequently, the probabilities were compared to the threshold obtained in the model validation to create a vector of binary classifications. This vector was compared with the vector of true labels. Any day with a treatment event could either have an associated alert or not. If an alert was present on a treatment day, it was classified as a true positive (TP), if not, a false-negative (FN); conversely, if an alert was present, that day was a false-positive (FP), and if an alert was not present, the day was a true negative (TN). This is demonstrated in Figure 4-4.

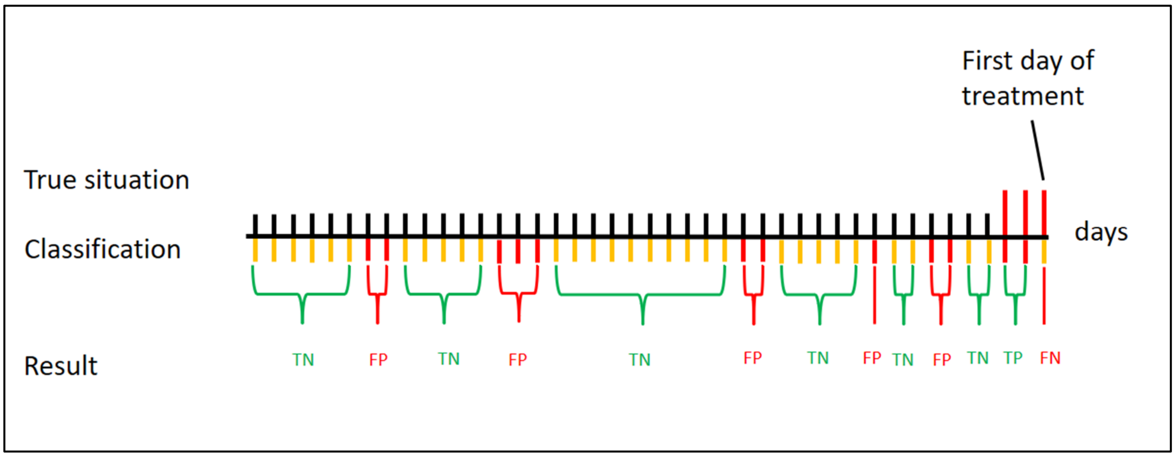

Figure 4-4. Example classification of days for a single cow with a treatment (last 3 days labeled positive). Yellow bars mean no alert, red bars mean alerts. Green brackets indicate true negative and true positive days, red brackets indicate false-positive and false-negative days.

All values of TP, FP, FN, and TP were summed up for all cows and used for the following calculations:

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = TN/(TN + FN)$$

Additionally, the block sensitivity was calculated. Here, a cow-day was considered TP if a positive classification was given on at least one of the three days prior to treatment, and FN if none of these three days were detected. This value is always expected to be higher than the sensitivity.

In addition, the positive predictive value (PPV, also referred to as "precision") was calculated and describes the percentage of correctly classified cows of all cows classified as "treated".

$$PPV = TP/(TP + FP)$$

The results were averaged for all four classification models per treatment category (mastitis or lameness) and per testing farm or per risk group. All results were presented as mean ± 95% confidence interval of the mean, which was calculated with the Python Statsmodels package [21]. Differences between farms and risk groups were performed using Welch's test due to the violated homogeneity of variance. For multiple comparisons in the post-hoc test, Dunnett-T3 was used. These tests were implemented within SPSS version 26.0 (IBM Corp, Armonk, NY, USA) with significant differences at $p < 0.05$.

3. Results

3.1. Validation Results

After the preprocessing steps described in Section 2.2, a total of 42,803 cow-days and 48,041 cow-days from 794 individual cows for the mastitis and lameness treatments classification, respectively, remained in the data (see Appendix A, Table A3). The 5-fold cross-validation of the classification models for farms A, B, and C on the data from the same farm led to the results shown in Table 4-2. The Random Forest feature importance obtained during this step is shown in Appendix A, Table A1 and Table A2.

The mean AUC was 0.73 for mastitis treatments and was lower for lameness treatments at 0.67. This was consistent with the results from Post et al. [17]. The sensitivity of 0.71 in both treatment categories resulted from the fixed probability threshold of the individual classification models. As expected, the block sensitivities for mastitis were higher at 0.92 for mastitis and 0.85 for lameness.

Table 4-2. Mean area under curve (AUC), sensitivity, block sensitivity, and specificity (± 95%-CI) for 5-fold-cross validation data. Training and testing were performed on the same farm.

Treatment	Farm	AUC[1]	Sen.[2]	Block Sen.[3]	Spe.[4]
Mastitis	A	0.70 ± 0.08^{b}	0.72 ± <0.01^{a}	0.93 ± 0.02^{ab}	0.59 ± 0.13^{b}
	B	0.80 ± 0.02^{a}	0.71 ± <0.01^{b}	0.89 ± 0.02^{b}	0.74 ± 0.04^{a}
	C	0.70 ± 0.04^{b}	0.72 ± <0.01^{a}	0.94 ± 0.03^{a}	0.59 ± 0.08^{b}
	Mean	0.73 ± 0.04	0.71 ± <0.01	0.92 ± 0.02	0.64 ± 0.06
Lameness	A	0.72 ± 0.02^{a}	0.71 ± <0.01^{c}	0.85 ± 0.03	0.59 ± 0.03^{a}
	B	0.66 ± 0.02^{b}	0.70 ± <0.01^{b}	0.83 ± 0.03	0.51 ± 0.04^{ab}
	C	0.63 ± 0.07^{b}	0.72 ± <0.01^{a}	0.86 ± 0.03	0.48 ± 0.09^{b}
	Mean	0.67 ± 0.03	0.71 ± <0.01	0.85 ± 0.01	0.53 ± 0.04

[1] area under ROC-curve; [2] sensitivity; [3] block sensitivity; [4] specificity; ^{a,b,c} superscript letters indicate significant differences at p ≤ 0.05 between farms within treatment categories.

Table 4-3 shows the results of the classification (AUC, sensitivity, block sensitivity, and specificity for mastitis and lameness treatments) of all combinations of training and testing farms. The mean AUC for mastitis treatments was higher at 0.72 than for lameness treatments with 0.61. The sensitivities, the level of which was set at 0.7 during validation by fixing the probability threshold values, could not reach this sensitivity value for all testing data sets, but the block sensitivity was on average 0.86 for mastitis treatments and 0.83 for hoof treatments and thus achieved the minimum sensitivity requirement. The mean AUC of 0.72 for mastitis treatments was similar compared to the results from the validation (0.73), whereas for lameness treatments the mean AUC was lower (0.61 compared to 0.67).

The AUCs obtained when using combined training data of two farms, as described at the end of Section 2.3, did not differ significantly from the results of the other approach (mean AUC over all test farms of 0.73 for udder treatments and 0.63 for hoof treatments), so they were not presented separately here.

3.2. Positive predictive values depending on the risk of occurrence of the treatments

The mean frequency of cow-days with mastitis and lameness treatments in the data sets was 3.6% and 5.6%, per cow per day, respectively, for all three farms combined (see Table 4-4). This relationship applied to the present approach resulted in the identification of animals at a higher risk in the following step (for definitions see Section 2.4). Table 4-4 shows the comparison of the frequency of occurrence for mastitis and lameness treatments in the entire farm data set to the partial data sets of the respective risk groups and risk time.

Table 4-3. Mean area under curve (AUC), sensitivity, block sensitivity, and specificity (± 95%-CI) for the classification of mastitis and lameness treatments of all cows. Models were trained on one farm and then tested on the two other respective farms.

	Farm					
Treatment	**Training**	**Test**	**AUC**[1]	**Sen.**[2]	**Block Sen.**[3]	**Spe.**[4]
Mastitis	A	B	0.78 ± 0.03^{a}	0.60 ± 0.19^{ab}	0.82 ± 0.13^{ab}	0.80 ± 0.23^{a}
		C	0.70 ± 0.05^{abc}	0.47 ± 0.18^{b}	0.72 ± 0.17^{ab}	0.80 ± 0.23^{a}
	B	A	0.73 ± 0.01^{b}	0.62 ± 0.14^{b}	0.86 ± 0.12^{ab}	0.72 ± 0.11^{ab}
		C	0.69 ± 0.02^{c}	0.55 ± 0.07^{b}	0.80 ± 0.08^{b}	0.71 ± 0.11^{ab}
	C	A	0.71 ± 0.06^{abc}	0.91 ± 0.07^{a}	1.00 ± <0.01^{a}	0.27 ± 0.20^{bc}
		B	0.74 ± 0.02^{ab}	0.89 ± 0.13^{a}	0.97 ± 0.04^{a}	0.30 ± 0.26^{b}
	Mean		0.72 ± 0.02	0.67 ± 0.08	0.86 ± 0.05	0.60 ± 0.11
Lameness	A	B	0.62 ± 0.02^{bc}	0.81 ± 0.12^{a}	0.90 ± 0.08^{ab}	0.28 ± 0.16^{bc}
		C	0.60 ± 0.02^{c}	0.84 ± 0.11^{a}	0.89 ± 0.08^{ab}	0.26 ± 0.13^{bc}
	B	A	0.66 ± 0.02^{ab}	0.74 ± 0.05^{a}	0.87 ± 0.05^{a}	0.45 ± 0.08^{abc}
		C	0.56 ± 0.02^{d}	0.64 ± 0.02^{ab}	0.76 ± 0.03^{b}	0.43 ± 0.06^{abc}
	C	A	0.66 ± 0.02^{a}	0.66 ± 0.11^{ab}	0.83 ± 0.13^{ab}	0.56 ± 0.13^{ab}
		B	0.57 ± 0.04^{d}	0.53 ± 0.09^{b}	0.73 ± 0.13^{ab}	0.58 ± 0.15^{a}
	Mean		0.61 ± 0.02	0.71 ± 0.05	0.83 ± 0.04	0.42 ± 0.06

[1] area under ROC-curve; [2] sensitivity; [3] block sensitivity; [4] specificity; ^{a,b,c} superscript letters indicate significant differences at $p \leq 0.05$ between farms within treatment categories.

Table 4-4. Frequency (%) of cow-days with a mastitis treatment and lameness treatment in the testing data for all cows (no specific risk group).

Treatment	Farm	All Cows	RGtreat-SC/PL	RGtreat-SC/SL	RGtreat-OC/SL	RG-SCC	RGtime-100
Mastitis	A	2.5	5.4	12.0	3.0	5.4	2.9
	B	4.8	8.1	13.6	6.5	9.6	4.9
	C	3.8	8.3	14.0	5.1	5.1	5.0
	All	3.6	7.8	13.5	5.7	8.2	4.5
Lameness	A	4.3	7.0	12.9	6.0	6.8	3.1
	B	5.5	8.5	13.2	6.6	6.4	5.2
	C	6.8	9.9	13.3	7.9	7.4	6.1
	All	5.6	8.5	13.2	6.6	6.7	5.0

It is noticeable that the highest increase in the frequency of occurrence or risk for both mastitis and lameness treatments was caused by restricting our analysis to animals that had already undergone a treatment of the same category in the same lactation (RGtreat-SC/SL). Here, the average increase compared to all cows was +9.5 percentage points to 13.5% risk for another mastitis treatment and +7.8 percentage points to 13.2% risk for another lameness treatment. Limitation to cows that had the same category of treatment (mastitis or lameness treatment) in the previous lactation (RGtreat-SC/PL) increased the mean risk for another treatment by 3.6 percentage points and 2.9 percentage points, respectively.

Likewise, a limitation to MR risk factors based on SCC increased the risk for corresponding mastitis treatments on average by 4.6 percentage points (RG-SCC). The grouping RGtreat-OC/SL produced only small increases of 1.6 percentage points and lower. Limiting the risk period to the first 100 DIM (early lactation, RGtime-100) led to inconsistent results. For the mastitis treatments, the risk for corresponding treatment was comparable to that in the total data set (+0.9 percentage points to 4.5%), whereas the risk for lameness treatments was lower in the early lactation of the three experimental farms than over the whole lactation (−0.6 percentage points to 5%). When only the first 60 DIM were considered as the risk period, it did not lead to a change in the frequency of occurrence compared to the limitation of the DIM to ≤ 100 (4.8% frequency of mastitis and 5.2% of claw treatments). The model data for this risk period were, therefore, not listed additionally hereafter.

Table 4-5. Comparison of classification results for mastitis and lameness treatments (mean area under the curve (AUC), block sensitivity, and specificity ± 95%-CI), as well as the total number of cow-days in the data and the treatment frequency in % in the testing data for all cows, as well as the risk groups.

Risk Group	Cow-days Total	Frequency of Treatment Days (%)	AUC[1]	Block Sen.[2]	Spe.[3]
		Mastitis			
All cows	42,803	4.1	0.72 ± 0.02^{a}	0.86 ± 0.05	0.60 ± 0.11
RGtreat-SC/PL	6,633	7.8	0.69 ± 0.02^{a}	0.83 ± 0.06	0.57 ± 0.11
RGtreat-SC/SL	4,251	13.5	0.65 ± 0.02^{b}	0.81 ± 0.06	0.57 ± 0.11
RGtreat-OC/SL	17,802	5.7	0.69 ± 0.02^{ab}	0.87 ± 0.03	0.63 ± 0.06
RG-SCC	10,414	8.2	0.71 ± 0.02^{a}	0.86 ± 0.05	0.58 ± 0.11
RGtime-100	18,289	4.5	0.73 ± 0.02^{a}	0.88 ± 0.06	0.60 ± 0.11
		Lameness			
All cows	48,041	5.6	0.61 ± 0.02^{a}	0.83 ± 0.04	0.42 ± 0.06
RGtreat-SC/PL	9,587	8.5	0.59 ± 0.02^{a}	0.83 ± 0.04	0.43 ± 0.06
RGtreat-SC/SL	8,417	13.2	0.55 ± 0.02^{b}	0.79 ± 0.04	0.40 ± 0.06
RGtreat-OC/SL	18,137	6.6	0.59 ± 0.02^{a}	0.72 ± 0.05	0.55 ± 0.04
RGtime-100	20,044	5.0	0.58 ± 0.01^{ab}	0.80 ± 0.04	0.42 ± 0.06

[1] Area Under ROC-Curve; [2] Block Sensitivity; [3] Specificity; ^{a,b,c} superscript letters indicate significant differences at $p \leq 0.05$ between farm combinations within treatment categories. RGtreat-SC/PL: at least one treatment of the same category in the previous lactation, RGtreat-SC/SL: at least one previous treatment of the same category in the same lactation, RGtreat-OC/SL: at least one previous treatment of another category in the same lactation, RG-SCC: high SCC in previous milk recording, RGtime-100: DIM ≤ 100.

Table 4-5 shows the validity criteria AUC, block sensitivity, and specificity per RG in comparison to all cows, while the comparison of the PPV is shown in Figure 4-5. It was noticeable that the value of the PPVs reflected the respective frequency of occurrence, i.e., the known low risk for treatment. The highest PPVs were achieved on average in the RGtreat-SC/SL with previous treatments of the same category in the same lactation; for mastitis treatments on average a PPV of 0.20 and for lameness treatments of 0.15. All other PPVs were significantly lower. However, it was noticeable that for AUC both in the classification of mastitis treatments (0.65) and lameness treatments (0.55), significantly lower values were obtained only in RGtreat-SC/SL compared to the test data of all cows. The sensitivity, block sensitivity, and specificity did not differ between the risk groups within the treatment categories; therefore, the sensitivity is not shown in Table 4-5. In the following paragraphs, the respective risk groups whose AUCs and PPVs showed statistically significant differences are described individually.

3.2.1. Predictive Value of the Models for the Respective Treatment Risk Groups (RGtreat)

Due to the higher frequencies of occurrence in the narrowed data to the risk group of animals already treated for mastitis or lameness in the past, the PPVs increased from 0.07 to 0.13 for mastitis treatments and from 0.07 to 0.10 for lameness treatments. The AUC of the models for RGtreat-SC/PL were comparable to the results for all cows and were 0.69 for udder and 0.59 for lameness treatments.

When applying the trained models to RGtreat-SC/SL (cows with at least one previous treatment of the same category in the same lactation), the higher PPVs of 0.20 for mastitis treatments and 0.15 for lameness treatments, compared to all other risk groups, stood out due to the highest frequency of days with treatment compared to the other risk groups. The AUC for predicting mastitis treatments of RGtreat-SC/SL was 0.65 and for lameness treatments 0.55, significantly lower than the results for the respective treatments in the data of all cow-days. At the same time, RGtreat-SC/SL was the group with the lowest number of cow-days in the test data (4251 d for the mastitis treatments and 8417 d for the lameness treatments, see Table 4-5). The combinations of block-sensitivities and specificities were in similar ranges as in RGtreat-SC/PL., i.e., animals with corresponding treatments in past lactation.

For RGtreat-OC/SL, the trained models were applied to the risk group of cows that had at least one treatment of another category (i.e., not also a treatment for mastitis, lameness, or metabolic disorders) in the same lactation. For both mastitis and lameness treatments, the AUC of 0.69 and 0.59 did not differ significantly from the AUC of the classification for all cow-days. As already shown in Table 4-4, this restriction did not lead to a significant increase in the frequency of days with treatment (+0.8% for mastitis treatments and +1.1% for lameness treatments), so the PPVs were in the range of 0.09 for mastitis treatments and 0.09 for lameness treatments.

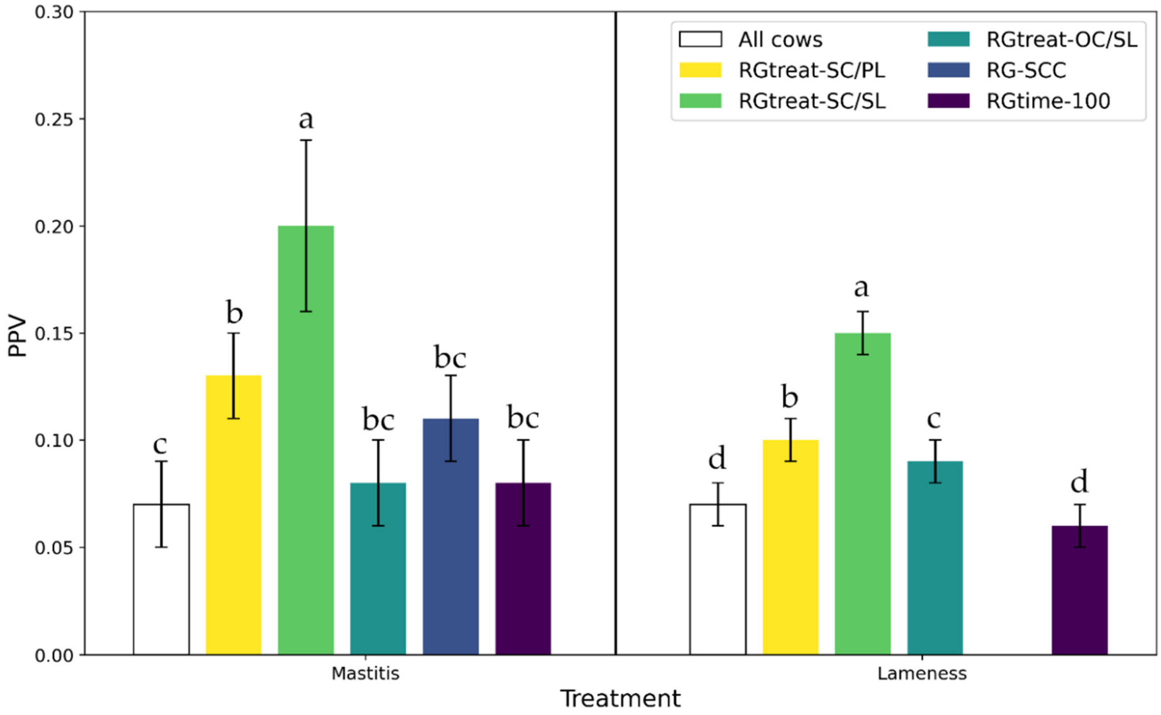

Figure 4-5. Mean positive predictive value (PPV) from classification results for mastitis and lameness treatments. Error bars indicate the 95%-CI, while letters a, b, c, d indicate significant differences at $p \leq 0.05$ between treatment categories. RGtreat-SC/PL: at least one treatment of the same category in the previous lactation, RGtreat-SC/SL: at least one previous treatment of the same category in the same lactation, RGtreat-OC/SL: at least one previous treatment of another category in the same lactation, RG-SCC: high SCC in previous milk recording, RGtime-100: DIM ≤ 100.

3.2.2. Predictive Value of the Models for the Risk Group According to the Information on SCC from Milk Recording (RG-SCC)

The SCC categories explained in Section 2.4 allowed the classification of the animals into the risk group RG-SCC. The mean values for AUC (0.71), block sensitivity, and specificity were comparable to those of the validation results. However, the mean PPV of 0.11 did not differ significantly from the PPV for all cows.

3.2.3. Predictive Value of Models in Early Lactation as Risk Time Period (RGtime)

However, as shown in Table 4-4, the frequency of mastitis treatments in the risk period of early lactation (≤ 100 DIM) in all three farms was comparable to the treatment frequency over the entire lactations. In terms of treatments for lameness, the risk of treatment was lower in the early lactation period than over the entire lactation period. Accordingly, the AUC values (0.73 for mastitis treatments and 0.58 for lameness treatments) and the PPV (0.08 for mastitis treatments and 0.06 for lameness treatments) showed no significant changes compared to the evaluation without time limitation. Reducing the risk period to only 60 DIM resulted in AUC values of 0.74 ± 0.03 for mastitis treatments and 0.60 ± 0.02 for lameness treatments, and a PPV of 0.07 ± 0.01 for mastitis treatments and 0.08 ± 0.01 for lameness treatments. Again, these values did not differ significantly.

4. Discussion

4.1. Validation

The aim of the validation was to test the applicability of the models developed by Post et al. [17] for the classification of mastitis and hoof treatments on data from two other experimental farms to simulate the use of trained models on other unknown datasets of additional farms.

The AUCs of 0.73 for mastitis treatments and 0.67 for lameness treatments obtained by cross-validation are comparable to values obtained in studies that have performed classifications with comparable sensors. For the detection of mastitis, a study with AMS data and additional cow information (parity, DIM, season, SCC history, and clinical mastitis history) over two years achieved a range of AUC values between 0.62–0.78 [7]. For the detection of lameness treatments, AUC values between 0.66–0.75 were achieved in a study by Kamphuis et al. [2] using data from a total of 4904 cows (from five farms) and the features live weight, activity and milk yield, and milking duration. In a further study on the detection of lameness treatments, an AUC of only 0.60 was obtained from a data set of 315 cows for a comparable time window of three days prior to each treatment [22]. It should be noted that an AUC of 0.70 and above describes a “strong model”, while a value of 0.60 and below describes only a “weak model” [23].

On average, higher AUC values are obtained for the classification of mastitis diseases or treatments than for the prediction of lameness treatments. This can be explained by the high importance of the feature “SCC of the last milk recording (MR)”. As already discussed [19], the feature “SCC” is directly related to the event “mastitis”. This could be confirmed on all three test farms. When comparing the individual models, it was noticeable that farms B and C showed a higher relative importance of the cell count with 0.22 compared to farm A with 0.13. This can be explained by the higher frequency of MR (B and C: weekly vs. A: monthly). The periods between the measured cell count and the mastitis treatments were shorter due to the weekly MR and the correlation was, therefore, more direct, whereas the monthly MR increased the probability of an intermediate healing or new infection, which is then not found in the data. For the same reason, the AUC values of the mastitis models were not significantly lower when the trained models were tested from one farm to the other farms (from 0.73 on the training farm to an average of 0.72 on the test farms). As shown in Table

4-3, the lower frequency of MR on farm A had no negative influence on the AUC when used for training the models, compared to farms B and C. In comparison, the trained models for classifying lameness treatments with AUC values averaging 0.67 are considered “weak models” and, therefore, their practical usefulness is questionable. This is further reinforced by the application of the trained models to other farms or their data, as the AUC values were significantly reduced to an average of 0.61. Since no available features in the detection of lameness treatments were directly or specifically related to the event, the operational differences are much more relevant for the model quality.

4.2. Positive Predictive Values Depending on the Probability of Occurrence of the Treatments

A common feature of the mentioned studies [6,7,22] is the high portion of false-positive alarms (also referred to as error rate). This value was 0.99 for mastitis and 0.89 for lameness treatments in the results from Miekley et al. [22]. The study of Steeneveld et al. [7] included 52 true positive and 3636 false-positive alarms, which led to an error rate of approximately 0.99. The PPV in our own study of 0.07, which corresponds to a 0.93 error rate, is due to the low frequency of days with treatment in the test data, which was 3.5% and 5.4% for mastitis and lameness treatments, respectively. In other studies, the ratio of treated to non-treated cows was artificially increased, i.e., the data were sampled, e.g., by pairing cows [5], by excluding unclear cases of lameness [24,25], or by considering shorter periods before treatment [2,4]. From the perspective of the developers of these models, these measures are justified because sensitivity and specificity are hardly influenced by the frequency of occurrence of the target trait. However, these results do not reflect the situation of the application on a practical farm. Furthermore, we could show that even different up- and downsampling methods for balancing training data during application to unknown, realistic data had no influence [17]. Thus, it becomes clear that, despite sufficient model quality, the frequency of which the event to be predicted occurs influences the magnitude of the predictive values substantially and thus the share of animals reported as positive. As known from medicine and other fields [8,9], the application of test procedures and, accordingly, algorithms in groups where the risk for the event to be predicted is higher, allows for improving the ratio between correct and false-positive reports, i.e., the PPV becomes higher. In the risk groups analyzed, the question arises whether this improves the ratio in such a way that an implementation of this approach can be recommended.

4.2.1. Classification of Cows with a Previous Treatment (RGtreat)

Our own results have shown that cows with mastitis or lameness treatment have a higher chance of needing to be treated again in the next lactation (RGtreat-SC/PL) or at a later stage of lactation (RGtreat-SC/SL). In other studies, an increased risk of further mastitis was found in cows that had already been infected with the cow-associated pathogen *Staphylococcus aureus* [26], as well as in cows with a past infection with the environment-associated pathogens *Streptococcus uberis* [26] and *Escherichia coli* [27], with 13% of all *E. coli* infected cows already had an infection in the same udder quarter. In [11], the odds ratio of mastitis was found to be up to 5.9 if at least one previous treatment was given in the current lactation. By narrowing the data to cows with a previous treatment in the same lactation, the frequency of occurrence in our own study was increased from 3.5% to 13.2% of days with treatment. This means a 3.7-fold higher risk of mastitis for this group. Another study found an odds ratio of 4.15 for mastitis incidence in the first 120 days of lactation for previous clinical mastitis [12]. These results suggest that cows or udder quarters are more likely to develop mastitis again [26,27]. However, in another study by Hammer et al. [28], no statistical correlation between the risk of mastitis and previous treatments that were more than 30 days old was found in 245 cases of mastitis.

An increased risk for subsequent treatments in the following lactation was also found by other authors. In a study of 402 cows, cows treated in the previous lactation were found to be 1.7 times more likely to develop subclinical mastitis in the first 60 days in the next lactation [29]. When restricted to animals treated in the last 60 days of the previous lactation, the risk there increased 4.9-fold. Another study with data from 350 Norwegian dairy herds and a total of 6046 cows in their second lactation [14] showed an increased risk (1.5-fold) when mastitis treatment was given in the first lactation. Limiting the risk to animals with mastitis treatment in the previous lactation (RGtreat-SC/PL) achieved a 2-fold increase in risk to 7.1% in our own study. This shows that cows with mastitis treatment also carry a higher risk into the next lactation due to individual susceptibility to pathogens or the persistence of a subclinical infection over the dry period [11]. However, this risk is reduced to some extent by the possibility of udder healing in the dry period through appropriate therapies [30], compared to the follow-up treatments within one lactation.

The risk that a cow will need to be treated again was also elevated for lameness treatments for both RGtreat-SC/PL and RGtreat-SC/SL. In another study with 600 cows over 44 months, a high range of positive odds ratios between 2.5–23 for all types of lameness diagnoses was found for the probability of a cow needing re-treatment [13]. A different study of over 7600 cows from 23 dairy farms found significant positive effects of prior lameness treatment for claw horn disruption lesions both at dry-off (2.5 times higher risk) and next lactation (twice the risk) [31]. In other studies, this association has also been established for treatments for sole ulcers, white line defects, and digital dermatitis [13,32]. This is the case when treatment of the clinical symptoms does not address the underlying cause sufficiently, e.g., a thinned digital cushion [13].

The AUCs of RGtreat-SC/PL and RGtreat-OC/SL did not differ significantly from those models applied to all cows. Only the AUCs after application in RGtreat-SC/SL showed significantly lower values in both treatment categories. This was due to the combination of low numbers of cow-days in the corresponding test data (see Appendix A Table A3) and the restriction of the test data to a subgroup with a different distribution of features for days with and without treatment than in the whole test data. This introduces a sampling bias into the classification, which has a negative effect on AUC, especially in small data sets [33,34]. At the same time, in this RGtreat-SC/SL, the risk of repeated treatment for mastitis or lameness was highest. Accordingly, PPVs in this RGtreat-SC/SL had the significantly highest values compared to the other groups. This means that they have the greatest potential for reducing false-positives compared to the other RGs, yet the PPVs were not in a range satisfactory for practical use, with 0.20 for mastitis and 0.15 for lameness treatments.

RGtreat-OC/SL narrowed the data down to cows that had already undergone a different treatment in the same lactation. A study on genetic correlations found a comparatively low correlation of 0.32 ± 0.07 between the occurrence of mastitis from DIM −10 to +50 and other treatments (fertility disorders, metabolic diseases, and lameness) in the period up to 100 DIM [35]. Another study by Hossein-Zadeh and Ardalan [12] found odds ratios for clinical mastitis in the first 120 DIM of 57,300 Holstein cows, 9.45 with previous retained placenta and 12.36 with previous milk fever. The association between the retained placenta and later clinical mastitis has before been quantified by [36] with a 1.5-fold higher risk for mild and 5.4-fold higher risk for severe mastitis, respectively. Acidosis can act as a trigger for laminitis, which then develops into lameness [37]. A study by Berge and Vertenten [38] with 131 Dutch farms found odds ratios for previous ketosis of 1.9 for mastitis treatments and 1.7 for lameness treatments in the rest of the lactation. Rowlands et al. [39] found a significant doubling of the frequency of interdigital dermatitis in cows with previous endometritis in the same lactation based on data with 2109 lactations, but the data showed no correlation between other previous diseases and mastitis. Our own results for RGtreat-OC/SL could only cause a small increase in the frequency of occurrence of mastitis and lameness treatments, and consequently no higher PPVs by limiting the animals to those treated against diseases from other disease categories (with otherwise comparable AUC values). Since the cows remained in this risk group for the remaining lactation, the effects of these pre-treatments were too small in relation to the total data at the daily level.

4.2.2. Classification of Cows with Increased SCC After Milk Recording (RG-SCC)

Several studies have investigated the association between increased SCC in MR and the subsequent occurrence of mastitis. Whist and Østerås [14] found a 1.9-fold higher risk of clinical mastitis for SCC > 200,000 cells/mL in the first MR after calving. The authors also found a 1.7-fold higher risk of developing mastitis in the second lactation with a geometric mean between 400,000–800,000 cells/mL of the last three MR cell counts before the second calving [14]. In a study by Steeneveld et al. [15], the relationship between the previous month's SCC and the geometric mean of all MR test days of the previous lactation with mastitis treatments was examined using data from almost 40,000 cows and 8,500 mastitis cases. The significant odds ratios here were 1.33 and 1.15 for elevated SCC (> 200,000 cells/mL) in the preceding MR and previous lactation on average, respectively, which signaled a slightly increased risk of a subsequent mastitis treatment.

RG-SCC showed a comparable AUC as an indicator of model quality, but a significantly lower PPV compared to RGtreat-SC/SL. The reason for this is that, whereas clinical symptoms were present at one time during pre-treatment, a SCC of > 100,000 cells/mL and thus a risk after MR is not necessarily associated with clinical symptoms, and therefore no treatment is performed. Thus, the limitation to this risk group and the application of the classification algorithms would not lead to any added value other than the animal listings themselves, which are conspicuous in the context of MR with regard to udder health.

4.2.3. Classification of Cows in Early Lactation (RGtime-100)

Only cow-days within the first 100 DIM were classified as this last risk group. It is known that treatments for mastitis are more common in early lactation [12,40]. The study by Hammer et al. [28] found in 245 cows that the odds ratio in cows over 100 DIM dropped to only 0.3 compared to the reference group between 10 and 20 DIM. However, this odds ratio was also only 0.4 between 30 and 100 DIM. The odds ratio for clinical mastitis decreased after the first month of lactation [15], but after the first three months (after about 100 DIM) the odds ratio was still 1.9 for primiparous and 3.6 for multiparous cows, compared to lactation month 8 and higher as reference. In terms of lameness treatments, in a study of

2100 cows over three years, these were most common between 61 and 150 DIM and least common between 16 and 60 DIM [41]. However, these data are from only one farm, so a farm effect cannot be excluded.

In our own study, the restriction only to animals in the first 100 DIM did not lead to an increase in the frequency of occurrence and thus had no effect on the PPV. The effects in the quoted studies often reported shorter time windows after calving with higher risk. This was also investigated in our own study (60 DIM) but did not lead to any change in the frequency of treatments. In line with the findings from the other risk groups, this narrowing of the data set also did not lead to any improvement in predictive values or false alarms.

4.3. General Discussion

Our results show that a limitation to risk groups can improve the PPVs of a daily detection of individual animals in need of treatment using sensor data up to 0.20 PPV (i.e., 80% error rate). However, the suitability for satisfactory practical use remains questionable. In contrast to the minimum requirements for the model quality criteria of the models that can be used in practice (i.e., a sensitivity of 0.70–0.80 and a specificity of 0.99 [42]), there are no recommendations for a minimum PPV to be achieved. A recent survey of practicing farmers' preferences for the performance of a lameness detection system included options for the percentage of false alarms from 0% to 15%, corresponding to a PPV of 0.85–1.00 [43]. Although not explicitly asked for tolerable, but rather for preferred values, this shows the discrepancy between user expectations and the actual percentage of false alarms. The study by Steeneveld et al. [7] on the reduction of false-positive mastitis alarms of an AMS showed that despite a very good test characteristic of their model of 0.70 sensitivity and 0.98 specificity and a reduction of false-positive alarms by up to 35 percentage points, the PPV was still only 0.03 due to the low frequency of cow days with mastitis in the data (227 out of 508,517). This shows that despite all optimization attempts, the known low risk of one treatment (i.e., a form of health data available in databases) per animal per day remains the main factor influencing the prediction. As already discussed in Van De Gucht et al. [42] and Zehner et al. [16], it is not possible to use classification models of sensor systems as the only tool to find the animals that need treatment or need special care.

5. Conclusions

The objectives of this study were to demonstrate the importance of applying classification models for cows in need of treatment to practical data sets and to highlight the importance of the low frequency of occurrence of the trait “treatment” in relation to individual cows and days, based on the assignment of cows to risk groups.

Within those risk groups, the frequency of occurrence of the target variable “treatment” and the respective PPV increased accordingly. This influence was largest when applying the algorithms to the risk group of cows with previous treatment in the same lactation, but even the highest achieved PPV of 0.20 is not sufficient for the prediction of mastitis and lameness treatments in practice. The critical factor influencing the prediction remains, despite all optimization variations with respect to model validity, the known low risk of a treatment per animal per day. It can be assumed that this also applies to other health or disease data. This requires rethinking and specific information about the fact that the detection of cows in need of treatment within a herd is not possible through sensor data and the corresponding algorithms only, but requires additional expert knowledge. However, if the responsible person already has certain animals in focus during the day, the existing animal-specific (sensor) data can be important decisional support.

Author Contributions: Investigation, C.P.; Project administration, U.M.; Supervision, C.R. and W.B.; Writing-original draft, C.P.; Writing-review & editing, U.M. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by funds of the German Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support program.

Acknowledgments: The authors gratefully acknowledge the assistance of the staff of the Frankenforst Research Station, University of Bonn, 53639 Königswinter, Germany, as well as the Educational and Research Center for Animal Husbandry, Hofgut Neumuehle, 67728 Muenchweiler an der Alsenz, Germany, and the Dairy Research Farm Karkendamm, Institute for Animal Breeding and Animal Husbandry Christian-Albrechts-University of Kiel, 24098 Kiel, Germany. We also thank the members of the Center of Integrated Dairy Research (CIDRe, University of Bonn, Germany).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Random Forest feature importance for mastitis treatments (Mean ± 95% CI), obtained during 5-fold cross validation.

Farm					
A		B		C	
Feature	Importance	Feature	Importance	Feature	Importance
Last milk recording SCC	0.13 ± 0.07	Last milk recording SCC	0.22 ± 0.03	Last milk recording SCC	0.22 ± 0.04
Highest milk flow p.m., slope	0.11 ± 0.04	Milk yield, RMdiff	0.14 ± 0.02	Conductivity a.m., RMdiff	0.14 ± 0.03
Conductivity p.m., slope	0.10 ± 0.04	Milk yield p.m., RMdiff	0.09 ± 0.01	Conductivity p.m., RMdiff	0.07 ± 0.02
Milking duration a.m.	0.06 ± 0.02	Milk yield, slope	0.08 ± 0.02	Day/night ratio activity, RM	0.05 ± 0.01
Conductivity p.m., RMdiff	0.06 ± 0.02	Milk yield p.m., slope	0.07 ± 0.01	Feed intake, RMdiff	0.05 ± 0.01
Milk yield, RMdiff	0.06 ± 0.03	Feeding time with intake	0.06 ± 0.01	Milk yield, RMdiff	0.05 ± 0.01
Feed intake, RMdiff	0.05 ± 0.01	Milk yield a.m., RMdiff	0.06 ± 0.01	Last milk recording lactose	0.05 ± 0.01
Feed intake, RMprev	0.05 ± 0.02	Conductivity p.m., RMdiff	0.05 ± 0.01	Feed intake, slope	0.04 ± 0.01
DIM	0.04 ± 0.01	Conductivity p.m., slope	0.04 ± 0.01	Conductivity p.m., slope	0.04 ± 0.01
Feed intake, slope	0.04 ± 0.01	Conductivity a.m., RMdiff	0.03 ± 0.00	Activity (Min), slope	0.04 ± 0.01
Milking duration a.m., diff	0.04 ± 0.02	Body weight, slope	0.02 ± 0.01	Feeding time with intake, slope	0.03 ± 0.01
Milk yield	0.04 ± 0.01	Feeding time with intake, slope	0.02 ± 0.01	Body weight, RMdiff	0.03 ± 0.00
Body weight	0.04 ± 0.01	Milking duration p.m., RMdiff	0.02 ± 0.00	Feed intake, d-1	0.03 ± 0.00
Conductivity a.m., RMdiff	0.04 ± 0.01	Highest milk flow a.m.	0.02 ± 0.00	Day/night ratio activity, RMprev	0.03 ± 0.01
Day/night ratio activity, RM	0.03 ± 0.01	Day/night ratio activity, RMprev	0.02 ± 0.01	Milk flow p.m., RMdiff	0.03 ± 0.01
Highest milk flow a.m.	0.02 ± 0.01	Milk yield	0.02 ± 0.01	Milking duration a.m., slope	0.03 ± 0.01
Highest milk flow a.m., RMdiff	0.02 ± 0.01	Feeding visit duration, RMprev	0.01 ± 0.01	Feeding visit duration, RMprev	0.02 ± 0.01
Activity (max), RMprev	0.02 ± 0.01	Day/night ratio activity, RM	0.01 ± 0.00	Activity Sum, RMprev	0.02 ± 0.01
Activity (Sum)	0.02 ± 0.01	Body weight, RMdiff	0.01 ± 0.00	Activity (Max), RM	0.02 ± 0.00
Activity (Sum) p.m., RMprev	0.02 ± 0.01	Feed intake, RMprev	0.01 ± 0.00	Activity (Sum) p.m., RM	0.01 ± 0.01

For variable descriptions, see [17].

Table A2. Random Forest feature importance for lameness treatments (Mean ± 95% CI), obtained during 5-fold cross validation.

		Farm			
A		**B**		**C**	
Feature	Importance	Feature	Importance	Feature	Importance
Feeding time with intake	0.13 ± 0.02	Feeding time with intake	0.18 ± 0.01	Feeding time with intake	0.11 ± 0.02
Feeding time with intake, slope	0.11 ± 0.03	Feed intake	0.12 ± 0.01	Feeding time with intake, slope	0.09 ± 0.03
Feeding visits	0.08 ± 0.01	Feeding time with intake, slope	0.11 ± 0.02	Feed intake, slope	0.07 ± 0.02
Body weight, slope	0.08 ± 0.02	Feeding visit duration, RMprev	0.06 ± 0.01	Last milk recording SCC	0.07 ± 0.04
Body weight, RMdiff	0.08 ± 0.03	Feed intake, slope	0.06 ± 0.02	Milk yield	0.07 ± 0.01
Feeding visit duration, RMdiff	0.07 ± 0.02	Last milk recording fat	0.06 ± 0.01	Feed intake, d-1	0.07 ± 0.02
Feed intake, slope	0.06 ± 0.00	Feed intake per visit, RMprev	0.04 ± 0.01	Body weight, slope	0.06 ± 0.02
Activity (Min), RM	0.06 ± 0.03	Feed intake, RMprev	0.04 ± 0.01	Highest milk flow a.m., RMdiff	0.04 ± 0.02
Feeding visit duration, slope	0.05 ± 0.02	Milk yield	0.04 ± 0.02	DIM	0.04 ± 0.01
Feeding visit duration, d-1	0.04 ± 0.02	Activity, slope	0.04 ± 0.01	Body weight	0.04 ± 0.01
Last milk recording lactose	0.04 ± 0.01	DIM	0.03 ± 0.01	Activity (Sum) p.m., RM	0.04 ± 0.01
Acitivity, 3 highest (Sum), slope	0.03 ± 0.02	Feed intake per visit, slope	0.03 ± 0.01	Conductivity a.m., RMdiff	0.04 ± 0.01
Feed intake, RMprev	0.03 ± 0.01	Last milk recording SCC	0.03 ± 0.01	Feeding visits, slope	0.04 ± 0.01
Activity (Max), RMprev	0.03 ± 0.01	Activity (Max), RMprev	0.03 ± 0.01	Activity (Min), RMprev	0.04 ± 0.01
DIM	0.02 ± 0.01	Day/night ratio activity, RMprev	0.03 ± 0.01	Feeding visit duration, RMprev	0.04 ± 0.01
Milk yield	0.02 ± 0.00	Day/night ratio activity, RM	0.02 ± 0.01	Day/night ratio activity, RMprev	0.03 ± 0.01
Milking duration a.m.	0.02 ± 0.01	Feeding visit duration, RMprev	0.02 ± 0.01	Day/night ratio activity, RM	0.03 ± 0.01
Conductivity p.m., slope	0.02 ± 0.01	Body weight, slope	0.02 ± 0.01	Last milk recording fat	0.03 ± 0.01
Day/night ratio activity, RM	0.01 ± 0.01	Activity (Min), slope	0.02 ± 0.00	Activity (Max), slope	0.03 ± 0.01
Day/night ratio activity, RMprev	0.01 ± 0.01	Feeding visit duration, slope	0.02 ± 0.01	Milking duration a.m., slope	0.02 ± 0.00

For variable descriptions, see [17].

Table A3. The number of cow-days in the test data for all cows, as well as in the respective risk groups.

			Risk Group				
Treatment	**Farm**	**All cows**	**RGtreat-SC/PL**	**RGtreat-SC/SL**	**RGtreat-OC/SL**	**RG-SCC**	**RGtime-100**
Mastitis	A	8.041	939	324	2.859	924	3.437
	B	25.474	4.687	3.286	11.974	7.055	11.312
	C	9.288	1.007	641	2.969	2.435	3.540
	Total	42.803	6.633	4.251	17.802	10.414	18.289
Lameness	A	9.155	1.596	1.049	3.385	1.037	3.853
	B	27.643	5.995	4.556	12.214	7.276	12.124
	C	11.243	1.996	2.812	2.538	2.865	4.067
	Total	48.041	9.587	8.417	18.137	11.178	20.044

RGtreat-SC/PL: at least one treatment of the same category in previous lactation, RGtreat-SC/SL: at least one previous treatment of the same category in the same lactation, RGtreat-OC/SL: at least one previous treatment of another category in the same lactation, RG-SCC: high SCC in previous milk recording, RGtime-100: DIM ≤ 100.

Table A4. The number of other treatments occurring before the classified treatment for cows in RGtreat-OC/SL. Cows in the group could have multiple previous treatments.

		Other Previous Treatments			
Treatment	**Farm**	**Mastitis**	**Lameness**	**Metabolic Disorder**	**Fertility Disorder**
Mastitis	A	-	45	64	21
	B	-	299	668	473
	C	-	109	57	39
Lameness	A	81	-	159	66
	B	396	-	671	529
	C	135	-	105	55

Chapter 4 References

1. Koeck, A.; Loker, S.; Miglior, F.; Kelton, D.; Jamrozik, J.; Schenkel, F. (**2014**): Genetic relationships of clinical mastitis, cystic ovaries, and lameness with milk yield and somatic cell score in first-lactation Canadian Holsteins. *Journal of Dairy Science* 97 (9), 5806–5813.

2. Kamphuis, C.; Frank, E.; Burke, J.; Verkerk, G.; Jago, J. (**2013**): Applying additive logistic regression to data derived from sensors monitoring behavioral and physiological characteristics of dairy cows to detect lameness. *Journal of Dairy Science* 96 (11), 7043–7053.

3. Jensen, D.; Hogeveen, H.; De Vries, A. (**2016**): Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis. *Journal of Dairy Science* 99 (9), 7344–7361.

4. Van Hertem, T.; Maltz, E.; Antler, A.; Romanini, C.; Viazzi, S.; Bahr, C.; Schlageter-Tello, A.; Lokhorst, C.; Berckmans, D.; Halachmi, I. (**2013**): Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *Journal of Dairy Science* 96 (7), 4286–4298.

5. Alsaaod, M.; Römer, C.; Kleinmanns, J.; Hendriksen, K.; Rose-Meierhöfer, S.; Plümer, L.; Büscher, W. (**2012**): Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. *Applied Animal Behaviour Science* 142 (3), 134–141.

6. Miekley, B.; Traulsen, I.; Krieter, J. (**2012**): Detection of mastitis and lameness in dairy cows using wavelet analysis. *Livestock Science* 148 (3), 227–236.

7. Steeneveld, W.; Van Der Gaag, L.; Ouweltjes, W.; Mollenhorst, H.; Hogeveen, H. (**2010**): Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *Journal of Dairy Science* 93 (6), 2559–2568.

8. Gigerenzer, G.; Hoffrage, U.; Ebert, A. (**1988**): AIDS counselling for low-risk clients. *AIDS Care* 10 (2), 197–211.

9. A Grimes, D.; Schulz, K. (**2002**): Uses and abuses of screening tests. *Lancet* 359 (9309), 881–884.

10. Yazdanpanah, Y.; Perelman, J.; DiLorenzo, M.; Alves, J.; Barros, H.; Mateus, C.; Pereira, J.; Mansinho, K.; Robine, M.; Park, J.-E.; et al. (**2013**): Routine HIV Screening in Portugal: Clinical Impact and Cost-Effectiveness. *PLOS one* 8 (12), e84173.

11. Berry, D.; Meaney, W. (**2005**): Cow factors affecting the risk of clinical mastitis. *Irish Journal of Agricultural and Food Research* 44, 147–156.

12. Hossein-Zadeh, N.; Ardalan, M. (**2011**): Cow-specific risk factors for retained placenta, metritis and clinical mastitis in Holstein cows. *Veterinary Research Communications* 35 (6), 345–354.

13. Green, L.; Huxley, J.; Banks, C.; Green, M. (**2014**): Temporal associations between low body condition, lameness and milk yield in a UK dairy herd. *Preventive Veterinary Medicine* 113 (1), 63–71.

14. Whist, A.; Østerås, O. (**2006**): Associations between somatic cell counts at calving or prior to drying-off and clinical mastitis in the remaining or subsequent lactation. *Journal of Dairy Research* 74 (1), 66–73.

15. Steeneveld, W.; Hogeveen, H.; Barkema, H.; Broek, J.; Huirne, R. (**2008**): The Influence of Cow Factors on the Incidence of Clinical Mastitis in Dairy Cows. *Journal of Dairy Science* 91 (4), 1391–1402.

16. Zehner, N.; Niederhauser, J.; Schick, M.; Umstatter, C. (**2019**): Development and validation of a predictive model for calving time based on sensor measurements of ingestive behavior in dairy cows. *Computers and Electronics in Agriculture* 161, 62–71.

17. Post, C.; Rietz, C.; Büscher, W.; Müller, U. (**2020**): Using Sensor Data to Detect Lameness and Mastitis Treatment Events in Dairy Cows: A Comparison of Classification Models. *Sensors* 20 (14), 3863.

18. Pedregosa, F.; Varoquaux, G.; Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (**2011**): Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

19. DLQ-Richtlinie 1.15: *Zur Definition und Berechnung von Kennzahlen zum Eutergesundheitsmonitoring in der Herde und von deren Vergleichswerten*. Available online: https://infothek.die-milchkontrolle.de/wp-content/uploads/2018/08/DLQ-Richtlinie-1.15-vom-17.11.2014.pdf (accessed on 28 January **2021**).

20. Schwarz, D.; Diesterbeck, U.; Failing, K.; König, S.; Brügemann, K.; Zschöck, M.; Wolter, W.; Czerny, C.-P. (**2010**): Somatic cell counts and bacteriological status in quarter foremilk samples of cows in Hesse, Germany—A longitudinal study. *Journal of Dairy Science* 93 (12), 5716–5728.

21. Seabold, S.; Perktold, J. (**2010**): Statsmodels: Econometric and Statistical Modeling with Python. In: *Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July* 2010, 92–96.

22. Miekley, B.; Traulsen, I.; Krieter, J. (**2013**): Principal component analysis for the early detection of mastitis and lameness in dairy cows. *Journal of Dairy Research* 80 (3), 335–343.

23. Kelleher, J.; MacNamee, B.; D'Arcy, A. (**2015**): *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*; MIT press: Cambridge, MA, USA.

24. Garcia, E.; Klaas, I.; Amigó, J.; Bro, R.; Enevoldsen, C. (**2014**): Lameness detection challenges in automated milking systems addressed with partial least squares discriminant analysis. *Journal of Dairy Science* 97 (12), 7476–7486.

25. De Mol, R.; Andre, G.; Bleumer, E.; Van Der Werf, J.; De Haas, Y.; Van Reenen, C. (**2013**): Applicability of day-to-day variation in behavior for the automated detection of lameness in dairy cows. *Journal of Dairy Science* 96 (6), 3703–3712.

26. Zadoks, R.; Allore, H.G.; Barkema, H.; Sampimon, O.; Wellenberg, G.; Gröhn, Y.; Schukken, Y.H. (**2001**): Cow- and Quarter-Level Risk Factors for Streptococcus uberis and Staphylococcus aureus Mastitis. *Journal of Dairy Science* 84 (12), 2649–2663.

27. Döpfer, D.; Barkema, H.; Lam, T.; Schukken, Y.; Gaastra, W. (**1999**): Recurrent clinical mastitis caused by Escherichia coli in dairy cows. *Journal of Dairy Science* 82 (1), 80–85.

28. Hammer, J.; Morton, J.; Kerrisk, K. (**2012**): Quarter-milking-, quarter-, udder- and lactation-level risk factors and indicators for clinical mastitis during lactation in pasture-fed dairy cows managed in an automatic milking system. *Australian Veterinary Journal* 90 (5), 167–174.

29. Pinedo, P.; Fleming, C.; Risco, C. (**2012**): Events occurring during the previous lactation, the dry period, and peripartum as risk factors for early lactation mastitis in cows receiving 2 different intramammary dry cow therapies. *Journal of Dairy Science* 95 (12), 7015–7026.

30. Eberhart, R. (**1986**): Management of Dry Cows to Reduce Mastitis. *Journal of Dairy Science* 69 (6), 1721–1732.

31. Foditsch, C.; Oikonomou, G.; Machado, V.; Bicalho, M.; Ganda, E.; Lima, S.; Rossi, R.; Ribeiro, B.; Kussler, A.; Bicalho, R. (**2016**): Lameness Prevalence and Risk Factors in Large Dairy Farms in Upstate New York. Model Development for the Prediction of Claw Horn Disruption Lesions. *PLOS one* 11 (1), e0146718.

32. Oikonomou, G.; Cook, N.; Bicalho, R. (**2013**): Sire predicted transmitting ability for conformation and yield traits and previous lactation incidence of foot lesions as risk factors for the incidence of foot lesions in Holstein cows. *Journal of Dairy Science* 96 (6), 3713–3722.

33. Zadrozny, B. (**2004**): Learning and evaluating classifiers under sample selection bias. In: *Proceedings of the 21st International Conference on Machine Learning*, Banff, AL, Canada, 4–8 July, 114.

34. Parker, B.; Günter, S.; Bedo, J. (**2007**): Stratification bias in low signal microarray studies. *BMC Bioinformatics*, 8 (1), 326.

35. Lassen, J.; Hansen, M.; Sørensen, M.; Aamand, G.; Christensen, L.; Madsen, P. (**2003**): Genetic relationship between body condition score, dairy character, mastitis, and diseases other than mastitis in first-parity Danish Holstein cows. *Journal of Dairy Science* 86 (11), 3730–3735.

36. Schukken, Y.; Erb, H.; Smith, R. (**1988**): The relationship between mastitis and retained placenta in a commercial population of Holstein dairy cows. *Preventive Veterinary Medicine* 5 (3), 181–190.

37. Westwood, C.; Bramley, E.; Lean, I. (**2003**): Review of the relationship between nutrition and lameness in pasture-fed dairy cattle. *New Zealand Veterinary Journal* 51 (5), 208–218.

38. Berge, A.; Vertenten, G. (**2014**): A field study to determine the prevalence, dairy herd management systems, and fresh cow clinical conditions associated with ketosis in western European dairy herds. *Journal of Dairy Science* 97 (4), 2145–2154.

39. Rowlands, G.; Lucey, S.; Russell, A. (**1986**): Susceptibility to disease in the dairy cow and its relationship with occurrences of other diseases in the current of preceding lactation. *Preventive Veterinary Medicine* 1986, 4 (3), 223–234.

40. Suriyasathaporn, W.; Schukken, Y.; Nielen, M.; Brand, A. (**2000**): Low Somatic Cell Count: A Risk Factor for Subsequent Clinical Mastitis in a Dairy Herd. *Journal of Dairy Science* 83 (6), 1248–1255.

41. Sanders, A.; Shearer, J.; De Vries, A. (**2009**): Seasonal incidence of lameness and risk factors associated with thin soles, white line disease, ulcers, and sole punctures in dairy cattle. *Journal of Dairy Science* 92 (7), 3165–3174.

42. Dominiak, K.; Kristensen, A. (**2017**): Prioritizing alarms from sensor-based detection models in livestock production—A review on model performance and alarm reducing methods. *Computers and Electronics in Agriculture* 133, 46–67.

43. Van De Gucht, T.; Saeys, W.; Van Nuffel, A.; Pluym, L.; Piccart, K.; Lauwers, L.; Vangeyte, J.; Van Weyenberg, S. (**2017**): Farmers' preferences for automatic lameness-detection systems in dairy cattle. *Journal of Dairy Science* 100 (7), 5746–5757.

5 General Discussion

The results of the two present studies (Chapters 3 and 4) showed that daily detection of the need for treatment for mastitis and lameness in dairy cows is generally possible using sensor data commonly available on practical farms. However, it also became clear that the quality of the classification is strongly related to the correlation of the features obtained from the sensors with the target variable, and furthermore that the low daily frequency of treatments has a strong influence on the number of false alarms and thus on the suitability of such a model for practice. In this chapter, the methodology of the two studies will be put in context to the existing literature and potentials for improvement will be discussed.

5.1 Importance of practicality of sensors and sensor systems

The development of prediction models intended for practical use requires, on the one hand, that the necessary sensors are already in widespread use on practical farms, or are at least close to market maturity; on the other hand, the data should have been generated in an environment that corresponds to a commercial farm as closely as possible.

The sensors used in our own studies are widely used on practical dairy farms. A non-representative survey in the Netherlands in 2015 showed a proportion of 39% of dairy farms are using at least one sensor (Steeneveld and Hogeveen 2015). There were large differences between farms with an AMS, and farms with a conventional milking system (CMS) in terms of sensors related to milk yield. 93% of AMS farms had an additional conductivity sensor, while this proportion was only 35% for CMS farms. In contrast, 70% of CMS farms had an activity measurement system available, while this was only the case for 41% of AMS farms (Steeneveld and Hogeveen 2015). Another survey in Italy found that 39% of farms use sensors in their milking system, and 15% use sensors for activity measurement (Lora et al. 2020). In a survey in the USA conducted by Borchers and Bewley (2015), the portion of farms using milking sensors (52%) and activity measurement (41%) was higher than in the European surveys, but this survey was also not representative and the herd size, with a median of 230 dairy cows, was higher than the median herd sizes of 90–123 cows (Steeneveld and Hogeveen 2015) and 68 cows (Lora et al. 2020) of the other two studies. This underlines that larger herds have a higher demand for sensors to aid herd management.

One exception among the sensors used in our own studies is the automatic recording of the amounts of roughage and water consumed by the cows. This technology has not yet played a role on practical dairy farms (Steeneveld and Hogeveen 2015; Lora et al. 2020; Borchers and Bewley 2015). A decisive criterion for the adaption of sensor technology is the expected financial benefit of the investment (Steeneveld and Hogeveen 2015); however, the necessary technology for animal-specific recording of feed intake is complex and expensive to install and maintain, and is therefore so far only found on experimental farms. In the three farms considered in our own studies, this system recorded data on the time of the individual feed and water visits, as well as their duration, in addition to the amount of feed and water intake. As a commercially interesting alternative for obtaining this information, technical solutions already exist with the help of which an estimation of the number and duration of feed and water visits of individual cows can be derived. For accelerometers on the ear and proximity sensors on the leg, Pearson correlation coefficients for feeding time (minutes per 1-h time block) of 0.88 and 0.93, respectively, were found in a study by Borchers et al. (2016) compared to visual observation of the cows (n=46 and 41, respectively). Here, the proximity sensor was used to determine whether a cow was at the feed bunk. A correlation coefficient of 0.88 for the determination of feeding time by accelerometer was also found by (Pereira et al. 2018). In another study, coefficients of determination (r^2) of 0.90 for feeding time but only 0.31 for the number of feed visits were determined by an accelerometer (Mattachini et al. 2016). These systems only register the mere presence of a cow at the feeding table or water trough, or its head tilts and movements interpreted as feeding movements, but no direct statement can be made about the amount of feed or water consumed, as the amount measured by the weighing troughs is missing as an additional plausibility check that a cow has actually consumed feed and water instead of just being near the trough. Results from the literature show that it is possible to derive this approximately, via a correlation with either the duration of time spent at the feeder (r=0.89) or the chewing movements (r=0.78) (Pahl et al. 2016). However, other authors came to the conclusion that the variability of this correlation between individual cows is too high to make a reliable quantitative statement about the amount of feed consumed (Leiber et al. 2016; Halachmi et al. 2016).

The results of the Random Forest Feature Importance in the two own studies (see Table 3-3 and Tables A1 and A2 of Chapter 4) have shown that the features derived from the daily amount of feed consumed were important for the classification, but especially for lameness the features derived from the number and duration of feed visits were also important. This suggests that commercially available sensors for recording individual cows’ feed visits also have the potential to be integrated into such models. On European dairy farms, sensor systems for recording feeding behaviour or animal location hardly play a role (Steeneveld and Hogeveen 2015; Lora et al. 2020), but in the USA 13% and 8% of farms reported already using technology for recording feed intake behavior and animal position and location, respectively (Borchers and Bewley 2015). It is interesting to note that on a scale of 1 (not useful) to 5 (useful), farmers rated feed intake behavior as relatively useful (mean rating of 4.3), while animal location was considered less useful (mean rating of 2.8) (Borchers and Bewley 2015). Thus, integrating animal location into multivariate statistical models to detect health disorders could potentially increase the popularity of tracking technology.

In addition to the availability of the required sensors, another criterion for the inclusion of the three farms in the own studies was that the conditions should correspond as closely as possible to those on a practical farm. These conditions refer to the husbandry (free movement in the barn, feeding of a TMR with ad libitum access), milking in the milking parlor and the management routine with regard to animal observation, including the recognition of cows in need of treatment. This is necessary, on the one hand, so that the data are comparable between the three farms, but also so that the statistical characteristics of the data roughly correspond to those on other practical farms. Most of the trials described in the literature are based on data from one single farm. Testing a model on data from other farms that are completely independent of the data used to train the model is a prerequisite for obtaining the most realistic estimation of performance in practical use as possible (Dominiak and Kristensen 2017). Otherwise, it would be impossible to infer from the results how differences in management or animal population affect the model. In Chapter 4, it was observed that the average performance of the models was weaker when used on data from a farm previously unknown to the model, especially for the hoof treatments (mean AUC of 0.61 compared to an AUC of 0.67 in the validation). This underlines the importance of validation that is independent from the training data and does also reflect data from a practical setting as close as possible. It also showed that the desired ratio of sensitivity and specificity could not always be achieved in the test data. This highlights the need for a user-adjustable threshold for the sensitivity of the predictive model.

5.2 Suitability of treatments as the target variable

As already described in Chapter 2.1.2, there are different ways of defining the target variable, which influence the quality and applicability of the resulting detection models. The implications of these influences will be discussed here. In our studies the target variable comprised all treatments with either a diagnosis for lameness or for mastitis, respectively. This approach was chosen because the analyzed dataset only contained records of treatments, but no further health examinations.

For lameness, all types of diagnoses were combined into a single target variable. However, within the different manifestations of hoof diseases there are differences in clinical symptoms and influence on the general condition of the cows. Digital dermatitis and heel horn erosion usually have little effect on cow gait (Beer et al. 2016). This can lead to a situation where a treatment for these diseases is carried out too late if the animal is only examined during daily routine. Despite this, few studies give details about the spectrum of diseases that cause lameness in their respective herd (Dutton-Regester et al. 2018). This could also have been improved in our own trials (Chapters 3 and 4) by listing all diagnoses and their respective frequencies. Not so much to improve the own statistical models in particular, but it could have led to an increased comparability to other studies.

Similar to the lameness treatments, in the own studies all mastitis treatments were aggregated into one target variable. It was assumed that treatment occurred at the presence of clinical symptoms and that the target variable therefore reflected as closely as possible the decision-making of practical farmers when detecting and treating mastitis. This definition is also found in other studies (Miekley et al. 2013a; Miekley et al. 2013b; Huybrechts et al. 2014; Steeneveld et al. 2010). As it is also the case with lameness, there may be variation within the target variable “mastitis” in terms of the underlying disease and the resulting signals in the sensor data. Therefore, it is also important to apply models to other herds previously unknown to the model, to check for a possible decrease in detection performance (Hogeveen et al. 2010). In our own study in Chapter 4, there was no drop in performance on data from other farms compared to the validation data from the same farm (mean AUC of 0.72 versus 0.73). However, a closer look at the results of the individual farms showed significant differences in AUC within the models. Nonetheless, it is not possible with the obtained results to discuss in more detail whether these differences were caused by divergent patterns in the features, or noise in the outcome variable, e.g., differences in the ability of farm staff to recognize clinical symptoms between farms (Dominiak and Kristensen 2017).

Observations for cows in need of treatment carried out by farm staff as their daily routine are prone to missed cases of disease unless these are clearly reflected in clinical signs (Rutten et al. 2013). In terms of the relationship between sensor data and the target variable, this means that in our study this could have caused a time lag between a subclinical health disorder and an actual treatment, which may have led to cases being classified as false negatives in the data, and consequently the model quality may have suffered. If only treatments are used as a target instead of an additional assessment of the locomotion score, it increases the risk that slightly or subclinically lame animals are seen as non-lame in the data (van Nuffel et al. 2015b). A system that uses treatment as a target will also only try to predict treatments. This is desirable if such a system is to be used to select cows for acutely needed treatments. Early detection though, which would be necessary for preventive treatment of subclinical diseases, cannot be achieved (Rutten et al. 2013).

The main advantage of using the treatments as the target variable in our own studies was that a large amount of data from several farms could be collected and analyzed in a short period of time, instead of requiring scientific staff to observe all cows as thoroughly as possible over the entire period covered by the data. Another potential benefit of using treatments would be the ability to design a detection system that allows users to enter new treatment data into a database system on their own, and the model to be constantly re-trained using this additional data. This could possibly lead to a better adaptation of the model to the individual farm conditions, but also bears the risk that users enter implausible or false positive data, or forget to enter treatments and thus generate false negative cases. At this point, no studies are known that investigate how the input of further training data in a practical scenario affects the classification models used in our own studies. The subject of self-learning models in this field is yet to be investigated.

5.3 Applicability of health disorder detection systems in practice

In the following section, it will be discussed how and to what extent detection systems for health disorders can relieve the users, how the results of the systems are perceived by the users, and how this can affect the perception of costs and benefits.

The presentation of the inner processes of sensor systems can take place on several levels, each with a different levels of complexity and user involvement (Rutten et al. 2013; Alsaaod et al. 2019): at the first level, only the values measured by the sensors are presented. The interpretation of these values lies completely with the user, an evaluation requires expert knowledge as well as knowledge about the individual animals, and is very time-consuming. Conversely, this presentation enables maximum transparency, as it does not conceal any, or only a few, internal processes from the users. The next level is the presentation of a list of probabilities as percentage values derived from sensor data with statistical models that show how likely a cow is to be in need of treatment (O'Leary et al. 2020). This display of percentage values facilitates interpretation and allows ranking and prioritization of alert cows (Rutten et al. 2013). In addition, the threshold at which an investigation or a treatment is initiated, and thus, ultimately, the ratio of sensitivity to specificity can be set here by users and adapted to their farm-specific mode of operation. Possible disadvantages, however, are that there is uncertainty about the exact meaning of the given probabilities, since these represent an abstract and not a measured value. It is conceivable that, over time, users could become accustomed to the system because what is actually an objective measurement is interpreted subjectively. There is a risk that the threshold for detection will be selected too low, resulting in a low sensitivity and thus endangering animal welfare. Treatments for cows that are mildly or intermittently lame are often delayed (Horseman et al. 2014) or treated only when symptoms worsen (van Nuffel et al. 2015b). Working time constraints are often cited by farmers, as working time is limited and hoof care competes with other activities for priority (Horseman et al. 2014). In our own experiments it was shown by model validation that different statistical models need different threshold values for their output probability to arrive at a similar ratio of sensitivity and specificity. Here, it would be necessary to check how this affects the intuition of the users when choosing their own threshold value, or at least be considered when implementing such a system in practice to ensure that a sensor system is working as intended. According to Rutten et al. (2013) however, no study exists yet that has specifically dealt with the validation of a system that presents a more differentiated classification, i.e., on the basis of probabilities.

Most existing studies present the output of their models to the user on a binary level: a single cow is either in need of treatment or not. The advantage for users here is the unambiguousness of the statement about the condition of an animal, which, at least at first glance, does not require any interpretation by the user and can thus be translated into an action (initiating a treatment or at least an inspection of the affected cows) even by less qualified working personnel within standardized working procedures. The disadvantage of this approach is that the unambiguousness of the binary classification is only superficial, because the underlying model still assumes different probabilities of the need for treatment, which are obscured by the binary output. Furthermore, as has been pointed out several times in this thesis, a large number of cows on a positive list is likely to be false positives, i.e., without additional expert knowledge on the part of the user, prioritization of cows is not possible, leading to unnecessary and time-consuming animal checks. There is consensus that the acceptance of these lameness detection systems in farming practice requires the highest possible PPV, as they lead to unnecessary selection of actually healthy animals, unnecessary work and thus cause costs (O'Leary et al. 2020; van Hertem et al. 2013; van Nuffel et al. 2015b). When it comes to the detection of clinical mastitis, farmers also prefer systems that achieve a high PPV and only present the more severe cases, i.e., those that in fact require a treatment (Mollenhorst et al. 2012).

In addition, it is also imaginable that a sensor system gives recommendations for action beyond a mere presentation of a list of conspicuous cows, or even acts automatically itself (Rutten et al. 2013). In the area of health disorders, this could be the suggestion of a particular type of treatment, the autonomous consultation of a veterinarian, or an automated action such as segregating individual cows into a selection area where they wait to be examined. In view of our results, such systems should only be implemented if both the sensitivity and specificity of the model are as high as possible, so that the consequences of a wrong classification for unnecessary labor, as well as animal welfare and health, are low. Examples of this would be the automatic adjustment of the amount of concentrate (Rutten et al. 2013) or an automatic separation of conspicuous milk in an AMS.

However, the reviewed research results have shown that the combination of sensitivity and specificity required for automated action in the event of health disorders, which was even demanded to be as high as 90% sensitivity and 99% specificity in the article by O'Leary et al. (2020), cannot be achieved by the majority of sensor systems. Therefore, instead of subsequently dismissing the developed systems as useless for practice, the output of these systems should instead be designed in such a way that, instead of suggesting that an alarm always means an immediate need for action, they instead provide a regular indication of conspicuous cows, and leave the interpretation of these results and thus the decision as to whether or not there is a need for treatment to the farmers.

5.4 Factors influencing costs and benefits of sensor systems

The success and adaptation of sensor systems for the detection of health disorders in dairy cows ultimately also depends on the ratio of costs and benefits for the individual farmer. It is difficult to find monetary evaluations of systems in the scientific literature that can be applied to German dairy farms. The data from literature is subject to fluctuations over time (inflation, volatility in producer prices) as well as geographic differences; in addition, technological progress is continuously being made and thus the costs and availability of sensor systems on the market also change. On the benefit side, it is difficult to quantify some factors, such as the workload of the farmer or the increase in animal welfare by the use of sensor systems, or costs of additional workload, or even reduced animal welfare, if no such systems are used.

Nevertheless, a comparison of possible factors should be made here, because subjectively perceived costs and benefits also play a role in the decision to invest in sensor systems, being that most farmers have difficulties estimating the costs of a lame cow (Van De Gucht et al. 2017) or a mastitis (Huijps et al. 2008). The costs of a sensor system contain many variables. One of them is whether the required sensors and software are already available on the farm or can be used for other purposes at the same time. For most farms, the acquisition of an AMS automatically comes with sensors for measuring milk quantity, electrical conductivity and milk color (Steeneveld and Hogeveen 2015). Pedometers are already used by many farms to monitor estrus (Steeneveld and Hogeveen 2015; Borchers and Bewley 2015). For these farms, the only additional investment is the purchase and maintenance of the additional software needed to detect the health disorders, at least as long as the underlying models are able to use the existing sensors in an unmodified way. In this respect, the models presented in Chapter 1.1 have different prerequisites. Systems for lameness detection that only need a single pedometer or accelerometer are financially attractive, as these technologies are already being used on many farms (Van De Gucht et al. 2017). Though, changes detected by these sensors are often non-specific and only point to a general welfare problem, rather than a specific disease (Stachowicz and Umstätter 2021). Systems that use two or more high-resolution accelerometers to describe gait asymmetry allow variables that are more specific to lameness, but are significantly more expensive as well (O'Leary et al. 2020) and might require modifications of the barn (Van De Gucht et al. 2017). The same would be true for lameness detection systems via computer vision, which depend on the installation and maintenance of specific cameras (Kang et al. 2021).

For sensors specifically related to mastitis, more specific indicators would be the automatic measurement the SCC, milk color, or milk temperature, which right now are mostly exclusively used by farms that already have an AMS installed (Steeneveld and Hogeveen 2015). A model that relies on these sensor data as features would accordingly either depend on the presence of an AMS, or require a very high willingness to invest in these systems seperately.

On the benefit side, there are the expenses and losses caused by diseases that can be prevented or mitigated by automated detection. The expenses include the treatment costs (veterinarian, medicines, additional labor from either the herd manager or farm staff), while the financial losses are reflected in the loss of milk (reduced production or milk that cannot be sold), the occurrence of secondary diseases and recurring treatments, as well as an increased risk of culling and thus costs for restocking the herd (Dolecheck and Bewley 2018). The share of losses in the overall costs, in particular reduced milk and reproductive performance and culled cows, is significantly higher than the expenditure on treatments, averaging 74%, as reported by Dolecheck and Bewley (2018). However, their meta-analysis does not include the costs of preventive measures, as they were not included in the studies mentioned. The real and perceived ratio of costs and benefits for sensor system are a deciding factor, and it is also worth noting that this ratio might be subject to change once the use of a detection system leads to better animal health and welfare, and the benefits decrease due to a lower risk of treatments that could be prevented. As a conclusion, the investment in sensor systems needed for detection models to work is an individual decision with a trade-off between being widely available, and being more specific to a certain condition. When the goal is practicability, models should aim to use as much available technology as possible, as was the case with the models used in our own studies.

5.5 Conclusions

The development of statistical models to detect health disorders in dairy cows is a very complex and diverse research topic, with a huge variation in technology and methodology. It is therefore advised to pay attention to the underlying data, the experimental conditions, the utilized statistical models, as well as the presented performance metrics when comparing different studies or planning to develop one’s own detection models. Of all approaches, models that base their predictions on data from widely used sensors have the highest potential to be adapted on practical farms, but only if their detection performance meets the farmers’ expectations. These models should always be validated on data that reflect a practical dairy farming setting to ensure comparable results to those obtained during model development. The own two studies have presented several challenges when developing such models: The use of retrospectively collected data for training, which affects the quality of the feature variables and target variables alike and allows for the processing of large data sets, but introduces potential noise when dealing with treatments as a target variable instead of planned clinical assessments. The use of sensor data that in retrospect is not specific enough to the health disorder in question. The differences in results from various machine learning models showed the importance of a model selection process that is customized to the given data and problem. The difficulties when dealing with a low daily prevalence of treatments as the target variable became apparent in the second study, and the attempt to pre-select cows with higher risk for a treatment improved model performance metrics, but still resulted in a low proportion of true positive cows on an alarm list, and the subsequent high false alarm rate is the main obstacle that prevents these detection models from being feasible for practical use. This challenge would further be intensified if an improved animal health leads to an even low daily prevalence. Together with the preferences of farmers for low numbers of false alarms and labor-reducing technology, this shows that future research should focus on the usability and possibly other approaches to presentation and usage of sensor data and model output as a decision support system that still requires interpretation and expert knowledge, rather than trying to develop an autonomous illness detection system.

General Discussion References

Alsaaod, M.; Fadul, M.; Steiner, A. (**2019**): Automatic lameness detection in cattle. *The Veterinary Journal* 246, 35–44.

Beer, G.; Alsaaod, M.; Starke, A.; Schuepbach-Regula, G.; Müller, H.; Kohler, P.; Steiner, A. (**2016**): Use of Extended Characteristics of Locomotion and Feeding Behavior for Automated Identification of Lame Dairy Cows. *PLOS one* 11 (5), e0155796.

Borchers, M.; Bewley, J. (**2015**): An assessment of producer precision dairy farming technology use, prepurchase considerations, and usefulness. *Journal of Dairy Science* 98 (6), 4198–4205.

Borchers, M.; Chang, Y.; Tsai, I.; Wadsworth, B.; Bewley, J. (**2016**): A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors. *Journal of Dairy Science* 99 (9), 7458–7466.

Dolecheck, K.; Bewley, J. (**2018**): Animal board invited review: Dairy cow lameness expenditures, losses and total cost. *Animal* 12 (7), 1462–1474.

Dominiak, K.; Kristensen, A. (**2017**): Prioritizing alarms from sensor-based detection models in livestock production - A review on model performance and alarm reducing methods. *Computers and Electronics in Agriculture* 133, 46–67.

Dutton-Regester, K.; Barnes, T.; Wright, J.; Alawneh, J.; Rabiee, A. (**2018**): A systematic review of tests for the detection and diagnosis of foot lesions causing lameness in dairy cows. *Preventive Veterinary Medicine* 149, 53–66.

Halachmi, I.; Ben Meir, Y.; Miron, J.; Maltz, E. (**2016**): Feeding behavior improves prediction of dairy cow voluntary feed intake but cannot serve as the sole indicator. *Animal* 10 (9), 1501–1506.

Hogeveen, H.; Kamphuis, C.; Steeneveld, W.; Mollenhorst, H. (**2010**): Sensors and Clinical Mastitis—The Quest for the Perfect Alert. *Sensors* 10 (9), 7991–8009.

Horseman, S.; Roe, E.; Huxley, J.; Bell, N.; Mason, C.; Whay, H. (**2014**): The use of in-depth interviews to understand the process of treating lame dairy cows from the farmers' perspective. *Animal Welfare* 23 (2), 157–165.

Huijps, K.; Lam, T.; Hogeveen, H. (**2008**): Costs of mastitis: facts and perception. *Journal of Dairy Research* 75 (1), 113–120.

Huybrechts, T.; Mertens, K.; De Baerdemaeker, J.; De Ketelaere, B.; Saeys, W. (**2014**): Early warnings from automatic milk yield monitoring with online synergistic control. *Journal of Dairy Science* 97 (6), 3371–3381.

Kang, X.; Zhang, X.; Liu, G. (**2021**): A Review: Development of Computer Vision-Based Lameness Detection for Dairy Cows and Discussion of the Practical Applications. *Sensors* 21 (3), 753.

Leiber, F.; Holinger, M.; Zehner, N.; Dorn, K.; Probst, J.; Spengler Neff, A. (**2016**): Intake estimation in dairy cows fed roughage-based diets: An approach based on chewing behaviour measurements. *Applied Animal Behaviour Science* 185, 9–14.

Lora, I.; Gottardo, F.; Contiero, B.; Zidi, A.; Magrin, L.; Cassandro, M.; Cozzi, G. (**2020**): A survey on sensor systems used in Italian dairy farms and comparison between performances of similar herds equipped or not equipped with sensors. *Journal of Dairy Science* 103 (11), 10264–10272.

Mattachini, G.; Riva, E.; Perazzolo, F.; Naldi, E.; Provolo, G. (**2016**): Monitoring feeding behaviour of dairy cows using accelerometers. *J. Agricult. Engineer.* 47 (1), 54.

Miekley, B.; Stamer, E.; Traulsen, I.; Krieter, J. (**2013a**): Implementation of multivariate cumulative sum control charts in mastitis and lameness monitoring. *Journal of Dairy Science* 96 (9), 5723–5733.

Miekley, B.; Traulsen, I.; Krieter, J. (**2013b**): Principal component analysis for the early detection of mastitis and lameness in dairy cows. *Journal of Dairy Research* 80 (3), 335–343.

Mollenhorst, H.; Rijkaart, L.; Hogeveen, H. (**2012**): Mastitis alert preferences of farmers milking with automatic milking systems. *Journal of Dairy Science* 95 (5), 2523–2530.

O'Leary, N.; Byrne, D.; O'Connor, A.; Shalloo, L. (**2020**): Invited review: Cattle lameness detection with accelerometers. *Journal of Dairy Science* 103 (5), 3895–3911.

Pahl, C.; Hartung, E.; Grothmann, A.; Mahlkow-Nerge, K.; Haeussermann, A. (**2016**): Suitability of feeding and chewing time for estimation of feed intake in dairy cows. *Animal* 10 (9), 1507–1512.

Pereira, G.; Heins, B.; Endres, M. (**2018**): Technical note: Validation of an ear-tag accelerometer sensor to determine rumination, eating, and activity behaviors of grazing dairy cattle. *Journal of Dairy Science* 101 (3), 2492–2495.

Rutten, C.; Velthuis, A.; Steeneveld, W.; Hogeveen, H. (**2013**): Invited review. Sensors to support health management on dairy farms. *Journal of Dairy Science* 96 (4), 1928–1952.

Stachowicz, J.; Umstätter, C. (**2021**): Do we automatically detect health- or general welfare-related issues? A framework. *Proceedings. Biological sciences* 288 (1950), 20210190.

Steeneveld, W.; Hogeveen, H. (**2015**): Characterization of Dutch dairy farms using sensor systems for cow management. *Journal of Dairy Science* 98 (1), 709–717.

Steeneveld, W.; van der Gaag, L. C.; Ouweltjes, W.; Mollenhorst, H.; Hogeveen, H. (**2010**): Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems. *Journal of Dairy Science* 93 (6), 2559–2568.

Van De Gucht, T.; Saeys, W.; Van Nuffel, A.; Pluym, L.; Piccart, K.; Lauwers, L.; Vangeyte, J.; Van Weyenberg, S. (**2017**): Farmers' preferences for automatic lameness-detection systems in dairy cattle. *Journal of Dairy Science* 100 (7), 5746–5757.

van Hertem, T.; Maltz, E.; Antler, A.; Romanini, C.; Viazzi, S.; Bahr, C.; Schlageter-Tello, A.; Lokhorst, C.; Berckmans, D.; Halachmi, I. (**2013**): Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *Journal of Dairy Science* 96 (7), 4286–4298.

van Nuffel, A.; Zwertvaegher, I.; van Weyenberg, S.; Pastell, M.; Thorup, V.; Bahr, C.; Sonck, B.; Saeys, W. (**2015**): Lameness Detection in Dairy Cows: Part 2. Use of Sensors to Automatically Register Changes in Locomotion or Behavior. *Animals* 5 (3), 861–885.

6 Summary

6.1 English Summary

Due to technological advancements and structural change, the use of digital support systems has become a standard in dairy farming. In order to ensure health and welfare of the cows, special attention must be paid to the detection of health disorders. In practical farming, sensors that monitor individual animals are used for decades, and many studies have researched their potential for developing detection models. The two most important disorders in dairy farming are mastitis and lameness, and the relationship between sensor data and these disorders has a significant influence on the quality of the models' predictions. Studies on mastitis detection focused primarily on milk parameters measured in the milking parlor or in an automated milking system, i.e., milk yield, milk contents, and electrical conductivity; because these parameters are directly affected by a mastitis to various extents. Studies that tried to detect lameness relied on measuring activity and behavior of cows. The measurement of electrical impulses of pedometers and accelerometers led to variables describing the number of steps and type of movement, e.g., walking, standing, lying. With these measurements, sequences over time are more informative than a single data point; here, feature extraction, i.e., generating new variables from the raw data like ratios and deviations, are being utilized.

Another influence on model development is the target variable, also referred to as the gold standard, since the models are trained using the statistical relationship between the sensor variable and the target variable. Different approaches were used for this in the studies considered: Recording of veterinary treatments, a fixed scoring of health, or an automatic recording of the target variable (e.g., somatic cell count). In each case, the choice of method has advantages and disadvantages in terms of reliability, labor required for recording, and transferability to practice, and makes it difficult to compare the results between studies.

The evaluation of these results is based on the comparison of the prediction with the gold standard via the diagnostic values sensitivity and specificity, as well as the positive and negative predictive value. The first two values depend primarily on the threshold for detection chosen by the model developer and are therefore only comparable between studies to a limited extent, which is why the ROC curve should be chosen here as a metric for model quality. A particularly critical metric for this dissertation was the positive predictive value, which, when transferred to a practical setting, determines the relative number of false alarms for the farmer. This is influenced primarily by the frequency of occurrence of the diseases or treatments, and was therefore given particular attention in the two studies published as part of this dissertation.

The goal of the first study was to develop machine learning models for predicting necessary mastitis and lameness treatments in Holstein cows based on data from sensors that are already established in the field. These included milking data, pedometer activity, feed and water intake, and live weight of 167 individual cows over a 40-month period, in addition to recorded treatments. Variables derived from these data were preselected using Random Forest Feature Importance, Pearson Correlation, and Sequential Forward Feature Selection methods and then used to train various machine learning models, including Logistic Regression (LR), Support Vector Machine, K-nearest Neighbors, Gaussian Naïve Bayes (GNB), Extra Trees (ET), and Random Forest (RF). These models were compared using sensitivity, block sensitivity (detection of a cow in need of treatment at least one of three days prior to treatment) and specificity, as well as area under the ROC curve (AUC). Furthermore, the effect of adjusting the frequency of occurrence of treatment in the training data was tested using random over- or under-sampling. The best model in the study was ET, with an AUC of 0.79 for mastitis treatments and 0.71 for lameness treatments without additional prior sampling, although the metrics for the GNB, LR, and RF models were not significantly lower. Over- and under-sampling had positive effects on the validation AUC during model training, but then failed to improve for the unsampled test data.

Using the best models from the first study, a second study was then used to examine the effects of a low frequency of treatment days (approximately 4-6%) on these models in practical use. Also, the formation of risk groups and risk times was tested with the goal of increasing the frequency in the data. For the study, data from two additional Holstein dairy cow farms with a similar timeframe (about 3 years) were used, which were comparable to each other in terms of housing type and sensor equipment. The criteria for forming the risk groups were earlier treatments in the current or previous lactation, information on somatic cell count from monthly milk recordings, and the first 100 days of lactation. This grouping increased the frequency of occurrence of treatments up to 13.5%. As a result, the positive predictive value increased significantly in the risk groups with a former treatment of the same disease in the same or previous lactation for both mastitis and lameness. The presence of treatment for a disease other than the one being predicted was able to increase the positive predictive value only for lameness. The positive predictive values increased in this way from the original 0.07 for both treatments to 0.2 for mastitis and 0.15 for lameness are nevertheless not sufficient by themselves to use these models in practical dairy farming to identify cows in need of treatment.

The sensor data used in the two studies were all collected under conditions very close to practical dairy farming. This proximity to practice is an important prerequisite for the commercial application of the developed prediction models. The own studies did not include all sensors used practically today, such as accelerometers for recording lying time, rumination sensors or position of the animal in the barn. However, when incorporating new sensors, care must always be taken to ensure that they also provide information specific to the disease being detected. The second study also showed that models should be validated on data from previously unknown farms, since the quality of the prediction usually decreases, or if possible be re-trained using data from the same farm. Furthermore, an adjustable threshold value of the model should be considered for practical use, based on which the ratio can be adjusted by the users themselves. The results obtained have shown that the models should not be used as the only tool to identify cows in need of treatment, but rather as an additional decision-making aid that complements the expert knowledge of the farmers.

6.2 Zusammenfassung

Durch den Strukturwandel in der Milchviehhaltung ist der Einsatz digitaler Assistenzsysteme zum Standard geworden. Der Erkennung von gesundheitlichen Problemen muss dabei besonderes Augenmerk geschenkt werden, um die Gesundheit und das Wohlbefinden der Kühe sicherzustellen. Die digitale Verarbeitung einzeltierbezogener Sensoren wird in der Praxis schon seit Jahrzehnten angewandt, und viele Studien beschäftigten sich mit der Eignung dieser Sensoren für die Entwicklung von Erkennungsmodellen für Gesundheitsprobleme. Die beiden in der Milchviehhaltung bedeutendsten und auch am häufigsten in diesen Studien untersuchten Erkrankungen sind Mastitis und Lahmheit. Die Auswahl der Sensordaten anhand deren Zusammenhang mit diesen Erkrankungen hat dabei einen entscheidenden Einfluss auf die Güte der Erkennungsmodelle. Versuche zur Mastitiserkennung fokussierten sich vor allem auf im Melkstand oder automatischen Melksystem erfasste Daten, u.a. Milchleistung, Milchzusammensetzung und elektrische Leitfähigkeit, da diese in verschiedenem Maße direkt von den Auswirkungen einer Mastitis beeinflusst werden, z.B. in Form von verringerter Milchmenge oder einer erhöhten Leitfähigkeit. Bei der Erkennung von Lahmheit stützten sich die Studien vorrangig auf die Messung von Aktivität und Verhalten. Über die Messung von Impulsen bzw. daraus abgeleiteten Schritten mit Pedometern und Accelerometern lassen sich Schlüsse auf das Verhalten der Kühe (Laufen, Stehen, Liegen) ziehen. Bei diesen Messungen sind die Verläufe über die Zeit aussagekräftiger als die einzelnen Messungen, sodass hier die sogenannte Feature Extraction häufig zum Einsatz kam, also das Generieren neuer Variablen aus den Rohdaten, wie z.B. Abweichungen und Verhältnisse zu vergangenen Messungen.

Einen weiteren Einfluss auf die Modellentwicklung hat die Zielvariable, auch als Goldstandard bezeichnet, da die Modelle mithilfe des statistischen Zusammenhangs zwischen Sensorvariablen und Zielvariable trainiert werden. In den betrachteten Studien wurden hierfür unterschiedliche Ansätze verfolgt: Aufzeichnung über vorgenommene veterinärmedizinische Behandlungen, eine festgelegte Bonitierung der Gesundheit, oder eine automatische Aufzeichnung der Zielvariable (z.B. somatische Zellzahl). Die Wahl der Methode hat jeweils Vor- und Nachteile bezüglich Reliabilität, Aufwand der Erfassung und Übertragbarkeit in die Praxis, und erschwert die Vergleichbarkeit der Ergebnisse der Studien.

Die Auswertung dieser Ergebnisse erfolgt dabei über den Vergleich der Vorhersage mit dem Goldstandard über die diagnostischen Kennzahlen Sensitivität, Spezifität, sowie dem positivem und negativem Vorhersagewert. Die beiden erstgenannten Werte hängen vor allem von dem vom Modellentwickler gewählten Grenzwert für die Erkennung ab und sind deshalb zwischen Studien nur bedingt zu vergleichen, weshalb hier die ROC-Kurve als Metrik für die Modellgüte gewählt werden sollte. Eine für diese Dissertation besonders kritische Kennzahl war der positive Vorhersagewert, der auf die Praxis übertragen die relative Anzahl der Fehlalarme bestimmt. Dieser wird vor allem von der Auftretenshäufigkeit der Erkrankungen bzw. Behandlungen beeinflusst, und wurde daher in den beiden im Rahmen dieser Dissertation veröffentlichten Studien besonders beachtet.

Das Ziel der ersten Studie war die Entwicklung von Machine Learning-Modellen zur Erkennung notwendiger Mastitis- und Lahmheitsbehandlungen bei Holstein-Kühen basierend auf Daten von bereits in der Praxis etablierten Sensoren. Diese enthielten neben den erfassten Behandlungen die Melkdaten, Pedometer-Aktivität, Futter- und Wasseraufnahme, sowie Lebendgewicht von 167 individuellen Kühen über einen Zeitraum von 40 Monaten. Die aus diesen Daten abgeleiteten Variablen wurden mithilfe der Methoden Random Forest Feature Importance, Pearson-Korrelation und Sequential Forward Feature Selection vorselektiert und dann zum Training diverser Machine Learning-Modelle genutzt, u.a. Logistische Regression (LR), Support Vector Machine, K-nearest Neighbors, Gaussian Naïve Bayes (GNB), Extra Trees (ET) und Random Forest (RF). Diese Modelle wurden mithilfe der Sensitivität, Block-Sensitivität (Erkennung einer behandlungsbedürftiger Kuh an mindestens einem von drei Tagen vor einer Behandlung) und Spezifität, sowie der Fläche unter der ROC-Kurve (AUC) verglichen. Des Weiteren wurde die Wirkung einer Anpassung der Auftretenshäufigkeit der Behandlung in den Trainingsdaten mittels Random Over- bzw. Undersampling getestet. Das beste Modell der Studie war ET mit einer AUC von 0,79 für Mastitisbehandlungen und 0,71 für Lahmheitsbehandlungen ohne zusätzliches vorheriges Sampling, wobei die Metriken für die Modelle GNB, LR und RF nicht signifikant niedriger waren. Das Over- und Undersampling hatte zwar positive Effekte auf die AUC beim Modelltraining, konnte dann bei den nicht gesampelten Testdaten aber keine Verbesserung mehr erzielen.

Mit den besten Modellen der ersten Studie wurde dann in einer zweiten Studie dazu verwendet, die Auswirkungen einer niedrigen Auftretenshäufigkeit von Behandlungstagen (ca. 4-6 %) auf diese Modelle in der praktischen Anwendung zu untersuchen; ebenfalls wurde die Bildung von Risikogruppen und -zeiten mit dem Ziel einer Erhöhung der Auftretenshäufigkeit getestet. Für die Studie wurden Daten von zwei zusätzlichen Holstein-Milchkuhbetrieben mit ähnlichem zeitlichem Umfang (ca. 3 Jahre) herangezogen, die von der Haltungsform und der Ausstattung mit Sensoren untereinander vergleichbar waren. Die Kriterien für die Bildung der Risikogruppen waren Vorbehandlungen in der laufenden oder vorherigen Laktation, Informationen über die somatische Zellzahl aus der Milchleistungsprüfung, und die ersten 100 Laktationstage. Durch diese Eingrenzung konnte die Auftretenshäufigkeit der Behandlungen auf bis zu 13,5 % erhöht werden. Der positive Vorhersagewert erhöhte sich infolgedessen signifikant in den Risikogruppen mit vorheriger Behandlung der gleichen Erkrankung in derselben oder der vorherigen Laktation sowohl für Mastitis als auch Lahmheit. Ein Vorhandensein der Behandlung einer anderen Erkrankung als die zu vorhersagende konnte den positiven Vorhersagewert nur für Lahmheit erhöhen. Die auf diese Weise erhöhten positiven Vorhersagewerte von ursprünglich 0,07 bei beiden Behandlungen auf 0,2 für Mastitis und auf 0,15 für Lahmheit reichen alleine dennoch nicht aus, um diese Modelle in der praktischen Milchviehhaltung zur Erkennung behandlungsbedürftiger Kühe einzusetzen.

Die in den beiden Studien genutzten Sensordaten wurden alle unter sehr praxisnahen Bedingungen erfasst. Diese Nähe zur Praxis ist eine wichtige Voraussetzung für die kommerzielle Anwendung der entwickelten Vorhersagemodelle. Die eigenen Studien beinhalteten nicht alle heutzutage praktisch genutzten Sensoren, wie z.B. Accelerometer zur Erfassung der Liegezeit, Wiederkausensoren oder Position des Tieres im Stall. Bei der Einbindung neuer Sensoren ist jedoch immer darauf zu achten, dass diese auch für die zu erkennende Erkrankung spezifische Informationen liefern. Auch hat die zweite Studie gezeigt, dass Modelle an Daten bisher unbekannter Betriebe validiert werden sollten, da die Güte der Vorhersage meistens abnimmt, oder wenn möglich anhand von Daten desselben Betriebes neu trainiert werden. Weiterhin sollte für den Praxiseinsatz ein einstellbarer Grenzwert des Modells in Betracht gezogen werden, anhand dessen das Verhältnis von den Nutzern selbst eingestellt werden kann. Die erzielten Ergebnisse haben gezeigt, dass die Modelle nicht zur selbstständigen Identifikation behandlungsbedürftiger Kühe genutzt werden können, und stattdessen als zusätzliche Entscheidungshilfe gesehen werden, die das Expertenwissen der Landwirte/-innen ergänzen.

7 Acknowledgements/Danksagung

Hiermit möchte ich mich ganz herzlich bei allen bedanken, die direkt oder indirekt zum Abschluss dieser Arbeit beigetragen haben.

Zu Anfang natürlich vielen Dank an Frau Prof. Sauerwein für die Aufnahme an die Uni Bonn und in die Arbeitsgruppe. Auch wenn ich mit meinem Dissertationsthema in der Abteilung ein bisschen exotisch war, gab es im Rahmen des PhD-Seminars immer konstruktive und hilfreiche Gespräche. Auch danke ich Herrn Prof. Büscher für die Übernahme des Korreferates und natürlich auch die Unterstützung als Co-Autor der beiden Veröffentlichungen und der Tagungsbeiträge.

Dann natürlich ein riesiges Dankeschön an Frau Dr. Müller, zunächst, dass Sie mich für das PaRADigMa-Projekt als Mitarbeiter ausgewählt haben, und dann für die stetige fachliche Unterstützung und Mitarbeit an den Vorarbeiten und Veröffentlichungen auch nach meiner Zeit in Bonn, was so sicherlich nicht selbstverständlich ist.

Für seinen Beitrag an den Veröffentlichungen auch ein herzliches Dankeschön an Prof. Rietz, der mir mit seinem scharfen Blick für statistische Zusammenhänge eine große Hilfe war. Ein Dank für die Zusammenarbeit im Projekt geht an Frau Dr. Treitel und Frau Otte für die sehr enge Zusammenarbeit, trotz aller Höhen und Tiefen und den Rückschlägen durch die Pandemie. Auch möchte ich an dieser Stelle dem Team von iDIGMA um Herrn Pörschmann danken. Von euch habe ich nochmal eine Menge über Programmierung und Machine Learning gelernt, und die Zusammenarbeit war essentiell für die Ergebnisse der Veröffentlichungen.

Ich danke allen MitarbeiterInnen und (ehemaligen) PhDs der Abteilung Physiologie für die tolle Zeit innerhalb und außerhalb der Uni, insbesondere Isabell, Inga, Thomas, Katharina, Barbara, Hannelore, Iris, Laura, Christina, Kathi, Morteza, Ruben, Taher, Yang, Rafaela und allen anderen. Ihr habt dazu beigetragen, dass mir meine Zeit in Bonn vor allem menschlich in Erinnerung bleiben wird.

Danke an Frau Prof. Traulsen und der Abteilung Systeme der Nutztierhaltung der Uni Göttingen dafür, dass ich auch neben meiner neuen Arbeitsstelle den nötigen Raum bekommen habe, um diese Dissertation zum Abschluss bringen zu können.

Und zum Schluss noch ein großes Dankeschön an meine Eltern, die mich auf meinem bisherigen Weg unterstützt haben und mir neben meiner Arbeit immer „Ferien auf dem Bauernhof“ ermöglicht haben, und auch meiner restlichen Familie für den Ansporn.

8 Publications and Related Work

Post, C.; Rietz, C.; Büscher, W.; Müller, U. (**2023**): Importance of low daily risk for the prediction of treatment events of individual dairy cows in need with sensor systems. Poster Presentation, *International Conference on Precision Dairy Farming.* Wien, 29.08.-02.09.2022.

Ghaffari, M.; Monneret, A.; Hammon, H.; Post, C.; Müller, U.; Frieten, D.; Gerbert, C.; Dusel, G.; Koch, C. (**2022**): Deep convolutional neural networks for the detection of respiratory disease and scours in preweaned dairy calves using data from automated milk feeders. *Journal of Dairy Science* 105 (12), 9882-9895. https://doi.org/10.3168/jds.2021-21547

Post, C.; Rietz, C.; Büscher, W.; Müller, U. (**2021**): The Importance of Low Daily Risk for the Prediction of Treatment Events of Individual Dairy Cows with Sensor Systems. *Sensors* 21 (4), 3863. https://doi.org/10.3390/s21041389

Post, C.; Rietz, C.; Büscher, W.; Müller, U. (**2020**): Using Sensor Data to Detect Lameness and Mastitis Treatment Events in Dairy Cows: A Comparison of Classification Models. *Sensors* 20 (14), 1389. https://doi.org/10.3390/s20143863

Ghaffari, M.; Jahanbekam, A.; Post, C.; Sadri, H.; Schuh, K.; Koch, C.; Sauerwein, H. (**2020**): Discovery of different metabotypes in overconditioned dairy cows by means of machine learning. *Journal of Dairy Science* 103 (10), 9604-9619. https://doi.org/10.3168/jds.2020-18661

Post, C.; Müller, U.; Büscher, W. (**2019**): Identifikation behandlungsbedürftiger Milchkühe mittels Sensordaten: Vergleich statistischer und Machine-Learning-Methoden zur Vorhersage von Klauenbehandlungen bei Milchkühen. *14. Tagung: Bau, Technik und Umwelt in der landwirtschaftlichen Nutztierhaltung*. Bonn, 24.-26.09.2019.