



University of
Nottingham

UK | CHINA | MALAYSIA

The application of machine learning to predict disease, production and reproduction outcomes from the transition period of dairy cattle

Efterpi Tsantila

Thesis Submitted to the University of Nottingham for the
degree of Doctor of Philosophy

March 2023

Abbreviations

ANN – Artificial Neural Networks

BCS – Body Condition Score

KNN – K-nearest neighbours

LDA – Left Displaced Abomasum

RF – Rumen Fill

RFM – Retained Foetal Membranes

SVM – Support Vector Machines

THI – Temperature Humidity Index

Abstract

Data collected under a transition period monitoring service, from 133 herds over the course of 2 years, were utilised in order to build predictive models for disease, production and reproductive outcomes. Both cow level and pen level variables were used as potential predictor variables, while a variety of methods including linear regression, decision tree, random forest, multiple adaptive regression splines (MARS) and artificial neural networks (ANNs) for continuous outcomes; and logistic regression, decision tree, random forest, ANNs, support vector machines (SVM) and naïve Bayes for binary outcomes. Models generating predictions on both the individual and the herd/quarter-year group level were produced.

Various health outcomes (occurrence or not of milk fever, LDA, RFM and metritis, as well as a collective disease status outcome) were explored. On the individual lactation level all models lacked predictive value; the best performing model was that for collective disease outcome, with a kappa value (measuring agreement between predicted and observed data) of 0.16, although accuracy was relatively high at 0.86. When building models on the herd/quarter-year level, the best performing model was for the milk fever outcome; predicted group prevalence of milk fever explained around 44% of variation in observed prevalence, suggesting relatively low predictiveness. Better prediction performance was revealed when individual lactation level model predictions were aggregated at herd-quarter-year level and compared with observed aggregated disease prevalences; just over two thirds (67%) of the variation in

observed outcome was explained by the aggregated predictions for occurrence of metritis.

Moving to the reproductive outcomes, probability of insemination success, as well as time from calving to successful insemination, were investigated. Kappa values for the former ranged from 0.04 to 0.17, while the R^2 value describing the relationship between aggregated predictions and actual aggregated values on the herd-quarter-year level was found to be 0.37. When building models on the aggregated level instead, the maximum R^2 value was found to be at 0.24 for the MARS model. Regarding the time to insemination outcome, the maximum R^2 value calculated was found just at 0.024 for the linear regression, indicating very low predictive value. Interestingly, while no strong predictive value was found in these models, inferential models were built for those same outcomes and found strong associations between insemination success and lactation number, calving month, as well as calf mortality; and between time to insemination and metritis, corrected protein percentage in milk, calving month and lactation number.

For the production outcomes, models for both the 305-day predicted milk yield and the daily residual milk yield (difference between observed yield for a given cow on a given day, and expected daily yield based on lactation curve shape for the appropriate parity in the cow's herd) were built. For the individual lactation level of the 305-day milk yield models, R^2 values were again relatively low, at around 0.1, with the exception of the random forest that had a value of 0.34. Similarly, when comparing aggregated predictions using the individual lactation models and actual aggregated values, the R^2 was as low as 0.024. Building models on a herd/quarter-year level yielded similar results with R^2

ranging from 0.12 to 0.39 for the linear regression and the random forest models respectively. For the daily residual milk yield outcome, the R^2 values of individual lactation models had a maximum value of 0.21 for the random forest model, while regarding the aggregated models the maximum value was at 0.134. When using the individual lactation level models to compare aggregated predictions with actual aggregated values the R^2 was found to be at 0.34. Not unlike our results on the reproductive outcomes, various strong inferential associations were identified for these outcomes, regardless of the predictive models' performance.

Since transition management is key to successful dairy farming, machine learning would be useful both in terms of predicting which individuals may get a negative outcome and possibly require enhanced observation or other preventive interventions, and also in providing a potential monitoring metric. The latter would mean that even if individual predictions are not good, knowing the predicted disease prevalence, insemination success or yield in each group's cows could be used as a measure of overall transition "success". Overall, very few of our models were predictive enough to be useful in either context most likely, but that could perhaps improve if we had other data available such as sensor data or history from previous lactations. The project as a whole provides a good example of why it is important to be cautious with choice of prediction performance metrics and avoid accuracy as the main measure in unbalanced data, and of how in many areas inferential models can find strongly significant associations but still generate very poor predictions when applied to new data.

Acknowledgements

I'd like to thank my four supervisors Chris Hudson, Laura Randall, Martin Green, and John Remnant for their support and guidance throughout the PhD. An extra special thank you to Chris who is the most patient and understanding supervisor a PhD student can ask for.

I'd also like to thank the University of Nottingham and Premier Nutrition for giving access to their data and funding my project.

I'm very grateful for my colleagues at the University of Nottingham, with whom I had the pleasure of working with (at least before Covid drove us apart), as well as all the friends I made along the way, especially Eli, Karla, Sophie, Veronica, Emma, Danny, Bobby and Charlie.

Of course, I'd like to thank my family Alice and Zizel, for always being there. Finally, a special thanks to my therapist Olga without whom I'd still be stuck. I should have met you sooner.

Contents

Abbreviations	2
Abstract.....	3
Acknowledgements	6
Chapter 1 - Introduction	22
1.1 Physiology of Transition	25
1.1.1 Immune system	25
1.1.2 Metabolism	26
1.2 Management	28
1.2.1 Body Condition Score	28
1.2.2 Rumen Fill Score	29
1.2.3 Housing.....	29
1.2.4 Heat Stress	30
1.3 Economic Impact.....	32
1.4 Statistical Methods	33
1.4.1 Explanatory vs Predictive Modelling	33
1.4.2 Building a predictive model using machine learning	33
1.4.3 Assessing the model's predictive power	36
1.5 Machine learning in the dairy industry	39
1.6 Conclusions.....	62
Chapter 2 – Data collection and methodology	64
2.1 Methods.....	64
2.1.1 Source of data	64
2.1.2 Datasets.....	67
2.1.3 Pre-processing and analysis.....	73
2.1.3.2 Inferential and Predictive analytics	74
2.1.3.4 Predictive Models	75

2.1.3.4.1 Binary Outcomes	78
2.1.3.4.2 Continuous Outcomes	83
2.2 Machine learning Approaches	84
2.2.1 Naive Bayes.....	84
2.2.2 Neural Networks	85
2.2.3 Support Vector Machines.....	87
2.2.4 Decision Trees	89
2.2.5 Random Forests	90
2.2.6 <i>K</i> -nearest neighbours.....	92
2.3 Inferential Models	93
2.3.1 Binary Outcomes	93
2.3.2 Continuous Outcomes	94
Chapter 3 - Descriptive statistics.....	95
3.1 Original Dataset.....	95
3.1.1 Disease Distribution	96
3.1.2 Cow level score variables	99
3.1.3 Pen level variables.....	102
3.2 Milk Records.....	108
3.3 Insemination and Event Records.....	111
3.4 Discussion	116
3.4.1 Conclusions	120
Chapter 4 – Prediction of Disease Status	121
4.1 Introduction.....	121
4.1.1 Periparturient diseases	121
4.1.2 Left displaced abomasum	122
4.1.3 Milk Fever	123
4.1.4 Mastitis.....	124

4.1.5 Lameness	125
4.1.6 Retained Foetal Membranes.....	126
4.1.7 Metritis	127
4.1.8 Twinning	128
4.2 Methods.....	130
4.2.1. Lactation Level.....	130
4.2.2 Lactations per herd/quarter-year level models.....	134
4.3 Results	136
4.3.1 Individual Disease Outcomes	136
4.3.1.1 Individual Lactation level	137
4.3.1.2 Individual lactation models predicting on aggregated lactations per herd/quarter-year	156
4.3.1.3 Models built using data aggregated at herd-quarter-year level	159
4.3.2 Collective Disease Status Outcome.....	164
4.3.2.1 Individual lactation disease models	164
4.3.2.2 Individual lactation disease model making predictions on an aggregated level.....	172
4.3.2.3 Aggregated Herd/quarter-year level models.....	174
4.4 Discussion	178
Chapter 5 – Prediction of reproductive outcomes	191
5.1 Introduction.....	191
5.2 Methods.....	197
5.2.1 Data preparation	197
5.2.2 Analysis	201
5.3 Results	205
5.3.1 Predictive models	205
5.3.1.1 Study A.....	205

5.3.1.2 Study B.....	214
5.3.1.3 Study C	219
5.3.2 Inferential Models	222
5.2.3.1 Study D	222
5.2.3.2 Study E.....	224
5.4 Discussion	230
Chapter 6 – Prediction of production Outcomes	239
6.1 Introduction.....	239
6.2 Methods.....	248
6.3 Results	252
6.3.1 Method A-Individual cow level and models	252
6.3.1.1 Outcome: Predicted 305-day lactation milk yield	252
6.3.1.2 Outcome: Residual Daily Milk Yield	259
6.3.2 Method B– Individual cow models used for predictions at herd- quarter-year level	267
6.3.2.1 Outcome: Predicted 305-day milk yield	267
6.3.2.2 Outcome: Residual milk yield.....	269
6.3.3 Method C – Herd-quarter-year level models	272
6.3.3.1 Outcome: Mean predicted 305-day milk yield	272
6.3.3.2 Outcome: Residual milk yield.....	276
6.4 Discussion	282
Chapter 7 – General Discussion	290
7.1 Summary of Results	290
7.2 Discussion	291
7.2.1 Predictiveness of Models	291
7.2.2 Predictive vs Inferential.....	293
7.2.3 Metrics	294

7.2.4 Limitations of the study	298
7.2.5 Possible Future Research.....	299
7.3 Conclusions.....	302
References.....	304
Appendices	359
Appendix I	359
Appendix II	361

Figures

Figure 3.1 The number of total lactations included in each one herd per year of recording.....	96
Figure 3.2 Overall prevalence of diseases/conditions across 13,244 lactations in 79 UK herds.....	97
Figure 3.3 Seasonal disease prevalence (%) of diseases/conditions across 13,244 lactations in 79 UK herds.....	98
Figure 3.4 Diseases and conditions prevalence distribution across 13,244 lactations in 79 UK farms.....	99
Figure 3.5 BCS distribution in dry and fresh cows throughout 13,244 lactations in 79 UK farms.....	101
Figure 3.6 BCS change from dry to fresh cows in data of 13,244 lactations in 79 UK farms.....	102
Figure 3.7 Stocking density distribution on 2,787 pens across 136 dairy cow herds.....	106
Figure 3.8 Monthly THI distribution in 2,787 pens across 136 herds.....	107
Figure 3.9 Daily milk yield distribution based on 564,962 milk recordings across 43,173 cows.....	109
Figure 3.10 Percentage of protein in milk based on 564,962 milk recordings across cows.....	43,173 110

Figure 3.11 Percentage of Butterfat in milk based on 564,962 milk recordings across	43,173
cows.....	111
Figure 3.12 Calving to first service interval, based on 54,443 first inseminations.....	113
Figure 3.13 Calving interval based on 41,186 lactations.....	114
Figure 3.14 Calving to conception distribution based on 43,507 lactations...	115
Figure 4.1 Scatterplot of aggregated predictions vs actual percentage of metritis diagnosis per herd/quarter-year using the Naïve Bayes model	158
Figure 4.2 Metric comparison of all lactation-level collective disease models after using up-sampling with 95% confidence intervals.....	171
Figure 4.3 Scatterplot of predictive vs actual collective disease diagnosis probability per herd/quarter-year.....	172
Figure 4.4 Pearson's correlation coefficients between aggregated predicted and observed outcomes (per herd-quarter-year) across model types for all individual disease outcomes and for collective disease status ("Disease_Status") using models built on individual lactation data.....	174
Figure 4.5 Histogram of collective disease diagnosis distribution per herd each quarter-year.....	175
Figure 4.6 Metric comparison of different models of the collective disease percentage outcome.....	177

Figure 5.1 Comparison of all metrics for all different methods predicting insemination success.....	212
Figure 5.2 Scatterplot of actual insemination success per group vs the predicted insemination success per group.....	213
Figure 5.3 Insemination success percentage per herd/month group distribution.....	216
Figure 5.4 Histogram of DIM at the time of conception.....	219
Figure 5.5 Kaplan-Meier survival curve of time to pregnancy in cows.....	224
Figure 5.6 Global Schoenfeld test and individual Schoenfeld tests for each independent variable included in the Cox Proportional Hazards Model with outcome time to pregnancy in cows conceiving at <100 DIM.....	228
Figure 5.7 Slopes of Hazards Ratios over time.....	229
Figure 6.1 R^2 , RMSE and MAE values of all models (excluding ANN) predicting 305-day milk yield, when applied on the test set.....	259
Figure 6.2 Distribution of milk yield residuals on the final dataset of a total 15,742 data points.....	260
Figure 6.3 R^2 , RMSE and MAE values of all models predicting daily residual milk yield on the test set.....	267
Figure 6.4 Scatter plot of predicted mean 305-day milk yield vs observed mean 305-day milk yield per herd-quarter-year group for the random forest model.....	269

Figure 6.5 Scatter plot of predicted residual milk yield vs observed residual milk yield per herd-quarter-year group for the random forest model.....	270
Figure 6.6 R^2 , RMSE and MAE values of all models predicting mean predicted 305-day milk yield per herd-quarter-year group with their 95% confidence intervals, as applied on the test set.....	276
Figure 6.7 R^2 , RMSE and MAE values of all models predicting mean daily residual milk yield per herd-quarter-year group on the test set.....	281
Figure A2.1 Rumen fill distribution based on 28,480 dry cows.....	361
Figure A2.2 Rumen fill distribution based on 43,185 fresh cows.....	362
Figure A2.3 Hock Hygiene distribution based on 12,847 lactations.....	362
Figure A2.4 Types of pens for both dry and fresh cows, based on 2,787 pens.....	363
Figure A2.5 Feed Fence space available per cow, based on 2,787 pens.....	363
Figure A2.6 Feed Fence space available separately per dry and fresh cows, based on 2,787 pens.....	364
Figure A2.7 Water Trough space available per cow, based on 2,787 pens.....	364
Figure A2.8 Water Trough space available separately per dry and fresh cows, based on 2,787 pens.....	365

Figure A2.9 Neck Rail Height available based on data on 2,787 pens.....365

Tables

Table 1.1 Kappa value interpretation (Vierra and Garrett, 2005).....	38
Table 2.1 Variables available in all the original datasets.....	68
Table 2.2 THI values and interpretation*	72
Table 2.3 Confusion matrix.....	81
Table 2.4 Definition of commonly used model metrics.....	82
Table 3.1 Variable distribution on cow comfort data on 2,787 pens across 136 dairy cow herds.....	104
Table 4.1 Potential predictor variables considered in models with the outcome of disease occurrence at lactation level.....	131
Table 4.2 Predictive variables considered in final analysis for second set of models.....	135
Table 4.3 Predictive variables used for machine learning models predicting individual disease outcomes at lactation level.....	136
Table 4.4 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Milk Fever outcomes, before up-sampling.....	138
Table 4.5 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting LDA outcomes, before up-sampling.....	140
Table 4.6 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting RFM outcomes, before up-sampling...	142

Table 4.7 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Metritis outcomes, before up-sampling.	144
Table 4.8 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Milk Fever outcomes, after up-sampling.....	148
Table 4.9 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting LDA outcomes, after up-sampling.....	150
Table 4.10 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting RFM outcomes, after up-sampling.....	152
Table 4.11 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Metritis outcomes, after up-sampling....	154
Table 4.12 Metrics of best performing models for each individual disease....	157
Table 4.13 Predictor variables included in final predictive models for various individual disease outcomes aggregated at herd-quarter-year level.....	160
Table 4.14 R2 values of all methods of individual disease percentage outcomes.....	163
Table 4.15 All metrics for collective disease outcome models as calculated on both the training and the test set, before up-sampling.....	166
Table 4.16 All metrics for collective disease outcome models as calculated on the training and the test set, after up-sampling.....	169
Table 4.17 Metrics of all methods of disease diagnosis percentage per herd per quarter-year on both the training and the test sets.....	176

Table 5.1 Potential Predictive Variables for Datasets W, X and Y.....	199
Table 5.2 Dataset Z Potential Predictive Variables.....	200
Table 5.3 Reproductive outcome Studies and Datasets used in each one...	201
Table 5.4 Comparison of Kappa values on the test set before and after the implementation of up-sampling.....	206
Table 5.5 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Insemination Success outcomes, before up-sampling.....	207
Table 5.6 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Insemination Success outcomes, after up-sampling.....	210
Table 5.7 Aggregated variables available for analysis, along with missing data.....	215
Table 5.8 Metric of all models predicting insemination success percentage per herd/month group on both the training and the test sets.....	218
Table 5.9 Metric of all models predicting DIM at conception for both the training and the test set.....	221
Table 5.10 Odds Ratios with 95% CI for mixed effects logistic regression model of insemination success.....	223

Table 5.11 Hazards Ratios from Cox Proportional Hazards model with outcome time to pregnancy in cows conceiving at <100 DIM.....	226
Table 6.1 VIF values for all possible predictive variables as calculated when fitting a linear regression model on the predicted 305 milk yield with all variables included.....	254
Table 6.2 R^2 , RMSE and MAE values of all models (excluding ANN) predicting 305-day milk yield, when applied on both the training and the test set.....	258
Table 6.3 VIF values for all possible predictive variables as calculated when fitting a linear regression model on the residual milk yield with all variables included.....	262
Table 6.4 R^2 , RMSE and MAE values of all models predicting milk yield residuals, when applied on both the training and the test set.....	266
Table 6.5 VIF values for all possible predictive variables as calculated when fitting a linear regression model on mean predicted 305-day milk yield per herd-quarter-year group with all independent variables included.....	273
Table 6.6 R^2 , RMSE and MAE values of all models predicting mean 305-day milk yield per herd-quarter-year group, when applied on both the training and the test set.....	275
Table 6.7 VIF values for all possible predictive variables as calculated when fitting a linear regression model on the mean milk yield residuals per herd-quarter-year group, with all independent variables included.....	278

Table 6.8 R^2 , RMSE and MAE values of all models predicting mean daily residual milk yield per herd-quarter-year group on both the training and the test set.....	280
--	-----

Chapter 1 - Introduction

The transition period is commonly defined as 3 weeks pre to 3 weeks post calving (Menta et al., 2022) and has been recognised as a phase of the utmost importance for the dairy cow (AlZahal et al., 2014, Esposito et al., 2014).

The majority of metabolic, as well as many infectious diseases in the cow are associated with the transition period. Examples of these include milk fever, ketosis, retained foetal membranes, metritis and displaced abomasum and are manifested mainly within the first 2 weeks postpartum (Caixeta and Omontese, 2021). There are however, other diseases such as endometritis that are clinically diagnosed at a later stage, yet they are linked to this period (Melendez and Risco, 2005). Metabolic disorders, infections of the mammary gland and reproductive disorders that manifest around this time, are also important issues, as their respective incidences have been reported to be from 7.8% to 16.8%, 2.8% to 12.6% and 6.7% to 19.2% (Wankhade et al., 2017). Therefore, since the importance of monitoring and addressing metabolic disorders like hypocalcaemia during the transition period has been emphasized to enhance the profitability of dairy herds, it is a phase of great interest for the dairy cow industry (Saed et al., 2020).

The transition period, being associated with such conditions, is the source of great economic losses due to efforts of preventing disease, managing and treating cows. However, at the same time our current understanding on how to prevent them is lacking (Eckel and Ametaj, 2016). In total, up to 50% of the cows develop either a metabolic or infectious disease during this period, which is attributed to a decline in immune function, and metabolic events such as

reduction in feed intake, negative energy balance, and insulin resistance due to calving (LeBlanc, 2010, Melendez and Risco, 2005, Moreira et al., 2018). By the end of the transition period milk yield, along with milk fat, protein and lactose increase much faster than feed intake, while at the same time in most management systems the cows' diet changes from mainly forage-based to being concentrate-rich. These demands for milk production along with the necessary nutritional adaptation predispose the cow to a negative energy balance (NEB) state (Bekuma, 2019, Bertoni et al., 2009). The dry matter intake (DMI) during the transition period can drop by 10-30% (Esposito et al., 2014), while the energy requirements rise (Imhasly et al., 2015). The overall profitability of the cow depends on a successful transition period, as nutritional and management factors can prevent them from producing milk to their full potential as well as cause decreased reproductive performance, increased treatment costs, and even a shortened lifespan for the cows (Hailemariam et al., 2014).

A great focus has been given by researchers into developing methods monitoring transition period cows (Lukas et al., 2015). As a result, techniques such as body condition scoring and blood or urine sampling have been developed to help provide insight on the transition cows' health status (Hachenberg et al., 2007, LeBlanc, 2010). Laboratory measurements most frequently used are serum non esterified fatty acids (NEFA) pre-partum, blood β -hydroxybutyrate (BHBA) post-partum and serum Ca the days around calving (Vergara et al., 2014). It can be argued, however, that blood and urine sample have the drawback of increasing labour and cost when routinely implemented on farm (Hachenberg et al., 2007). Other more low cost

monitoring methods include body condition scoring, which is associated with energy balance, and locomotion scoring, which was found to be linked with post-partum disease (Calderon and Cook, 2011, Cook, 2003, Espejo and Endres, 2007, Hoedemaker et al., 2009, Ingvarlsen, 2006, Machado et al., 2011). In addition, risk factors such as milk yield during the previous lactation, dry period and gestation length, twinning, stillbirths and dystocia can also provide useful information on the cows' expected health and performance, however they have not been used collectively into predictive models (Fleischer et al., 2001, Ingvarlsen, 2006, LeBlanc et al., 2006). Lukas et al (2015) developed 3 transition period monitors that used daily milk yield in order to evaluate the success of the transition period on both herd and individual cow level. A recent study (Wisnieski et al., 2019) has used a variety of predictive models using various biomarkers to predict metabolic stress in the transition period and has suggested that predictive modelling should be applied to other outcomes, including culling rates, reproductive and milk production outcomes and could potentially be used on farm through monitoring applications. It is without doubt, that such models can predictive models can be of practical use to the farmers when used as management tools and it is the primary goal of this project to investigate this potential, develop and present practical models that can be then used on-farm to ensure an improved transition period for the cows.

1.1 Physiology of Transition

1.1.1 Immune system

There is a reduction in immune function around the calving period (Overton and Waldron, 2004). The cause of this phenomenon is multifactorial and it results in the cows being particularly susceptible to invading pathogens and infectious diseases such as mastitis, during the periparturient period (Sordillo, 2016).

The aetiology of immunosuppression around calving is complex. Maternal immune responses are naturally suppressed up until calving, in order to prevent a reaction against the allogeneic conceptus. Among the factors that cause this immunosuppressive response is progesterone secretion and regulatory immune cell differentiation (Esposito et al., 2014). Furthermore, the periparturient period is characterized by intense lipolysis in adipose tissues, leading to the release of free fatty acids into circulation, which can impact immune function (Contreras et al., 2017). Metabolic profiling during the periparturient period has shown changes in serum concentrations of macro minerals and a drop in feed intake, which may contribute to immunological dysfunction (Kabir et al., 2022). Additionally, the association between prepartum feeding behaviour and periparturient health disorders highlights the importance of nutrition in maintaining immune function during this critical period (Luchterhand et al., 2016). The severe negative energy balance experienced by dairy cows during early lactation, due to insufficient dry matter intake to meet the demands of high milk production, has been linked to impaired immune function (Gümen et al., 2011, Gross et al., 2013).

The reported negative impact of ketosis on immune responses could be linked to the effects of fatty liver on the immune function (Overton and Waldron, 2004). Leukocytes in ketotic cows lack in chemotactic differentials and leucocytes in general show limited chemotactic capacity in a ketotic environment (Esposito et al., 2014). Immune function deficiencies have also been reported in diseases other than ketosis. Kimura et al. (2002) also reported declined chemotactic capacity and cellular killing function in neutrophils in cows with retained placenta.

1.1.2 Metabolism

The role of the liver in adapting metabolic pathways and supporting lactation is key to a successful transition period. Dietary supplements of folic acid and vitamin B12 on the metabolism of dairy cows in early lactation, showing improved metabolic efficiency and liver function (Graulet et al., 2007). Furthermore, Li et al. (2020) emphasized the crucial role of the liver in metabolic adaptation to support pregnancy and lactation through nutrient coordination and interconversion, especially during the transition from late gestation to early lactation, while Ringseis et al. (2014) discussed the molecular mechanisms underlying liver-associated diseases in transition dairy cows, underscoring the importance of understanding these mechanisms to prevent liver disorders and enhance production.

It was proposed by Allen and Bradford (2009) that the oxidation of fuels that occurs in the liver elevates adenosine triphosphate concentrations that trigger the hepatic vagus nerve and send a message to the brain's feeding centre to reduce feed intake. These fuels include fatty acids, propionate, lactate and amino acids, therefore NEFA mobilization, which is common during the

transition period, may cause the DMI decrease. NEFA mobilization is in turn amplified by a decrease in plasma insulin concentration and insulin sensitivity by up to 50% (Wankhade et al., 2017).

Accumulation of triglycerides within the hepatocytes can lead to fatty liver syndrome or ketosis (Eckel and Ametaj, 2016) and is considered to be caused by the negative energy balance (NEB) state during the early stages post-partum (Melendez and Risco, 2005). Fatty liver is characterized by the storage of triglycerides within hepatocytes (Eckel and Ametaj, 2016). At least half of the cows could be experiencing subclinical ketosis during the first month after calving (Esposito et al., 2014), while ketosis has been associated with conditions such as metritis and displacement of the abomasum (LeBlanc, 2010). Decreased DMI seemed to significantly influence the development of both milk fever and retained foetal membranes (Kimura et al., 2006). Therefore, maintaining appropriate body weight during the transition period is pivotal for the cow's health and performance. To do so it is important to minimize all factors that could affect feed intake. For instance, moving and regrouping cows, especially around calving is shown to reduce DMI and subsequently delay calving, increase calf mortality and the incidence of retained placenta (Schirmann et al., 2011, Nordlund and Cook, 2004, von Keyserlingk et al., 2008).

1.2 Management

1.2.1 Body Condition Score

The Body Condition Score (BCS) reflects the nutritional status of the cow and therefore the stage of lactation, but can also be affected by a number of cow-level factors, such as parity, age, season of calving and genetics, as well as on a herd-level, such as stocking rate, and type of diet (Berry et al., 2003, Berry et al., 2006, Butler, 2014, Coffey et al., 2004, Koenen et al., 2001, Macdonald et al., 2008, Lean et al., 2022, McCarthy et al., 2007, Pryce et al., 2001, Pryce and Harris, 2006, Roche et al., 2006, Roche et al., 2007).

BCS can be a very quick and effective monitoring tool for nutritional management and health outcomes (LeBlanc, 2010, Melendez and Risco, 2005). Low scores correspond to emaciation and high scores to obesity (Roche et al., 2009). Although traditionally a BCS of 3.5 to 3.75 (using a 1 to 5 scale) was considered ideal at dry-off, more recent studies have suggested that 3.0 to 3.2 or even lower than 3.0 is a more efficient aim (Contreras et al., 2004, LeBlanc, 2010, Melendez and Risco, 2005, Overton and Waldron, 2004). This could potentially be attributed to the link between decreased DMI and a high BCS (Esposito et al., 2014, Hayirli et al., 2002, Overton and Waldron, 2004). Overall, studies suggest that a BCS lower than the traditional 3.5 - 4 is preferred in order to have the desired transition period outcomes (Overton and Waldron, 2004). As the loss of body condition in the weeks postpartum is related to NEB (Roche et al., 2009) the BCS score is a useful, easy to apply on-farm health indicator (Danicke et al., 2018). Reducing BCS or body weight of dry cows is not recommended at any point of the dry period (Melendez and

Risco, 2005). Higher BCS used to be the target, since it has been established that it is associated with greater milk yield potential (Zahrazadeh et al 2017, Berry et al., 2007; Jamali Emam Gheise et al., 2017), however it is also linked with higher serum NEFA levels and increased incidence of reproductive and health issues (Berry et al., 2007, Jamali Emam Gheise et al., 2017, Zahrazadeh et al., 2018). BCS score may not have a very high sensitivity and specificity when used on its own to predict health and production outcomes (LeBlanc, 2010), however when combined with other predictors it could potentially be a very useful predictive tool.

1.2.2 Rumen Fill Score

Feed intake monitoring can be useful to assess the cow's energy status, however it can also be difficult to perform on a routinely basis on commercial dairy farms. A simple and feasible method of assessing feed intake is the Rumen Fill Score (RFS) as they are shown to be associated (Burfeind et al., 2010). Burfeind et al. (2010) also suggested that RFS should be routinely measured at the same time of day in order to draw conclusions on the DMI. Kawashima et al. (2016) further supported the association between RFS and feed intake and suggested that RFS did not change in dry cows until close to the calving date. Overall, RFS during the close-up dry period can be used as a predictor of metabolic status and consequently health and production outcomes.

1.2.3 Housing

The comfort of the cows has been recognised as an important factor contributing to optimization of milk production (Schirmann et al., 2011, Wilkes et al., 2008). Among the indicator of cow comfort, the lying-down behaviour has

been reported to be pivotal (Broucek et al., 2017). In a study investigating the effect of reduced competition for feeding and lying space on the health and immune function it was found that cows kept in lower stocking density pens had improved blood metabolites both pre- and post-calving (Miltenburg et al., 2018). Recommendations state that overcrowding at transition pens should be avoided with 80% cows to stalls and the recommended feeding space for cows is a minimum of 76cm and even wider for cows at late stages of gestation (Miltenburg et al., 2018, Nordlund et al., 2006). The association between the housing environment of transition cows and metabolic health has not been thoroughly examined, however two existing studies support no increase in RFM and metritis incidence, or blood metabolite levels (Silva et al., 2014, Silva et al., 2013).

1.2.4 Heat Stress

Heat stress in cows leads to decreased milk yield and reproductive function, as well as a worsening health status (Lamp et al., 2015). The thermal neutral zone for dairy cattle is generally between 5 and 20°C (NRC, 2001), therefore high-yielding dairy cows start experiencing heat stress in temperatures above 21°C (Hahn, 1999). Despite taking measures such as installing intensive cooling systems, the economic toll on farms in the USA was estimated to approximate 897 million dollars yearly (St-Pierre et al., 2003). A measure frequently used to assess the presence of heat stress in livestock is the Temperature-Humidity index (THI) (Ansari-Mahyari et al., 2019, Polsky and von Keyserlingk, 2017).

The most common index in cattle is calculated as:

$$THI = (1.8 \cdot T + 32) - (0.55 - 0.0055 \cdot H)$$

where T is the dry light bulb temperature (°C) and H is the relative humidity of the air (%) (Díaz et al., 2017).

Dairy cattle may start experiencing heat stress at indexes over 68 to 72, with some variations across different climatic regions (Díaz et al., 2017, Polsky and von Keyserlingk, 2017). Heat stress has been reported to decrease DMI, negatively affect milk production, reproductive performance and to be a major risk factor for lameness (Polsky and von Keyserlingk, 2017). The mechanism through which heat stress affects lameness is not known (Polsky and von Keyserlingk, 2017), but it is speculated to be either via increasing the standing times (Cook et al., 2007) or through changes in nutrient metabolism caused by the decline of DMI (Cook et al., 2004). Vitali et al. (2009) reported that for a minimum index of 70 the risk of death in dairy cows starts to increase.

1.3 Economic Impact

Even though research has focused on transition period of dairy cows it is still a challenging area with great economic losses for the farmers (Overton and Waldron, 2004). A study in Minnesota indicated that 25% of cows removed from the herd left during the first 60 DIM with an additional percentage leaving due to difficulties associated with transition period difficulties (Godden et al., 2003). Decrease of reproductive performance and milk yield are the most significant sources of economic loss for the farmers due to the cost of treatments and increased culling rate (Grohn et al., 2003, Melendez and Risco, 2005). This decrease is linked with periparturient disease (Melendez and Risco, 2005). In a recent study, the cost of RFM in the United States was calculated as a total \$386, \$287 attributed to milk yield reduction, \$73 attributed to an increase in days open, \$25 for an increase in the risk of disease and \$1 for an increase in culling risk (Gohary and LeBlanc, 2018). For displaced abomasum, the cost per diagnosis may reach up to \$700 due to direct and indirect costs (Caixeta et al., 2018). The estimated cost of milk fever in the United Kingdom was estimated at £220 or \$343 (Saborío-Montero et al., 2017). Even though it was previously considered that twinning in dairy cows was profitable due to an increase in milk production it is now controversial due to the losses from a higher incidence of dystocia, RFM and stillbirths (Cabrera & Fricke, 2021). The frequency of twinning in dairy cattle has been estimated to be approximately 5%, resulting in annual losses to the industry ranging from \$22.5 to \$112.5 million, assuming a US national herd of 9 million cows (Lett & Kirkpatrick, 2018). It is, therefore, evident that preventing transition period related issues before they manifest would greatly impact the farmers' profits.

1.4 Statistical Methods

1.4.1 Explanatory vs Predictive Modelling

A great number of studies on transition period diseases have focused on identifying risk factors for disease outcomes, suggesting a possible causal link between the two (Daros et al., 2017, Huzzey et al., 2007, LeBlanc et al., 2004, LeBlanc et al., 2005). This method is called explanatory modelling and it aims to interpret the outcome utilising the independent, or explanatory, variables, without however attempting to make predictions about said outcomes (Shmueli, 2010). In summary, explanatory modelling involves applying statistical techniques to evaluate causal hypotheses, where the underlying factors are believed to drive the observed effect (Sainani, 2014, Shmueli, 2010, Vergara et al, 2014). Models that seek to forecast specific outcomes using given predictors are called predictive models. The model building process differs between the two (Sainani, 2014) and the resulting models often differ in variables and predictive value (Shmueli, 2010). An area that focuses on building predictive models is that of machine learning. Veterinary epidemiology utilises both explanatory and predictive modelling in its research (Froud et al., 2017, Vergara et a., 2014)

1.4 2 Building a predictive model using machine learning

Machine learning approaches in data science are increasingly popular methods of identifying patterns in data (Biffani et al., 2017). Their main purpose is making predictions on new unobserved datasets, while as it gets exposed to more data the algorithm is adapted and improved (Hudson et al., 2018).

Machine learning techniques are steadily becoming more widely used over the past years, along with the advent of “Big Data”. Biffani et al. (2017) demonstrated that in just 16 years the number of publications related to machine learning have increased drastically, from 10,690 in 2000 to 1,211,400 in 2016, and even though the peak rate was between 2011 and 2013 it steadily continues to increase.

The techniques are categorised in various ways in different areas, but a distinction between supervised and unsupervised methods is widely recognised (Lanier et al. 2020, Patel & Jhaveri, 2015, Moujahid et al., 2018). The first of these aims to predict chosen outcomes based on various variables, while the second aims to identify clusters in the data without a specified outcome. Some research questions fall naturally into one of those categories, however there are instances where it is logistically difficult to collect the response variable, a combination of the two may be used (James et al., 2014).

A way of differentiating between the various supervised methods is whether they can be used to model quantitative (numerical) or qualitative (categorical/factor/binary) responses (James et al., 2014). The former are referred to as regression methods, while the latter as classification methods (Yang et al., 2022). However, the distinction is not always clear as quantitative responses coded as binary can be handled with classification techniques and similarly, binary qualitative responses can be modelled using logistic regression, which is considered a regression method due to the fact that it estimates class probabilities. Furthermore, there are some statistical methods that can handle both qualitative and quantitative responses, such as k-nearest neighbours (James et al., 2014).

Machine learning algorithms offer great flexibility with regards to problems of multicollinearity, missing values, or complex interactions among variables (Kuhn and Johnson, 2013). A potential issue with some of these techniques is that they may be affected by noise in the data. Modelling this noise can lead to overfitting the data and it results into models that have low accuracy when predicting responses on new datasets. As the model fits too closely to the existing data, it may follow their pattern too closely resulting in an overly complicated model that cannot perform well on new observations as it is basically built on idiosyncrasies of the original dataset (James et al., 2014, Yeom et al., 2018). One way to overcome this issue is by using robust ways of evaluating the models' performance, involving resampling methods, such as *k*-fold cross-validation and bootstrapping (Hartono and Ongko, 2022, Kernbach and Staartjes, 2021, Kuhn and Johnson, 2013). The theory behind it is to split the dataset into two parts (usually multiple times), build the model using the first part (train set) and then use the algorithm developed to make predictions on the second part (test set). Afterwards, the model is evaluated based on the comparison of the predictions with the true values of the test dataset. For *k*-fold cross validation the dataset is split randomly into *k* parts of equal size and each time a model is fit for the entirety of the dataset excluding one part (fold) that acts as the test set (Ayranci et al., 2021, Baykan & Yilmaz, 2011). The estimates of performance for all the models are then summarized (Kuhn and Johnson, 2013). The difference in bootstrapping is that each data point is taken with replacement, meaning that it can be selected multiple times when sampling and the final bootstrapping sample is the same size as the original data set (Kuhn and Johnson, 2013 Waitman et al., 2003). The process can be computationally

expensive, however with the advances in computing power over the past years, this is becoming less of an issue and these techniques have become a pivotal tool in the practical applications of machine learning methods (James et al., 2014).

In some models there are parameters whose optimum value cannot be calculated by an analytical formula, such as the choice of k in k -nearest algorithms. The selection of inappropriate values for these parameters may result to overfitting the data. There are many approaches to defining the optimum value for a parameter and the process is called parameter tuning (Kuhn and Johnson, 2013, Yang and Shami, 2020). There are several approaches to deciding the most appropriate parameters, generally by defining a set of possible values, generating reliable estimates of model utility across said values, then selecting the optimal settings. Many researchers opt to complete this process manually, however in order to do so a clear understanding of the different parameters for each corresponding machine learning method is needed (Abreu, 2019). After determining a set of values we can get the estimates of model performance using resampling methods, which are then aggregated into a performance profile to help choose the final parameters, which will be used for the model building (Kuhn and Johnson, 2013).

1.4.3 Assessing the model's predictive power

There are many methods to access model accuracy, meaning how well the predictions match the actual observed data. In the regression setting one such measure of accuracy is the mean squared error (MSE), given by the square of the difference between the observed and predicted values over the total

number of observations. The closer the predictions are to the true values the smaller the MSE will be (James et al., 2014, Joham et al., 2012). For classification problems, a common way of describing model performance is the confusion matrix (cross-tabulation of the observed and predicted classes of the data, indicating the true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP)), where we can calculate metrics such as the overall accuracy ($(TP + TN) / (TP + TN + FP + FN)$), the Kappa statistic, sensitivity ($TP / (TP + FN)$), specificity ($TN / (TN + FP)$), positive predictive value (PPV) ($TP / (TP + FP)$) and negative predictive value (NPV) ($TN / (TN + FN)$) (Kuhn and Johnson, 2013). The Kappa statistic in particular is a measure designed to assess the agreement between two raters, assuming that a proportion of the agreement can be due to chance alone (Warrens, 2010). It can be calculated as the difference between observed accuracy and expected accuracy (based on marginal totals of the confusion matrix) over the difference between 1 and the expected accuracy, with values ranging from -1 (total disagreement) to 1 (total agreement) (Kuhn and Johnson, 2013). The interpretation of Kappa has been described by Landis and Koch (1977) and can be found in Table 1.1. In contrast with other metrics such as accuracy, kappa takes into account the prevalence of the outcome (Viera and Garrett, 2005), meaning that it can be of particular interest in datasets where the outcome is rare.

Table 1.1 Kappa value interpretation (Vierra and Garrett, 2005)

Kappa	Agreement
<0	Less than chance agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Almost perfect agreement

Another method to evaluate the class probabilities and access the sensitivity and specificity thresholds is the receiver operating characteristic (ROC) curves. The combinations of specificity and sensitivity for different cut-off points are plotted against each other and in the resulting plot the area under the receiver operating characteristic (AUROC) can be calculated as a measure of predictive value for the model (Ho, 2017, Kuhn and Johnson, 2013). Values closer to 1 are optimal, whereas those close to 0.5 indicate that the model has no predictive value (Lin et al., 2021). A disadvantage of this metric is that there can be a loss of information when evaluating models, as the shape of the curve might be a better way of comparing models instead of just reporting the AUROC (Kuhn and Johnson, 2013).

1.5 Machine learning in the dairy industry

It has been emphasized that clinicians in farm animal practice, need to focus on consulting their clients on farm management in order to prevent disease, rather than just offering treatment (Hudson et al., 2018). Machine learning techniques have recently started being used in veterinary medicine to help improve farm management. There have been some recent studies trying to utilize various methods in order to develop predictive models and explore different aspects of the dairy cow's health and performance.

Most papers have focused on developing predictive models for reproductive outcomes. In a study by Caraviello et al. (2006) an alternating decision tree model was developed to predict first-service conception rate by the frequency of hoof trimming, type of bedding in dry pens, restraint system and duration of the voluntary waiting period. The model correctly classified 75.6% of the records, using 10-fold cross-validation, with 99.3% of the incorrectly classified being false negatives and the AUROC was 0.68. Using a similar model, they also identified variables such as bunk space per cow, number of cows in maternity the pen, BCS, strategy for using clean-up bulls, temperature of thawing semen and milk yield at first service informative predictors for insemination outcomes at 150 DIM. This model correctly classified 71.4% of cows, using 10-fold cross-validation, with 16.3% of the incorrectly classified being false negatives and the AUROC was reported to be 0.73. Fenlon et al. (2017a) also tried to predict the probability of conception in heifers, with an overall prediction accuracy between 77.1% and 78.9%. However due to low specificity the models were not successful at identifying failed services and they were thought to be of little predictive value. Machine learning methods have

also been used to determine the time-to-calving (Miller et al., 2020) with models for dairy cows showing an increase of predictive performance up to 4 hours before calving (Matthew's correlation coefficient increasing from 0.06 to 0.14) and the highest AUROC, sensitivity and specificity combination 2 hours before calving (95.4%, 91.3% and 93.5% respectively). Borchers et al. (2017) used techniques, such as random forests, linear discriminant, and neural network analyses, along with precision technology to predict the time of calving, with the method yielding the best sensitivity/specificity combination being ANN (100% sensitivity and 86.8% specificity when the variables were summarized at the daily level, 82.8% sensitivity and 80.4% specificity when analysing bihourly increments). In a study with similar outcomes, Keceli et al. (2020) utilised activity and behavioural data providing models with sensitivity, specificity, PPV and NPV for the day before calving reaching 100%. Vázquez-Diosdado et al. (2023) also investigated calving prediction using sensors with the best results being achieved when inputting data from 2 days prior to calving (87.81 % accuracy, 92.99 % specificity, 75.84 % sensitivity, 82.99 % PPV, 78.85 % F-score, and 90.02 % NPV). In another study by Fenlon et al. (2017b), four machine learning methods were compared to identify the most suitable model for predicting calving difficulty in dairy heifers and cows. Using the AUROC, which for all models ranged from 0.64 to 0.79 they reported that all models had good discriminatory power with ANN and multinomial regression performing best (75% cases correctly classified). Avizheh et al. (2023) utilised historical data, also for the prediction of calving difficulty, producing models with low AUROC and F1-score due to an imbalanced dataset. Sampling methods were found to improve the metrics, however they remained in overall low levels (F1-

score ranging between 0.38 to 0.42). Brand et al. (2021) utilised milk spectral data to predict pregnancy, reporting a model with a sensitivity of 0.89, a specificity of 0.86, and prediction accuracy of 0.88. A few recent studies have used similar models to predict insemination outcomes (Hempstalk et al., 2015, Shahinfar et al., 2014). Zaborski et al. (2018) utilised a range of techniques, such as ANN and logistic regression, to identify dystocia in Holstein-Friesian cows, with the maximum overall accuracy being 0.589 for heifers (using a multivariate adaptive regression splines model) and 0.649 for cows (using a ANN model). Dolecheck et al. (2015) explored oestrus detection using random forests, linear discriminant analysis and neural networks, with the overall accuracy for all models ranging from 91.0% to 100.0%. Higaki et al. (2019) also tried to detect oestrus with ANNs, SVMs and decision trees utilising vaginal temperature and conductivity data, with an ANN model having the numerically (but not statistically) higher values of sensitivity and precision (both at 0.94). Cairo et al. (2020) also focused on the prediction of oestrus using behavioural data, reporting high values of accuracy. Another similar study (Hemalatha et al., 2021) utilized milk parameter data also reporting high accuracy, as well as precision, recall, specificity and F1 score, while Schweinzer et al. (2019) used accelerometer data to build a model with over 90% sensitivity, specificity, PPV and NPV. Another study (Wang et al., 2020) used accelerometer as well as location data for their predictions with their best performing model being a neural network predicting within a 30-minute time window (sensitivity = 99.36%, specificity = 53.33%, PPV = 95.76%, NPV = 93.72%, accuracy = 95.36%, F1 = 97.51%). Romadhonny et al. (2019) built a classification model for oestrus reporting over 80% accuracy, however the model only correctly classified

oestrus being late at a rate of 6.4%. In an attempt to classify bovine semen based on mineral imbalances Aguiar et al. (2012) managed to almost halve the predictors needed for the classification, with the highest accuracy model being at 97.25%. Grzesiak et al. (2010) presented models detecting artificial insemination difficulties with an AUROC value of almost 0.9. Bates and Saldias (2019) demonstrated a comparison of regression and machine learning methods by building models predicting the 21-day submission rate in dairy cows. The study concluded that no significant differences in predictive power were found and that even though models had a good enough AUROC (0.68-0.73) the positive outcomes had a better chance at being predicted than the negative outcomes. Keshavarzi et al. (2020) built predictive models detecting abortion incidence, with a mean AUROC of 0.863 and F1 score of 0.520, which showcased an improvement after sampling methods (AUROC 0.893 and F1 0.610 when up-sampling/AUROC 0.897 and F1 0.626 when down-sampling).

Another part of the research has primarily investigated milk production. Murphy et al. (2014) compared 3 different predictive models that focused on predicting milk production, with the reported root mean square error (RMSE) being $\leq 12.03\%$. Njubi et al. (2010) also explored the same area, presenting ANN models with an estimated accuracy of 79%, that predicted next month and first lactation 305-day milk yield of Holstein-Friesian cows in Kenya. ANNs which are among the most common method used in animal sciences for various models, were also used in other papers to develop algorithms predicting milk yield (Gianola et al., 2011, Grzesiak et al., 2006) and breeding values in dairy cattle (Shahinfar et al., 2012). Sefeedpari et al. (2015) focused on milk yield forecast in Iranian farms utilising energy consumption, producing models with

R^2 values ranging between 0.65 and 0.93. Zegler et al. (2020) investigated potential pasture milk production using regression trees and finding that out of the variables explored, the ones most associated with this outcome were improved legume cover, residual sward height, and non-improved grass cover. Nguyen et al. (2020) analysed the associations between fat/protein content and milk yield, stating that while their random forest model achieved the best performance with an average R^2 value of 0.734, their SVM model followed closely with a value of 0.712 and significantly less computational time, making the latter overall more efficient. Dallago et al. (2019) explored the prediction of first day milk yield in heifers and provided three different models with less than 4kg MSE, out of which the ANN was considered the best. Important milk metabolites that could be used to predict milk traits have also been identified in a study by Melzer et al. (2013). Frizzarin et al. (2021) also focused on the prediction of milk traits utilising both regression and classification methods and using milk spectra. Fuentes et al. (2020) analysed feed, weight and weather data to develop models predicting milk yield, protein and fat content as well as concentrated feed intake, with the model for all cows achieving a correlation coefficient of 0.86 and slope of 0.74. Muniz et al. (2020) built linear regression and ANN models predicting lactose, protein, fat and solids-non-fat parameters in milk, with ANNs achieving overall lower bias. Pietersma et al. (2003) conducted lactation curve analysis, presenting classification tree models with different levels of intensity when it came to outlier removal, achieving sensitivities of 52%, 68% and 92% for each increasing level of intensity. More recently, Anglart et al. (2020) focused their research on the prediction of monthly composite somatic cell count, concluding to some MSE disparity

among the different predictor variable setups (0.09 to 0.17 for the generalized additive model). Ji et al. (2022) explored various production measures, such as the daily milk yield, fat and protein content in milk, as well as frequency of individual cow milking during the next 28 days, proposing models with good results ($R^2 > 0.90$ and overall accuracy $> 80\%$). Farah et al. (2021) explored the prediction of milk adulteration showcasing that their optimal model was a random forest with 100% on the training set and 88.5% accuracy on the test set. In a study with a similar aim, Neto et al. (2019) utilized spectral milk data and proposed a neural networks model with 98.76% classification accuracy. Conde et al. (2020) also investigated milk adulteration, this time with the addition of whey, and provided an ANN model with 15 hidden layers to which the most influential variables were the milk's fat content and density.

Amongst the health outcomes, mastitis appears to be among the most frequently explored with predictive models. Kim and Heald (1999) compared decision tree classification of mastitis with culture diagnosis and estimated a 58-61% accuracy for the former. Several studies have utilised a variety of methods, such as decision trees, SVMs and ANN to investigate clinical and/or subclinical mastitis (Ebrahimie et al., 2018b, Ebrahimie et al., 2018a, Kamphuis et al., 2010, Kamphuis et al., 2008, Luo et al., 2023, Mammadova et al., 2013, Panchal et al., 2016, Sharifi et al., 2018). Sun et al. (2010) also attempted to identify mastitis from data collected by automatic milking systems, using cluster analysis. The correct classification rate of the models generated ranged from 86.9% to 91.6%. Dhoble et al. (2019) attempted predictions utilising cytometric fingerprints. Their four outcomes included identifying the source cow of the milk sample, distinguishing pathogens in infected samples and recognizing healthy

samples, determining the lactation stage of the sample, and gauging the severity of infection. The best models proposed for all four outcomes had an accuracy of over 99%. Post et al. (2020) and Post et al. (2021) focused on the classification of mastitis, as well as lameness, utilising a vast variety of methods, including logistic regression, SVM, k-nearest neighbours, naive bayes, decision trees and random forests, while emphasising the effect of imbalanced datasets on the metrics. Maciel-Guerra et al. (2021) presented models on the success of the treatment of mastitis caused by *Streptococcus uberis* reporting an accuracy of 92.2% and kappa of 84.1%. Doupbrate et al. (2019) debated the reliability of person vs machine-based hygiene scores for the teat. While investigating *Staphylococcus aureus* antibiotic resistance Esener et al. (2021) used ten different machine learning methods including SVM, logistic regression, naïve Bayes and MLP neural networks. Hyde et al. (2020) presented random forest models with 98% accuracy, 86% PPV and 99% NPV (when distinguishing between environmental and contagious diagnoses) and 78% accuracy, 76% PPV and 81% NPV (when distinguishing between environmental dry period and environmental lactation period diagnoses). Srikkok et al. (2020) took advantage of the presence of mRNA in milk to determine infection and reported models with AUROC ranging between 0.77 and 0.89. Regarding the environmental and contagious distinction of mastitis caused by *Streptococcus uberis*, Esener et al. (2018) presented models with high accuracy and kappa for an individual farm classifier and a global classifier after cross-validation, that however decreased after external validation (70.67% accuracy and 0.34 kappa). When investigating the possibility of subclinical mastitis prediction Ebrahimie et al. (2021) proposed the use of a classification

based on associations (CBA) model which utilises scaled data and generates rules that define sub-groups in complex datasets, thus increasing the generalisability of the model. A paper by Hassan et al. (2009) presented both supervised and unsupervised ANNs to help detect mastitis pathogens based on alterations in milk parameters, with unsupervised ANNS yielding overall greater sensitivity and specificity. A few years prior, Heald et al. (2000), had also developed an ANN model for mastitis detection, that was reported to offer a greater predictive value compared to classical statistical methods, the classification rate ranging from 57 to 71%. Even further back, two studies (Nielen et al., 1995a, Nielen et al., 1995b) used neural network models alongside logistic regression models for the detection of both clinical and subclinical mastitis.

Lameness has also been a significant part of machine learning based research. Shahinfar et al. (2021) also attempted to predict lameness, using a naïve Bayes model, amongst other methods, with an AUROC of 0.66 and an F1 value of 27%. Warner et al. (2020) proposed the use of machine learning for the prediction of lameness with their best performing model achieving an AUROC of 0.76, a sensitivity of 0.54 and specificity of 0.94. Volkmann et al. (2021) focused on the identification of claw lesions by analysing the acoustics of the animals' gait, while Barney et al. (2023) utilised computer vision to detect lameness. Haladjian et al. (2018) also investigated lameness and proposed a motion sensor with 91.1% accuracy, while Shrestha et al. (2018) suggested radar sensing with over 85% accuracy in dairy cows. Alsaad et al. (2012) proposed the use of pedometers recording activity, lying time, and temperature, with the resulted model having an accuracy of 81% for non-lame cows and 72%

for lame ones. In yet another study using sensor data, Taneja et al. (2020) proposed a model able to identify lame animals up to 3 days before visual confirmation with an accuracy of 87%. Boghart et al. (2021) presented a model using behavioural metrics, milk production and animal characteristics, with 85% AUROC. After investigating the possibility of prediction of digital dermatitis, Cernek et al. (2020) reported a model with 71% accuracy and 0.51 Cohen's kappa before external validation and 88% accuracy and 0.36 kappa value after.

In recent years, metritis has been the subject of several machine learning studies. Vidal et al. (2023) provided some models with high F1 scores, utilising sensor data from accelerometers. Risvanli et al., 2024 used a sensor measuring intrauterine gases, and provided models with high accuracy (71.22%), precision (64.4%) as well as recall (71.2%). De Oliveira et al. (2021) investigated the treatment success of metritis, presenting models with high F1 (0.81), sensitivity (0.85) and PPV (0.78), but low specificity (0.39) and NPV (0.50). Another study (Sadeghi et al., 2022) presented models predicting subclinical endometritis by interpreting polymorphonuclear leukocyte proportions. Finally, Merenda et al. (2020) attempted to predict metritis, acute metritis, along with success and failure of treatment. The models for metritis and acute metritis produced had fair AUROC (0.82 and 0.87 respectively) with reasonable specificity and sensitivity, however the model for acute metritis had low PPV (0.30) while that of metritis was fair (0.60).

Various other diseases have also been the subject of more recent machine learning studies. Lasser et al. (2021) utilised a number of different algorithms and reported different metrics, including the F1, precision, recall, specificity, and accuracy in an attempt to predict anoestrus, ovarian cysts, lameness, ketosis,

periparturient hypocalcaemia, metritis, chronic mastitis, as well as acute mastitis. Wagner et al. (2020) developed models using behavioural data to identify cases of sub-acute ruminal acidosis from a sample of 14 diseased cows and 14 controls. Their best performing one, was a KNN model with 12 hours of prediction achieving a PPV of 0.83 and a NPV of 0.66. In another study using behavioural data (Cantor et al., 2022) this time for the prediction of respiratory disease in calves, the produced KNN model returned accuracies up to 95% when predicting clinical disease and 52%-90% when predicting pre-clinical disease. Sturm et al. (2020) proposed models for the prediction of ketosis, which while having fair accuracy and NPV (0.72 and 0.92 respectively) had low kappa, F1-score and PPV (0.28, 0.43 and 0.32 respectively). Wang et al. (2023) explored the possibility of utilising explicit and implicit features found in text records in the prediction of disease which was broadly classified in 7 classes (rumen indigestion, rumen bulging, atonia proventriculorum, ketosis, epidemic fever, oesophageal obstruction and ruminal acidosis). Their suggested model had a F1-score of 94.89%. Reporting a collective disease outcome has also been attempted by Hernandez et al. (2021), accompanied by poor metrics (sensitivity = 61.74%, PPV = 59.99%) with high standard deviation (15.99% and 26.20% respectively). Even more recently, Zhou et al. (2022) proposed models for a disease outcome including digestive disorders, lameness, mastitis, and metritis, utilising potential predictive variables, such as the season, days in milking, parity, age at the time of disorders, milk yield, activity, rumination time, and electrical conductivity of milk. Although three of their models produced high metrics (AUROC 81.58%-92.86%) their sensitivity ranged between 48% and 85%, which was concluded to be low for the purpose of the study. Dineva and

Atanasova (2023) also proposed a general disease outcome, using a variable with three classes as assessed by a veterinarian (Healthy-cow is not in any discomfort, Unhealthy-any disease state, including those caused by cold or heat stress, Suspect-presence of sufficient conditions for the occurrence of a disease state in the animal, not yet manifested). Their worst performing model was a Naïve Bayes model with 0.62 accuracy, 0.52 recall and 0.53 precision, while the best one was a random forest classifier with 0.95 accuracy, 0.95 recall and 0.97 precision. Lardy et al. (2023) investigated various conditions in Holstein cows, several diseases being amongst them, however instead of binning them in a binary outcome they included them all in a multi-class variable. The conditions included oestrus event, calving, lameness, mastitis, acidosis, inflammatory reaction to lipopolysaccharide injection (LPS), accidents (such as Injuries, retained placenta and vaginal laceration), other disease, mixing and disturbance (such as Handling for vaccination, oestrus synchronisation, anthelmintic cure, claw trimming and relocation). For their predictive variables they utilised sensors measuring the distribution of the activity level in 24-hour time series. While highly specific, the random forest model returned low accuracy for all classes (44.4% to 5.2%). This however improved when considering the presence of at least one 24-hour time series classified correctly before the event, after multiple are recorded (acidosis 85.6%, oestrus 72.4%, calving 74.0%, other diseases 78.2%, lameness 66.3%, mastitis 56.6%, mixing 54.6%, LPS 45.1%, disturbances 40.9%, and accident 10%). Despite the sensitivity issues, the random forest model was able to differentiate between the events with good results.

Several genetic studies have also used machine learning. In 100-year a review, Weigel et al. (2017) described how researchers have recently started to use machine learning techniques alongside more traditional methods, in order to develop models for genetic selection. In fact, a number of studies have used machine learning for genomic predictions (Biffani et al., 2017, Ehret et al., 2015, Gonzalez-Recio et al., 2010, Yao et al., 2013). Yao et al. (2016) while looking into genomic prediction, suggest that a self-training algorithm incorporated into an SVM prediction model can enhance the accuracy of said prediction by gathering additional genomic data from animals lacking measured phenotypes. In another paper, genomic prediction methods have been compared in order to determine the most accurate, while in another study by the same author the value of imputation in genetic studies was explored (Jiménez-Montero et al., 2013a, Jiménez-Montero et al., 2013b). Rodriguez et al. (2019) attempted a genome-wide classification in order to identify high-producing cows by training decision trees and ANN algorithms and getting a mean prediction of 92.4% and 82.19% respectively.

A few behavioural studies have also utilized these techniques. Williams et al. (2016) used machine learning methods to develop a behavioural model of the pasture-based dairy cow, yielding an accuracy of 85%, false positive rate of 10% and AUROC of 0.87. Similarly, other studies (Benaissa et al., 2017, Martiskainen et al., 2009, Smith et al., 2016, Tamura et al., 2019, Vázquez Diosdado et al., 2015) have used decision trees, *k*-nearest neighbours, naïve Bayes and Support Vector Machine methods to classify behavioural data collected by sensor technology. Chelotti et al. (2018) in particular, presented methods for identifying and classifying jaw movements in grazing cows, with

the best model achieving accuracy, precision and recall of over 90%. In a more recent study (Shen et al., 2020) accelerometer data were input in KNN, SVM and ANN algorithms to predict ingestive-related behaviours with good success. In a similar study only a year prior Benaissa et al., (2019a) had also assessed feeding and ruminating behaviour, reporting models with high precision, sensitivity and specificity for both outcomes. The same researchers (Benaissa et al., 2019b) also compared neck and leg mounted sensors, determining that the optimal position depends on the behaviour than needs to be monitored. Dutta et al. (2015) utilised ensemble methods for behaviour prediction, with the best model achieving 96% accuracy, 97% sensitivity, 89% specificity, 89% F1 score and 9% false discovery rate. Riaboff et al. (2020) presented 4 models, classifying 6 behaviours with accuracies ranging between 0.95 and 0.98 and Cohen's Kappa ranging between 0.91 and 0.96. Using 8 surveillance cameras and ANN methods Salau and Krieter attempted to segment animal behaviour with overall high 'averaged precision score' but moderate 'averaged recall score'. Williams et al. (2019) explored both base learner and ensemble methods to predict behaviour with logistic regression being the best overall model for the former (accuracy 0.90; sensitivity 0.88; specificity 0.92; precision 0.92; F1-score 0.90). The ensemble methods produced overall similar measures. Balasso et al. (2021) produced ANN, SVM, KNN and extreme boosting algorithm models, in an attempt to identify posture and behaviour with positive results. Hunter et al. (2021) developed ANN and random forest algorithms using neck muscle activity and heart rate data aiming to differentiate between sleep stages, with the best model being the ANN achieving an AUROC of 92.5%. Carslake et al. (2021) used collar-based sensors collecting signal

data in calves to correctly identify lying and standing posture while also classifying locomotor play, self-grooming, active lying, inactive lying and different feeding behaviours. The model with the worst performance was that of active lying (90% accuracy, 64% sensitivity and 69% precision), while the one for locomotor play achieved 98.98% sensitivity, 99.73% specificity and 99.23% precision. Ren et al. (2021) analysed social interactions with each of the 6 classes of the outcome reaching accuracies ranging between 72.73% and 92.16%. Chen et al. (2020) aimed to recognize feeding behaviour using neural networks, achieving accuracies of up to 89.5%.

A few studies have taken a different approach and rather than focusing on animal traits, they aimed to predict farms' energy consumption. Shine et al. (2019) utilized an SVM algorithm and empirical data from 56 to investigate electricity consumption. They found, among other outcomes, that the model could predict yearly electricity consumption within 10.4%, with a correlation coefficient at 0.97. A year prior the same research group had investigated water and electricity consumption on pasture-based farms with a SVM model predicting electricity use within 12%, while a random forest model predicting water consumption within 38% (Shine et al. 2018a), while in another study that year with the same goal (Shine et al. 2018b) they presented a multiple linear regression predicting electricity and water use within 26% and 49% respectively. Sefeedpari et al. (2013) used an ANN method and data from 50 farms to present a model focusing on energy input and output with an R^2 of 0.88 and RMSE of 0.015. In another study a year later they (Sefeedpari et al., 2014) modelled fossil fuel as well as electricity consumption with an R^2 metric of 0.79. Todde et al. (2017) expanded not only on the prediction of energy use,

but on that of related emissions and costs as well, with the models' MSE values ranging at below 15%.

The estimation of body weight in dairy cows has also been the subject of multiple papers, especially in the most recent years. Two recent studies (Nagy et al., 2023, Siachos et al., 2024) have utilised various algorithms to measure BCS and both used Cohen's Kappa to assess their results. Huang et al. (2019) also proposed a method of BCS estimation with 98.46% classification accuracy, while Cevik (2020) achieved 78% accuracy. In a similar study, Zhao et al. (2020) proposed several models for BCS approximation. Amongst those models were a classification decision tree which while predicting with around 60% accuracy for each class 95% of the predictions were within a 0.25 score difference, and linear regression and ANN models with over 80% R^2 . In another similar study Rodriguez Alvarez et al. (2018) presented classification models for BCS and found that while when predicting the exact score the precision, recall and F1 were 0.40, 0.40 and 0.39 respectively, when predicting within a 0.25 score difference the metrics increased to 0.79, 0.78, 0.77 respectively and even further at 0.94, 0.94 and 0.94 when predicting within a 0.5 range. A year later the same research group (Rodriguez Alvarez et al., 2019) expanded on their research adding more classification methods and slightly improving their results, with the accuracy of BCS estimations within a 0.25-unit difference from actual reaching 82%, while overall accuracy within a 0.50-unit difference achieving 97%. Tedde et al. (2021b) also attempted to approximate body weight, but without the use of the BCS scale and they produced models with an RMSE ranging between 52 and 56kg.

Studies focusing on different research areas have also demonstrated the use of machine learning methods. Pastell and Kujala (2007) have developed a probabilistic neural network model with 96.2% correctly classified cases that focused on lameness detection, whereas Dórea et al. (2018) compared partial least square models to ANN models when trying to predict dairy cows' feed intake using milk spectra, finding ANNs superiors possibly indicating a non-linear relationship between predictors and outcome. Predicting metabolic stress in the transition period was also explored with a variety of methods by Wisnieski et al. (2019). Craninx et al. (2008) investigated the use of ANNs for the prediction of rumen proportions of volatile fatty acids showcasing a model with an RMSE of just 2.76% which did not however outperform the multi-linear regression model. Shafiullah et al. (2019) used sensor technology to identify the sufficiency of herbage allowance presenting models with 88% AUROC and overall high metrics. Nikoloski et al. (2019) built numerous tree models for nutrient uptake and herbage production with R^2 values ranging between 0.64 and 0.78. Tedde et al. (2021a) explored the prediction of dry matter intake using milk spectral data, among other parameters, and producing a regression and an ANN model with RMSE of 3.27kg and 3.25kg respectively. In another study, Fu et al. (2021) implemented a kernel extreme machine learning technique to approximate the cows' digestible energy and energy digestibility, producing R^2 values of almost 90%. Becker et al. (2021) utilised a variety of pen-level as well as cow-level variables to focus on the prediction of heat-stress, with high accuracy. Ji et al. (2020) also investigated heat stress by using a decision tree model with 79-94% accuracy. Also on the subject of heat stress, Gorczyca, and Gebremedhin (2020) aimed to predict respiratory rate, skin temperature and

vaginal temperature, and the RMSE of their models was 9.695 respirations per minute, 0.334 °C and 0.434 °C respectively. Pacheco et al. (2020) proposed ANN models for the prediction of respiratory rate and rectal temperature with R^2 values at 0.74 and 0.71 respectively, classifying thermal stress with 83% and 84% accuracy, again respectively. Chung et al. (2020) explored the use of implanted sensors to approximate vaginal temperature building models with an RMSE of 0.081 °C. Mota et al. (2021) investigated the predictive ability of models identifying phenotypic characteristics that are difficult to measure, such as κ -CN in milk and blood BHB. Dettmann et al. (2020) proposed the use of milk fatty acid profiles for the estimation of bodyweight change in cows post calving, showcasing a model with an R^2 value of 0.94 that after external validation dropped to 0.31. Cernek et al. (2020) while trying to identify digital dermatitis reported models with accuracy of up to 88% but only “fair” Kappa. Paratuberculosis diagnosis through ELISA has also been the subject of recent research (Imada et al., 2024) utilising decision trees as well as random forest models. Bovine tuberculosis has also been the subject of machine learning research (Denholm et al., 2020) by using milk spectral data as predictive variables and building models with sensitivity and specificity of up to 0.96 and 0.94 respectively. Multiple studies explored the prediction of hyperketonemia (Bonfatti et al., 2019, Luke et al., 2019, Pralle et al., 2018, Walleser et al., 2023) using milk spectra and utilising mainly partial least squares regression, reporting good sensitivity and specificity. Sturm et al. (2020) reported models aiming to predict subclinical ketosis, with the best performing model producing high accuracy (0.725) and NPV (0.922), while also having low PPV (0.322) and F1 (0.435). Van der Heide et al., (2019) produced regression, naïve Bayes and

random forest models to evaluate the prediction of survival to second lactation in heifers, with only one model achieving an AUROC higher than 0.7. In an attempt to improve this result, they (van der Heide et al. 2020) produced four different ensemble methods that ultimately did not result in greater performance, the maximum precision value being at 0.250. Salau et al. (2021) used KNN and ANN algorithms to identify and classify body parts, reporting accuracies reaching 0.976, and precision and recall ranging from 0.84 to 1 and 0.83 to 1 respectively. Nir et al. (2018) opted to use machine learning to estimate heifer height and body mass, with R^2 values starting from 94.6% to 98.5%. John Wallace et al. (2019) investigated rumen metabolism, diet and host characteristics, using ridge regression and random forests, with propionate predictions reaching R^2 of 0.9 in some farms, while methane emissions reaching values of 0.4. Hempel et al. (2020) focused solely on methane emissions, building models for 27 scenarios with R^2 values ranging between 0.394 and 0.664. Genedi and Ogejo (2021) aimed to predict manure temperature during storage using weather data, time and manure depth above the sensor as inputs and creating models with R^2 values of over 0.97. Ghaffari et al. (2019) worked towards metabolic profile prediction in dairy cow serum, by using sequential minimal optimization, random forest, alternating decision tree, and naïve Bayes–updatable methods. In another study about metabolic status Xu et al. (2019) identified a Random Forest (error rate from 12.4 to 22.6%) and a SVM (error rate from 12.4 to 20.9%) model as the best performing ones. In a study assessing a model processing digital images to categorise teat cleanliness (Doughrate et al., 2019), it was found that the accuracy reached within each class of the model was 90% or higher. Salzer et al. (2021)

conducted an experiment in an attempt to predict mild pain in cows, producing classification models with overall high accuracies. Oehm et al. (2022) implemented clustering analysis by inputting milk production and other cow-level data and the resulted clusters identified infection by *Fasciola hepatica* and *Ostertagia ostertagi* with great accuracy. Finally, Probo et al. (2018) used decision tree models and random forests alongside more traditional survival analysis, to investigate the association between metabolic diseases and the culling rate in high-yielding cows, reporting milk fever as the most influential factor.

Shine and Murphy (2021) conducted a systematic review of papers applying machine learning techniques to dairy industry related issues dating from 1999 to 2021. Amongst the most frequently used methods were tree-based algorithms (25% in 1999-2017, 26% in 2018-2021) and regression-based algorithms (22% in 1999-2017, 17% in 2018-2021), with ANNs showing an increase in popularity (16% 1999-2018, 25% in 2018-2021). Almost half the studies (48%) leveraged sensor data when developing their models. Furthermore, cow characteristics (34%), milk properties (37%), calving data (23%), and lactation information (19%) were commonly used as features. This was followed by meteorological data (14%), dietary and feeding practices (10%), farm characteristics (16%), milking parameters (10%), soil properties (1%), and various other variables (7%). Concerning the outcomes, they reported that a great number of research prior to 2018 focused on animal husbandry (35%), with that recently being decreased to 14% and replaced with physiology and health outcomes (38%). In fact, they identified that the number of papers addressing these outcomes had been increased 7 times since 2018.

A small subset of research was also dedicated to feeding (6% both prior and after 2018). A smaller systematic review (Cockburn, 2020) included papers from January 2015 to June 2020. For the most common among physiological and health outcomes they mentioned BCS, lameness, heat stress, mastitis, metabolic status and infectious disease, while among other popular outcomes they included reproduction, behavioural and feeding outcomes. They conclude that despite the abundance of available research, most tested algorithms have not performed adequately for dependable implementation in practical settings, which they speculate may be attributed to subpar training data.

An interesting aspect is the metrics the studies use to report their results regarding the classification models. A lot of models relied on sensitivity combined either with specificity (Nielen et al., 2015a, Nielen et al., 2015b, Hassan et al., 2009, Sun et al., 2010, Kamphuis et al., 2015, Mammadova et al., 2013, Panchal et al., 2016, Fenlon et al., 2017b, Post et al., 2020, Becker et al., 2021, Lasser et al., 2021, Lardy et al., 2023, Srikok et al., 2020, Volkman et al., 2021, Esener et al., 2021, Sadeghi et al., 2022, Imada et al., 2024, Vergara et al., 2014, Miller et al., 2020, Warner et al., 2020), precision (Imada et al., 2024, Barney et al., 2023, Esener et al., 2018, Hunter et al., 2021, Lasser et al., 2021, Higaki et al., 2019, Fenlon et al., 2017b, Benaissa et al., 2017, Martiskainer et al., 2009, Hernandez et al., 2021, Rodriguez Alvarez et al., 2018, Rodriguez Alvarez et al., 2019, Wang et al., 2023) or both (Benaissa et al., 2019a, Carslake et al., 2021, Ghaffari et al., 2019, Merenda et al., 2020, de Oliveira et al., 2021, Keceli et al., 2020, Shen et al., 2020, Xu et al., 2021). Salau and Krieter (2020) on the other hand, reported only precision and recall as averaged metrics, while Pietersma et al. (2003) based model evaluation on

sensitivity and false positive rate, which is the inverse of specificity. Accuracy was among the most reported metrics, either as a sole metric (Pastell and Kujala, 2007, Aguias et al., 2012, Chen et al., 2020, Cevik, 2020, Jiménez-Montero et al., 2013, Dolechek et al., 2015, Ebrahimie et al., 2018a, Ebrahimie et al., 2021, Farah et al., 2021, Zaborski et al., 2018, Tamura et al., 2019, Douphrate et al., 2019, Njubi et al., 2010, Sturm et al., 2020, Romadhonny et al., 2019, Rodriguez et al., 2019, Taneja et al., 2020, Zhao et al., 2020) alongside AUROC (Dhoble et al., 2019, van der Heide et al., 2019, Hunter et al., 2021, Neto et al., 2019, Srikok et al., 2020, Williams et al., 2016), most recently Kappa (Cerneek et al., 2020, Balasso et al., 2021, Esener et al., 2018, Esener et al., 2021, Riaboff et al., 2020, Sadeghi et al., 2022), both (Boghart et al., 2021, Shen et al., 2020), balanced accuracy (Ji et al., 2020), with sensitivity and specificity (Haladjian et al., 2018, Ji et al., 2020), with precision and recall (Benaissa et al., 2019b, Chelotti et al., 2018, Risvanli et al., 2024, Salau et al., 2021), PPV and NPV (Salzer et al., 2021), in a few other studies with a combination of sensitivity, specificity, PPV and NPV (Cairo et al., 2020, Denholm et al., 2020, Schweinzer et al., 2019, Wang et al., 2020) and finally in some studies with a combination of metrics that also include the F1 score (Carslake et al., 2021, Cantor et al., 2022, Dineva and Atanasova, 2023, Dutta et al., 2015, Ghaffari et al., 2019, Hemalatha et al., 2021, Hyde et al., 2020, Luo et al., 2023, Shafiullah et al., 2019, Sturm et al., 2020, Vázquez-Diosdado et al., 2023, Williams et al., 2019, Wang et al., 2023). There were a few studies reporting only accuracy (Douphrate et al., 2019, Huang et al., 2019, Li et al., 2022, Ren et al., 2021), however they displayed the accuracy of each individual class separately, thus addressing any possible class imbalances, while

Shrestha et al. (2018) and Pacheko et al. (2020) included the confusion matrix. Alsaad et al. (2012) also displayed individual class accuracy, along with precision. Wagner et al. (2020) reported only the PPV and NPV of their models. Post et al. (2021) demonstrated the impact of PPV specifically on practical applications. AUROC has also been reported in various studies (Avizheh et al., 2023, Shahinfar et al., 2014, Williams et al., 2016, Panchal et al., 2016, Wisnieski et al., 2019, Post et al., 2020, Shahinfar et al., 2021, Imada et al., 2024, Vergara et al., 2014, Merenda et al., 2020, Post et al., 2020, Grzesiak et al., 2010, Keshavarzi et al., 2020, Miller et al., 2020, Warner et al., 2020) and, especially in more recent research, Cohen's Kappa has been used to evaluate the models' predictive values (Hassan et al., 2009, Balasso et al., 2021, Hyde et al., 2020, Maciel-Guerra et al., 2021, Esener et al., 2021, Volkmann et al., 2021, Sadeghi et al., 2022, Nagy et al., 2023, Barney et al., 2023, Siachos et al., 2024, Sturm et al., 2020, Imada et al., 2024). Some mostly recent studies also rely on or at least include F1-score in the assessment of the predictive value of their models (Avizheh et al., 2023, Hunter et al., 2021, Keshavarzi et al., 2020, Sturm et al., 2021, de Oliveira et al., 2021, Rodriguez Alvarez et al., 2018, Rodriguez Alvarez et al., 2019, Smith et al., 2016, Vidal et al., 2023, Wang et al., 2020). Bates and Saldias (2019) included all the aforementioned metrics in their reporting, discussing the impact of different prediction rates between classes for their specific outcome, while van der Heide et al. 2020 reported a combination of recall, precision, balanced accuracy and AUROC. Shine and Murphy (2021) determined that in 85 studies centred on classification problems, the most frequently utilized evaluation metric was classification accuracy (77%), followed by recall (66%), specificity (49%), PPV (48%), F1

Score (27%), AUROC (26%), NPV (15%), Cohen's Kappa (12%), false positives (FP) (9%), and false negatives (FN) (6%).

Although there are a variety of studies to have explored some aspects of machine-learning applications, there is still room for further research to develop more algorithms with the use of potentially more practical predictors. No study has investigated all the possible different methods of model building to determine the one with the most fitting results. Furthermore, there are areas, including specific periparturient disease with significant economic impact on the dairy farms that have not been adequately explored by predictive models. As Wisnieski et al. (2019) suggested, predictive modelling could be used for practical on-farm applications by predicting a variety of outcomes ranging from health to productive and reproductive performance and culling rates. Especially with the relatively widespread use of sensor technology on farms, such as on-animal sensors for activity monitoring or milking systems collecting data during the milking process, large quantities of data are now easy to collect, so these models can have practical on-farm application (Hudson et al., 2018).

1.6 Conclusions

As established above, the transition period management can be critical for the dairy farms. Difficulties in the demanding adaptations of this phase can lead to imbalances, resulting in health issues, mainly during the early lactation. Targeted management can help prevent such issues and minimize losses. Predictive models and machine learning can play a key part, by identifying disease and other potential problems in high-risk individuals timely in order to effectively tackle the issues.

Blood and urine metabolites, as accurate as they may be, are often an impractical way to routinely identify and prevent potential health and production issues on-farm. Alternative, non-invasive methods using frequently collected data that are already available can be proved to be practical and helpful for the farmers. A few recent studies have tried to address this topic, using machine learning methods and precision technology, however they have only briefly touched the surface of what predictive modelling can achieve. No other study has had access to a dataset of this size that contains information on a great variety of outcomes, as well as predictors that can easily be collected on farms on a routinely basis.

Having access to a unique dataset with both individual cow and farm level data and a variety of outcomes represents a unique opportunity for this project to implement innovative learning algorithms to predict and prevent disease and production issues. The data allows examination of various aspects regarding the transition period and utilization of a variety of different methods and techniques to determine the best model with the best possible predictive value. The results from this research can benefit the dairy cow industry worldwide, as

predictive models could potentially be used on-farm and have a great impact on decision-making.

Therefore, the aims of this study are:

Develop predictive models of peri-parturient disease in dairy cows using various cow-level and herd-level variables.

Develop predictive models of production outcomes in dairy cows using various cow-level and herd-level variables.

Develop predictive models of reproduction outcomes in dairy cows using various cow-level and herd-level variables.

Chapter 2 – Data collection and methodology

2.1 Methods

2.1.1 Source of data

All data was provided by a commercial dairy cow feed and consultancy organisation and was collected as part of a transition cow monitoring service. The aim of the service was to evaluate cow health during the transition period and offer advice to farmers in order to identify areas of transition period management that might need improvement. An assessor visited each farm enrolled in the service once a month to collect cow- and pen-level data from pre- and post-calving cows. A total of six different assessors collected the data from the farms. Assessors held calibration sessions (generally twice per year) where they evaluated and scored cows together to minimise variation between scorers.

Cow level data were collected from 15th April 2016, while pen level data collection was added to the service from 1st October 2016. In the datasets provided both cow and pen level data were recorded until October 2018. The data collected included the date, the cow number, whether the cow was fresh or dry (i.e. pre- or post-calving), whether she was a heifer or not, whether she was found by the assessor for scoring, the BCS (scale 1 to 5 with 0.25 point intervals), rumen fill (scale 1 to 5 with 1 point intervals measured when the cow was standing), hock hygiene (scale 1 to 5 with 1 point intervals), lameness (score >3 was classified as lame). For fresh cows, milk fever (yes/no), LDA (yes/no), RFM (yes/no), calf mortality (yes/no), twinning (yes/no), metritis

(yes/no), daily milk yield (in litres), protein (%), butterfat (%), cell count and drying-off cell count, all as recorded in farm records were also included.

The BCS scale used is the most widely utilised for dairy cows in the UK is a five-point system. This scale evaluates the fat reserves of dairy cows, with scores ranging from 1 (very thin) to 5 (very fat), and includes 0.25-point increments for more precise assessment (Edmonson et al., 1989). Rumen fill scoring involves a visual assessment from behind and slightly to the left of the cow. The focus is on the left sublumbar fossa and flank of the dairy cow (Atkinson, 2009, Bramley et al., 2013, Burfeind et al., 2010, Zaaijer and Noordhuisen, 2003). The assessment should be performed when the cow is standing with all four digits on a flat plane and there is no visible rumen contraction (Burfeind et al., 2010). The score ranges from 1 to 5. For score 1 the rumen appears empty, with a sunken area on the left side. For score 2: there is light fill and the hollow is less pronounced. For score 3 there is moderate fill and the hollow is barely visible. For score 4 there is good fill and the flank is almost flat. And finally, for score 5 the rumen is full, and the flank is bulging (Zaaijer and Noordhuisen, 2003).

Lameness was evaluated using the AHDB (Agriculture and Horticulture Development Board) Dairy Mobility Scoring system, which employs a 4-point scale: 0 = good mobility (not lame), 1 = imperfect mobility, 2 = impaired mobility (lame), and 3 = severely impaired mobility (severely lame) (Gleerup et al., 2015). Since only animals that were scored as severely lame were identified in the company's recording system's binary interpretation of lameness, it is acknowledged that these cases are mostly non-reversible. It would be of more interest to farmers to be able to identify earlier, more treatable cases. Therefore,

lameness was not considered as an outcome but only as a predictive variable for modelling purposes. For other diseases, farm records were used based on diagnosis from the farm personnel; in general using the principles described in the paragraph below.

For uterine bacterial disease, timing relative to calving was also used to define diagnosis; with metritis defined as occurring within the first 3 weeks after calving (Eckel and Ametaj, 2016) and endometritis later than this (Sheldon et al., 2009). However, as will be described in section 2.1.2. we only included the post-partum assessments that were taken up to 21 days post-partum for our analysis, meaning that a possible metritis diagnosis could not have happened after that, avoiding inclusion of any endometritis diagnoses. Identification of retained foetal membranes was completed by visual examination 24 hours postpartum.

For the pen level, the data collected included the type of cows in the pen (dry or fresh), the name of the pen, the type of pen (straw yard or cubicles), the pen length (in metres), pen width (in metres) and pen area (in square metres), feed fence space (in metres), water trough space (in metres), neck rail height (in metres), the number of cows in the pen, the number of cows waiting, as well as the number of cows not waiting, the time period the pen was evaluated (month and year), the water and feed availability (both as binary variables), the temperature (in °C), humidity (%) and number of cubicles. Binary subjective ratings (satisfactory/unsatisfactory) for quality of the bedding, air, feed, water, and light were also recorded.

Data provided was fully anonymised, such that no individuals or businesses could be identified. The project was subject to the University of Nottingham School of Veterinary Medicine and Science ethical review process (approval number 2197 180130).

2.1.2 Datasets

All data cleaning and analysis was completed using R version 3.5.1 (R Core Team, 2018). The data included two main datasets, the first containing information on the individual cow level (cow scores), while the second on the pen level (cow comfort). Additional farm records were also provided from milk recording organizations, which included data routinely collected on the farms via this route, such as daily milk yields each test day (generally monthly) and insemination and calving records.

The cow scores dataset consisted of separate observations for dry and fresh (post-calving) cows, with multiple recordings per cow per transition event, as the monitoring process was repeated every month. There were a total of 23 variables available (Table 2.1). The cow comfort dataset included data on individual pens, with 26 variables available (Table 2.1).

Table 2.2 Variables available in all the original datasets

Cow	Cow	Farm records			
Variables	Comfort				
	Variables	Cow level	Insemination level	Test-day level	Event level
Farm ID	Farm ID	Farm ID	Lactation Number	Daily Milk Yield	Event
Period	Period	Animal ID	Lactation ID	Protein	Event ID
Date	Type of cows	Ear Tag	Insemination Date	Butterfat	DIM at event
Assessor	Pen Name	Date on Farm	DIM at insemination	Milk ID	Event Date
Cow Number	Type of Pen	Date of Birth	Insemination ID	Date of Milk recording	
Dry or Fresh	Pen length	Date of Exit		DIM at Milk recording	
Heifer	Pen width	Calving Dates			

Found for Pen area
assessment
nt

BCS Feed Fence
 Space

Rumen Water Trough
Fill Space

Hock Neck Rail
Hygiene Height

Lameness Number of
 Cows in Pen

Milk Fever Cows waiting

LDA Cows not
 waiting

RFM Feed
 available

Calf Water
mortality available

Twinning Temperature

Metritis Humidity

Daily	Milk	Bedding
Yield		quality
Protein	Air quality	
Butterfat	Feed quality	
Cell count	Water quality	
Drying-off	Light quality	
cell count	Cubicle count	
	Pen score	

The farm records contained additional information collected by milk recording organizations (Table 2.1). The events that could be recorded in this format included calving, insemination, drying-off, mastitis, lameness, general health issues categorized as “sick”, positive and negative pregnancy diagnosis, abortion, a decision to not breed the cow any longer (DNB), and whether the cow was sold or culled. Out of those, the calving, insemination, positive and negative pregnancy diagnoses, abortion, DNB, mastitis diagnosis and sold or culled cows were included in the final analysis (other events were recorded inconsistently across herds).

The milk variables in the cow scores dataset (milk yield, protein, butterfat, cell count and drying-off cell count) were not used, as these were captured by the assessors from the farm’s milk recording data (which was directly available here, so these variables were gathered from the milk recording dataset rather

than the assessor data). As a next step, additional variables were created. A cow ID was created using the farm ID and cow number, in order to differentiate between the same cow numbers used by different farms. The datasets were amalgamated and restructured so that each unit (line) of data represented a transition or calving event for a given cow; with pre-calving variables (e.g. BCS, pen stocking density) included alongside post-calving variables (e.g. BCS, occurrence of periparturient disease). In order to select a “best” pre-calving score for each calving event, the closest scoring event to 20 days pre-calving was chosen. Similarly, the post-calving scoring occasion closest to 25 DIM was used as the representative post-calving scoring event for a given transition event.

Additional variables on stocking density, feed and water space per cow, month and season of recording were calculated. A variable for the overall subjective quality of the pen environment was added, as a combination of the variables describing the quality of feed, water, light, bedding and air (ordinal, scale 0-5). The temperature-humidity index was also calculated, using the formula given by NRC (2001):







$$THI = (1.8 \times T + 32) - (0.55 - 0.0055 \times H)$$

where T is the dry light bulb temperature (°C) and H is the relative humidity of the air (%). The threshold for heat stress in dairy cows is reported to be between scores 68 and 71, mild heat stress between 72 and 78, moderate between 79 and 88 and severe between 89 and 98, while dairy cattle cannot survive in values above 99 (Table 2.2)

Table 2. 2 THI values and interpretation *

Temp		% Relative Humidity																	
F	C	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
72	22.2	65	66	66	67	67	67	68	68	69	69	69	70	70	70	71	71	72	72
73	22.8	66	66	67	67	68	68	68	69	69	80	80	71	71	71	72	72	73	73
74	23.3	67	67	67	68	68	69	69	70	70	70	71	71	72	72	73	73	74	74
75	23.9	67	68	68	68	69	69	70	70	71	71	72	72	73	73	74	74	75	75
76	24.4	68	68	69	69	70	70	71	71	72	72	73	73	74	74	75	75	76	76
77	25.0	68	69	69	70	70	71	71	72	72	73	72	74	74	75	75	76	76	77
78	25.6	69	69	70	70	71	71	72	73	73	74	74	75	75	76	76	77	77	77
79	26.1	69	70	70	71	71	72	73	73	74	74	75	76	76	77	77	78	78	79
80	26.7	70	70	71	72	72	73	73	74	74	75	76	76	77	78	78	79	79	80
81	27.2	70	71	72	72	73	73	74	75	75	76	77	77	78	78	79	80	80	81
82	27.8	71	71	72	73	73	74	75	75	76	77	77	78	79	79	80	81	81	82
83	28.3	71	72	73	74	75	75	75	76	77	78	78	79	80	80	81	82	82	83
84	28.9	72	73	73	74	75	75	76	77	78	78	79	80	80	81	82	83	83	84
85	29.4	72	73	74	75	75	76	77	78	79	79	80	81	81	82	83	84	84	85
86	30.0	73	74	74	75	76	77	78	78	79	80	81	81	82	83	84	84	85	86
87	30.6	73	74	75	76	77	77	78	79	80	81	81	82	83	84	85	85	86	87
88	31.1	74	75	75	76	77	78	79	80	81	81	82	83	84	85	86	86	87	88
89	31.7	75	75	76	77	78	79	79	80	81	82	83	84	85	86	86	87	89	89
90	32.2	75	76	77	78	79	79	80	81	82	83	84	85	86	86	87	88	89	90
91	32.8	76	76	77	78	79	80	81	82	83	84	85	86	86	87	88	89	90	91

92	33.3	76	77	78	79	80	81	82	83	84	85	85	86	87	88	89	90	91	92
93	33.9	77	78	79	80	80	81	82	83	84	85	86	87	88	88	90	91	92	93
94	34.4	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
95	35.0	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
96	35.6	78	79	80	81	82	83	85	86	87	88	89	90	91	92	93	94	95	96
97	36.1	79	80	81	82	83	84	85	86	87	88	89	91	92	93	94	95	96	97
98	36.7	80	80	82	83	84	85	86	87	88	89	90	91	93	94	95	96	97	98
99	37.2	80	81	82	83	85	86	87	88	89	90	91	92	93	94	96	97	98	99
100	37.8	81	82	83	84	85	86	87	88	90	91	92	93	94	95	97	98	99	100
101	38.3	81	82	83	86	86	87	88	89	90	92	93	94	95	96	97	99	100	101
102	38.9	82	83	84	85	86	87	89	90	91	92	94	95	96	97	98	99	101	102
103	39.4	82	83	86	86	87	88	89	91	92	94	95	96	97	98	100	101	102	103
104	40.0	83	84	85	86	88	88	90	91	93	94	95	96	97	99	100	101	103	104
105	40.6	83	84	86	87	88	89	91	92	93	96	96	97	98	99	100	101	104	105
106	41.1	84	85	86	88	89	90	91	93	94	95	97	98	99	101	102	103	105	106

*  No Heat Stress (<68),  Light Heat Stress (68-71),  Moderate Heat Stress (72-78),  Severe Heat Stress (79-89),  Life-Threatening Heat Stress (90 – 98),  Dead Heat Stress (>99)

2.1.3 Pre-processing and analysis

Overall disease prevalence and disease prevalence per farm were calculated, as well as the proportion of cows waiting in pens per farm. These were plotted against the stocking density, feed fence space per cow and water trough space per cow.

Both predictive and inferential approaches were used to evaluate associations between the available predictor variables and the various outcomes of interest (disease, reproduction and production); outcome variables are described in more detail in the relevant chapters. These two approaches are briefly discussed below.

When it came to missing data, while we considered many approaches that would substitute or approximate missing values in the end none proved valuable to our analysis. This was due to some variables with missing variables had a very high proportion of lactations missing. Therefore, it would not be useful to impute them, for example, to impute them. Hence we decided that variables where the majority of the datapoints were missing were unusable in the analysis while for the rest of the data we deleted rows with missing data when fitting the model. This left a dataset with a high proportion of “complete cases” (no missing variables in a lactation), while retaining the majority of its total size.

Scaling of predictor variables was also considered (e.g. min-max scaling that put all values for a given variable on a scale of 0 to 1). It was explored in each set of predictive models. However, as it was found to not substantially improve model performance metrics, it was decided to report results without any scaling so that values such as the RMSE and MAE can be more easily interpretable and comparable to their inferential model equivalents.

2.1.3.2 Inferential and Predictive analytics

The main difference between the two methodologies lies in their purpose. In general terms, inferential analysis aims to uncover potential causal

associations between variables, whereas predictive analysis focuses on predicting future outcomes, without necessarily gaining any insights as to what variables may influence the predicted outcome (Meeker, 2017). This differentiation in end goals results in various differences in the process of making the models. For example, during predictive modelling the dataset is usually divided into smaller parts, with a holdout subset of the data being used after the model is built as validation of the model's performance. The metrics used to evaluate each model also differ, with accuracy or kappa primarily applying to predictive models (and measuring overall model predictiveness), whereas odds ratios (OR) more commonly used in inferential models (as measures of relative influence of each predictor variable in the model). More details on how we approached each category and the specific methodologies we used are explained below. Some inferential modelling techniques (such as logistic and linear regression) are also used in predictive modelling, and likewise there are approaches that allow the contribution of each predictor variable to a given predictive model to be evaluated.

2.1.3.4 Predictive Models

A variety of predictive models were used for each outcome of interest, with the primary aim of determining which produced most accurate predictions.

To ensure that all results could be reproducible, even after sampling for the test and training data split, a random seed was set at 23. A sample code for the predictive model fitting is presented in appendix 1.

In some cases we had to consider the possibility of data leakage. Overly optimistic results may stem from data leakage, which occurs when information not available at prediction time is used during model training (Yagis et al., 2021). Data leakage can result from target leakage or incorrect data splitting. For instance, leakage may happen if feature selection is based on the entire dataset prior to cross-validation (Reunanen, 2003, Varma and Simon, 2006), allowing the target variable of test samples to inadvertently enhance the learning process. Incorrect data splitting can also cause leakage, such as when data augmentation is performed before separating the test set from training data. In this scenario, augmented data from the same original image might appear in both training and test sets, leading to artificially inflated performance (Wen et al., 2020). Another form of train-test contamination involves using the same test set to optimize training hyperparameters and evaluate model performance (Varma and Simon, 2006). Information leakage can also occur with longitudinal data if future information leaks into past data. An especially problematic form of data leakage happens when target information inadvertently becomes part of the input data (Yagis et al., 2021).

In order to avoid data leakage, the dataset was separated into train and test data making sure that the grouping variables used did not overlap in between datasets. More specifically, the original dataset as a whole was grouped into either herd/month or herd/trimester groups, depending on the occasion. It was then split 80% to 20% on the condition that datapoints that belonged in the same group were all included into one of the two parts. That would mean that once a specific herd/time was “drafted” to be included in the 80% training set all the data points included in that particular group would automatically be

assigned to the training set. That way, variables that were used for training a model and may have been the same within herd/time group were not presented to the model again for predictions, which would lead to the model “cheating” the answer through data leakage. The 80% of the data was then used for training the models, while the remaining 20% was utilised to evaluate predictions. Once the models were finalised based on the 80% of the data, they were tested on the remaining 20%. The predictive outcomes were compared to the actual ones and a Pearson’s correlation coefficient looking at the relationship of the predicted vs the actual outcome per group was then determined to investigate the predictiveness of the models.

Data leakage is an issue that may accidentally occur during data preparation. Usually it is a subtle, unnoticed availability of test data information to the model in the training dataset (Brownlee, 2020). Essentially, future datapoints are becoming available in the past which gives the model an “unfair” advantage as the predictions are not actual predictions but actual known information (Zheng, 2018).

2.1.3.4.1 Binary Outcomes

Machine learning techniques used for binary outcome variables included logistic regression, decision trees, random forest, artificial neural networks, naïve Bayes, support vector machines and k-nearest neighbours. Additional information on these algorithms is provided in the next section.

Forward selection was used in all approaches to determine the final variables that would be included in the models. As a measure of evaluation, we used the kappa value which is the difference between the observed agreement and expected agreement by chance. After adding each variable to the model, the change in kappa value was used to assess whether the variable improved the predictiveness of the algorithm and should therefore be included in the final analysis. Any variables that contributed to any increase of kappa were considered eligible for inclusion.

A 10 fold cross-validation was used in all predictive model analysis to measure predictiveness of the models using the “caret” package (Kuhn, 2008). Some algorithms include tuning parameters that can be selected and altered to potentially improve the predictive value of the model. One such example is the prune parameter for the random forest methods, which sets the number for maximum decision trees in the forest. For our analyses, tuning parameters for all algorithms and their various combinations were evaluated automatically by the “caret” package during model building and those that produced the best kappa value were chosen. The kappa values were interpreted using the thresholds stated in Viera and Garrett (2005) with values lower than 0 indicating less than chance agreement, values between 0.02 and 0.20 slight agreement,

values between 0.21 and 0.40 fair agreement, values between 0.41 and 0.60 moderate agreement, values between 0.61 and 0.80 substantial agreement, and finally values between 0.81 and 0.99 almost perfect agreement.

For situations where there was substantial imbalance in a binary outcome variable (such that one class was much more common than the other), various sampling methods were considered. Sampling methods are techniques commonly used when dealing with imbalanced datasets, with the aim of “rebalancing” the outcome variable, to minimise the risk of the predominant class having undue influence on the model, which can result in poor predictive performance in the minority class. The sampling method of choice for our studies was up-sampling, meaning randomly duplicating the minority class until both classes were of equal size (Aghdam, 2017).

The rest of the metrics that were calculated and compared between the models were accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), kappa, balanced accuracy (the average of the individual accuracies per class), detection rate, F1 (the harmonic mean of the sensitivity and specificity) and area under the receiver operating characteristic curve (AUROC). Each metric is calculated using the confusion matrix (Table 2.3); definitions for each metric are shown in Table 2.4.

Table 2.3 Confusion matrix

	1	0
	(Predicted)	(Predicted)
1		
(Observed)	True Positive (TP)	False Negative (FN)
0		
(Observed)	False Positive (FP)	True Negative (TN)

Table 2.4 Definition of commonly used model metrics

Metric	Formula
Accuracy	$\frac{TP + TN}{TP + FP + FN + TN}$
Sensitivity/True Positive Rate/Recall	$\frac{TP}{TP + FN}$
Specificity/True Negative Rate	$\frac{TN}{TN + FP}$
Positive Predictive Value/Precision	$\frac{TP}{TP + FP}$
Negative Predictive value	$\frac{TN}{TN + FN}$
Kappa	$\frac{(\text{observed accuracy} - \text{expected accuracy})}{(1 - \text{expected accuracy})}$
Balanced accuracy	$\frac{(TP/TP + FP) + (TN/TN + FN)}{2}$
F1	$\frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}$
Detection rate	$\frac{TP}{TP + FN}$

TP + FP + FN + TN

2.1.3.4.2 Continuous Outcomes

The models that were used to predict continuous outcomes were linear regression, artificial neural networks, multivariate adaptive regression spline (MARS), decision trees and random forest. Additional information on these algorithms is provided in the next section.

The process of modelling was similar to that of the binary outcome models, the main difference being the metrics used to evaluate performance. The main metric in this case was the R^2 , with others such as the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) complementing it. The R^2 determines the percentage of the variation in the outcome variable that can be explained by model predictions. RMSE is essentially the standard deviation of the model residuals, indicating how far from the regression line data points lie. Similarly, MAE tells us how far from the truth our predictions are on average. The exact interpretation of RMSE as well as MAE values are depended on the actual values of the outcome and its range. Both MAE and RMSE are considered better the closer they are to 0, while R^2 is considered best the closer it is to 1.

2.2 Machine learning Approaches

The following are some common machine-learning techniques and that were used during this project, along with a short description about the pros and cons of each method.

2.2.1 Naive Bayes

Naive Bayes is a classification method based on Bayes' theorem, which describes the probability of a hypothesis given the evidence. In the context of classification, it calculates the probability of one class given the predictive variables. The formula for the theorem is:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Where: - $P(A|B)$ is the probability of hypothesis A given the evidence B. - $P(B|A)$ is the probability of evidence B given the hypothesis A. - $P(A)$ and $P(B)$ are the probabilities of A and B independently (Pawlak, 2003). It, therefore, combines prior probability as well as conditional probability in a formula used for the probability calculation of each class (Bramer, 2007).

The method simplifies the Theorem by assuming independence amongst the predictors, regardless of any actual correlations (Williams et al., 2016, Yang and Webb, 2001). This means that the presence of a particular feature in a class is independent of the presence of other features. For some studies this assumption may be unrealistic, however it often outperforms more complex methods in efficiency (Benaissa et al., 2017). Another advantage is that it is not computationally challenging, even for large datasets (Kuhn and Johnson, 2013). Naive Bayes calculates the posterior probability of each class given the features of a data point using Bayes' theorem in order to classify it. Whichever

class has the highest posterior probability is assigned to said data point (Pazhanikumar and Aswathi, 2020). This method has numerous advantages, including the ability to handle missing values, analyse both continuous and discrete data, speed, efficiency and robustness when it comes to irrelevant features (Jollyta et al. 2019). It has also been shown to be robust when handling imbalanced datasets (Somasundaram and Reddy, 2018).

2.2.2 Neural Networks

Artificial neural networks (ANN) are powerful machine-learning methods inspired by the structure of the brain (Haykin, 1998). A neural network is a system of interconnected artificial neurons (Kearns and Vazirani, 1994). Each neuron is characterized by a weighted sum of input values represented as an inner product plus a bias value, which is then passed through an activation function, such as a linear function or a sigmoid function (Raiko et al., 2012). The outcome is determined by linear combinations of the predictors, called hidden variables or units. A neural network is a multi-layer network with three layers, the input layer, hidden layer and output layer (Li and Wang, 2018). The linear combinations are then transformed by a non-linear function which is achieved by increasing the size of the hidden layer and allows neural networks to model complex non-linear relationships (Murphy et al., 2014). Due to this ability, however, they tend to overfit models, there are however ways to address this issue, such as applying weights as a penalization method (Kuhn and Johnson, 2013). The feed-forward back-propagation network (FFBP) is the most commonly employed neural network architecture. The backpropagation algorithm includes two phases, the feedforward phase, where the external input data at the input nodes is propagated forward to calculate the output signal at

the output nodes, and a backward phase, where adjustments to the connection weights are made based on the discrepancies between the calculated and observed output signals (Alizadeh et al. 2011, Chen, 2018). In neural networks, a significant behaviour is the weighted sum of node states, which is highlighted as crucial for information expression and encoding in several biological studies (Jazayeri and Movshon, 2006, Majaj et al. ,2015, Schnitzer and Meister, 2003), while also considered a vital and fundamental step during the operation of artificial neural networks (Fei et al., 2018). Individual nodes, on the other hand, can only express limited and coarse information. Therefore, studying and controlling the weighted sum of node states is essential for the operation of neural networks. Another important aspect of ANN is the loss function. The loss function measures the distance between the model's output and the actual value in a neural network. A smaller loss function value indicates that the model's output is closer to the real data, thereby increasing the model's accuracy (Viju, 2021). Commonly used loss functions include the mean absolute error loss function and SVM (Iida and Kiya, 2019).

Recently, it has been suggested that the training of deep neural networks demonstrates a spectral bias (Rahaman et al., 2019, Xu, 2018), meaning that low frequencies are learned more quickly during training via stochastic gradient descent. This bias is proposed as a mechanism that steers networks toward low-complexity solutions (Rahaman et al., 2019).

Neural networks can accommodate the non-linearity, uncertainty, and complexity of control systems, demonstrating strong robustness and adaptability (Chen and Ji, 2016). However, these models tend to be referred to as black box. A black box system is characterized by having an unknown

topology and/or parameters, typically interpreted through its input and output signals (Valdivia et al., 2009). Generally, the outputs are the response to stimuli or excitation applied to the black box in the form of input values or vectors. Estimating the topology and parameters of such a system, given only the input and output values, is a challenging problem (Rojas-Duenas et al., 2020). Therefore, the interpretation of how such models work can be challenging.

2.2.3 Support Vector Machines

Support Vector Machines (SVMs) are a category of flexible and robust modelling methods (Kuhn and Johnson, 2013). They are used for classification, based on identifying the most appropriate hyperplane that divides the data into two by employing a kernel function (Resheff et al., 2014). The boundary between the classes does not have to be linear and the method can be expanded to include more than two classes (James et al., 2014). The SVM adeptly and efficiently manages these two types of data:

For linearly separable data: where an optimal hyperplane can be delineated to distinguish between the two classes using training data. This hyperplane can be described by the equation:

$$x_i \cdot w + b = 0$$

where w represents the weight vector, b is the bias (or $-b$ is the threshold), and x_i denotes an observation. We can establish two additional hyperplanes, H_1 and H_2 , which are parallel to the separating hyperplane. The space between these two planes is known as the SVM's margin. The objective is to determine the optimal hyperplane that maximizes the margin while maintaining equidistance from both H_1 and H_2 (Rejab et al., 2014).

SVM can also be employed to distinguish between classes that cannot be separated with a linear classifier. In such instances, the initial observations are transformed into a feature space, which may be of high dimensionality or even infinite, using non-linear functions known as feature functions ϕ . Within this new space, a linear classifier can effectively separate the classes (Rejab et al., 2014).

SVM, being a kernel-based method, supports the use of several kernels that satisfy the Mercer condition such as Gaussian, polynomial, wavelet, and others (Smola et al., 1998). Its performance heavily relies on the appropriate selection of the parameter values, including the kernel function, kernel parameter values, and the regularization parameter, among others (Cristianini and Shawe-Taylor, 2000)

SVM relies on maximizing the margin, which is the distance between the hyperplane and the nearest data points from each class, and minimizing structural risk (Vapnik, 2013). Given its effectiveness with small datasets, SVM stands as an innovative approach for analysing microarray data (Guckiran et al., 2019).

SVM has been widely used in the machine learning field due as it adequately handles high-dimensional data, and it possesses robust generalization properties, as well as the ability to establish the classifier architecture once the kernel function and parameters are selected by the user (Vapnik, 2013). The disadvantage of the method is that it requires careful parameter tuning in order to classify the data points correctly, however once the parameters are set correctly it can perform very well for a variety of classification problems,

including non-linearly separated classes (James et al., 2014). Another limitation is that while SVM offers effective solutions for both linear and nonlinear data, it does not inherently incorporate new information provided over time. To address this constraint, modified versions of SVM have been proposed (Bordes et al., 2005, Cauwenberghs and Poggio, 2000).

2.2.4 Decision Trees

Decision trees are amongst the most widely used machine learning methods, as they are easy to interpret (Shahinfar et al., 2014). Decision trees are straightforward classifiers composed of decision nodes organized in a tree structure. New observations pass by internal nodes, split into branches and reach the leaves, which are the final classes of the model (Shahinfar et al., 2014). In more detail, each decision node corresponds to a predicate or test on the query. Evaluating a decision tree involves traversing the tree (Wu et al., 2016). The decision tree is created by recursively partitioning the training data using a splitting attribute until all records in each partition belong to the same class. The splitting attribute is selected based on the value of a node splitting measure (Chandra and Paul Varghese, 2009).

Issues of overfitting and complexity in resulting trees have highlighted the need for pruning procedures. It has been argued that simplifying trees by removing parts that do not contribute to classification accuracy can improve the performance of nearly all decision trees (Garcia-Almanza and Tsang, 2006).

They have the advantage of being applicable to both regression and classification problems (James et al., 2014). They are thought to mimic human decision-making and they are easily presented graphically, which makes them

ideal for communicating results (James et al., 2014). Not only are they easily interpretable, but they can be directly converted into if-then-else rules (Wei and Hsu, 2008). Decision-tree algorithms are highly efficient, capable of processing a large number of records with numerous fields while maintaining predictable response times (Krishnan et al. 1999). Furthermore, they are considered a robust method of dealing with missing data (Shahinfar et al., 2014). However, they generally do not have the same level of accuracy as other methods (James et al., 2014) while also more prone to overfitting. Moreover, decision-tree algorithms typically handle only one attribute at a time, disregarding dependencies among attributes, which are common in real-life datasets (Wei and Hsu, 2008).

2.2.5 Random Forests

Random forests are a powerful ensemble learning method that consists of multiple decision trees. They employ bootstrap aggregating, also known as bagging, to create multiple models, resulting to enhanced prediction accuracy (Breiman, 2001). Bagging involves randomly selecting examples from the training set to grow each tree, without replacement (Breiman, 1996). Another example of random vectors that influence the development of each tree is the random split selection, in which a split is chosen randomly from among the K best splits at each node (Dietterich, 2000). Feature randomness in building each individual tree is also employed to create an uncorrelated forest of trees, whose combined predictions are more accurate than those of any single tree (Rigatti, 2017). A subset of features is selected at random in each iteration of building a tree, making them more robust than decision trees yet computationally efficient at the same time. By utilising these techniques, a

group of low-correlated decision trees is constructed. A class prediction is produced by each individual tree, and the final prediction of the random forest is determined by the class with the most votes (Li, 2023). Random forests entail several hyperparameters that control the structure of each tree, such as the minimal node size required for a split, as well as the structure and size of the forest, including the total number of trees. Additionally, it manages the randomness by determining the number of variables considered as candidate splitting variables at each split (m_{try}) and the sampling scheme used to generate the datasets on which the trees are built (Probst et al., 2019).

Random forests can deal with high-dimensional data and missing data (Shahinfar et al., 2014). It has been demonstrated that the random forest algorithm exhibits high accuracy as well as robustness, has good tolerance for noise and outliers, can manage high-dimensional datasets, and is resistant to both overfitting and underfitting (Fang et al., 2011). It can determine each variable's weights and efficiently evaluate their importance and role in the model, all while maintaining good generalizability (Ouyang and Chen, 2020). In terms of drawbacks, random forest classifiers may underperform on highly complex and nonlinear problems (Fawagrah et al., 2014). Additionally, random forest classifiers can be computationally intensive and slow in making predictions, particularly with large datasets and numerous trees in the forest (Schonlau and Yuyan Zu, 2020). Moreover, they may struggle with unbalanced data, where one class is underrepresented (Breiman, 2001, Schonlau and Yuyan Zu, 2020), potentially resulting in biased outcomes favouring the majority class. Lastly, while random forest classifiers can provide feature importance,

they are not as interpretable as models like linear regression, where the coefficients have a clear and direct meaning.

2.2.6 K-nearest neighbours

K-nearest neighbour (KNN) is the simplest non-parametric machine learning approach used for classification (James et al., 2014). It is an instance-based learning method, utilized to generate candidate labels, with improvements made by weighting votes based on the similarities between an instance and its neighbours (Qu et al., 2011). It can use both linear and non-linear boundaries to separate the data (Kuhn and Johnson, 2013). In KNN, determining the *K* value and conducting nearest neighbour queries are two crucial issues. Nearest neighbour queries can be addressed using various distance measurement functions. In more detail, this method classifies the cases according to the status of the majority of their nearest data points, usually “nearest” being defined as the closest Euclidean distance, Mahalanobis distance, Manhattan distance, and cosine similarity (Benaissa et al., 2017, Kuhn and Johnson, 2013). For calculating the *K* value, the prevalent methods include expert settings or cross-validation techniques (Zhang and Li, 2021). For classification, the predicted class for the test instance is determined through a majority vote among its *k* neighbours in the training set (Schlemmer et al., 2014), while for regression tasks, the algorithm calculates the average of the *K* nearest neighbours' target values. In regression on the other hand, the property value is determined by averaging the values of its *k* nearest neighbours (Jing et al., 2016).

The benefits of KNN encompass its simplicity in comprehension and result interpretation, and its suitability for nonlinear data. Furthermore, it is resilient to

noisy training data and can be used for multi-class classification (Kramer, 2013). Additional advantages of KNN are its independence from any particular data distribution, the use of local information, and the ease of interpreting outcomes (Kiyak et al. 2021). KNN methods are also more flexible than linear regression, however they are not as easily interpretable and they do tend to underperform compared to the latter when there is a small number of observations per predictor (James et al., 2014). Regarding the downsides, KNN can be computationally expensive for large datasets, as it requires calculating distances for all data points (Kepa and Szymanski, 2015). Furthermore, in settings of high dimensionality, it is impacted by nuisance (noninformative) features and suffers from the "curse of dimensionality" (Aggarwal et al., 2001, Lu et al., 2013, Radovanovic et al., 2010).

2.3 Inferential Models

2.3.1 Binary Outcomes

A mixed effects logistic regression model was used, to take into account the fact that the animals were clustered into farms. As most farms had one dry pen and one fresh pen (if at all) only one random effect "level" (farm) was used. Univariable analysis was conducted using all the available potential predictor variables, to determine the ones that appeared to be associated with the outcome and could potentially be included in the multivariable model. Further multivariable analysis was conducted including all the variables that met the threshold for the univariable correlation with the outcome. Backwards elimination was applied, meaning variables that in the multivariable analysis

had a p-value higher than a threshold of 0.05 were eliminated one by one, starting from the highest value, and the model parameters estimated again until all remaining variables have a p-value below the threshold. The fit of the models were assessed using Hosmer-Lemeshow test for logistic regression and standard residual plots for linear regression.

And finally, the odds ratios were calculated as described in Appendix 1, which concluded the analysis.

2.3.2 Continuous Outcomes

For the continuous outcomes a mixed effect linear regression model was fitted. Similarly to the mixed effects logistic regression models, the farm was added as the random effect and they were built using a backwards elimination method.

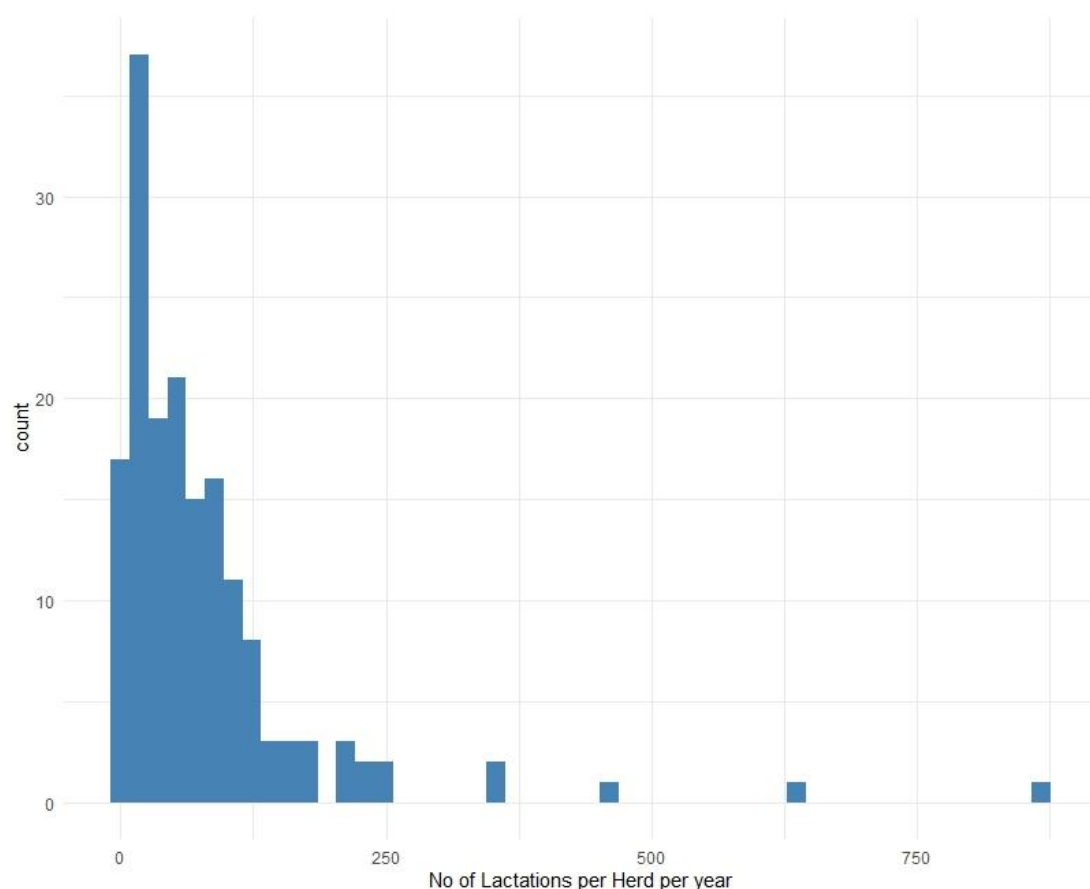
For time-based outcomes, survival analysis was conducted and a Cox proportional hazards model was fitted (Harrell, 2001). The models were fitted using backwards selection.

Chapter 3 - Descriptive statistics

3.1 Original Dataset

The cow scores dataset consisted of 71,665 observations collected from 15th April 2016 until 24th October 2018. The total number of cows included in the dataset was 32,867 from 133 farms. After removing duplicate data, the total observations dropped to 68,029. The number cows having at least one recording pre-partum was 20,733 with a total of 27,659 unique recordings. Similarly, the number of cows with at least one recording post-partum was 27,901 with a total of 40,370 unique recordings. After filtering the date differences so that only pre- and post-partum recordings that matched the same calving date were included, the total number of lactations that had both pre- and post- partum scores was 13,244, with 11,007 cows from 79 different farms. The difference between the pre- and post- partum scoring dates ranged from a minimum of 5 days to a maximum of 77 with a median of 31 days (Q1 = 15 days, Q3 = 47 days). Lactation number ranged from 1 to 14 with a median of 3; 2,726 cows were in parity 1 (20.6%). As a proxy for herd size, the number of lactations per herd per year was measured and the resulting variable had a minimum value of 1, a median of 56 (Q1 = 21.5, Q3 = 92.5) and a maximum of 868 (Figure 3.1).

Figure 3.1 The number of total lactations included in each one herd per year of recording



3.1.1 Disease Distribution

The incidence of diseases and conditions was 3.7% for lameness in dry cows, 3.2% for lameness in fresh cows, 3.0% for milk fever, 1.0% for LDA, 4.0%, for RFM, 2.9%, for calf mortality, 2.4% for twinning and 5.2% for metritis (Figure 3.2). The seasonal disease incidence when plotted suggests some potential patterns, such as peak of RFM and metritis in spring and milk fever in winter (Figure 3.3). There are some clear outliers in the distribution of disease prevalences across farms (Figure 3.4), these could generally be attributed to the small numbers of cows scored in specific farms.

Figure 3.2 Overall incidence of diseases/conditions across 13,244 lactations in 79 UK herds

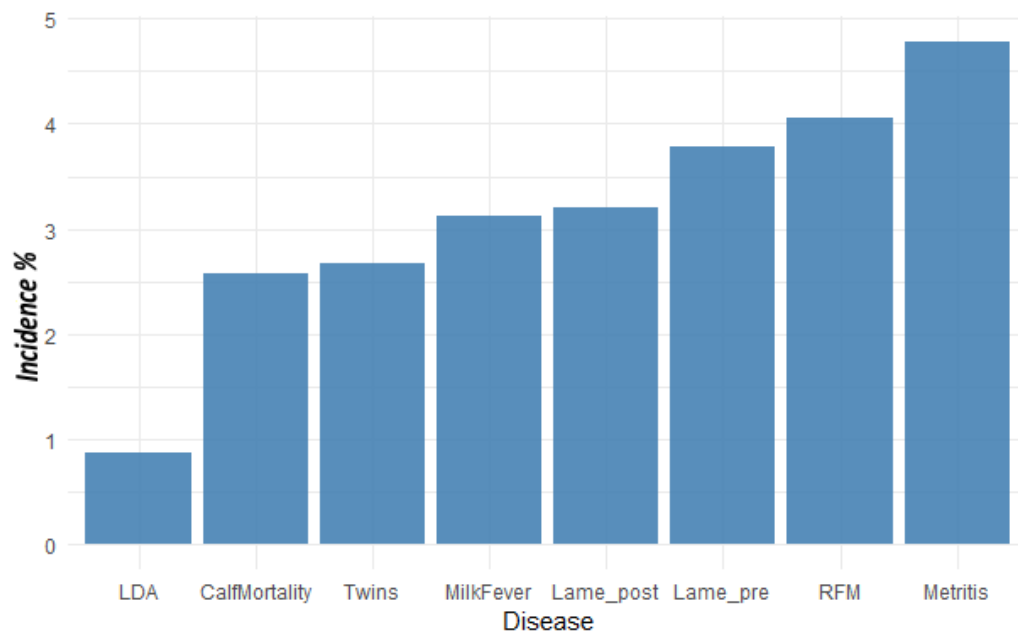


Figure 3.3 Seasonal disease incidence (%) of diseases/conditions across 13,244 lactations in 79 UK herds

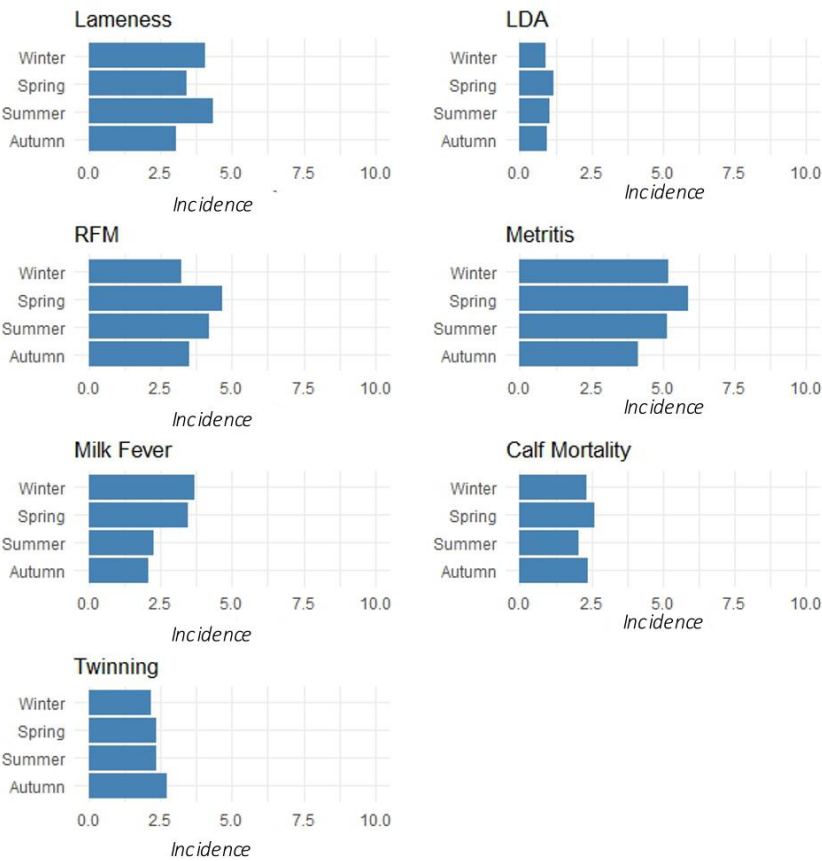
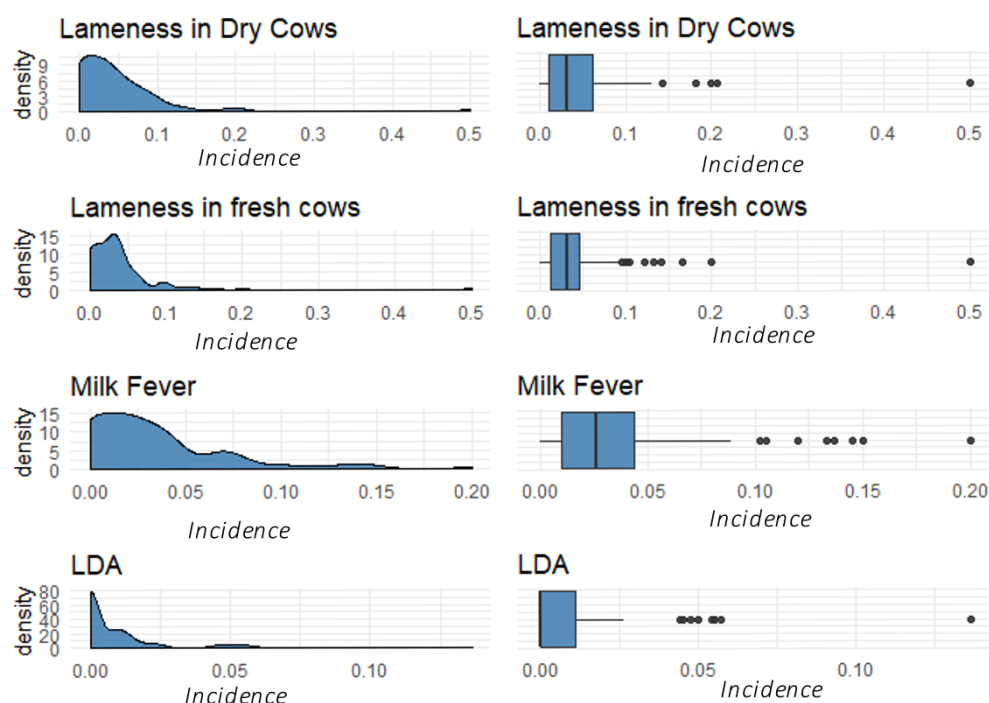


Figure 3.4 Diseases and conditions incidence distribution across 13,244 lactations in 79 UK farms



3.1.2 Cow level score variables

The BCS for dry cows ranged from 1.5 to 5 with a median of 3.75 (Q1 = 2.75, Q3 = 4) and 85 missing values (0.6%), whereas for the fresh cows it ranged from 1.0 to 4.5 with a median of 2.75 (Q1 = 2.5, Q3 = 3) and 1139 missing values (8.6%) (Figure 3.5). Overall dry cows have a higher BCS compared to fresh cows, which is also reflected on the BCS change. For this variable the missing values were 9% of the total dataset. In general, there was a drop in BCS, however not too severe, with the median being at -0.25 (minimum -2.25, Q1 = -0.50, Q3 = -0.25, maximum 1) (Figure 3.6). It is interesting to note that

the shape of the distribution from dry to fresh changes, with a lot of the dispersion in dry cows being pulled towards to median in fresh cows. Which is evidently translated as the drop in BCS witnessed in BCS change.

Figure 3.5 BCS distribution in dry and fresh cows throughout 13,244 lactations in 79 UK farms

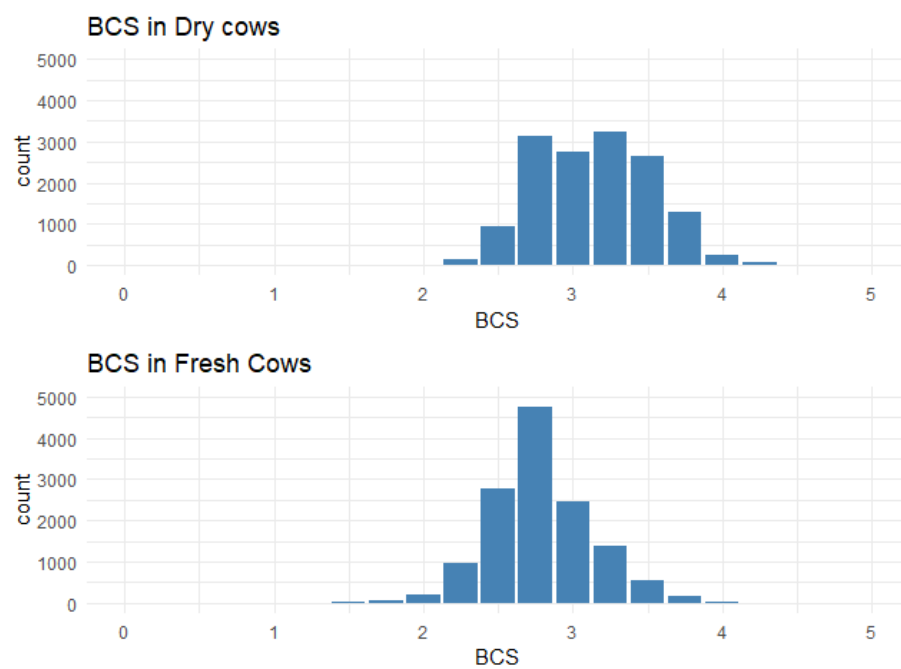
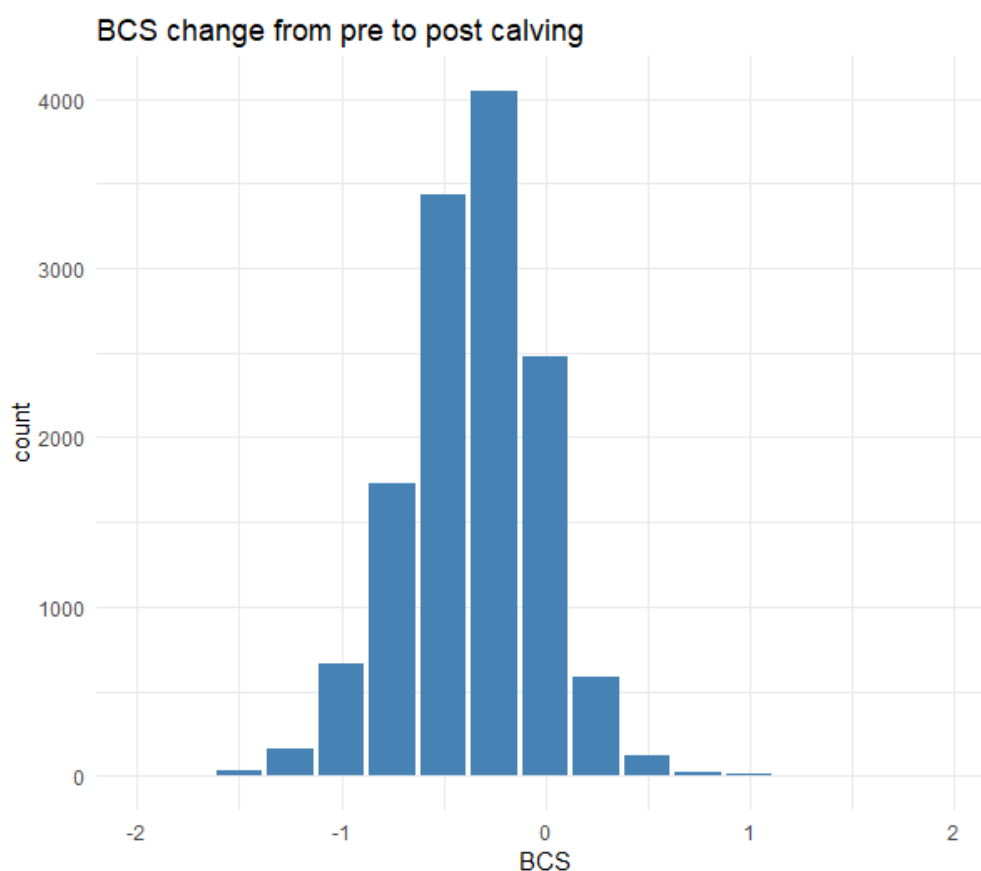


Figure 3.6 BCS change from dry to fresh cows in data of 13,244 lactations in 79 UK farms



Rumen fill scores had a median of 4 (Q1 = 3, Q3 = 4) and 94 missing values (0.7%) for dry cows, and 3 (Q1 = 2, Q3 = 3) for fresh cows and 1,149 missing values (8.6%) (Figures A2.1 – A2.2), and hock hygiene with a median of 3 (Q1 = 2, Q3 = 4) and 1,157 missing values (8.7%) (Figure A2.3).

3.1.3 Pen level variables

The number of farms included in the cow comfort dataset, prior to data cleaning, was 136, with a total of 2,761 distinct pen recordings. After including only the farms that matched the final version of the cow scores dataset the number of observations dropped to 1,923, from a total of 67 farms. The number of dry pen

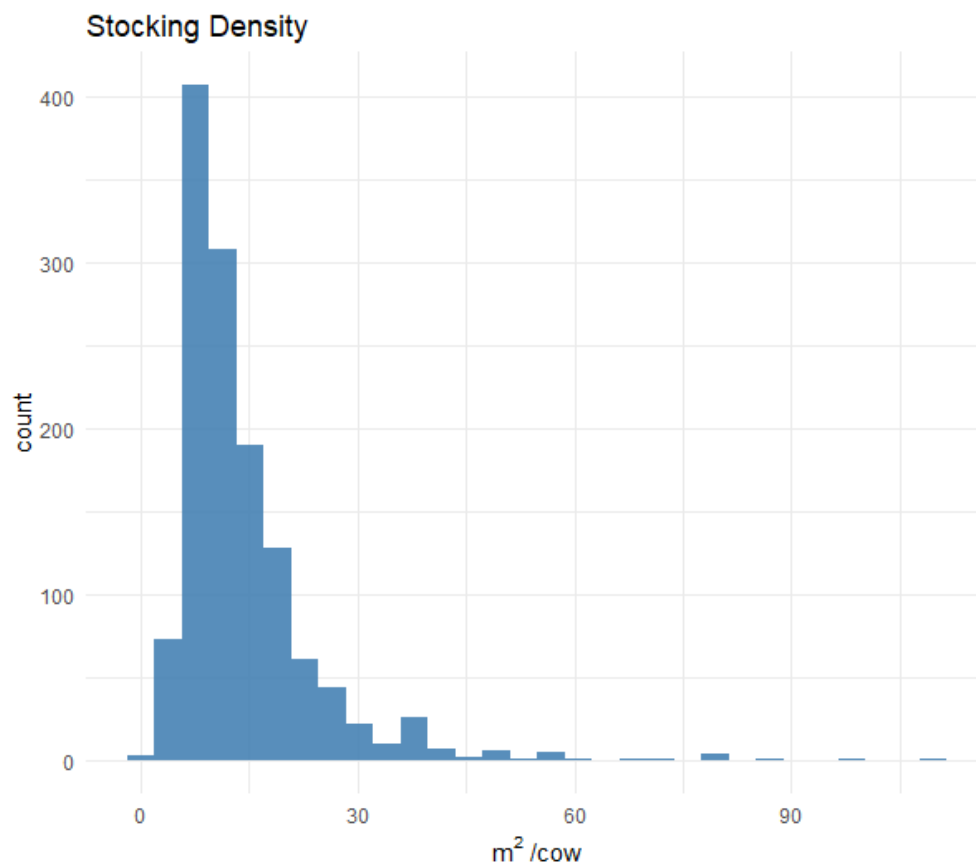
observations was 1,741 (90.1%) compared to 182 observations on fresh cow pens. This disparity was explained by scoring practices – pen scores were only assessed for fresh cows in herds which had a dedicated fresh cow (early lactation) group. The distribution of the potential quantitative predictors included in the cow comfort dataset is shown in Table 3.1. Additionally, a more detailed distribution of stocking density is displayed in Figure 3.7.

Table 3.1 Variable distribution on cow comfort data on 2,787 pens across 136 dairy cow herds

	Minimum	Q1	Median	Q3	Maximum	Missing data
Pen Area (m ²)	20.5	78.3	110.0	181.4	420.0	397 (20.6%)
Cows in Pen	1	6	10	18	81	514 (26.6%)
Feed Fence Space (m)	1.9	9	12.0	20	66	39 (2.0%)
Water Trough Space (m)	0.3	1	1.5	2	5.4	88 (4.6%)
Neck Rail Height (m)	0.5	1.2	1.3	1.5	3.0	183 (9.5%)
Stocking density (m ² /cow)	1.7	8.47	13.7	18.6	110.0	728 (37.9%)
Feed fence space per cow (m)	0.1	0.76	1.1	1.63	11.0	537 (28.0%)

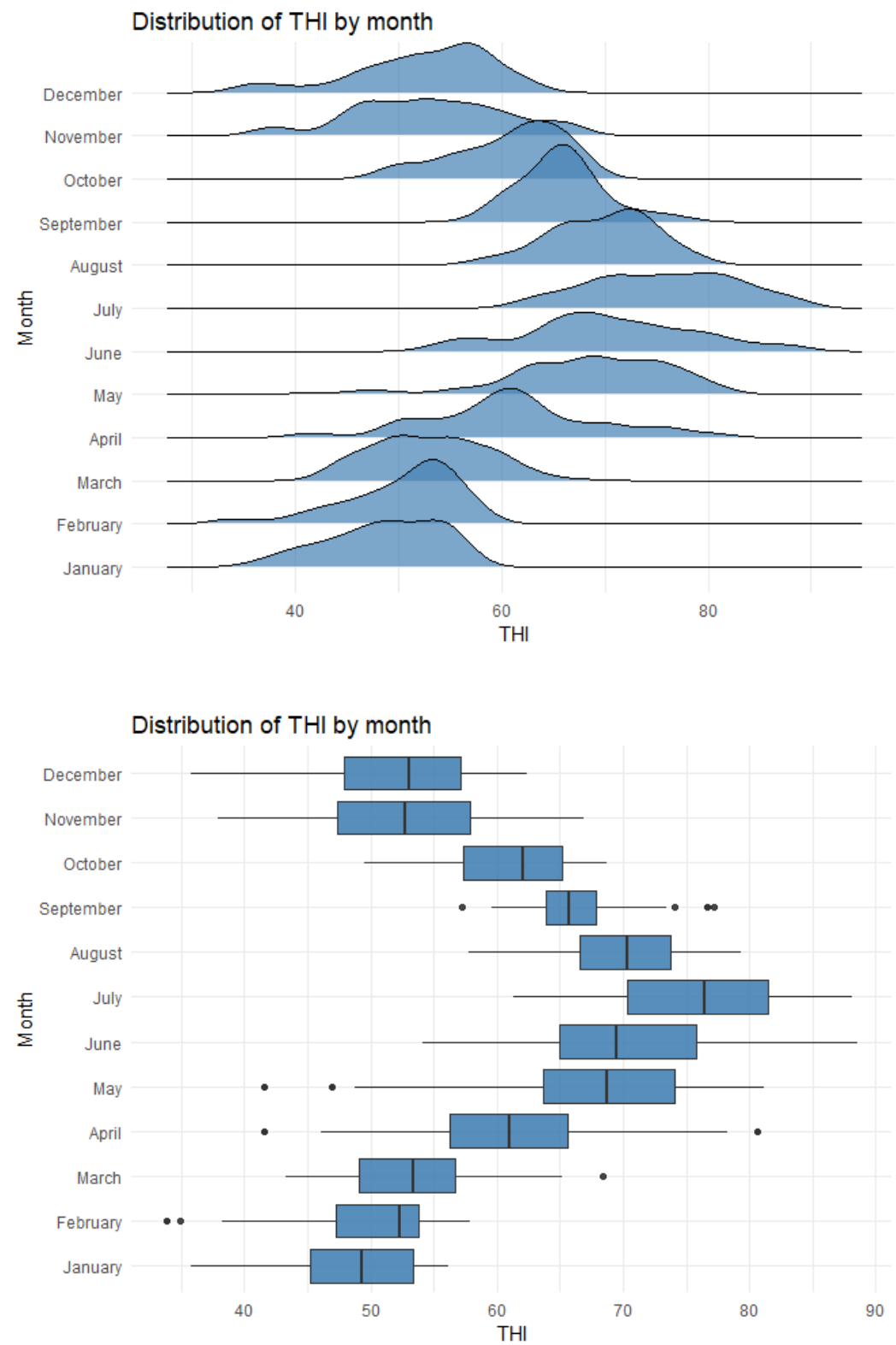
Water trough	0.02	0.08	0.13	0.2	7.22	551
space per						(28.7%)
cow						
Temperature	0.7	11.9	16.7	20.9	31.1	470
(°C)						(24.3%)
Humidity (%)	0.6	0.46	58.0	0.68	93.0	470
						(24.3%)

Figure 3.7 Stocking density distribution on 2,787 pens across 136 dairy cow herds



The distribution of THI by month revealed that cows in some pens in the dataset start experiencing heat stress as soon as April, about half have a THI > 72 in July and for some pens these conditions persist until September (Figure 3.8).

Figure 3.8 Monthly THI distribution in 2,787 pens across 136 herds



Dry cow pens were mostly straw yard pens (84.2%), while the fresh cow pens were more equally distributed between straw yard (52.2%) and cubicles

(47.8%) (Figure A2.4). Detailed distributions of neck rail height, feed fence space per cow and water trough space per cow, as well as their distribution for fresh and dry cows separately are also provided (Figure A2.5 – A2.9) and no obvious outliers were identified.

3.2 Milk Records

In the milking records, daily milk yield's first quartile laid at 24 L the median at 30.9L, and the third quartile at 38.3L, with 72,207 missing data (3.4%) (Figure 3.9). The percentage of protein in the milk had a minimum value of 1.0%, a maximum of 7.0% and a median of 3.28% (Q1 = 3.06%, Q3 = 3.54%), with a total of 73,863 missing data (3.5%) and the percentage of butterfat ranged from 1.27% to 10.0% with a median of 4.03% Q1 = 3.54%, Q3 = 4.57%) and similarly to protein percentage 73,863 missing values (3.5%) (Figures 3.10 - 3.11).

Figure 3.9 Daily milk yield distribution based on 564,962 milk recordings across 43,173 cows

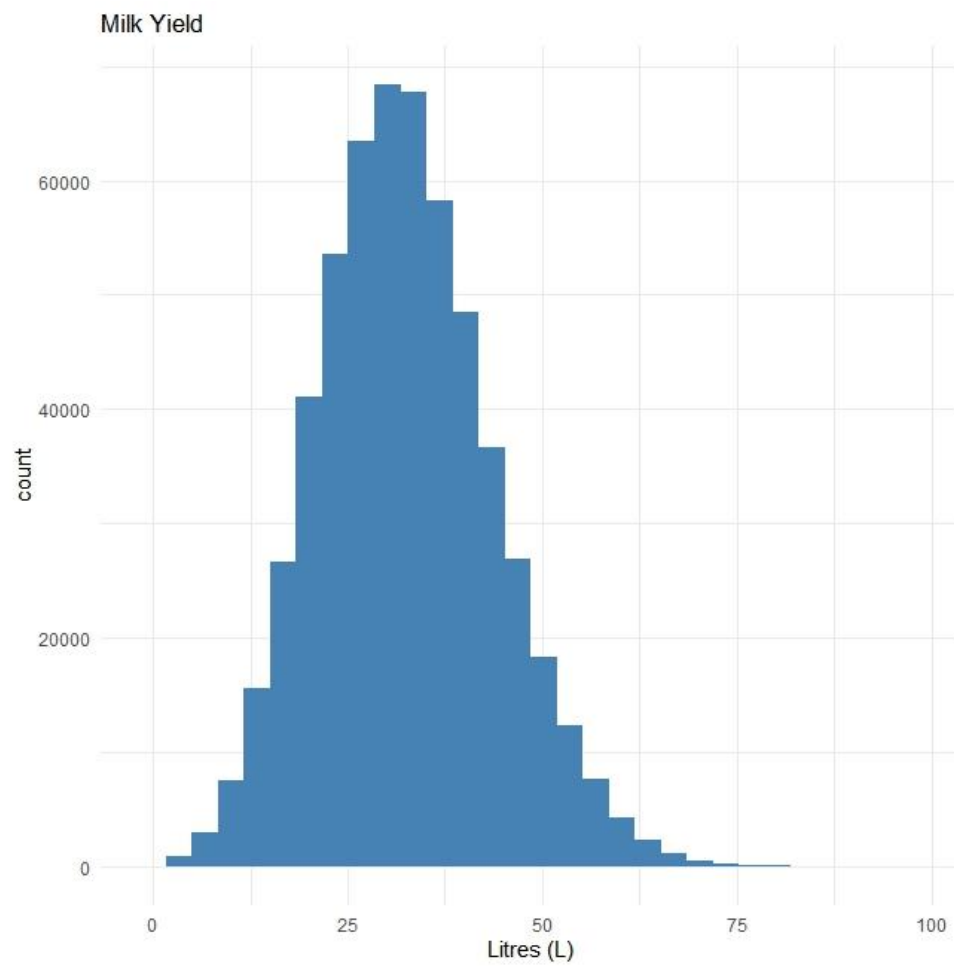


Figure 3.10 Percentage of protein in milk based on 564,962 milk recordings across 43,173 cows

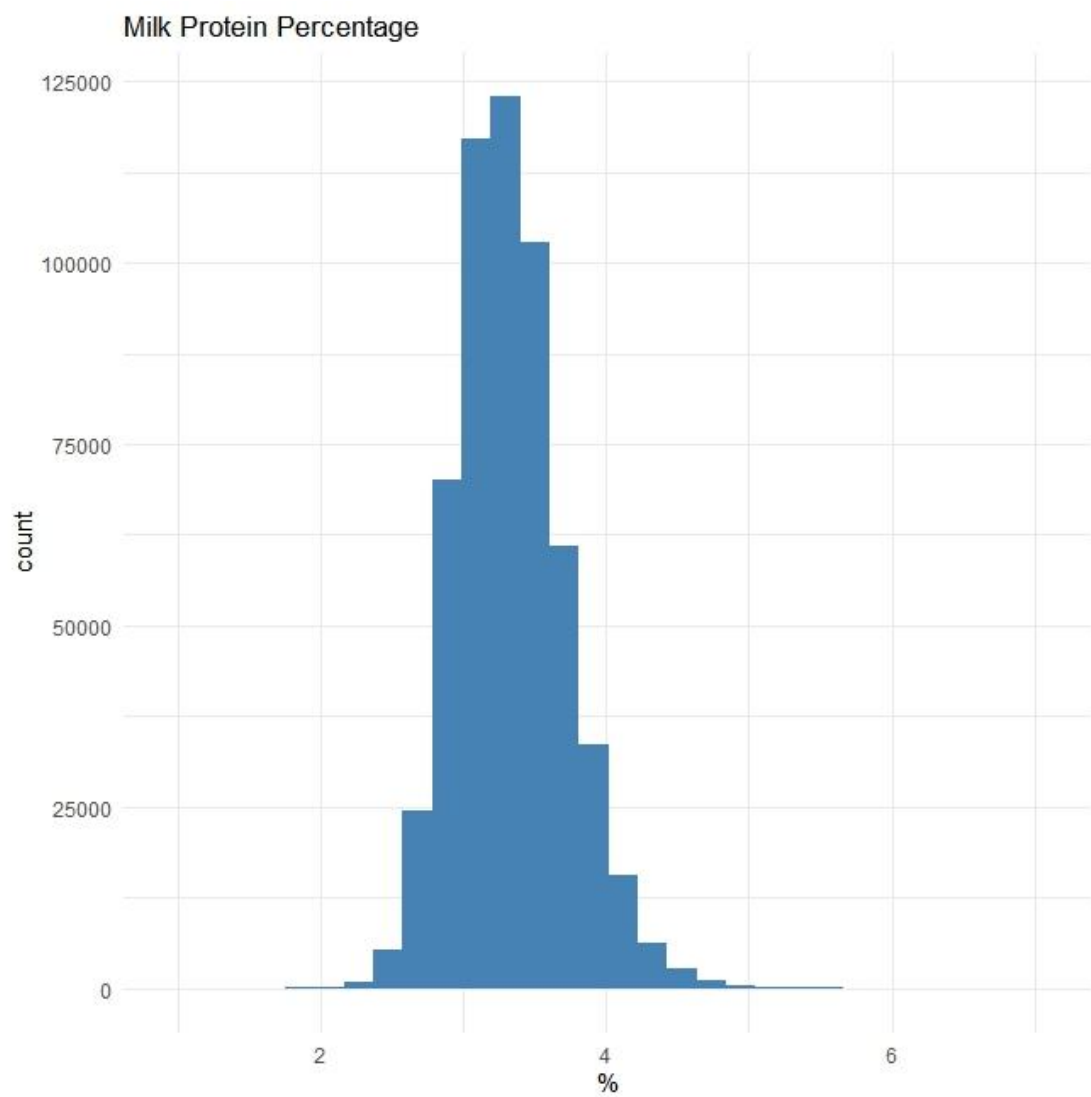
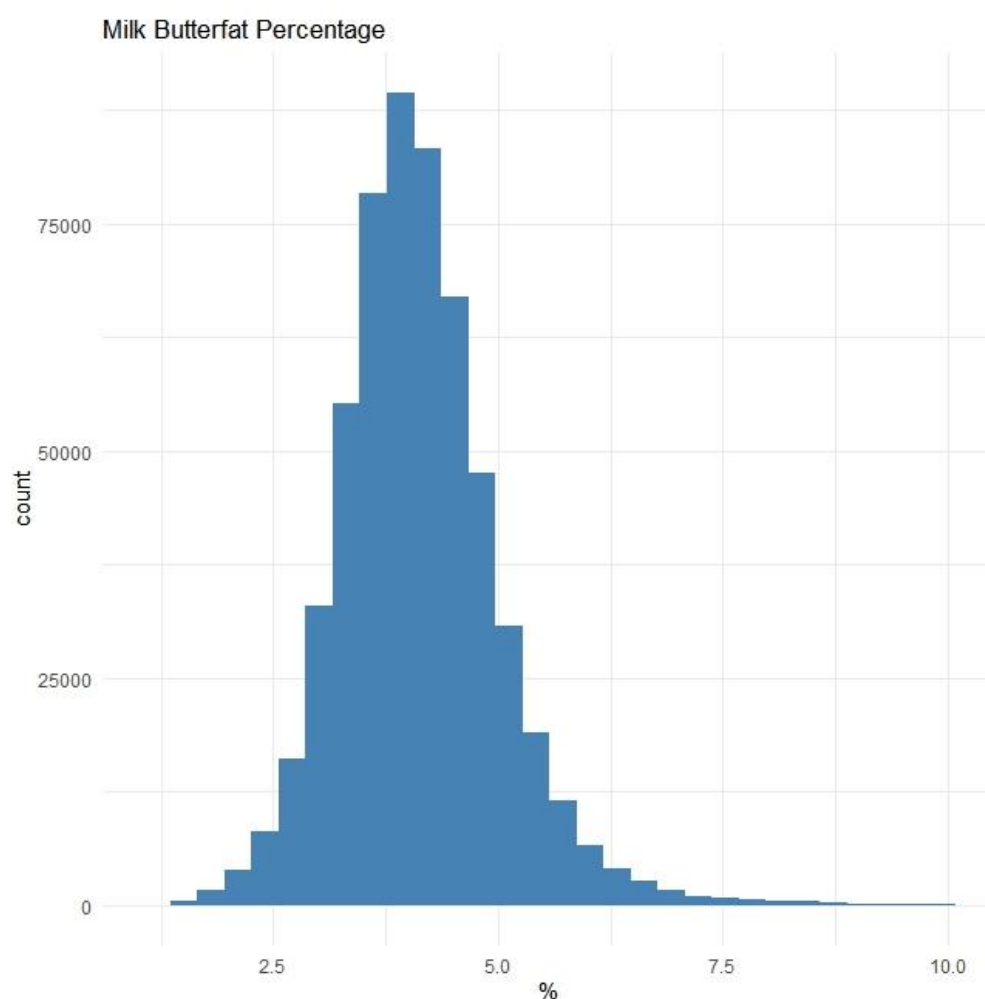


Figure 3.11 Percentage of Butterfat in milk based on 564,962 milk recordings across 43,173 cows



3.3 Insemination and Event Records

The event records dataset had 1,415 abortion events, 20,740 “Do not Breed” events, 193,339 dry-off events, 21,971 mastitis events, 13,365 negative pregnancy diagnoses, 59,286 positive pregnancy diagnoses, 20,740 DNB decisions, 61,641 selling events and 9,707 cows culled.

132,585 data points on insemination data were available on a separate dataset. The median interval of calving to first service was found to be 69 days, with the first quartile being 54 days and the 3rd quartile 82 days (Figure 3.12). In total

39,301 out of 54,655 lactations (71.9%) had at least one service before 80 DIM. The inter calving period, after removing outlier and extreme values ranged from 301 days to 699 days with a median of 376 days (Q1 = 347 days, Q3 = 390 days) (Figure 3.13). The optimal calving interval is in fact at one year (365 days) (NADIS, 2022a) with herds aiming to get as close to that as possible. Thus, this measure indicated that our herds tend to be very well managed. Calving to first service interval, after removing extreme values that were below 20 days and 300 days, had a minimum period of 20 days, a maximum of 300 days and a median of 66 days (Q1 = 39 days, Q3 = 120 days). The calving to conception interval, after also removing extreme values (less than 20 days and more than 400 days) ranged from 20 to 400 days with a median of 95 (Q1 = 58 days, Q3 = 172 days) (Figure 3.14). The recommendations for the calving to first service and calving to conception intervals are 65 and 95 for non-seasonal, higher-yielding herds respectively (NADIS, 2022b), with about half the herds present in this study achieving these measures.

Figure 3.12 Calving to first service interval, based on 54,443 first inseminations

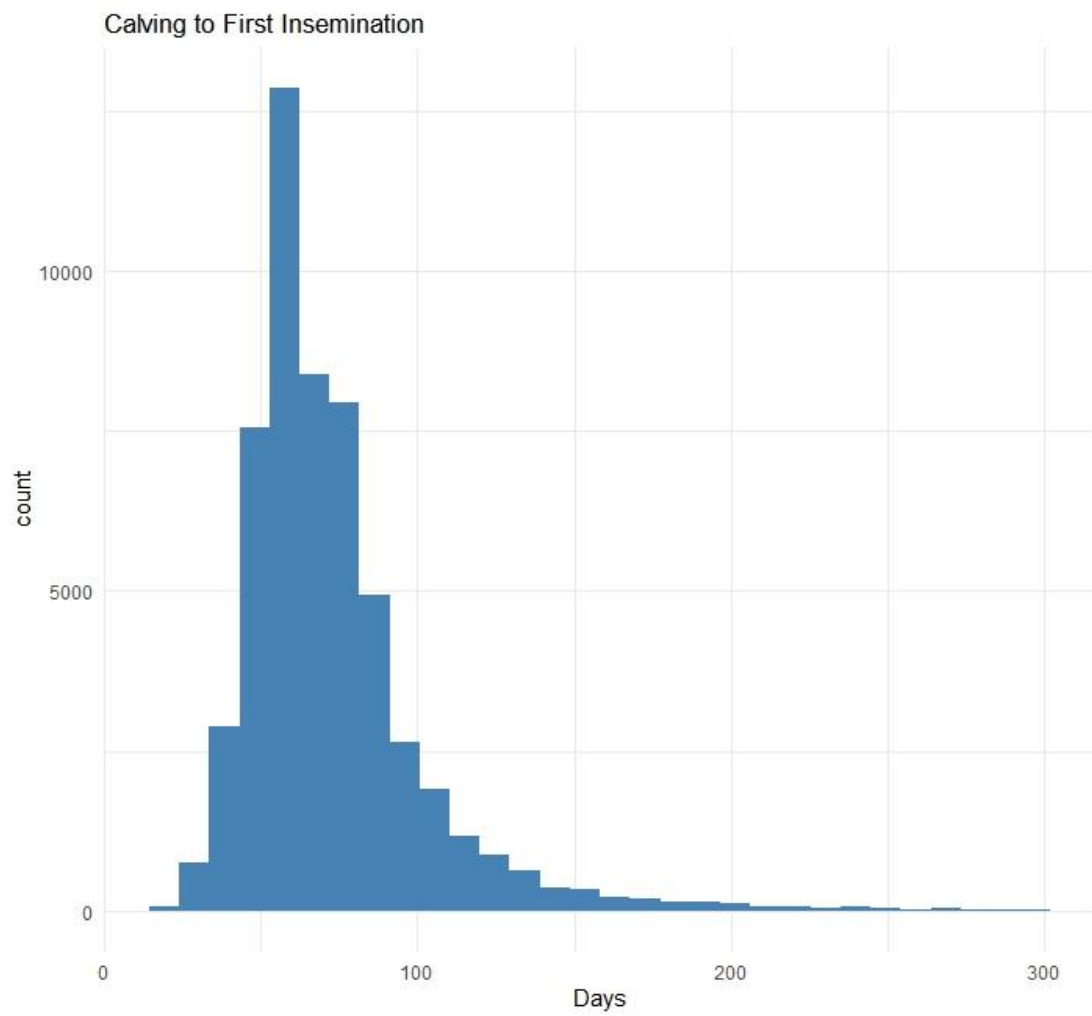


Figure 3.13 Calving interval based on 41,186 lactations

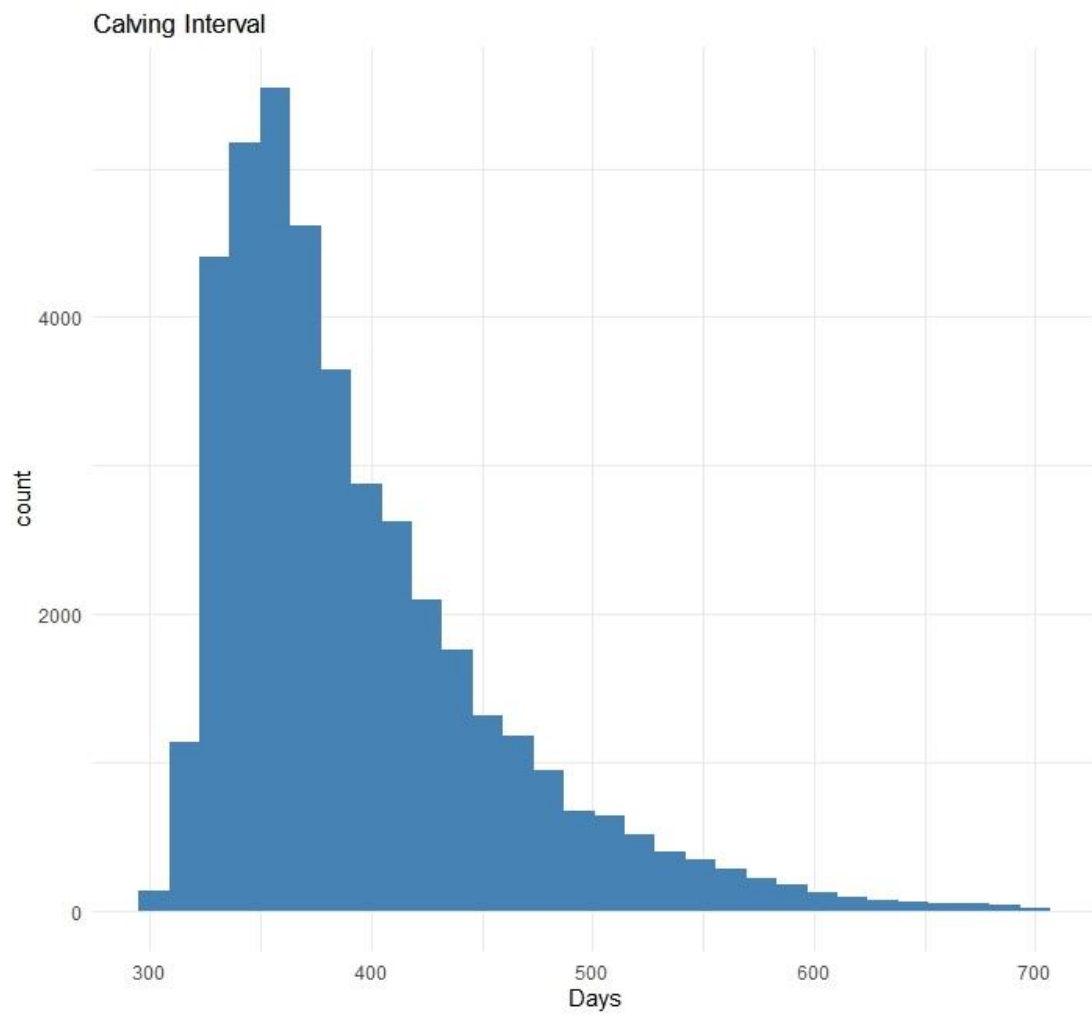
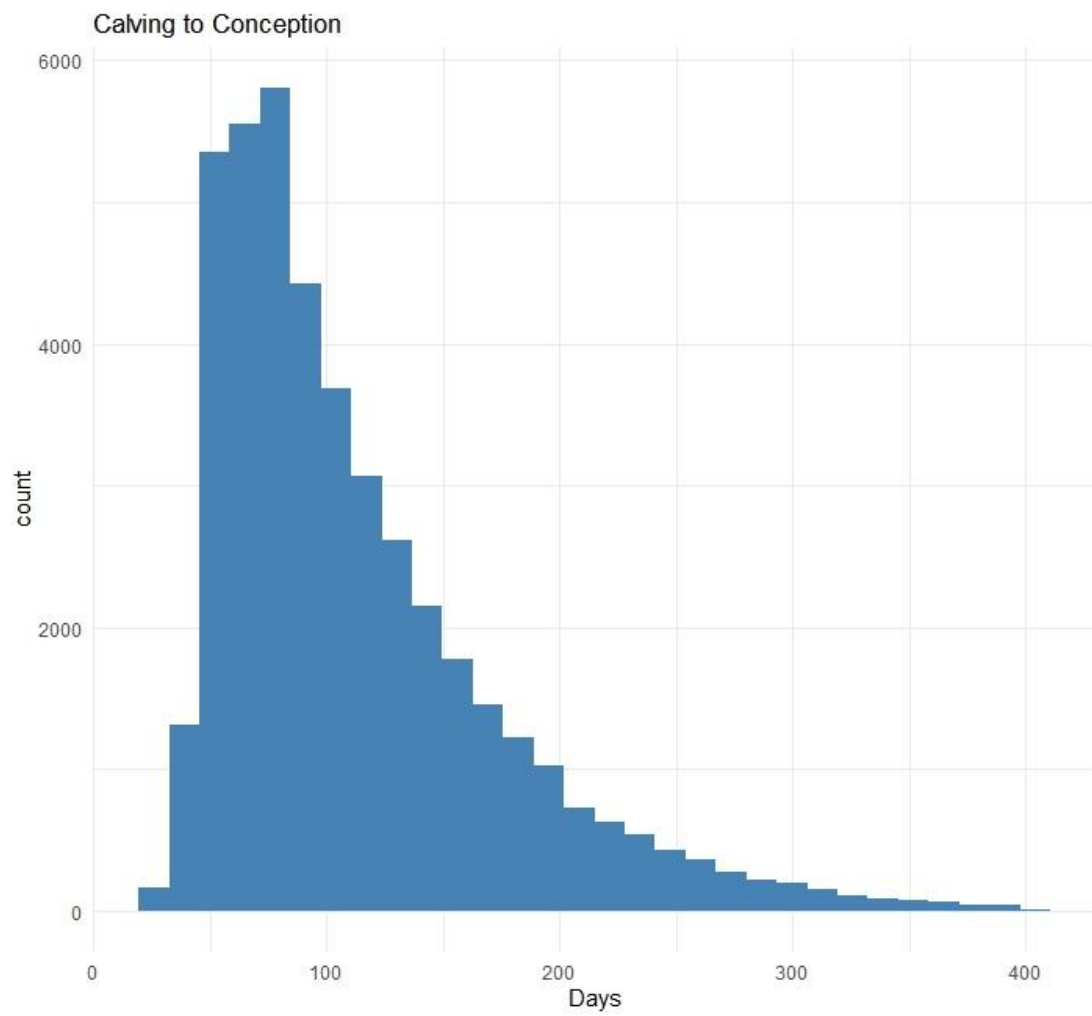


Figure 3.14 Calving to conception distribution based on 43,507 lactations



3.4 Discussion

Various incidence rates have been reported for LDA, with Melendez and Risco (2005) reporting a rate of 1.7% but Le Blanc et al. (2005) reporting higher rates (5-7%). This study found LDA incidence of just below 1% when averaged across all herds, but with many herds not reporting any LDAs whilst several had incidence rates of up to 5%. There is a number of possible explanations, as firstly this difference could be accounted to a regional variance. The two studies included American farms with higher milk yield and hence higher risk. Another reason could be that our farms were particularly well-managed as they were paying a service to specifically help with transition period management. Finally, and perhaps most likely, it is possible that our farms are underreporting LDA incidence. Milk fever has reportedly ranged around similar levels, from 5% to 7% (Goff, 2008, Roche, 2003), compared to the present thesis' herds at 4%. Again, this could be a genuine difference due to the American herds' higher yield and risk but could also be attributed to underreporting from our herds' part. Retained Foetal Membranes appear to have a slightly higher incidence at 8.6% while metritis seems to be even higher at 10.1% (Melendez and Risco, 2005). Once again, our study reported a lower RFM and metritis incidence of 4-5%. Metritis in particular is often poorly defined, with a lot of endometritis cases (which occurs before 21 DIM) potentially classified as metritis, which may very well be the case in our dataset as TMS did not clearly define the criteria for metritis diagnosis. It is safe to assume that Melendez and Risco, (2005) used the correct definition however it is also possible that their screening was more proactive and hence identifying much more cases than those that would have been discovered spontaneously by our farmers. Prevalence of lameness

reported in the UK is typically higher than the post-calving diseases discussed above, with recent estimates between 21% and 36.8% (Clarkson et al., 1996, Leach et al., 2010, Barker et al., 2010, Griffiths et al., 2018, Randall et al., 2019). Lameness in our dataset ranged a lot lower at just over 3%. Again, there could be a number of reasons for this. The present study is only looking cows very close to calving and although a lot of the changes that lead to lameness occur around calving than later in lactation, it still takes time for the actual lameness to develop; therefore it is possible that lameness is lower around calving compared to later stages of lactation. Furthermore, TMS scoring aimed at detecting the severely lame cows (with scores of 3), meaning that the system was not sensitive enough to capture the entire lame population. And of course, as stated previously, there is the possibility that due to convenience sampling the herds in our study are overall “better” with lower incidences, although in the case of lameness the incidence rate is perhaps too low for this to be a plausible explanation. What these differences in incidence could mean for this thesis is that, in the event of a successful predictive model for disease outcomes it would make it difficult to guarantee the same successful results without external validation. The models would still be useful for herds similar to ours and for future studies of course, but widely adopted use would have to be more carefully examined.

Suggested target BCSs are 3.0–3.25 at dry off and calving, with BCS loss of over 0.5-0.75 during early lactation considered sub-optimal (AHDB, 2023). Pre partum BCS for our dataset appeared to be exactly at that cut-off (median 3.25) and while the median BCS post-partum fell by a half point (2.75) the median change was just at -0.25 points. This contradiction of having a median of 0.25

BCS change, yet a 0.5-point difference in BCS pre and post medians could be explained by the shape of the distributions, as BCS is more evenly distributed between 2.75 and 3.75, with BCS post being more centralised around 2.75, hence the differences being more evenly distributed between -0.5 and 0.25.

It is reported that on average 0.6 m of feed fence space is needed per cow in order to avoid negative impacts on feeding behaviour and intakes (Krawczel and Lee, 2019). In our dataset the median was at 1.1m so well above that, however the minimum was just at 0.1m, so it is highly likely that we had a number of outliers. THI of course depends highly on geographic location and season as well as the time of day the assessor conducted the scoring. In the event that assessors visited farms early in the morning high temperatures and THI would be less likely and since the time of visit is not reflected in the dataset, we have no way of knowing the effects of the possible underestimation. The negative effects of heat stress in the livestock industry have been well established, and they are especially important considering that animals are expected to perform on a variety of geographical locations and not just in places where the climate is optimal for their breeds (St-Pierre et al., 2003). In our data, it is notable that over half (75%) of the herd seem to be experiencing heat stress conditions (THI over 68-71) in July with a significant portion of them being under heat stress conditions at some point from May until August. This raises caution as global climate change is also becoming a more urgent issue, as with the rise in temperatures cows might experience even harsher heat stress conditions that would result in rapid economic losses. Gauly and Ammer (2020), focusing on the temperate climate of Central Europe, reported on the effects climate change and heat stress has already had on dairy livestock and even looked in

more depth on further effects that would happen in the future, mentioning health and production losses. Adhikari et al. (2022) looked into THI of dairy animals in two different locations in Hawaii, using historical data from 1920 to 2019 as well as future estimates and came to the conclusion especially in one of the sites the cows were experiencing heat stress of over THI 72 for the entirety of summer and did not drop below 68 until winter. Estimated temperature increases were calculated at 1.3 to 1.8°C by mid-century and 1.6 by 3.1°C by end century, for both sites. They also suggested that in order to maintain sustainability due to future THI estimations, relocation of dairy farms to areas with lower temperatures, as well as selecting animals with suitable genetic characteristics.

The total milk yield per cow per year in the UK has been averaged to 8,100 lts (AHBD, 2023), which amounts to 26.5 L per cow per day. The median for herds in this study was higher at 30.9 L. Milk protein percentage has ranged, over the past 5 years, between 3.2% and 3.5%, while for the same time period milk butterfat has varied between 3.9% and 4.4% (DEFRA, 2022). In our data we had a median of 3.28% and 4.03% for protein and butterfat respectively. Therefore, our averages have been well within that range.

Calving interval is recommended to be as close to a year as possible (Herring, 2014). The median for our data was 376 days, quite close to this recommendation. However, we did have a few extreme values on both ends of the spectrum, which could indicate a few errors in the recording of calving dates. However, as the majority of the data seems centralized around that median this did not appear to raise a serious issue regarding their quality.

Using the proxy of lactations per herd per year, we determined that the median herd size was 56 cows. This figure is significantly below the average UK herd size at the time of data collection, which was reported to be 125 cows (Minnaert et al., 2018). The most likely explanation for this discrepancy is that the proxy measure was inaccurate, indicating that only a fraction of the cows per herd were scored. This introduces potential bias into our data, as we lack information on whether the selection of the cows was random, which means that higher yielding cows were more likely to be selected, our herd averages would appear inflated. Any potential predictive models of this study would still be of value, but further external validation will be needed to determine whether these potential models can be used in a commercial setting.

3.4.1 Conclusions

The herds in this study showed disease incidences generally below reported ranges from other nations, lameness incidence very much below nationally reported averages, cows spending a perhaps surprising amount of time in heat stress throughout summer, generally higher levels of production and high quality in terms of reproduction standards. Overall, this suggests that more intensive and potentially larger in size, most likely housed year-round (at least pre- and post-calving). There is also quite possibly some selection bias towards better managed herds, which is not surprising considering they are investing in a service to monitor and trying to refine their transition cow management.

Chapter 4 – Prediction of Disease Status

4.1 Introduction

4.1.1 Periparturient diseases

Most of the periparturient diseases of dairy cattle are results of the metabolic and immunological imbalances described in Chapter 1 (Melendez and Risco, 2005). Amongst the most important diseases are milk fever (clinical hypocalcemia), retained foetal membranes (RFM), metritis, displaced abomasum, mastitis and lameness (Goff and Horst, 1997b, Melendez and Risco, 2005). Generally, these diseases are mostly affected by management; with the exception of lameness and ketosis, they present low heritability (Van Dorp et al., 1998).

A number of studies have investigated the relationships between periparturient diseases and even though the case definitions for the diseases have not always been consistent, the results appear to be similar among the studies (Melendez and Risco, 2005). Milk fever was found to have a positive association with parity (Curtis et al., 1985, Erb et al., 1985), RFM with milk fever, parity, dystocia and twinning (Correa et al., 1993, Curtis et al., 1985, Erb et al., 1985), metritis with milk fever, RFM, left displaced abomasum and dystocia (Correa et al., 1993, Curtis et al., 1985, Erb et al., 1985, Melendez et al., 2003), left displaced abomasum with ketosis and milk fever (Correa et al., 1993, Curtis et al., 1985), ketosis with left displaced abomasum, RFM and milk fever (Curtis et al., 1985), while metritis was found to be negatively associated with parity (Melendez et al., 2003).

Management of transition period cows should be equally focused on maintaining physiological functions, immunological functions, normal calcium levels, helping the rumen adapt to high energy diets, as well as optimizing cow comfort, maintaining appropriate BCS and providing calving assistance when needed. Whenever those standards are not met the risk of the transition period cow developing a periparturient disease during the postpartum period is increased (Goff and Horst, 1997b, Goff et al., 1996, Risco et al., 1994).

4.1.2 Left displaced abomasum

Displacement of the abomasum is a multifactorial condition that occurs in dairy cows possibly due to decreased rumen fill and abomasal atony (LeBlanc et al., 2005, Shaver, 1997, Wittek et al., 2004). Left-displaced abomasum (LDA) occurs primarily in high yielding cows after calving (Geishauser, 1995), and may not cause apparent clinical signs (Van Winden et al., 2002). Reported incidence of LDA ranges from 0.3% to 6.3% with a median of 1.7% (Melendez and Risco, 2005). LeBlanc et al. (2005) suggested that the incidence was rising over the previous decade from 1% – 2% to 5% - 7%. Caixeta et al. (2018) reported an incidence of 3.5% among dairy herds in the United States. Amongst the reported risk factors are poor rumen fill, high-concentrate diets, hypocalcemia, high BCS at calving, season, inadequate feed space and limited availability of fresh feed, early parity and the presence of other conditions, such as fatty liver, milk fever, twinning, dystocia, retained placenta, metritis and mastitis (Caixeta et al., 2018, Esposito et al., 2014, Shaver, 1997, LeBlanc et al., 2005, Cameron et al., 1998).

4.1.3 Milk Fever

Hypocalcaemia at calving is a common phenomenon in cows, as at the beginning of the lactation period the demand for lactation places a substantial burden on calcium homeostasis (Goff and Horst, 1997b, Horst et al., 1994). At times when the drop in calcium concentrations is severe, the function of muscles and nerves cannot be supported, which results in parturient paresis, also known as milk fever (Goff and Horst, 1997b, Goff and Horst, 1997a). Where the level of hypocalcaemia is less severe, impacts on smooth muscle function can result in increased disease risks even where clinical milk fever is not evident. This is known as subclinical hypocalcaemia.

It has been reported that 5 to 10% of cows are affected after calving, with up to 15% of these not responding to treatment (Eckel and Ametaj, 2016). Milk fever is associated with increased incidence of retained placenta and mastitis, which could be attributed to loss of muscle tone, immunosuppression due to intracellular calcium drop (Kimura et al., 2006) or a combination of the two (Bradford et al., 2015, Goff and Horst, 1997a). Furthermore, reduced DMI plays an important part in the causal pathways of both milk fever and retained placenta (Bradford et al., 2015). The postpartum feed intake decline in cows with milk fever is more severe compared to those without, and negative energy balance (NEB) is thought to follow a similar trend (Goff and Horst, 1997a, Marquardt et al., 1977). It also reduces rumen fill, the depth of rumen fibre mat and abomasal contractility, all of which can contribute to displacement of the abomasum (Goff and Horst, 1997b).

In a recent study, it was found that milk fever was the disease most strongly associated with culling risk within the first 120 days in milk (DIM) (Probo et al.,

2018). Dohoo and Wayne Martin (1984), Grohn et al. (1998), as well as Milian-Suazo et al. (1988) all had similar findings, with increased culling risk in cows with milk fever. Kelton et al. (1998) gathered 33 citations dating from 1979 to 1995 that reported a lactational incidence risk of milk fever ranging from 0.03% to 22.3% with a median of 6.5%. Goff (2008) reported an incidence of 5-7% in cows in confinement whereas Roche (2003) reported a 5% incidence rate in grazing cows. Potential risk factors for milk fever are age, prepartum diet with a high dietary cation anion difference (DACD), breed (with Jersey and Guernsey cattle at increased risk), milk yield, presence of other diseases and previous history of milk fever (Saborío-Montero et al., 2017).

4.1.4 Mastitis

Mastitis is an inflammation of the udder, which can be caused by either Gram-positive or Gram negative bacteria and vary in severity (Eckel and Ametaj, 2016). The case definition for clinical mastitis in cows is an animal with one or more quarters producing visually abnormal milk, with or without any other systemic symptoms (Kelton et al., 1998). Clinical mastitis is one of the most prevalent diseases in dairy cattle that can occur at any time with peak incidence around 30 to 50 days in milk (DIM) and has a great economic impact on many farms (Rollin et al., 2015, Zahrazadeh et al., 2018). A survey conducted in England and Wales indicated the mean annual incidence of clinical mastitis at 47 cases per 100 cows per year when collected from historic farm records, whereas it was reported as high as 71 cases per 100 cows per year when using dates of milk samples submitted for bacteriological analysis as part of the study (Bradley et al., 2007). Irreversible damage of the mammary tissue during the inflammation is what causes the majority of economic losses associated with

mastitis (Eckel and Ametaj, 2016). Studies have found that mastitis is amongst the most influential factors when producers make culling decisions (Grohn et al., 1998, Probo et al., 2018).

4.1.5 Lameness

Lameness is defined as decreased mobility; a number of mobility scoring systems exist to allow more objective assessment at individual and group level. The reports of mean prevalence of lameness in the UK vary between 21% (Clarkson et al., 1996), 36% (Leach et al., 2010) and 36.8% (Barker et al., 2010), with more recent reports being at 31.6% (Griffiths et al., 2018) and 30.1% (Randall et al., 2019). Lameness along with infertility and mastitis have been identified as the 3 diseases most associated with increased culling rates (Eckel and Ametaj, 2016). It is considered as one of the most costly disease in dairy farms with the total cost, depending on lameness definition and expenditures-losses included, having been estimated up to over US\$300 per case (Dolecheck and Bewley, 2018). Lower milk yield as well as worse reproductive performance has also been associated with lameness (Machado et al., 2010).

As explained by Sepulveda-Varas et al. (2018) there are studies supporting the association between the metabolic and behavioural changes that occur during the transition period and the development of claw horn lesions. Low BCS after calving in particular, has been targeted as a potential risk factor of lameness (Green et al., 2014, Hoedemaker et al., 2009, Newsome et al., 2017). In a recent study however, it was highlighted that it is the loss rate of BCS rather than the BCS itself that affected lesion development (Sepulveda-Varas et al., 2018). Moreover, NEB was found to be associated with poor hoof health in

primiparous cows during the transition period (Sepulveda-Varas et al., 2018). Lameness during the dry period have been found to have increased risk of postpartum disease and increased culling rates (Calderon and Cook, 2011, Hoedemaker et al., 2009, Machado et al., 2011, Vergara et al., 2014) and lameness has a well-established link with ruminal acidosis (Eckel and Ametaj, 2016). Vergara et al. (2014) indicated in their study that monitoring locomotion score could potentially be useful in explanatory models investigating postpartum health issues.

4.1.6 Retained Foetal Membranes

After calving, the immune response plays a key role in severing the cotyledon-caruncle attachment and detachment of membranes from maternal tissue. Failure of the immune system to complete this process within 24 hours leads to a condition defined as retained foetal membranes (RFM) (LeBlanc, 2008). Decreased motility of the uterus is generally not seen to be an underlying cause of RFM, as affected cows appear to have normal, if not increased uterine motility in the days following calving (Frazer, 2005, LeBlanc, 2008).

The case definition according to Kelton et al. (1998) is observing foetal membranes at the vulva, vagina or uterus by vaginal examination at more than 24 hours after calving, and reported incidence rates range from 1.3% to 39.2% with a median of 8.6% (Melendez and Risco, 2005). The average duration the membranes are retained in cows with RFM was reported to be 7 days (LeBlanc, 2008). NEB seemingly has a role in the pathogenesis of RFM, likely through impairment of immune function (Goff and Horst, 1997b, LeBlanc, 2008). More specifically, cows with higher NEFA concentrations (a marker of NEB) were found have 80% greater risk of developing RFM (LeBlanc et al., 2004).

Pregnancy rate in cows with RFM was found to be reduced by 15% compared to healthy animals (Fourichon et al., 2000). Loss of milk production appears to be an issue only for those cases that lead to clinical metritis (Fourichon et al., 1999), while culling was not found to have a significant association with RFM (Grohn et al., 1998).

4.1.7 Metritis

Uterine involution starts immediately after calving and is a complex process (Sheldon, 2004), which seems to naturally involve bacterial invasion (Chapwanya et al., 2012). Considering the immune suppression that occurs postpartum, there are favourable factors for the development of uterine disease during this time (Azawi, 2008, Mallard et al., 1998).

Inflammation of the uterus is defined as metritis and it can cause systemic symptoms, such as fever, dullness, decreased appetite, elevated heart rate, presence of watery or purulent discharge from the uterus and a decrease in milk production (Sheldon et al., 2006a). It occurs within the first 3 weeks after calving, usually within the first 10 days (Eckel and Ametaj, 2016).

Conditions such as RFM, maceration of the foetus and difficulties during calving may increase the risk of metritis (Foldi et al., 2006, Sheldon et al., 2006a, Sheldon et al., 2006b), with RFM being the most important risk factor with an odds ratio of approximately 6 (Correa et al., 1993, Curtis et al., 1985, Erb et al., 1985). It has been reported that 25 – 50% of cows with RFM progress to clinical metritis (LeBlanc, 2008). Decreased DMI has also been associated with the disease (Huzzey et al., 2009, Huzzey et al., 2007). The reported median incidence of metritis was at 10.1%, ranging from 2.2% to 37.3% (Melendez and

Risco, 2005). It has an impact on reproductive performance and was reported to increase the calving to first oestrus interval by 6.9 days, the calving to first insemination interval by 7.3 days, the first to last insemination interval to 15.4 days, the calving to conception interval by 18 days and the number of inseminations until conception by 0.2 (Bruun et al., 2002, C. Bartlett et al., 1986). The risk of culling was 1.3 time higher in cows with metritis compared to those without and were more likely due to the decreased reproductive performance rather than the disease itself (C. Bartlett et al., 1986, Lewis, 1997).

Appropriate transition period management, including proper nutrition during the dry period in order to maintain optimum BCS and a sanitary environment, may help prevent metritis (Lewis, 1997). Furthermore, as mentioned above competent function of the immune system is vital, therefore events such as uterine trauma, dystocia and manual removal of RFM that can lead to a declined phagocytic activity of neutrophils may also predispose the cows to metritis (Cai et al., 1994).

4.1.8 Twinning

Twinning, though once sought in order to increase the milk production per cow it is now not a desired attribute in dairy herds (Cai et al., 1994, Correa et al., 1993, Curtis et al., 1985, Lewis, 1997). Although twinning is not a transition “disease”, it has been described as a risk factor linked with other periparturient diseases (Probo et al., 2018). It can cause a decline a reproductive performance, as cows with twins were shown to have a much higher risk of early pregnancy loss (3 to 9 times), increased calving to conception intervals and culling rates (Bicalho et al., 2007). Culling rates before 120 DIM, in particular, were reported to be almost double in cows with twins, compared with

singletons, at 16.1% (Andreu-Vazquez et al., 2012). Probo et al. (2018) reported a high culling rate before 120 DIM for cows with twins at 30.3%. Twinning has also been associated with metritis (Lewis, 1997) and its incidence was reportedly ranging from 9 to 12%, having a substantial economic impact on dairy herds (Silva del Rio et al., 2007).

Establishing and quantifying disease risk factors remains important as there is very little recent evidence in this field, particularly from modern UK dairy systems. Furthermore, additional application of predictive modelling has potential to contribute in early notification and implementation of management measures that can reduce financial losses; this has not been widely explored for post-calving disease outcomes.

4.2 Methods

The outcomes considered for analysis included both the individual diseases (Milk Fever, LDA, RFM and Metritis) and a collective Disease Status outcome. The latter was defined as either negative if the cow had not been recorded positive for any disease for that lactation, or positive if she had been marked as disease positive for at least one out of the four diseases during that lactation.

The types of predictive models that were investigated for each outcome were logistic regression, decision trees, random forests, support vector machines, artificial neural networks and naïve Bayes.

Two sets of models were considered when analyzing disease as a collective outcome. The first set was focused on predicting disease at lactation level, while the second set aimed to predict disease risk aggregated across groups of lactations. The source, initial data preparation steps, and descriptive statistics on the dataset are described in chapter 2.1.3.4 and chapter 3.

4.2.1. Lactation Level

For the first set of models, the units of data were individual lactations, with presence or absence of disease in that lactation as the outcome, and the lactation-level predictor variables listed in Table 4.1 used as potential predictors. The total number of lactations was 12,863.

Table 4.1 Potential predictor variables considered in models with the outcome of disease occurrence at lactation level

Variable	Missing data
Rumen Fill in the pre-calving cow	86 (0.007%)
Hock Hygiene	1151 (0.089%)
Neck Rail Height in the pre-calving pen	5421 (43.41%)
THI in the pre-calving pen	5450 (42.42%)
Feed Fence space per cow in the pre-calving pen	5507 (42.87%)
Water Trough space per cow in the pre-calving pen	5527 (43.02 %)
Good Bedding quality in the pre-calving pen	4763 (37.07%)
Good Light Quality in the pre-calving pen	4763 (37.07%)
Good Feed Quality in the pre-calving pen	4763 (37.07%)
Good Water Quality in the pre-calving pen	4763 (37.07%)
Good Air Quality in the pre-calving pen	4763 (37.07%)
Feed available in the pre-calving pen	4763 (37.07%)
Water available in the pre-calving pen	4763 (37.07%)
BCS pre-calving	77 (0.006%)
BCS change pre- to post-calving	1201 (0.09%)

Month of pre-calving recording	0 (0.00%)
Lactation number	0 (0.00%)
Calf Mortality	0 (0.00%)
Twining	0 (0.00%)
Mean Milk Yield of previous lactation	644 (0.05%)
Mean Protein % in milk in the previous lactation	644 (0.05%)
Mean Butterfat% in milk in the previous lactation	644 (0.05%)
Stocking density in pre-calving pen	5494 (0.43%)
Lameness in the pre-calving cow	0 (0.00%)

General outline of model building is described in more detail in chapter 2.3.4.1, As the dataset was imbalanced, the absence of disease diagnosis being substantially more common compared to its presence, kappa was the metric used to assess model predictiveness and furthermore, up-sampling was implemented (further explanation on up-sampling was provided in Chapter 2). Since kappa was the metric of choice it was also used during parameter tuning with the parameters being automatically chosen by the caret package (Kuhn, 2008). The package automatically tested various parameter values and whichever one provided the largest kappa was declared as the most optimal

In addition to using the models built on the individual lactation level to make predictions on each individual lactation, it was decided to investigate whether accuracy of predictions is improved when data were aggregated at herd/quarter-year level (i.e. predicting the incidence risk of disease across all the lactations in a given herd beginning in a given quarter-year). Quarter-year periods were January to March, April to June, July to September and October to December. The process of model building described above was repeated with a holdout dataset. 80% of the original dataset was used for model building, while 20% was kept for external cross-validation. In addition, the test and train datasets were split based on farm/quarter-year group to ensure that pen-level information that would remain constant within each group did not contribute to data leakage when building the original models. In order to avoid dealing with groups with a low number of observations, the dataset was filtered to include only groups with a group count of 10 or higher. After training these models they were used to produce predictions on the holdout dataset aggregated by herd/quarter-year/year group and these predictions were then compared with the actual values. To evaluate the comparisons the predicted and actual outcomes for each data point were graphed using a scatterplot and the variation of the outcome explained by the predictions was assessed using an R^2 . The values were determined by fitting a linear regression model. High values indicated highly explained variation, meaning that the predictions aggregated did approximate the aggregated values and can potentially lead to meaningful predictions on a herd/quarter-year level.

4.2.2 Lactations per herd/quarter-year level models

A further set of models was built using data aggregated at herd-quarter-year level. The units of data for this dataset therefore represented all lactations in a given herd beginning in a given quarter-year period. The outcome variable for each unit was the proportion of disease status positive lactations in that herd-quarter-year, with potential predictor variables aggregated in the same way (calculated as either a mean across cows/measurement-occasions for continuous, or proportions for binary predictors).

Herd/quarter-year groups with less than 10 lactation recordings were removed from the analysis. In total 79 groups were removed (18.7%). The final dataset size for this part of the analysis was 343 data points. The number and percentage of missing data for all potential predictor variables is shown in Table 4.2.

Table 4.2 Predictive variables considered in final analysis for second set of models

Variable	Missing Data (%)
Metritis percentage per herd/month	0 (0.0)
Milk Fever percentage per herd/month/	0 (0.0)
Twinning percentage per herd/month	0 (0.0)
Feed Fence Space per cow	193 (45.7)
Mean Neck Rail height	210 (49.8)
Mean BCS pre calving	37 (8.8)
Mean BCS change	230 (54.5)
Mean Lactation Number	0 (0.00)
Mean Milk Yield	22 (5.2)
THI pre calving	240 (56.8)
Water Trough Space per cow	192 (45.5)
Rumen Fill pre calving	37 (8.8)
Rumen Fill post calving	220 (52.1)

4.3 Results

4.3.1 Individual Disease Outcomes

The variables used for fitting the models are presented in table 4.3.

Table 4.3 Predictive variables used for machine learning models predicting individual disease outcomes at lactation level

Milk Fever	LDA	RFM	Metritis
Rumen Fill post-partum	Rumen Fill pre-partum	Rumen Fill pre-partum	Rumen Fill pre-partum
Lactation Number	Rumen Fill post-partum	Rumen Fill post-partum	Rumen Fill post-partum
Hock Hygiene Score	Lactation Number	Lactation Number	Lactation Number
BCS change	Hock Hygiene Score	Hock Hygiene Score	Hock Hygiene Score
BCS pre partum	BCS change	BCS change	BCS change
Twinning	BCS pre-partum	BCS pre-partum	BCS pre-partum
Calf Mortality	Twinning	Twinning	Twinning
Mean 305 Milk Yield of Previous lactation	Calf Mortality	Calf Mortality	Calf Mortality

	Mean 305 Milk
	Yield of Previous
	lactation
	Neck Rail Height
	in dry pen

4.3.1.1 Individual Lactation level

Before up-sampling, a lot of algorithms returned sensitivity values of 0 and PPV either was 0 as well due to the lack of True Positives or could not be computed due to the lack of False Positives, while specificity values were at 100%. Even for the rest of the models, sensitivity remained very close to 0, indicating the inability of those models to properly predict the diseased class. Kappa values were all consistently low for all diseases. All results are shown in tables 4,4-4.7.

This was the result of an imbalanced dataset, due to the low frequency of disease and thus a sampling method was implemented to improve predictions.

Table 4.4 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Milk Fever outcomes, before up-sampling

<u>Milk Fever</u>							
<u>Training Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.972	0.972	0.969	0.824	0.971	0.943	0.971
Kappa	0	0	0.025	0.028	0	0.050	0.023
Sensitivity	0	0	0.016	0.022	0	0.074	0.012
Specificity	1	1	0.998	0.999	0.999	0.968	0.999
PPV	-	-	0.300	0.523	0	0.061	0.350
NPV	0.972	0.972	0.972	0.972	0.972	0.973	0.972
AUROC	0.842	0.5	0.783	0.824	0.645	0.834	0.595
Detection	0	0	0	0	0	0.002	0
Rate							
Balanced	0.5	0.5	0.507	0.511	0.499	0.521	0.502
Accuracy							
F1	-	-	0.074	0.083	-	0.066	0.062
<u>Test Set</u>							

	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.982	0.982	0.892	0.881	0.981	0.981	0.979
Kappa	0	0	0.029	0	0	0.001	0.003
Sensitivity	0	0	0.036	0	0	0	0
Specificity	1	1	0.982	0.999	0.999	0.999	0.997
PPV	-	-	0.181	0	0	0	0
NPV	0.981	0.981	0.906	0.981	0.981	0.981	0.981
AUROC	0.842	0.5	0.509	0.5	0.5	0.5	0.5
Detection	0	0	0.003	0	0	0	0
Rate							
Balanced	0.5	0.5	0.509	0.499	0.499	0.499	0.498
Accuracy							
F1	-	-	0.060	-	-	-	-

Table 4.5 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting LDA outcomes, before up-sampling

<u>LDA</u>							
<u>Training Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.991	0.991	0.991	0.991	0.991	0.991	0.991
Kappa	0	0	0	0.013	0	0	0
Sensitivity	0	0	0	0.01	0	0	0
Specificity	1	1	1	0.999	1	1	1
PPV	-	-	-	0.125	-	-	-
NPV	0.991	0.991	0.991	0.991	0.991	0.991	0.991
AUROC	0.749	0.5	0.574	0.611	0.505	0.705	0.536
Detection	0	0	0	0	0	0	0
Rate							
Balanced	0.5	0.5	0.5	0.504	0.5	0.5	0.5
Accuracy							
F1	-	-	-	0.142	-	-	-
<u>Test Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	

Accuracy	0.992	0.992	0.992	0.992	0.992	0.992	0.992
Kappa	0	0	0	0	0	0	0
Sensitivity	0	0	0	0	0	0	0
Specificity	1	1	1	1	1	1	1
PPV	-	-	-	-	-	-	-
NPV	0.992	0.992	0.992	0.992	0.992	0.992	0.992
AUROC	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Detection	0	0	0	0	0	0	0
Rate							
Balanced	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Accuracy							
F1	-	-	-	-	-	-	-

Table 4.6 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting RFM outcomes, before up-sampling

<u>RFM</u>							
Training Set							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.960	0.961	0.958	0.961	0.961	0.935	0.961
Kappa	0.030	0	0.078	0.033	0	0.176	0.048
Sensitivity	0.013	0	0.040	0.026	0	0.224	0.028
Specificity	0.999	1	0.995	0.999	1	0.963	0.998
PPV	0.404	-	0.290	0.619	0	0.199	0.410
NPV	0.961	1	0.962	0.962	1	0.968	0.962
AUROC	0.689	0.5	0.630	0.635	0.503	0.679	0.621
Detection	0	0	0.001	0.001	0	0.008	0.001
Rate							
Balanced	0.506	0.5	0.517	0.513	0.499	0.594	0.513
Accuracy							
F1	0.050	-	0.076	0.098	-	0.196	0.066
Test Set							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	

Accuracy	0.961	0.962	0.957	0.961	0.962	0.936	0.962
Kappa	0.001	0	0.028	0.037	0	0.190	0.067
Sensitivity	0	0	0.022	0.022	0	0.240	0.037
Specificity	0.999	1	0.994	0.998	1	0.963	0.998
							9
PPV	0	-	0.142	0.333	0	0.207	0.714
NPV	0.961	0.962	0.962	0.962	0.962	0.969	0.963
AUROC	0.501	0.5	0.526	0.521	0.503	0.681	0.544
Detection	0	0	0.001	0.001	0	0.009	0.001
Rate							
Balanced	0.499	0.5	0.503	0.505	0.499	0.602	0.518
Accuracy							
F1	-	-	0.038	0.042	-	0.222	0.071

Table 4.7 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Metritis outcomes, before up-sampling

<u>Metritis</u>							
Training Set							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.947	0.947	0.943	0.947	0.947	0.919	0.947
Kappa	0	0	0.015	0.003	0	0.112	0
Sensitivity	0	0	0.015	0.003	0	0.141	0
Specificity	0.999	1	0.994	0.999	1	0.962	0.999
PPV	0	-	0.095	0.055	-	0.167	0
NPV	0.947	0.947	0.948	0.948	0.948	0.953	0.947
AUROC	0.646	0.5	0.615	0.621	0.557	0.633	0.561
Detection	0	0	0.001	0	0	0.007	0
Rate							
Balanced	0.499	0.5	0.504	0.5	0.5	0.551	0.499
Accuracy							
F1	-	-	0.065	0.055	-	0.151	-
Test Set							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	

Accuracy	0.959	0.961	0.960	0.961	0.961	0.926	0.961
Kappa	0	0	0.021	0	0	0.028	0
Sensitivity	0	0	0.013	0	0	0.067	0
Specificity	0.998	1	0.998	1	1	0.961	1
PPV	0	-	0.250	-	-	0.065	0
NPV	0.961	0.961	0.961	0.961	0.961	0.962	0.961
AUROC	0.613	0.5	0.608	0.5	0.551	0.601	0.510
Detection	0	0	0	0	0	0.002	0
Rate							
Balanced	0.499	0.5	0.505	0.5	0.5	0.514	0.5
Accuracy							
F1	-	-	0.025	-	-	0.066	-

After up-sampling the overall picture slightly changed for all four diseases. For Milk Fever accuracy on the test set was high at 0.733 at the lowest (logistic regression) and 0.929 at the highest (decision tree). Sensitivity had a quite wide range starting from very low values (0.200 for the decision tree) to quite adequate ones (0.850 for the logistic regression). Specificity on the other hand ranged from medium to quite high values (0.731 logistic regression - 0.943 decision tree). Two of the most telling metrics were the PPV and NPV with the former scoring extremely low (Highest being 0.070 for the SVM) and latter scoring extremely high (Lowest 0.984, again, for the decision tree). Detection

rate and F1 both producing very low results across all models, while Balanced Accuracy ranged from very close to the baseline of 0.5 (0.571 for the decision tree) to medium values (highest 0.802 for the SVM). AUROC ranged from low values at 0.602 for the decision tree to moderately high at 0.834 for the SVM. All metrics for the milk fever models after up-sampling are presented on table 4.8.

The rest of the diseases followed a similar pattern. LDA accuracy for the test set ranged from reasonable as to extremely high (0.693 for logistic regression to 0.983 for Naive Bayes). Sensitivity was low, at below 0.555 for all models, while specificity was generally relatively high for all algorithms. PPV stayed consistently low (highest being naïve bayes at 0.029) and NPV consistently high (lowest being KNN and Naïve Bayes at 0.992). Once again F1 and detection rate ranged very low for the entirety of the methods, with ANN yielding the best results for the former (0.046) and logistic regression and random forest for the latter (0.004 for both). Balanced accuracy ranged overall low (0.513 for Naïve Bayes to 0.657 for random forest), while AUROC values ranged along the same values (0.506 for KNN to 0.753 for ANN). All metrics for LDA models after up-sampling are shown in table 4.9. Metrics for RFM on the test set followed the same trends after up-sampling, with accuracy ranging from moderate to high (0.651 for the KNN to 0.953 for the Naïve Bayes), sensitivity having a wide range but remaining at low values (0.127 for the Naïve Bayes to 0.601 for logistic regression), specificity being relatively high (0.657 to 0.985 for KNN and Naïve Bayes respectively), PPV being extremely low (highest being 0.261 for the Naïve Bayes), NPV being extremely high (lowest at 0.978 for the ANN), detection rate and F1 being consistently low the best results being 0.171

and 0.022 for Naïve Bayes and ANN respectively), balanced accuracy ranging low from 0.556 for the Naïve Bayes to 0.657 for the ANN, and finally AUROC ranging only slightly higher at 0.591 to 0.698 for KNN and random forest respectively. Metrics for all RFM models after up-sampling are available on table 4.10. Lastly, metrics for metritis, once again, displayed a similar behaviour. Accuracy for the test set scored from mediocre (0.628 for the logistic ANN) to very high (0.942 for Naïve Bayes). Sensitivity, while having a wide range, remained low (0.094 for Naïve Bayes to 0.513 for ANN), with specificity being mediocre to high (0.633 for ANN to 0.976 for Naïve Bayes). PPV being overall low (highest 0.137 for Naïve Bayes), in contrast to NPV (lowest at 0.961 for decision tree/KNN). Detection rate and F1 remained low across all methods (highest at 0.019 for ANN for the former and 0.112 for Naïve Bayes for the latter). Balanced accuracy ranged between 0.505 for the decision tree and 0.582 for the logistic regression, while AUROC scores ranged along similar values, from 0.505 for the KNN and 0.622 for Naïve Bayes. All metrics for metritis models after up-sampling are presented on table 4.11.

Table 4.8 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Milk Fever outcomes, after up-sampling

<u>Milk Fever</u>							
<u>Training Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.746	0.944	0.796	0.782	0.813	0.627	0.846
Kappa	0.106	0.116	0.181	0.110	0.115	0.050	0.074
Sensitivity	0.796	0.145	0.456	0.674	0.668	0.709	0.383
Specificity	0.750	0.967	0.798	0.785	0.816	0.625	0.860
PPV	0.084	0.111	0.093	0.084	0.096	0.053	0.072
NPV	0.992	0.975	0.990	0.988	0.988	0.987	0.979
AUROC	0.839	0.668	0.826	0.793	0.751	0.623	0.622
Detection	0.271	0.060	0.224	0.257	0.019	0.393	0.184
Rate							
Balanced	0.773	0.556	0.762	0.729	0.742	0.667	0.622
Accuracy							
F1	0.153	0.143	0.133	0.139	0.160	0.240	0.084
<u>Test Set</u>							

	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.733	0.929	0.786	0.743	0.805	0.920	0.774
Kappa	0.071	0.067	0.086	0.065	0.099	0.068	0.030
Sensitivity	0.850	0.200	0.783	0.766	0.800	0.233	0.416
Specificity	0.731	0.943	0.786	0.743	0.805	0.932	0.781
PPV	0.054	0.060	0.063	0.051	0.070	0.059	0.033
NPV	0.996	0.984	0.994	0.994	0.995	0.985	0.986
AUROC	0.801	0.602	0.814	0.790	0.834	0.619	0.632
Detection	0.278	0.003	0.014	0.265	0.014	0.004	0.007
Rate							
Balanced	0.790	0.571	0.784	0.755	0.802	0.583	0.598
Accuracy							
F1	0.103	0.093	0.116	0.097	0.129	0.095	0.062

Table 4.9 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting LDA outcomes, after up-sampling

<u>LDA</u>							
<u>Training Set</u>							
	Logistic	Decisio	Random	ANN	SVM	Naïve	KNN
	Regression	n Tree	Forest			Bayes	
Accuracy	0.719	0.948	0.814	0.625	0.833	0.912	0.895
Kappa	0.016	0.012	0.010	0.026	0.014	0.021	0.014
Sensitivity	0.697	0.123	0.330	0.624	0.324	0.063	0.114
Specificity	0.717	0.934	0.818	0.781	0.837	0.917	0.901
PPV	0.017	0.014	0.014	0.017	0.016	0.031	0.009
NPV	0.996	0.992	0.993	0.997	0.993	0.993	0.991
AUROC	0.725	0.675	0.677	0.754	0.674	0.723	0.508
Detection	0.005	0.052	0.188	0.381	0.169	0.090	0.106
Rate							
Balanced	0.695	0.519	0.574	0.613	0.581	0.520	0.508
Accuracy							
F1	0.034	0.033	0.027	0.254	0.028	0.078	0.027
<u>Training Set</u>							

	Logistic Regression	Decisio n Tree	Random Forest	ANN	SVM	Naïve Bayes	KNN
Accuracy	0.693	0.936	0.794	0.861	0.859	0.983	0.906
Kappa	0.012	0.029	0.023	0.013	0.032	0.024	0.009
Sensitivity	0.555	0.185	0.518	0.259	0.444	0.037	0.148
Specificity	0.694	0.942	0.797	0.866	0.862	0.990	0.912
PPV	0.013	0.024	0.019	0.014	0.024	0.029	0.012
NPV	0.995	0.993	0.995	0.993	0.995	0.992	0.992
AUROC	0.727	0.670	0.671	0.753	0.677	0.719	0.506
Detection Rate	0.004	0.001	0.004	0.002	0.003	0	0.001
Balanced Accuracy	0.625	0.563	0.657	0.562	0.653	0.513	0.530
F1	0.027	0.043	0.037	0.028	0.046	0.032	0.023

Table 4.10 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting RFM outcomes, after up-sampling

<u>RFM</u>							
<u>Training Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.735	0.793	0.770	0.754	0.728	0.935	0.658
Kappa	0.065	0.068	0.079	0.071	0.065	0.183	0.038
Sensitivity	0.520	0.397	0.455	0.510	0.517	0.224	0.508
Specificity	0.744	0.809	0.783	0.780	0.737	0.963	0.664
PPV	0.075	0.080	0.078	0.087	0.073	0.202	0.057
NPV	0.974	0.970	0.972	0.975	0.974	0.968	0.971
AUROC	0.689	0.652	0.674	0.706	0.657	0.682	0.626
Detection	0.295	0.222	0.247	0.020	0.292	0.074	0.362
Rate							
Balanced	0.632	0.603	0.619	0.657	0.627	0.594	0.586
Accuracy							
F1	0.128	0.124	0.139	0.147	0.161	0.210	0.212
<u>Test Set</u>							

	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.707	0.763	0.773	0.709	0.750	0.953	0.651
Kappa	0.067	0.059	0.077	0.073	0.075	0.150	0.028
Sensitivity	0.578	0.428	0.473	0.601	0.518	0.127	0.481
Specificity	0.712	0.777	0.785	0.713	0.759	0.985	0.657
PPV	0.073	0.070	0.080	0.076	0.078	0.261	0.052
NPV	0.977	0.971	0.974	0.978	0.975	0.966	0.969
AUROC	0.656	0.630	0.649	0.698	0.655	0.591	0.600
Detection	0.021	0.016	0.017	0.022	0.019	0.004	0.018
Rate							
Balanced	0.645	0.602	0.629	0.657	0.639	0.556	0.569
Accuracy							
F1	0.130	0.121	0.137	0.135	0.136	0.171	0.094

Table 4.11 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Metritis outcomes, after up-sampling

<u>Metritis</u>							
<u>Training Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.634	0.890	0.825	0.652	0.684	0.911	0.645
Kappa	0.053	0.072	0.079	0.057	0.044	0.148	0.026
Sensitivity	0.558	0.216	0.287	0.562	0.478	0.170	0.423
Specificity	0.638	0.899	0.855	0.657	0.695	0.951	0.657
PPV	0.078	0.105	0.098	0.083	0.079	0.170	0.063
NPV	0.963	0.954	0.956	0.964	0.960	0.954	0.954
AUROC	0.645	0.587	0.627	0.640	0.624	0.658	0.535
Detection	0.395	0.106	0.189	0.377	0.341	0.098	0.377
Rate							
Balanced	0.586	0.550	0.530	0.609	0.577	0.554	0.534
Accuracy							
F1	0.233	0.133	0.111	0.134	0.130	0.179	0.107
<u>Test Set</u>							

	Logistic Regression	Decision Tree	Random Forest	ANN	SVM	Naïve Bayes	KNN
Accuracy	0.695	0.872	0.797	0.628	0.708	0.942	0.687
Kappa	0.038	0.006	0.008	0.028	0.015	0.083	0
Sensitivity	0.459	0.108	0.202	0.513	0.337	0.094	0.297
Specificity	0.704	0.903	0.821	0.633	0.723	0.976	0.707
PPV	0.058	0.043	0.043	0.053	0.046	0.137	0.038
NPV	0.970	0.961	0.962	0.970	0.964	0.964	0.961
AUROC	0.601	0.555	0.590	0.619	0.599	0.622	0.505
Detection Rate	0.017	0.004	0.007	0.019	0.013	0.003	0.011
Balanced Accuracy	0.582	0.505	0.511	0.573	0.530	0.535	0.5
F1	0.104	0.061	0.071	0.096	0.082	0.112	0.068

After up-sampling, kappa values overall improved compared to fitting the models without the use of sampling methods.

Regardless of improvements, the values were still consistently poor for all models across all disease outcomes, though there was some improvement compared to the models trained without a sampling method. All ranged between 0.01 and 0.20, showing only slight agreement of predictions and actual values (Viera and Garrett, 2005) for all methods across all 4 diseases.

4.3.1.2 Individual lactation models predicting on aggregated lactations per herd/quarter-year

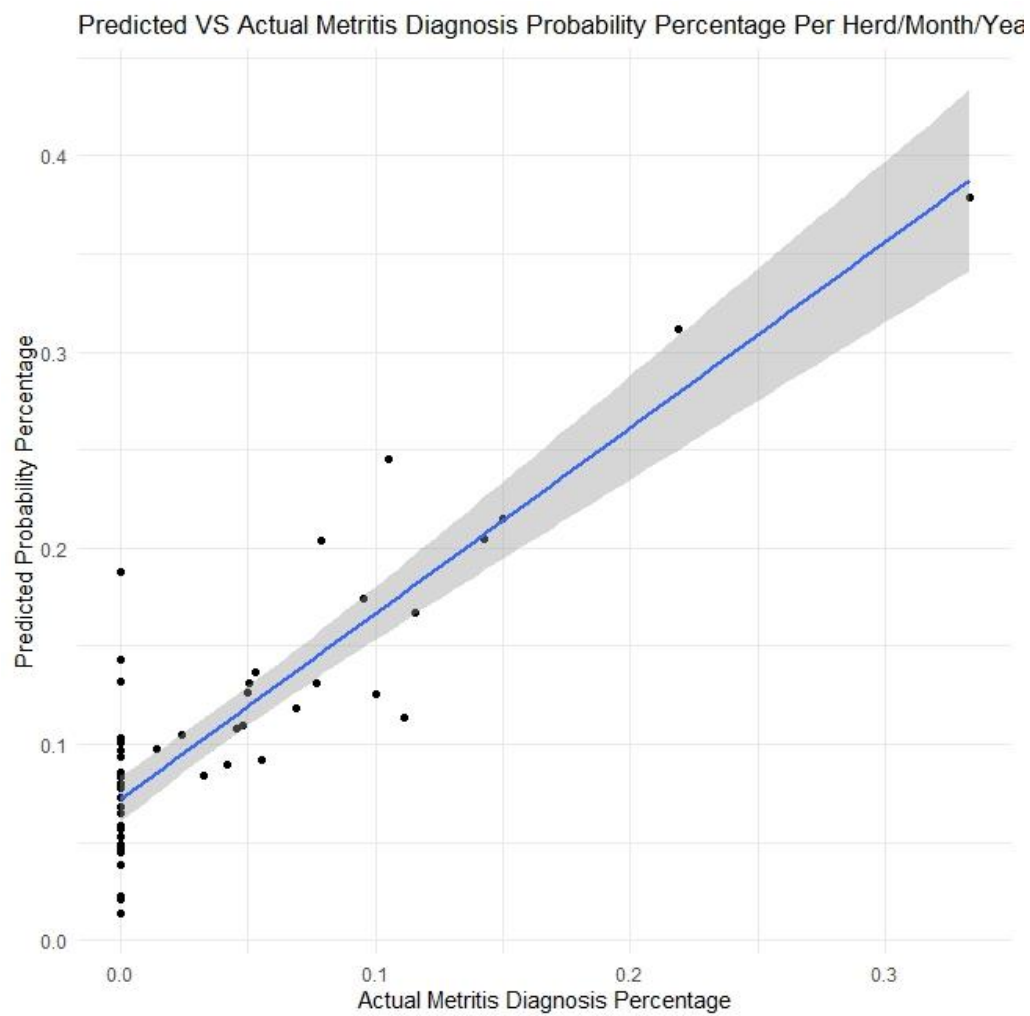
The best performing models for each disease were selected using the best Kappa value to be used for further analysis. The best models for each disease along with their metrics are available on table 4.12.

Table 4.12 Metrics of best performing models for each individual disease

Metric	Milk Fever	LDA	RFM	Metritis
	(SVM)	(SVM)	(Naïve Bayes)	(Naïve Bayes)
Accuracy	0.805	0.859	0.953	0.942
Kappa	0.099	0.032	0.150	0.083
Sensitivity	0.800	0.444	0.127	0.094
Specificity	0.805	0.862	0.985	0.976
PPV	0.070	0.024	0.261	0.137
NPV	0.995	0.995	0.966	0.964
AUROC	0.834	0.677	0.591	0.622
Detection Rate	0.014	0.003	0.004	0.003
Balanced Accuracy	0.802	0.653	0.556	0.535
F1	0.129	0.046	0.171	0.112

The aggregated predictions (at herd-quarter-year level) produced from these models were graphed against the observed outcome values, and the R^2 values were 18.2% for Milk Fever and 5.7% for LDA both using SVM, 14.5% for RFM using Naïve Bayes and 66.9% for metritis also using Naïve Bayes (Figure 4.1).

4.1 Scatterplot of aggregated predictions vs actual percentage of metritis diagnosis
per herd/quarter-year using the Naïve Bayes model



4.3.1.3 Models built using data aggregated at herd-quarter-year level

Predictor variables improving model performance for each of the disease outcomes using the aggregated dataset (whereby each unit of data represented aggregated disease outcomes for all quarter-year lactations in a given herd in a given quarter-year) are shown in Table 4.13.

Table 4.13 Predictor variables included in final predictive models for various individual disease outcomes aggregated at herd-quarter-year level

Milk Fever %	LDA %	RFM %	Metritis %
Mean Rumen Fill pre-partum	Mean Rumen Fill pre-partum	Mean Rumen Fill pre-partum	Mean Rumen Fill pre-partum
Mean Lactation Number	Mean Rumen Fill post-partum	Mean Rumen Fill post-partum	Mean Rumen Fill post-partum
Mean Hock Hygiene Score	Mean Lactation Number	Mean Lactation Number	Mean Lactation Number
Mean BCS change	Mean Hock Hygiene Score	Mean BCS change	Mean BCS change
Mean BCS pre partum	Mean BCS change	Mean BCS pre- partum	Mean BCS pre- partum
% Twinning	Mean BCS pre- partum	% Twinning	% Twinning
% Calf Mortality	% Twinning	% Calf Mortality	% Calf Mortality
Mean 305 Milk Yield of Previous lactation for the group	% Calf Mortality	Mean THI pre- partum	Mean THI pre- partum

Mean THI pre-partum	Mean 305 Milk Yield of Previous lactation for the group	Mean Feed Fence Space per cow in dry pen	Mean Feed Fence Space per cow in dry pen
Mean Feed Fence Space per cow in dry pen	Mean THI pre-partum	Mean Water Trough Space per cow in dry pen	Mean Water Trough Space per cow in dry pen
Mean Water Trough Space per cow in dry pen	Mean Feed Fence Space per cow in dry pen	Mean Neck Rail Height in dry pen	Mean Neck Rail Height in dry pen
	Mean Water Trough Space per cow in dry pen	Month of pre-partum recording	Month of pre-partum recording

The mean percentage of cows diagnosed with milk fever per herd per quarter-year was 3.4% with a median of 0.3% (maximum 61.1% and minimum 0%). For LDA the mean percentages per herd/month/quarter-year were lower, at 1.0% with a median of 0% (maximum 18.8% and minimum 0%). For RFM the mean was 4.3% (median 3.0%, minimum 0% and maximum 40.0%), while for metritis it was at 4.5% with a median of 2.4% (minimum 0% and maximum 42.9%).

R^2 values for models predicting incidence rates of individual diseases and built on aggregated data are shown in Table 4.14.

Table 4.14 R^2 values of all methods of individual disease percentage outcomes

Training Set				
	%Milk	%LDA	%RFM	%Metritis
	Fever			
Linear	0.261	0.202	0.288	0.209
Regression				
Decision Tree	0.390	0.216	0.208	0.139
Random Forest	0.321	0.210	0.211	0.262
ANN	0.263	0.213	0.172	0.340
MARS	0.478	0.118	0.337	0.220
Test Set				
	%Milk	%LDA	%RFM	%Metritis
	Fever			
Linear	0.241	0.186	0.274	0.196
Regression				
Decision Tree	0.375	0.200	0.192	0.133
Random Forest	0.304	0.201	0.198	0.247
ANN	0.249	0.201	0.159	0.324
MARS	0.443	0.104	0.323	0.202

The best performing models for each disease outcome based on the R^2 on the test set were ANN for both LDA and metritis, and MARS for both milk fever and RFM.

4.3.2 Collective Disease Status Outcome

4.3.2.1 Individual lactation disease models

For the models using the individual cow lactations as units of data, the combination of variables that produced the most predictive models were the rumen fill pre and post-partum, neck rail height in pre-calving pens, lactation number, hock hygiene, BCS change, BCS pre-calving, THI in the pre-calving pen, stocking density in the pre-calving pen, mean milk yield in the previous lactation, calf mortality and twinning.

The results of our analysis for the collective disease outcome before up-sampling resembled that of the individual disease analysis (Table 4.15). Accuracy values were high (0.866-0.880), similar to NPV (0.880-0.891), with specificity showing near perfect values (0.966-1), while sensitivity and PPV ranged to very low values (0-0.135 and 0.190-0.468). Similarly to the individual disease outcomes this was a result of dataset imbalance, as evidenced by the low kappa (0-0.136) and balanced accuracy (0.500-0.550). The F1 score was also low (0.041-0.192) and AUROC ranged from the baseline of 0.5 to a low 0.679. The effect of the imbalanced dataset was so great that the decision tree model predicted every datapoint as “Healthy”, hence achieving a sensitivity of 0, specificity of 1, kappa of 0 and not being able to compute the PPV and F1-score, thus earning the spot of the worst performing model. The rest of the methods, while not being so absolute did not show significant improvements, with the best overall algorithm being the Naïve Bayes, both in terms of kappa

with a value of 0.136, and in terms of sensitivity and specificity trade off (0.135 and 0.966 respectively) and F1 (0.192). It also achieved the highest balanced accuracy at 0.550. Nevertheless, the Naïve Bayes still had very low performance.

Table 4.15 All metrics for collective disease outcome models as calculated on both the training and the test set, before up-sampling

<i>Training Set</i>							
	Logistic Regression	Decision Tree	Random Forest	ANN	SVM	Naïve Bayes	KNN
Accuracy	0.880	0.880	0.875	0.877	0.879	0.866	0.878
Kappa	0.054	0	0.100	0.067	0.033	0.136	0.003
Sensitivity	0.038	0	0.083	0.051	0.023	0.135	0.004
Specificity	0.995	1	0.983	0.990	0.996	0.966	0.997
PPV	0.462	-	0.407	0.464	0.468	0.348	0.190
NPV	0.883	0.880	0.887	0.884	0.882	0.891	0.880
AUROC	0.657	0.5	0.679	0.634	0.572	0.640	0.527
Balanced Accuracy	0.516	0.5	0.533	0.520	0.509	0.550	0.501
F1	0.087	-	0.136	0.091	0.054	0.192	0.041
Detection Rate	0.004	0	0.010	0.006	0.002	0.016	0.0005
<i>Test Set</i>							
	Logistic Regression	Decision Tree	Random Forest	ANN	SVM	Naïve Bayes	KNN

Accuracy	0.901	0.904	0.882	0.898	0.904	0.883	0.886
Kappa	0.023	0	0.029	0.002	0.028	0.058	-0.006
Sensitivity	0.018	0	0.036	0.009	0.018	0.072	0.018
Specificity	0.995	1	0.982	0.992	0.998	0.969	0.977
PPV	0.285	-	0.181	0.111	0.500	0.200	0.080
NPV	0.905	0.904	0.906	0.904	0.905	0.908	0.904
AUROC	0.636	0.5	0.631	0.650	0.603	0.633	0.557
Balanced Accuracy	0.506	0.5	0.509	0.501	0.508	0.520	0.498
F1	0.034	-	0.060	0.016	0.035	0.106	0.029
Detection Rate	0.001	0	0.003	0.001	0.003	0.006	0.001

After up-sampling the overall accuracy appeared to be from fair to good across all methods (ranging from 66.3% to 85.8%) with the exception of KNN where accuracy was low (0.511), while the kappa values were consistently low, with the random forest, while having the highest value at 0.172, still being below the 0.2 benchmark for achieving anything more than slight agreement between predictions and actual values (Viera and Garrett, 2005).

Sensitivity, while having a wide range, remained low with the highest score being 0.548 for the KNN. Specificity also had a wide range, starting from 0.506 for the KNN and peaking high at 0.950 for the Naïve Bayes. PPV remained

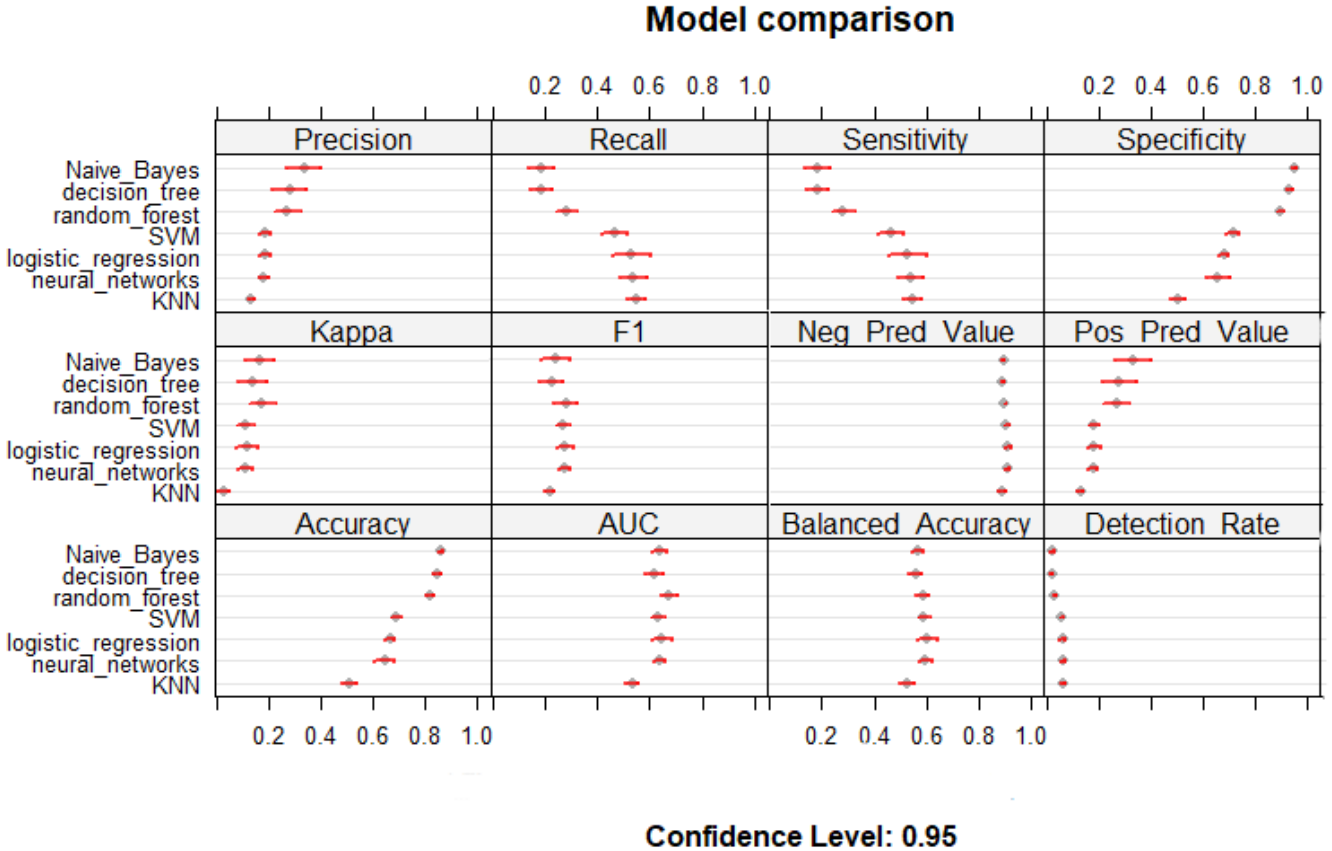
consistently low across all methods, the highest value being at 0.335 for the Naïve Bayes, while NPV ranged high, with KNN having the lowest value at 0.890. Both F1 and Detection Rate values were low across all algorithms (highest at 0.274 for the random forest and 0.066 for KNN respectively). Balanced accuracy ranged close to the 0.5 baseline (0.527 for KNN to 0.605 for logistic regression), with AUROC displaying similar values (0.531 to 0.646 for KNN and logistic regression respectively). The metrics of all methods on both the training and the test set are presented in Table 4.16 and the values of all the models applied on the test set are also graphically shown in Figure 4.2.

Table 4.16 All metrics for collective disease outcome models as calculated on the training and the test set, after up-sampling

<i>Training Set</i>							
	Logistic Regression	Decision Tree	Random Forest	ANN	SVM	Naïve Bayes	KNN
Accuracy	0.663	0.844	0.820	0.643	0.685	0.858	0.511
Kappa	0.116	0.137	0.172	0.106	0.109	0.166	0.023
Sensitivity	0.528	0.187	0.281	0.537	0.466	0.185	0.548
Specificity	0.682	0.934	0.894	0.658	0.715	0.950	0.506
PPV	0.184	0.280	0.270	0.178	0.183	0.335	0.132
NPV	0.914	0.893	0.901	0.912	0.907	0.895	0.890
AUROC	0.646	0.615	0.674	0.637	0.630	0.637	0.531
Balanced Accuracy	0.605	0.558	0.587	0.598	0.591	0.566	0.527
F1	0.273	0.218	0.274	0.267	0.263	0.236	0.211
Detection Rate	0.063	0.022	0.033	0.065	0.055	0.021	0.066
<i>Test Set</i>							
	Logistic Regression	Decision Tree	Random Forest	ANN	SVM	Naïve Bayes	KNN

Accuracy	0.718	0.867	0.814	0.690	0.673	0.876	0.473
Kappa	0.144	0.039	0.082	0.108	0.078	0.024	0.011
Sensitivity	0.536	0.081	0.218	0.509	0.463	0.054	0.572
Specificity	0.737	0.950	0.877	0.709	0.696	0.963	0.462
PPV	0.177	0.147	0.158	0.156	0.138	0.136	0.101
NPV	0.937	0.907	0.913	0.931	0.924	0.905	0.910
AUROC	0.664	0.636	0.635	0.658	0.639	0.644	0.595
Balanced Accuracy	0.636	0.515	0.548	0.609	0.579	0.509	0.517
F1	0.266	0.105	0.183	0.239	0.213	0.077	0.172
Detection Rate	0.051	0.007	0.020	0.048	0.044	0.005	0.054

Figure 4.2 Metric comparison of all lactation-level collective disease models after using up-sampling with 95% confidence intervals

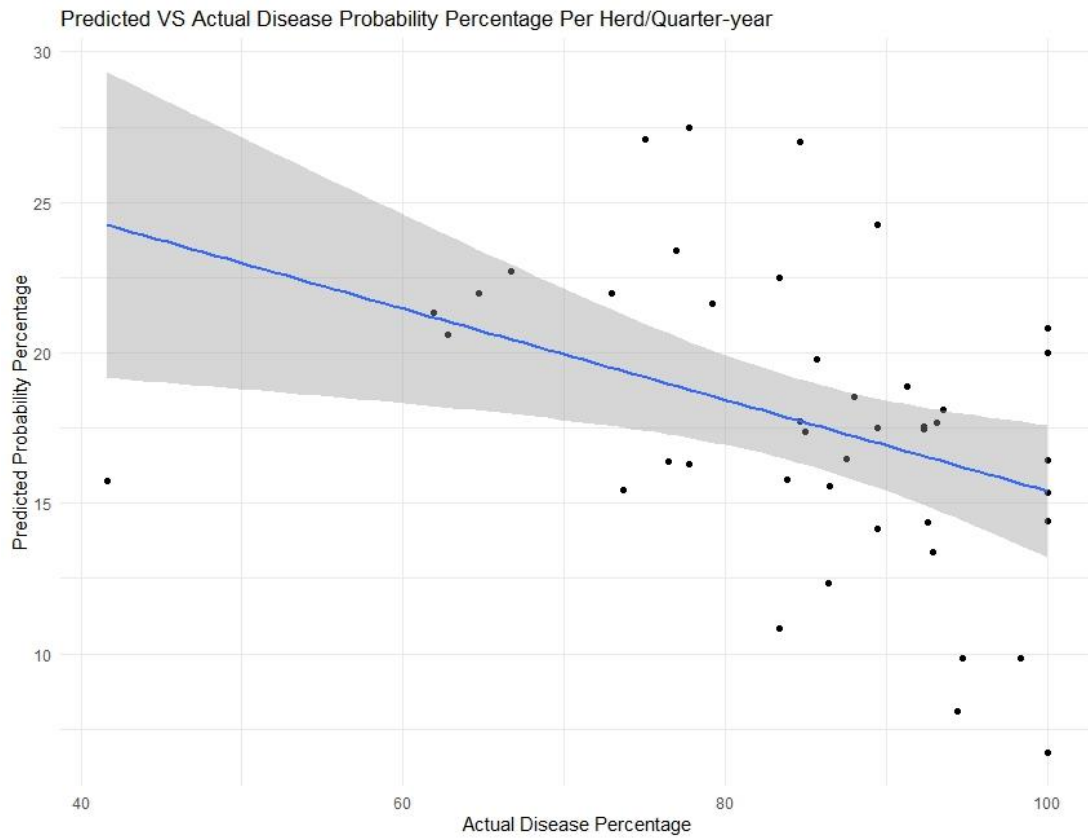


4.3.2.2 Individual lactation disease model making predictions on an aggregated level

The same set of models were used to generate predictions from the holdout dataset. The total number of herd/quarter-year groups used were 461 with 368 of them being used for training (9298 lactations total) and the rest being used for testing (93 groups of 1056 lactations). The minimum group was 10 as set, the median 29 (mean = 46.06) and the maximum 181.

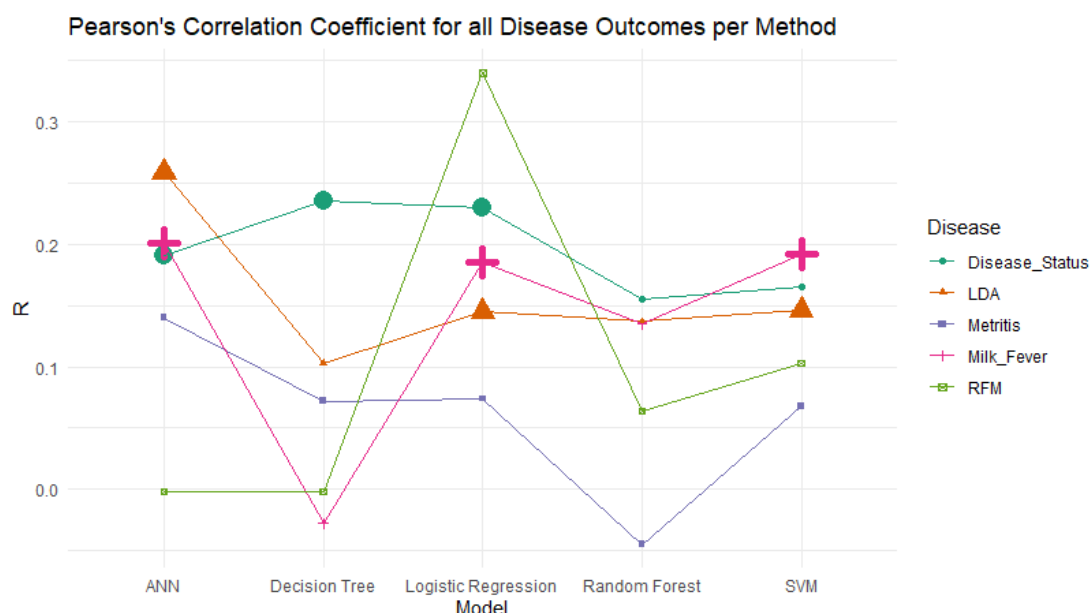
The linear association of predicted and actual values for data aggregated at herd-quarter-year level was analysed. The R^2 value describing the proportions of explained variation between actual and predicted values for logistic regression was 44.5%, indicating that about half of the aggregated outcome's variation can be explained by aggregating the predictions (Figure 4.3). As for the rest of the models it ranged to 23.61% for decision tree, 27.60% for random forest, 25.63% for ANN, 30.09% for SVM and 10.57% for Naïve Bayes.

Figure 4.3 Scatterplot of predictive vs actual collective disease diagnosis probability per herd/quarter-year



There was no substantial improvement in model performance using an overall disease status outcome compared to the models predicting occurrence of an individual disease (Figure 4.4).

Figure 4.4 Pearson's correlation coefficients between aggregated predicted and observed outcomes (per herd-quarter-year) across model types for all individual disease outcomes and for collective disease status ("Disease_Status") using models built on individual lactation data,

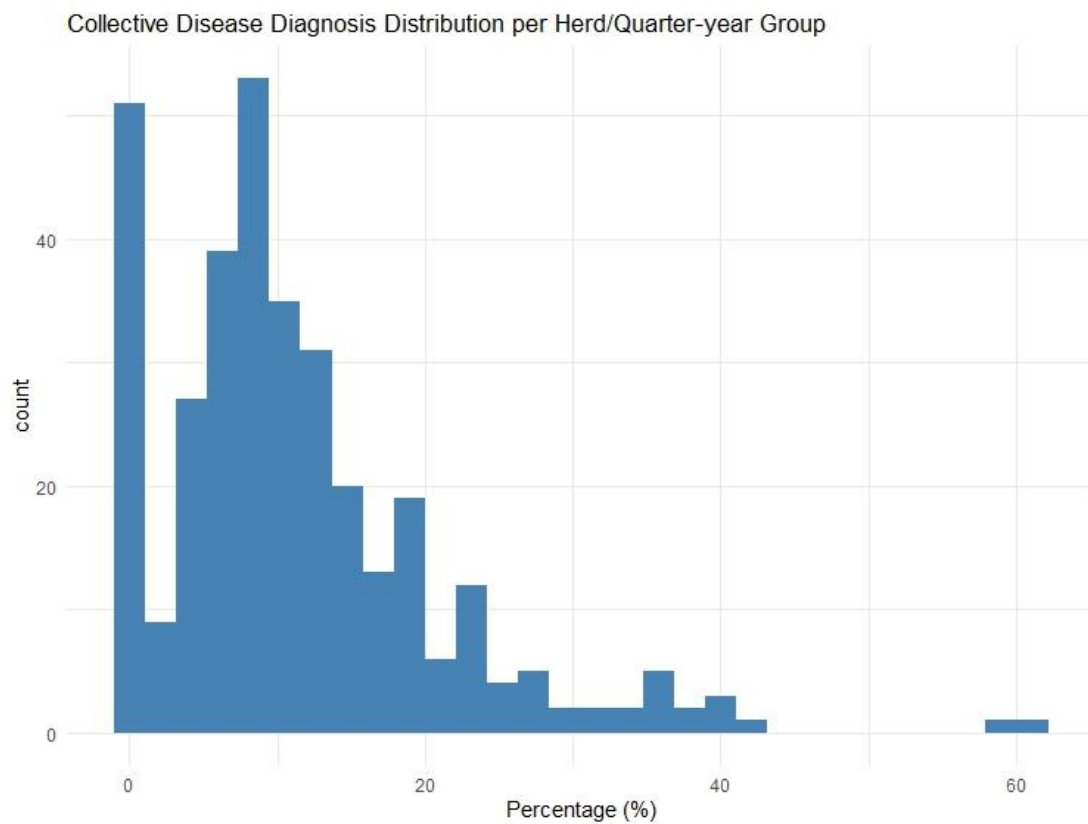


4.3.2.3 Aggregated Herd/quarter-year level models

For the models built using the herd/quarter-year as a unit of data, the final predictive variables that were included in the models were the mean rumen fill per group both pre and post-partum, the neck rail height, water trough space, feed fence space and THI per group in dry pens, the month pre calving, the mean BCS pre partum as well as the mean BCS change pre and post calving, the mean 305 milk yield per group for the previous lactation, and finally the percentages of calf mortality and twinning, again, per group.

The mean percentage of cows diagnosed with at least one disease per herd-quarter-year was 10.1% with a median of 8.8% (minimum 0% and maximum 75%). (Figure 4.5)

Figure 4.5 Histogram of collective disease diagnosis distribution per herd each quarter-year

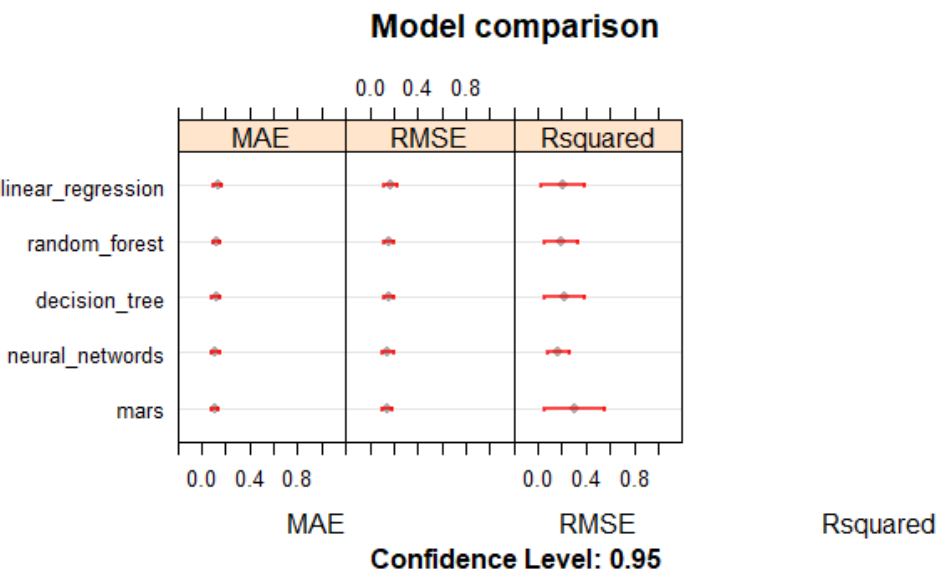


The metrics of the resulted aggregated level models are displayed in table 4.17 and Figure 4.6. R^2 on the test set was consistently low with the highest yielded value being 32% for the MARS model, meaning that the models were at best able to explain up to nearly a third of the variation of the outcome.

Table 4.17 Metrics of all methods of disease diagnosis percentage per herd per quarter-year on both the training and the test sets

Training Set			
	RMSE	R ²	MAE
Linear	0.160	0.230	0.126
Regression			
Decision Tree	0.149	0.228	0.110
Random Forest	0.150	0.109	0.110
Artificial Neural Networks	0.167	0.251	0.121
MARS	0.131	0.343	0.122
Test Set			
	RMSE	R ²	MAE
Linear	0.175	0.211	0.133
Regression			
Decision Tree	0.157	0.222	0.119
Random Forest	0.158	0.199	0.122
Artificial Neural Networks	0.176	0.239	0.132
MARS	0.142	0.320	0.129

Figure 4.6 Metric comparison of different models of the collective disease percentage outcome on the test set



4.4 Discussion

For this chapter, individual and collective disease outcomes were investigated on both an individual lactation and an aggregated herd/quarter-year level. Various predictive models were built further looking into the possible effect of the machine learning method used on the overall performance. With the kappa value as guide, it became clear that all models produced were of little predictive value and could not provide a model with reliable predictions. Up-sampling, while improving the overall metrics was not enough to increase the kappa values to an acceptable threshold. When it came to individual disease outcomes on a lactation level, the kappa values of all methods in all diseases did not exceed the maximum of 0.122 which was the SVM method for milk fever. Using the individual lactation models for aggregated predictions did not produce meaningful R^2 values (5.7% for LDA to 18.2% for milk fever both for the SVM model), for all diseases except for metritis (66.9% for the Naïve Bayes model). The same value for the collective disease outcome was at 44.5%. Finally, for the aggregated models the R^2 ranged low from 20.1% to 36.5% for LDA and milk fever respectively. When it came to the collective disease outcome, again the individual lactation models produced low kappa values with the highest one being 0.172 for the random forest model. Further using the model for aggregated predictions had similar negative results with the correlation coefficient being non-statistically significant. Moving to models built on an aggregated level also did not improve the results, with the R^2 ranging low from 0.199 to 0.302 for the random forest and the MARS models respectively.

There was a wide variety of predictor variables available for predictive modelling and a wide variety of methods and techniques was utilised. Yet no

meaningful predictive models for post-calving disease, neither individual nor collective, were created. This does not clash with any existing research, as according to the current research as well as the systematic review conducted by Slob et al. (2020) the papers that looked into using machine learning for dairy cow disease outcomes mainly focused on mastitis, which was not a part of the outcomes used for this thesis.

These results also highlight that, in cases of unbalanced outcomes, even models with relatively low predictive power can exhibit high accuracy metrics. For instance, the random forest, which achieved one of the highest accuracies along with balanced PPV/NPV and kappa values, still lacked strong predictive capabilities. A closer examination of metrics, particularly kappa and balanced accuracy—which account for outcome imbalances—revealed that none of the models demonstrated adequate predictive power. Although up-sampling appeared to enhance model performance, the resulting algorithms were still unlikely to be sufficiently predictive for practical use. Very unbalanced outcomes are likely to become quite common in terms of predictive modelling. One such example is oestrus detection, since most of the times most cows are not even in heat when it occurs. Post et al. (2021) while focusing on the prediction of mastitis and lameness, which were beyond the scope of our study, did emphasise the effects of an imbalanced dataset on the metrics and therefore practical application of a model. Their models in particular, while producing fair values of sensitivity and specificity, also had very low PPV, similar to a lot of our individual disease models. They argued that these results, even though sometimes warranted by a developer standpoint, do not accurately represent the real-world application on a practical farm. Post et al. (2020) even

showcased that different up- and down-sampling techniques for balancing training data had no impact when applied to unknown, realistic datasets, again reinforcing our findings. Sturm et al. 2020 also dealt with a health outcome that was not part of our study (subclinical ketosis), however their results are worth reporting as they were dealing with an imbalanced dataset as well. The best performing model had some reasonable metrics with accuracy at 0.725, sensitivity at 0.669, specificity at 0.736 and NPV at 0.922, accompanied with some quite low, with F1 at 0.435 and PPV at 0.322, once again showcasing the effects of a minority class. Avizheh et al. (2023) also implemented sampling methods (down-sampling and cost-effective method) in order to improve metrics and found that the AUROC of their models did not show improvement, while the F-score showed an average difference of 0.031 while the Roc curve did not reveal any improvement in predictive power. On the other hand, Keshavarzi et al. (2020) when dealing with an imbalanced dataset for the prediction of abortion incidence, reported that both up-sampling and down-sampling methods were found to improve predictiveness. On average the F1-score had an average difference of 0.106 and 0.088, while AUROC showed differences of 0.897 and 0.893 for down-sampling and up-sampling respectively. It was also noted that while rules, trees and functions showed significant increase in metrics, Naïve Bayes models did not. These results seem to be in agreement with our findings, as the average difference of F1 in milk fever was 0.077, for RFM 0.071 and for metritis 0.053, only including the methods that were able to compute a F1 value pre-sampling. The only model that could compute an F1-score for LDA before up-sampling was actually the ANN that showed a numerical increase in F1-value, of 0.112. The differences

in AUROC were -0.014 for milk fever, 0.079 for LDA, 0.057 for RFM and 0.026 for metritis, so for all disease except milk fever there was an overall numerical increase. The average numerical increase in kappa pre and post sampling was 0.090 for milk fever, 0.013 for LDA, 0.032 for RFM and 0.051 for metritis.

Accuracy was moderate across all models, even after applying resampling techniques, underscoring its potential to be misleading. Sensitivity and specificity, when considered alone, also failed to consistently reflect the models' true performance. In the case of the random forest, decision tree, and naïve Bayes algorithms, while specificity was high, sensitivity was notably low, highlighting the issue of class imbalance. However, in other algorithms, there was a better balance between the two, with sensitivity around 0.6, masking the models' inability to make accurate predictions. This demonstrates that these metrics can vary significantly between methods and are insufficient to fully evaluate model performance. Since accuracy, sensitivity, and specificity are frequently cited in the literature, this presents concerns. These issues were evident in our study, where the kappa values across all models were below the 0.4 threshold required for moderate agreement (Viera & Garrett, 2005), reinforcing the idea that accuracy, even when combined with sensitivity and specificity, is inadequate for evaluating predictive models—particularly in imbalanced datasets, as seen in our case with a disease prevalence of only 10.98%.

Brodersen et al. (2010) proposed that balanced accuracy should replace overall accuracy, particularly for addressing imbalanced datasets. The findings of this study support that perspective, as balanced accuracy across all models

hovered just above 50%. This can be attributed to the algorithms' tendency to predict the majority class for all instances. Balanced accuracy accounts for both classes equally by calculating their individual accuracies and averaging them. In imbalanced datasets, it is common for overall accuracy to be maximized by misclassifying all instances of the minority class, with all predictions assigned to the majority class. In such cases, the accuracy of the majority and minority classes would be 1 and 0, respectively, resulting in a balanced accuracy of 0.5. In our study, the fact that balanced accuracy is close to 0.5, while overall accuracy remains high for most models, highlights an issue with class imbalance.

Viera and Garrett (2005) describe how kappa is influenced by the prevalence of the disease examined, hence the frequency of its class in the dataset. At the same time, from the literature it is evident that in the past and even in some recent studies there is a preference in reporting the overall accuracy, which can be misleading in the case of imbalanced datasets, and not complement it with additional metrics that take the proportion of the classes into account (Table 1). A lot of contemporary studies include the AUROC, it has been argued that it is not an appropriate metric in every situation and it can in fact provide an overly optimistic assessment of a model's performance in imbalanced datasets, as it may be inflated by the model's success in classifying the majority class while ignoring the minority class (Lobo et al., 2008, King et al., 2021, Hancock et al., 2023, Bednarski et al., 2022). From our results it is evident that kappa is one of the metrics most affected by this situation. In fact, the maximum value of kappa was just 0.16 for the random forest model even after up-sampling, pointing to

the issue of the model's actual predictive performance perhaps more effectively than other metrics.

Another issue highlighted in our study is the challenge of algorithm selection. In imbalanced datasets, if models are chosen primarily based on overall accuracy, those with low predictive power may be favoured, as high overall accuracy can be easily achieved by misclassifying the minority class. A notable example is the naive Bayes model in both collective disease outcomes and individual lactation models, which exhibits one of the highest overall accuracies but also the lowest balanced accuracy and a very low kappa. In reality, its predictive power does not surpass that of the logistic regression model, which has a much lower overall accuracy. This illustrates that relying on inappropriate metrics can lead to misleading model comparisons.

It is clear from the literature that machine learning algorithm reporting in the veterinary epidemiology field has evolved from mainly using accuracy as the headline metric to utilising and interpreting more complex metrics. This work further supports the more widespread adoption of alternative metrics such as balanced accuracy or kappa, especially where the outcome being predicted is either very common or very rare.

The results for the aggregated outcomes milk fever, LDA, RFM, metritis and collective disease percentage per quarter-year were similar to that of the binary models, as the R^2 of all aggregated prediction models was lower than 0.5 and for most cases much lower than that. Nonetheless, when it came to using the individual level models for making predictions on an aggregated level the results varied. For most disease the individual level models failed to make

meaningful aggregated predictions. However, the results for metritis indicated that the aggregated predictions explained almost 67.0% of the total variation of the aggregated outcome, resulting in a fair R^2 value. The same method for the collective disease status resulted in a lower value of 44.5%, which though higher than most other diseases, does not even explain half of the variation of the outcome and it would be safe to assume that the increase in value compared to LDA, milk fever and RFM was probably an effect of the influence of metritis. From these results it is evident that the use of the individual level model predictions on an aggregated level is possible when it comes to metritis, since it presented a moderate value of R^2 and it outperformed both the individual level model making predictions on the individual level and the fully aggregated model.

As stated above the majority of existing research that aimed in utilizing machine learning to predict disease outcomes mainly focused on mastitis (Ebrahimie et al., 2018b, Ebrahimie et al., 2018a, Hassan et al., 2009, Kamphuis et al., 2010, Kamphuis et al., 2008, Mammadova et al., 2013, Panchal et al., 2016, Sharifi et al., 2018, Slob et al., 2020, Sun et al., 2010). Common predictors for mastitis appear to be milk parameters (Hassan et al., 2009, Sun et al., 2010) or in some cases genetic data (Sharifi et al., 2018). In one systematic review (Slob et al., 2020) that did include papers with both health and production outcomes, although mostly health outcomes (21/38), most papers included milk parameters (66%) or milk properties (58%) as independent variables, with only a few (29%) having calving information and/or cow characteristics and just a fifth of them (21%) using lactation information. Other have also reported the use of machine learning for the detection of lameness (Pastell and Kujala, 2007) as

well as ketosis (Slob et al., 2020). A number of studies have focused on milk spectra for disease prediction. Hernandez et al. (2021) presented a collective disease outcome that included lameness, mastitis, reproductive disorders, calving disorders and finally other ailments. They presented various metrics, however they focused on the sensitivity and the PPV as, similar to our data, they had only a fraction of positive diagnoses. Their best model produced low values of these two metrics (61.74% and 59.99% respectively), which were also accompanied by very high values of standard deviation (15.99% and 26.20% respectively). Therefore, while there might be an indication that milk MIR spectra might be better predictors and that a collective disease outcome might be predictable as a collective measure, these results need to be interpreted cautiously. Multiple studies explored the prediction of hyperketonemia (Bonfatti et al., 2019, Luke et al., 2019, Pralle et al., 2018, Walleser et al., 2023) using milk spectra reporting good specificity and sensitivity. Franceschini et al. (2022) utilised unsupervised methods to create clusters of animals with possible metabolic disease and general health status, using milk spectra as well, and their results were promising. All these studies indicate that milk spectra might be a viable predictor for health status, however it was not considered in our study as it was not routinely analysed on most herds in the dataset, with our focus targeting more easily routinely gathered data. Another recent study (Lasser et al., 2021) focused on predicting disease such as metritis and periparturient hypocalcaemia, reporting various metrics such as F1, specificity, sensitivity, precision and accuracy. The F1 for metritis ranged within moderate values from 0.521 to 0.606, with periparturient hypocalcaemia following with similar, if not slightly lower, values from 0.482 to 0.548. Vidal et al. (2023)

focused on the prediction of metritis as well, providing some models with high F1 scores, but they utilised sensor data from accelerometers. An even more recent study (Risvanli et al., 2024) also aimed to predict metritis, using a sensor measuring intrauterine gases, providing models with high accuracy (71.22%), precision (64.4%) as well as recall (71.2%). However, this study examined uteri collected from abattoirs and not live cows so potential practical application could yield different results. De Oliveira et al. (2021) also investigated metritis, but rather than its diagnosis they attempted to predict its treatment success, presenting models with high F1 (0.81), sensitivity (0.85) and PPV (0.78), but low specificity (0.39) and NPV (0.50). Finally, Merenda et al. (2020) attempted to predict metritis, acute metritis, along with success and failure of treatment. The predictive variables used were routinely available data, BCS and behavioural data. The models for metritis and acute metritis produced had fair AUROC (0.82 and 0.87 respectively) with reasonable specificity and sensitivity, however the model for acute metritis had low PPV (0.30) while metritis' was fair (0.60). The metritis model scored considerably better than even the best of our metritis' models. One possible explanation is that their predictive variables were overall more appropriate for prediction. Their routinely gathered data included information such as indigestion and California Mastitis Test results which we did not have access to. However, it is also possible that the fact they only data from 2 farms, as opposed to our 79, could lead to some degree of overfitting. The study also lacked external validation. In addition to Hernandez et al. (2023), Zhou et al. (2022) also explored a collective disease outcome. They utilized variables such as milk yield, physical activity, rumination time changes, and the electrical conductivity of milk, while the out of the multiple diseases they

included in their outcome the one in common with our study was metritis. The model presented as best, a decision tree, achieved the highest F1, precision and AUROC (0.787 92.86%, and 0.908, respectively), however it lacked in sensitivity (68.42%). In comparison, our models never achieved such values, with the highest F1 score for the collective disease outcome being 0.274, highest precision 33.5% and highest AUROC 0.674, while the highest sensitivity was at 54.8%. This could be an indication that the combination of milk variables and behavioural data collected through the automated monitoring system Zhou et al. (2022) used were in fact better predictors for the combination of diseases selected. Nevertheless, they emphasized that the low sensitivity is an indication that their algorithms still required improvement to be properly utilized. It should also be noted that they used a control group, eliminating the class imbalance naturally occurring on farms. However, having a limited number of cows in their dataset (131 sick and 149 control cows) in addition to the fact that they were sampled from only 2 farms, as opposed to our 12,863 from 79 farms. might have led to overfitting of the algorithm which occurs when the data is not variable enough. Dineva and Atanasova (2023) also opted for a general disease outcome, however their definition was even broader as the data was collected from health diaries and disease was defined as any kind of condition, including those caused by heat or cold stress. A 'suspect' group was added as an outcome, to include the presence of condition in cows that have not yet manifested as disease. The predictors included age, lactation number, DIM, daily milk yield, weather data, month and week and were all collected through IoT (Internet of Things) devices. Their best model was a random forest classifier with 95.4% sensitivity, 97% PPV and 95.4% accuracy. They also used

sampling methods as they were dealing with an imbalanced dataset. Once again, it is possible that variables such as the weather-related data that were not available in our analysis, did in fact improve predictiveness, however similar to Zhou et al. (2022), Dineva and Atanasova (2023) collected data from only one farm and 120 cows, eliminating any farm variability that could affect predictions. In fact, not only were the data collected from one farm but the month and week variables were added as predictors as well as opposed to the approach taken in our study where the herd-month groups were distinctly separated in training and test groups in order to avoid data leakage. Nevertheless, it should be acknowledged that both Zhou et al. (2022) and Dineva and Atanasova (2023) had small sample sizes, due to the fact that they utilised new sensor technology to collect their data, therefore a wide implementation was logistically difficult, and that part of the difference in our models' predictiveness could be attributed to the better quality of predictor variables an automated system could potentially collect compared to a human-driven system with potential subjectivities, however small.

There are a number of limitations in our study, such as the number of missing data in the predictors and potential under-reporting of certain diseases. When it comes to the impact on predictive modelling, predictors with missing data can result into great loss of information if used in a model, as only complete cases can be included. For this reason, we excluded variables such as the rumen fill post-partum with great numbers of missing data. For the predictive models, the possible under-reporting in farms, during the data collection, could mean that the model does not provide accurate predictions as an entire subset of the outcome is ignored. However, the incidence of the disease was so low that the

percentage which would be overseen would be most likely negligible. The overall quality of the predictive variables might also lack in certain areas, considering that they were recorded by human assessors, who however well trained might insert a level of subjectivity or variability in their assessments. Furthermore, the choice of variables collected were not in any way specifically collected for their presumable correlation with a certain outcome but were rather monitored in an already established program as general measures of farm management. However, since the purpose of the study was to showcase whether it is possible to make predictions using easily available information that farmers may already have the results are still valuable. It is evident that most relevant studies have shown the best results when using special equipment such as sensors and automated monitoring systems (Hernandez et al., 2021, Bonfatti et al., 2019, Luke et al., 2019, Pralle et al., 2018, Walleser et al., 2023, Franceschini et al., 2022, Vidal et al., 2023, Risvanli et al., 2024, Zhou et al., 2022). Therefore, there is an indication that perhaps the most advanced use of technology for variable collection might be the more valuable approach.

A further limitation could be the convenience sampling from farms with similar characteristics, which would result into poor generalisability of the models. Nonetheless, if we acknowledge that the results only apply in UK farms with high productivity the models and their results can still be of use for other similar farms. Furthermore, the use of data from 79 farms in the final models would still introduce some variability and probably better generalizability compared to Merenda et al. (2020) and Zhou et al. (2022) with data from 1 and 2 farms respectively. Nonetheless, our study lacks external validation which would

ultimately verify any possible highly predictive model, adding another limitation to our research.

Chapter 5 – Prediction of reproductive outcomes

5.1 Introduction

Fertility is a key driver of profitability in dairy farming systems (Cabrera, 2014), and there is extensive evidence that events around “transition” (i.e. occurring pre- and post-calving) have associations with subsequent reproductive performance (Roche et al. 2017). Microbial infections that are established postpartum can lead to decreased reproductive function in females (Dobson et al., 2008, Gautam et al., 2009, Kasimanickam et al., 2004). Furthermore, the NEB that occurs during this period is also linked with a decline in reproductive performance by increasing the calving to first ovulation interval and decreasing pregnancy rates at first service (Butler and Smith, 1989). Factors associated with transition cow management, such as nutrition, appropriate BCS maintenance, udder health, calving difficulties, reproductive diseases and cow comfort seem to influence reproductive performance (Caraviello et al., 2006, Lucy, 2001, Schefers et al., 2010). Other factors that are reported to be linked with reproduction include genetics, milk yield, heat stress, NEB, timing of insemination and the presence of reproductive or other diseases (Shahinfar et al., 2014). Clinical mastitis and high somatic cell count recording have also been associated with decreased reproductive performance (Hudson et al., 2012). Poor reproductive performance has been reported to be the first reason for culling, much higher than low milk production (42% and 25% of cows culled, respectively) (Coleman et al., 1985). Gröhn and Rajala-Schultz (2000) also reported that the primary deciding factor when it comes to culling was

pregnancy status, with milk yield being the second one. NEB especially seems to have a key role in determining reproductive performance (Hudson, 2011). Cows with elevated BHBA concentrations during the first or second week postpartum were 20% less likely to get pregnant after first service, and up to 50% less likely if BHBA levels were high on both weeks (Ospina et al., 2013). The relationships between various proxy measures for degree of NEB (such as those based on BCS) and reproductive performance have been established, with a common finding being that with increased NEB there is a delay to the resumption of ovarian cyclicity (Butler and Smith, 1989, Gümen et al., 2005).

Interpreting the extensive body of literature on factors affecting reproductive performance in dairy cows is hindered by the wide variety of fertility outcome measures. These can broadly be categorized as rate- or interval-based measures. Recently, 21-day pregnancy rate has emerged as the key reproductive metric (at least for year-round calving herds). This is defined as the proportion of “eligible” cows (those past the herd’s voluntary wait period, not marked as selected to cull and not already pregnant) which become pregnant every 21 days. This is determined largely by the 21-day insemination rate (proportion of eligible cows inseminated every 21 days) and the conception risk (proportion of inseminations which lead to a pregnancy). Interval-based measures (such as time from calving to first insemination, establishment of pregnancy or subsequent calving) are also used.

The majority of papers utilizing machine learning methods on dairy cow derived data have in fact dealt with reproduction outcomes; including models to predict first service conception rate, the probability of conception in heifers, time of calving, calving difficulty and/or dystocia in both heifers and cows, oestrus

detection and prediction, insemination success and semen quality (Caraviello et al. 2006, Fenlon et al. 2017a, b , Borchers et al. 2007, Grzesiak et al., 2010, Higaki et al. 2019, Hempstalk et al., 2015, Shahinfar et al., 2014, Dolecheck et al. 2015, Aguiar et al. 2012, Bates and Saldias, 2019, Romadhonny et al., 2019, Cairo et al., 2020, Keceli et al., 2020, Keshavarzi et al., 2020, Miller et al., 2020, Avizheh et al., 2023, Vázquez-Diosdado et al., 2023, Hemalatha et al., 2021, Schweinzer et al., 2019, Wang et al., 2020).

Caraviello et al. (2006) used the accuracy and AUROC to report the predictive ability of their models. Namely, they focused on building models on first service conception rate and pregnant status at 150 DIM, with an accuracy of 75.6% and 71.4% and AUROC of 0.68 and 0.73 respectively. Machine learning techniques have been applied to estimate the time-to-calving (Miller et al., 2020), with models for dairy cows demonstrating improved predictive accuracy up to 4 hours before calving. During this period, Matthew's correlation coefficient increased from 0.06 to 0.14. The optimal combination of AUROC, sensitivity, and specificity occurred 2 hours before calving, reaching 95.4%, 91.3%, and 93.5%, respectively. After evaluating various techniques, Borchers et al. (2017) produced ANN models focused on predicting the time of calving both on daily and bihourly intervals, with the first one having a sensitivity of 100% and specificity of 86.8% and the latter a sensitivity of 82.8% and specificity of 80.4%. In a study with comparable findings, Keceli et al. (2020) used activity and behavioural data to develop models that achieved 100% sensitivity, specificity, PPV, and NPV for predicting calving on the day before. Similarly, Vázquez-Diosdado et al. (2023) explored calving prediction using sensor data, with the best performance obtained using data from 2 days before

calving, achieving 87.81% accuracy, 92.99% specificity, 75.84% sensitivity, 82.99% PPV, 78.85% F-score, and 90.02% NPV. Another study which investigated the prediction of calving difficulties reported AUROC values between 0.64 and 0.89 (Fenlon et al., 2017b), while a second one attempting to predict dystocia (Zaborski et al., 2018) reported accuracy up to 0.589 for a MARS model focused on heifers and 0.649 for an ANN model focused on cows. Avizheh et al. (2023) used historical data to predict calving difficulty, but the resulting models showed low AUROC and F1-scores due to an imbalanced dataset. Although sampling methods improved these metrics, they remained relatively low, with F1-scores ranging from 0.38 to 0.42. In contrast, Brand et al. (2021) employed milk spectral data for pregnancy prediction, achieving a model with a sensitivity of 0.89, specificity of 0.86, and overall accuracy of 0.88. Two studies researched into predicting oestrus detection, with one producing models with very high accuracy, lying between 91% and 100% for all models (Dolecheck et al., 2015), while the other one produced less conclusive results, demonstrating only a numerical difference in sensitivity and precision, both at 94% (Higaki et al., 2019). Cairo et al. (2020) also focused on predicting oestrus using behavioural data, reporting high accuracy. Similarly, Hemalatha et al. (2021) used milk parameter data, achieving high accuracy along with strong precision, recall, specificity, and F1 scores. Schweinzer et al. (2019) developed a model using accelerometer data, which demonstrated over 90% sensitivity, specificity, PPV, and NPV. Wang et al. (2020) incorporated both accelerometer and location data, with their best-performing neural network model predicting within a 30-minute time window, achieving 99.36% sensitivity, 53.33% specificity, 95.76% PPV, 93.72% NPV, 95.36% accuracy, and an F1 score of

97.51%. Romadhonny et al. (2019) created an oestrus classification model with over 80% accuracy, but it only correctly identified late oestrus 6.4% of the time. Two studies (Hempstalk et al., 2015, Shahinfar et al., 2014) looked into predicting the likelihood of conception success per insemination and reported AUROC values ranging between 0.487 and 0.675 for the one, and accuracy between 72.3% and 73.6% with AUROC between 0.73 and 0.75. Another study trying to predict the probability of conception, specifically on heifers, however even though their models had an overall accuracy between 77.1% and 78.9%, however the consistently low specificity deemed them as of low predictive value (Fenlon et al., 2017a). Grzesiak et al. (2010) developed models for detecting artificial insemination difficulties, achieving an AUROC of nearly 0.9. Bates and Saldias (2019) compared regression and machine learning methods by creating models to predict the 21-day submission rate in dairy cows. Their study found no significant differences in predictive performance between the methods, and while the AUROC values were reasonably strong (0.68-0.73), positive outcomes were more accurately predicted than negative ones. Finally, Keshavarzi et al. (2020) created models to predict abortion incidence, with a mean AUROC of 0.863 and an F1 score of 0.520. After applying sampling methods, performance improved, with AUROC values reaching 0.893 and 0.897, and F1 scores improving to 0.610 (up-sampling) and 0.626 (down-sampling).

Only 3 studies have focused on the prediction of an individual insemination outcome. Fenlon et al. (2017a) only focused on the prediction of their outcome only in heifers, hence leaving room for further research on a similar outcome for the rest of the herd. Hempstalk et al. (2015) did focus on the entirety of the

herd with a total of 7 herds included, using herd and cow level variables and producing AUROC as the evaluating metric. Similarly, Shahinfar et al. (2014) utilised cow level data from 26 dairy farms and reported the AUROC and accuracy with moderate results. Our study further builds upon these two papers, as it draws data from 133 herds making it the largest within its scope. Furthermore, it incorporates the reporting of a wide variety of metrics, with a special focus on Kappa for the evaluation of its models, which is a more appropriate metric compared to accuracy and AUROC in the case of an imbalanced dataset. Finally, our study focuses on multiple outcomes both binary and continuous, such as the insemination success, the mean percentage of insemination success per herd per month, the mean DIM at conception, which has not been studied prior, and it also includes inferential models, providing a more well-rounded approach in regards to reproduction.

Therefore, the aim of this chapter is to provide predictive modelling for the insemination success, the mean percentage of insemination success per herd per month and the mean DIM at conception, as well as inferential models for the insemination success and time to pregnancy.

5.2 Methods

5.2.1 Data preparation

Lactations beginning with calving events between October 2016 and October 2018 were considered eligible for inclusion in this study; calving events earlier than this window would not have had any related scoring data and the end date was set to ensure that outcomes were known for almost all inseminations. For each insemination recorded in these lactations, the outcome was determined using the following set of rules. An insemination was considered successful if:

- 1) if there was a positive pregnancy diagnosis before the next insemination, *or*
- 2) there was no positive pregnancy diagnosis recorded in the lactation, and this insemination was the one closest to day 282 before the next calving

An insemination was categorized as unsuccessful if:

- 1) there was a next serve before a positive pregnancy diagnosis, *or*
- 2) there was a negative pregnancy diagnosis before the next serve, *or*
- 3) the cow was culled more than 90 days after the insemination, with no fertility events recorded prior to culling, *or*
- 4) no positive pregnancy diagnosis was recorded in the lactation, and the subsequent calving date is outside of a range of 282 +/- 25 days from the insemination date (i.e. subsequent calving is unlikely to relate to this insemination)

In every other scenario the outcome was defined as unknown.

Inseminations between 25 and 100 DIM were included, on the basis that these were most likely to be influenced by events around transition. Pen level data were not included at this part of the analysis as the relevant information was missing for over 50% of the cows.

Two separate datasets were then created. The first one was using a binary outcome of insemination success with the unit of data being each insemination (Dataset X), while the second one was using the continuous outcome of days from calving to successful insemination with the unit of data being each lactation. (Datasets W and Y). All variables present in each dataset are shown in table 5.1.

Table 5.1 Potential Predictive Variables for Datasets W, X and Y

Categorical	Continuous
Milk Fever	305 Milk Yield
RFM	Corrected Protein % in milk
LDA	Corrected Butterfat % in milk
Metritis	
Calf Mortality	
Twinning	
Rumen Fill pre-partum	
Rumen Fill post-partum	
BCS pre partum	
BCS change	
Lactation number	

A third dataset was created by grouping the insemination level dataset by herd-month (Dataset Z). That allowed the aggregation of several variables, displayed in table 5.2 All the aforementioned variables were averaged to either what percentage of animals inseminated per herd per month per year had the relevant characteristic (for binary variables) or what was the mean value for all cows per herd per month per year (for continuous variables).

Table 5.2 Dataset Z Potential Predictive Variables

% of value for inseminated cows per herd/month	Mean of value for inseminated cows per herd/month
Milk Fever	Rumen Fill pre-partum
RFM	Rumen Fill post-partum
LDA	BCS pre-partum
Metritis	BCS change
Calf Mortality	Lactation number
Twinning	305 Milk Yield
	Corrected Protein % in milk
	Corrected Butterfat % in milk

Similarly for the outcome a percentage of successful inseminations of all cows per herd-month was calculated, making the units of data of the dataset the herd-month with the outcome being the proportion of inseminations in each herd-month that were successful. Data points with less than 10 insemination events were excluded from the dataset.

5.2.2 Analysis

Analysis for a total of 5 studies was conducted for the insemination outcomes.

The studies used one of the 3 aforementioned datasets each, as demonstrated on table 5.3.

Table 5.3 Reproductive outcome Studies and Datasets used in each one

Study	Dataset (Outcome)	Type
A	X (Insemination Success)	Predictive
B	Z (% of insemination success per herd/month)	Predictive
C	W (DIM at conception)	Predictive
D	X (Insemination Success)	Inferential
E	Y (Time to insemination success)	Inferential

The predictive algorithms used for study A included the ones mentioned in previous chapters for binary outcomes: logistic regression, decision trees, random forest, ANNs and SVM. A 10-fold cross-validation was used, and assessment of model performance primarily based on the kappa metric. Furthermore, sampling methods and most specifically up-sampling was implemented to assess any improvement in model performance in the case of

an imbalanced dataset. Further details on how and whether up-sampling affected our results is presented in chapter 5.3.1.1, while more information on model building, having binary outcomes, is presented in chapter 2.1.3.4.1.

For study B, having a continuous outcome, the modelling methods used were linear regression, decision trees, random forest, ANNs and MARS. A 10-fold cross-validation was used again, with R^2 being the primary metric used for assessment.

Study C also used continuous outcome (DIM at the time of conception). The modelling methods were the same as in study B: linear regression, decision trees, random forest, ANNs and MARS with a 10-fold cross validation and R^2 as the main metric. The dataset W used was the same as dataset X used in study A with the main difference being that only one data point per lactation was kept. The data point kept in each lactation the successful insemination, if one occurred for that lactation. If no successful inseminations took place the data point was removed from the analysis. Successful inseminations occurring after more than 300 DIM were also removed. More information on building models, having continuous outcomes, is presented in chapter 2.1.3.4.2.

All predictive models, both binary and continuous were fit using the caret package in R (Kuhn, 2005) with a 10-fold cross validation. After acquiring the metrics of all methods for each outcome, they were compared to each other (per outcome) and the best performing model with the best kappa (binary) or R^2 (continuous) was determined.

The last part of the analysis was inferential, rather than predictive analysis and was included in order to assess the differences between predictive and

inferential models and to evaluate potential indicators of effect size and direction.

Study D was similar to study A in terms of the dataset and the outcome used (insemination success, dataset X), however instead of predictive modelling, inferential modelling was applied. A generalized mixed effects model was built using backwards elimination. The random effect added in the model was the farm ID.

Study E was a survival analysis to determine the “risk of insemination success” and conducted using the “survival” package in R (Therneau, 2022). A Kaplan-Meier estimator and a Cox Proportional Hazards model were built. The outcome used was similar to that of study C. The proportional hazards assumption was tested using a Schoenfeld test. This approach to this outcome is looking at a bigger picture in terms of biology and management, as time to conception would also depend on the cow being detected to be in heat and inseminated, whereas on the other studies we just focused on the success of a certain insemination after it was already performed.

The dataset (Y) was very similar to that of study C (dataset W) and was derived from dataset A. Only one insemination was kept per lactation and that was either the successful insemination (if one occurred during said lactation) or the last insemination (if no successful inseminations took place in that lactation). Again, inseminations (either successful or unsuccessful) that occurred later than 300 DIM were removed.

Backwards elimination was implemented to remove unnecessary variables from the model. Any variable that produced a p-value larger than 0.05 was

excluded. The `aareg()` function from the “survival” package (Therneau, 2022) was then used to visualize the change of the different covariates over time.

5.3 Results

5.3.1 Predictive models

5.3.1.1 Study A

The final dataset consisted of 9239 data points, each point being an insemination. The total number of unsuccessful inseminations was 5996 (64.9%), while the successful ones were 3088 (33.4%). A total of 155 data points (1.7%) had no assigned outcome and were excluded from the analysis.

The potentially predictive variables that were considered in the model after forward selection were metritis, milk fever, predicted milk yield, corrected protein percentage in milk, residual milk yield, service number, hock hygiene both pre and post calving, lameness both pre and post calving, rumen fill both pre and post-partum, twinning, lactation number, left displaced abomasum, calf mortality, retained foetal membranes and calving month.

After producing models both with and without the use of sampling methods, it was determined that up-sampling improved their performance as the numerical value of kappa throughout all methods increased. In fact, before up-sampling the kappa value for logistic regression, decision tree, ANN, SVM and Naïve Bayes was 0. A comparison of kappa values on the test set before and after up-sampling is presented on table 5.4. More detailed information on the metrics of all models before up-sampling, both on the train and the test set, is presented on table 5.5. In addition to kappa, sensitivity and detection rate was 0 for logistic regression, decision tree, ANN, SVM and Naïve Bayes, while specificity was 1

and PPV as well as F1 could not be computed, indicating that the imbalance of the dataset affected the computations in a similar way as in the models presented in chapter 4.3.1.1.

Table 5.4 Comparison of Kappa values on the test set before and after the implementation of up-sampling.

Method	Kappa before up-sampling	Kappa after up-sampling
Logistic regression	0	0.059
Decision Tree	0	0.054
Random Forest	0.176	0.174
ANN	0	0.054
SVM	0	0.068
Naïve Bayes	0	0.038
KNN	0.061	0.052

Table 5.5 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Insemination Success outcomes, before up-sampling

<u>Insemination Success</u>							
<u>Training Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.650	0.650	0.621	0.650	0.650	0.650	0.645
Kappa	0	0	0.179	0	0	0	0.078
Sensitivity	0	0	0.404	0	0	0	0.289
Specificity	1	1	0.729	1	1	1	0.783
PPV	-	-	0.392	-	-	-	0.408
NPV	0.650	0.650	0.783	0.650	0.650	0.650	0.680
AUROC	0.549	0.5	0.591	0.5	0.5	0.5	0.550
Detection	0	0	0.161	0	0	0	0.098
Rate							
Balanced	0.5	0.5	0.550	0.5	0.5	0.5	0.536
Accuracy							
F1	-	-	0.458	-	-	-	0.338
<u>Test Set</u>							

	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.652	0.652	0.713	0.652	0.652	0.652	0.638
Kappa	0	0	0.176	0	0	0	0.061
Sensitivity	0	0	0.389	0	0	0	0.218
Specificity	1	1	0.701	1	1	1	0.799
PPV	-	-	0.327	-	-	-	0.392
NPV	0.652	0.652	0.800	0.652	0.652	0.652	0.701
AUROC	0.541	0.5	0.572	0.5	0.5	0.5	0.543
Detection	0	0	0.158	0	0	0	0.087
Rate							
Balanced	0.5	0.5	0.543	0.5	0.5	0.5	0.530
Accuracy							
F1	-	-	0.437	-	-	-	0.303

After up-sampling, the metrics and the kappa value in particular, showed at least numerical improvements. On the training set the accuracy of all models ranged from 43% (Naive Bayes) to 62% (random forest), Kappa from 0.04 (Naive Bayes) to 0.18 (random forest), balanced accuracy from 0.52 (Naive Bayes) to 0.59 (random forest) and AUROC from 0.542 (decision tree) to 0.647 (random forest). Both sensitivity and specificity had a wide range with the former being from 0.415 (random forest) to 0.804 (Naive Bayes) and the latter

being from 0.243 (Naïve Bayes) to 0.703 (random forest). PPV and NPV were overall similar for all models with the best ones being 0.419 (random forest) and 0.752 (random forest) respectively and the worst being 0.699 (logistic regression) and 0.752 (random forest) respectively. Detection rate and F1 were low for all models, the highest ones being 0.273 and 0.495 respectively, both for the Naïve Bayes.

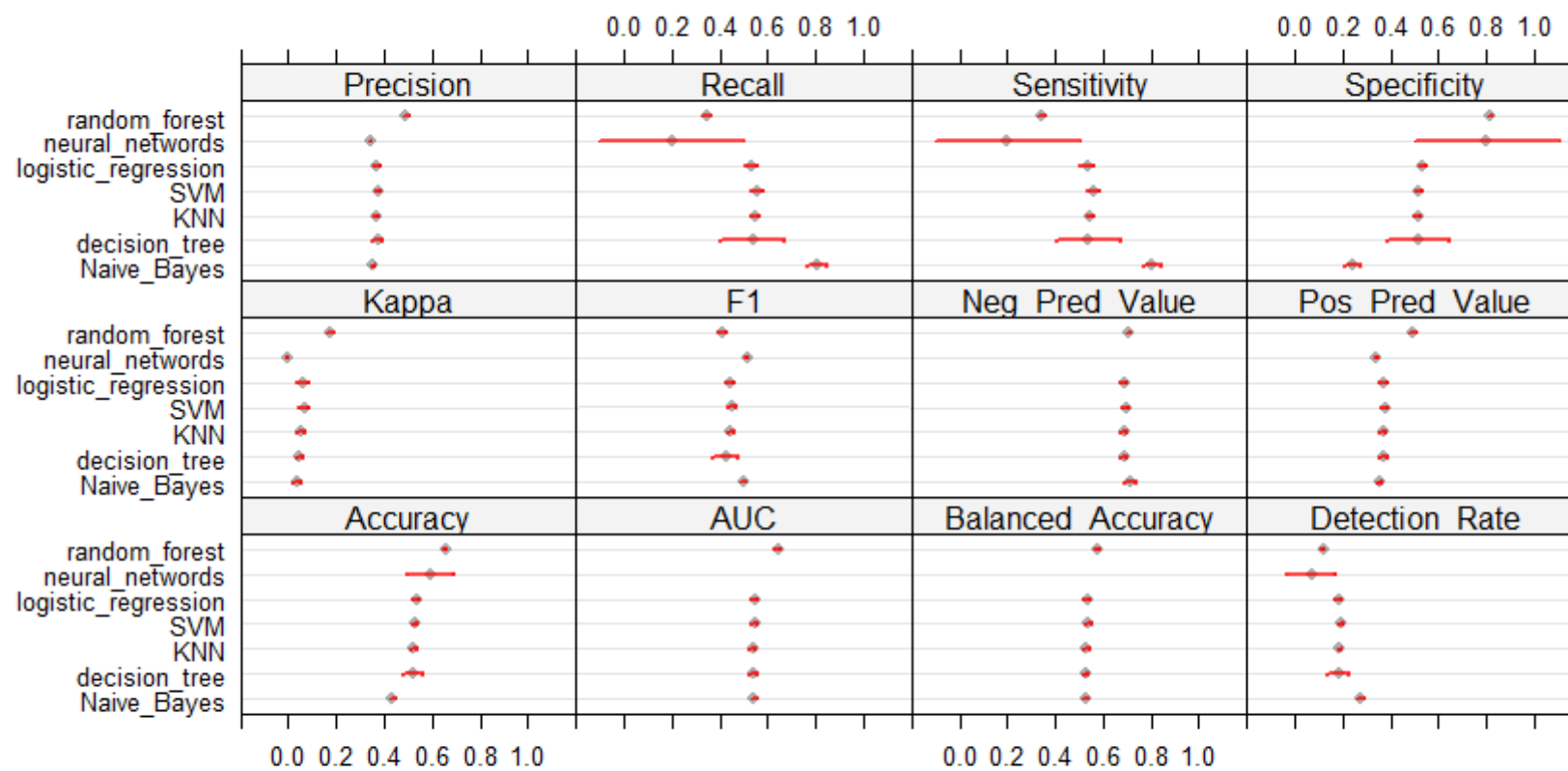
Results on the test set were very similar with kappa being low ranging from 0.038 (Naïve Bayes) to 0.174 (random forest), accuracy ranging from 0.436 (Naïve Bayes) to 0.621 (random forest) and AUROC ranging from 0.542 (decision tree) to 0.647 (random forest). Sensitivity as well as specificity had a wide range, from 0.493 (random forest) to 0.805 (Naïve Bayes) for the former and 0.244 (Naïve Bayes) to 0.687 (random forest) for the latter. PPV was consistently low with the highest at 0.419 for the random forest, while NPV had moderate to high values with the lowest one at 0.686 for the KNN. Detection rate, balanced accuracy and F1 were all consistently low, with the highest values being 0.200 (decision tree), 0.590 (random forest) and 0.493 (Naïve Bayes) respectively. Detailed information on the metrics of all models both on the training and the test set are presented on table 5.6. The metrics of all models on the test set are also graphically shown in Figure 5.1.

Table 5.6 All metrics of all machine learning models, as calculated on both the training and the test sets, predicting Insemination Success outcomes, after up-sampling

<u>Insemination Success</u>							
<u>Training Set</u>							
	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.547	0.514	0.621	0.515	0.531	0.436	0.523
Kappa	0.063	0.059	0.183	0.060	0.071	0.043	0.058
Sensitivity	0.512	0.570	0.415	0.561	0.551	0.804	0.539
Specificity	0.543	0.489	0.703	0.501	0.530	0.243	0.526
PPV	0.329	0.349	0.419	0.351	0.368	0.349	0.356
NPV	0.699	0.714	0.752	0.708	0.725	0.720	0.710
AUROC	0.550	0.542	0.647	0.559	0.555	0.543	0.544
Detection	0.192	0.206	0.172	0.203	0.194	0.273	0.200
Rate							
Balanced	0.540	0.543	0.594	0.537	0.551	0.529	0.535
Accuracy							
F1	0.441	0.454	0.477	0.453	0.459	0.495	0.447
<u>Test Set</u>							

	Logistic	Decision	Random	ANN	SVM	Naïve	KNN
	Regression	Tree	Forest			Bayes	
Accuracy	0.533	0.514	0.621	0.515	0.531	0.436	0.523
Kappa	0.059	0.056	0.174	0.054	0.068	0.038	0,052
Sensitivity	0.532	0.587	0.493	0.578	0.558	0.805	0.547
Specificity	0.533	0.477	0.687	0.482	0.517	0.244	0.511
PPV	0.371	0.367	0.449	0.367	0.374	0.355	0.367
NPV	0.688	0.694	0.724	0.689	0.694	0.712	0.686
AUROC	0.545	0.539	0.630	0.542	0.547	0.544	0.539
Detection	0.181	0.200	0.168	0.197	0.190	0.274	0.186
Rate							
Balanced	0.533	0.532	0.590	0.530	0.538	0.525	0.529
Accuracy							
F1	0.437	0.448	0.470	0.447	0.448	0.493	0.439

Figure 5.1 Comparison of all metrics for all different methods predicting insemination success after upsampling, on the test set.

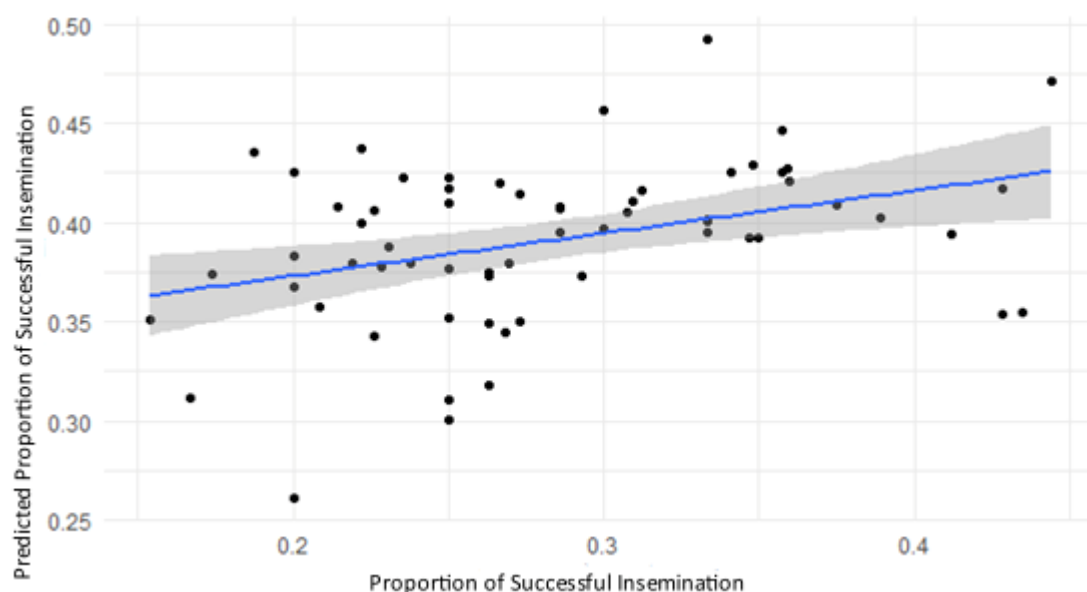


Confidence Level: 0.95

Since from the results above the random forest was the best performing model, it was further investigated on whether it could make predictions on a herd-month level, instead of an individual lactation level.

A total of 685 herd-month groups were identified, with 20% of those (137 groups) being separated as a test dataset. After applying the exclusion filter in order to only include groups with insemination success within 15% and 45% the test set was reduced to 64 groups. When testing the predictions, the difference in the probability difference in each group ranged from -8.0% (less likely to predict a positive outcome compared to the actual percentage of positive outcomes) to 24.2% (more likely to predict a positive outcome compared to the actual percentage of positive outcomes). Variable importance indicated predicted 305 milk yield as the most valuable predictor, with residual daily milk yield and corrected protein percentage following it closely.

Figure 5.2 Scatterplot of actual insemination success per group vs the predicted insemination success per group



The R^2 describing the relationship shown in the scatterplot (Figure 5.2) was found to be 36.5%, meaning that the averaged predictions explain over a third of the variation of the averaged insemination success per group.

5.3.1.2 Study B

The initial size of the dataset after the aggregation of the variables as well as the outcome was 268 herd/quarter-years. 42 farms were included in the analysis in total. The missing data of all potentially predictive variables are listed at table 5.7.

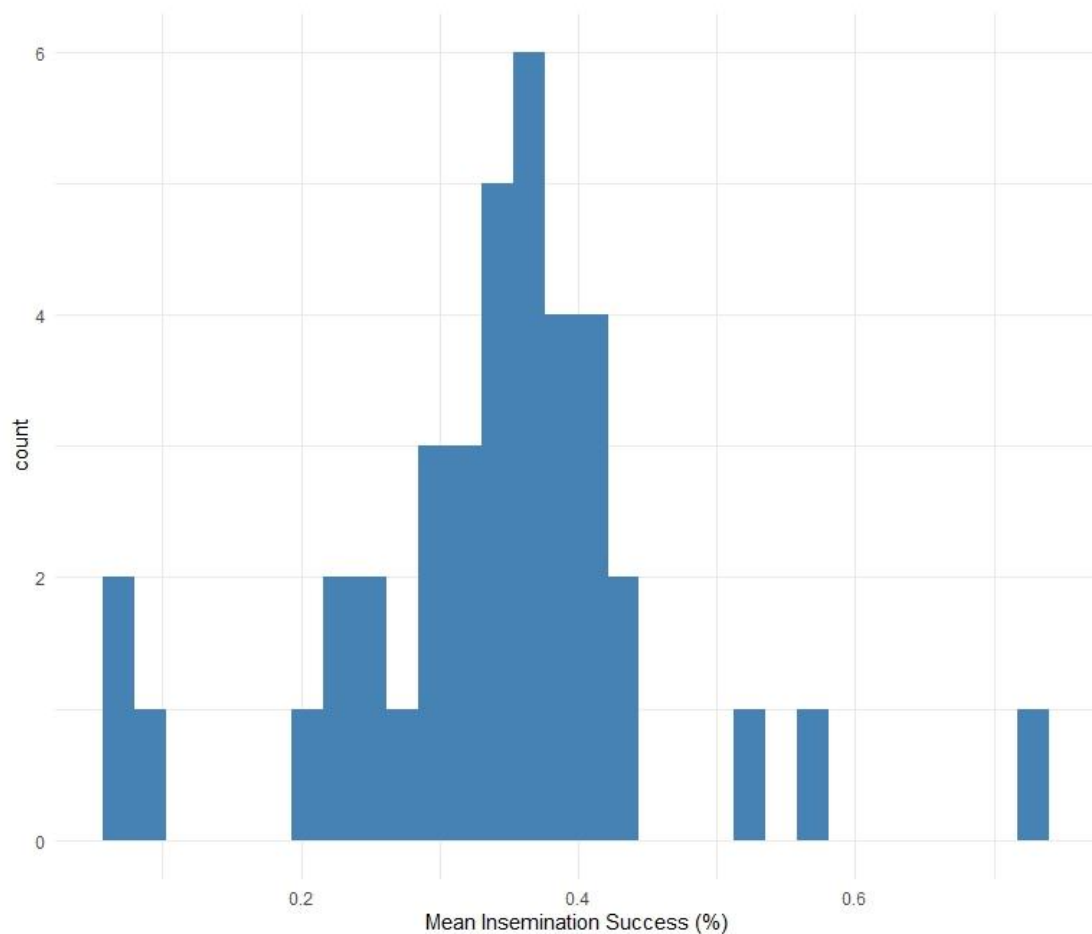
Table 5.7 Aggregated variables available for analysis, along with missing data.

Variable	Missing Data (%)
Metritis percentage per herd/month/year	8 (1.2)
Milk Fever percentage per herd/month/year	8 (1.2)
Twinning percentage per herd/month/year	8 (1.2)
Calf Mortality percentage per herd/month/year	8 (1.2)
Service number average	0 (0.00)
Mean rumen fill pre calving	0 (0.00)
Mean BCS	0 (0.00)
Mean BCS change	0 (0.00)
Mean Lactation Number	0 (0.00)
Mean predicted Milk Yield	36 (5.4)
Mean residual Milk Yield	36 (5.4)
Mean corrected Protein percentage	32 (4.8)

Mean corrected Butterfat percentage	32 (4.8)
Mean Butterfat Protein ratio	32 (4.8)

The mean percentage of insemination success per herd per month was at 34.3%, with the least successful months having as low as 6.7% insemination success rate and the most successful ones having as high as a 72.7% rate (Figure 5.3).

Figure 5.3 Insemination success percentage per herd/month group distribution



The results of all models are shown in table 5.6. On the training set the R^2 values ranged low with the highest being the MARS model at 0.252. On the test set the RMSE remained at below 12.4% except for the MARS model which had a much higher RMSE at 34.5%. Similarly, MAE was below 10.1% for all models except the MARS model with an MAE of 33.0%. After applying the models on the test sets the results were comparable. all R^2 values were found to be under 25%, with the highest performing model (MARS) being at 0.235. RMSE and MAE were below 13.6% and 10.8% respectively, with the exception of the MARS model which produced an RMSE and MAE of 38.1% and 35.8% respectively. All metrics are shown in table 5.8.

Table 5.8 Metrics of all models predicting insemination success percentage per herd/month group on both the training and the test sets.

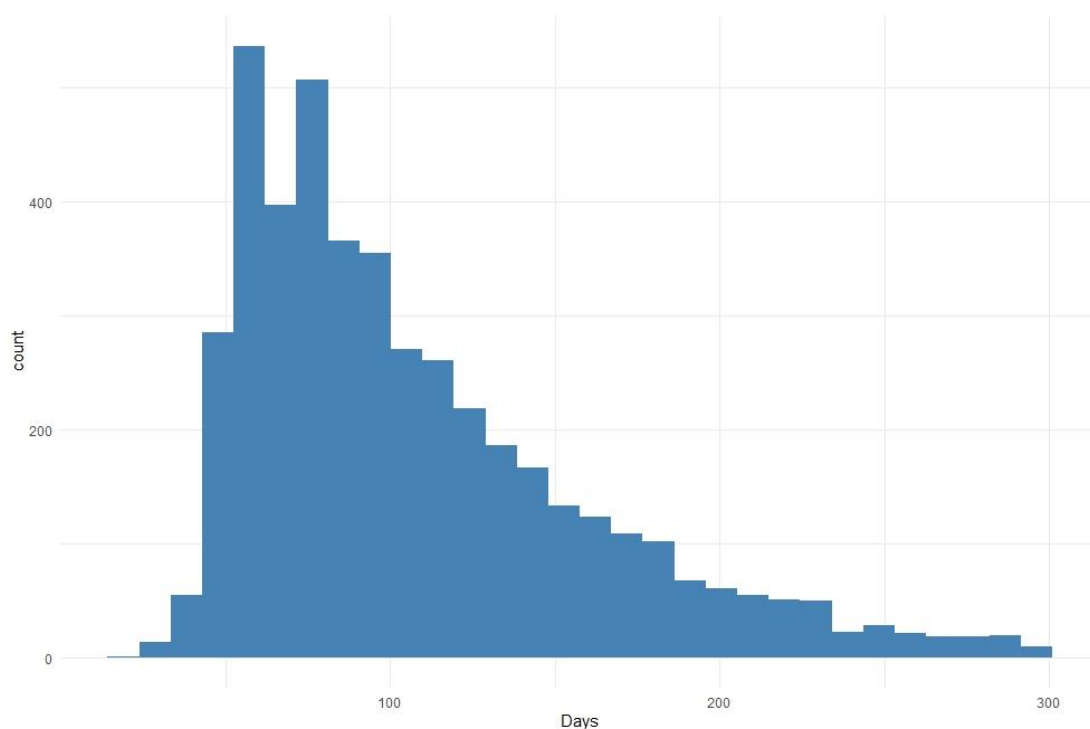
Training Set				
		RMSE	R ²	MAE
Linear Regression		0.124	0.082	0.101
Decision Tree		0.119	0.149	0.089
Random Forest		0.122	0.240	0.090
Artificial Neural Networks		0.345	0.219	0.330
MARS		0.103	0.252	0.092
Test Set				
		RMSE	R ²	MAE
Linear Regression		0.136	0.074	0.108
Decision Tree		0.124	0.133	0.097
Random Forest		0.129	0.223	0.097
Artificial Neural Networks		0.381	0.204	0.358
MARS		0.126	0.235	0.100

5.3.1.3 Study C

The final dataset after keeping only successful inseminations was comprised of 4,500 lactations. The potentially predictive variables that were considered in the model after forward selection were the same as mentioned in study A, with the final variables included in the model being metritis, LDA, RFM, milk fever, corrected percentage of protein in milk, the lactation number, BCS post-partum, rumen fill again both pre- and post-partum and finally the calving month.

The mean DIM at conception was 107.4, with a median of 93, a minimum of 23 and a maximum of 300 (DIM of 301 or larger had been removed during preparation of the dataset) (Figure 5.4)

Figure 5.4 Histogram of DIM at the time of conception



On the training set, the R^2 values were consistently very low, with the highest being 3.0% for the linear regression. RMSE and MAE values were high with the former ranging from 51.0 DIM (random forest) to 116.7 DIM (ANN) and the latter ranging from 39.8 DIM (random forest) to 104.2 (ANN). The R^2 values of all models on the test set were also very low, ranging from 0.1% for the ANN model to 2.4% for the linear regression model. RMSE values were correspondingly high and ranged from 51.8 to 118.7 DIM for the random forest and ANN models respectively. Similarly, MAE values ranged from a minimum of 40.8 DIM for the random forest to a maximum of 106.5 DIM for the ANN. It should be noted that the ANN model was an outlier especially when it came to RMSE and MAE values – other models had RMSE values of around 51.8-52.3 DIM while MAE ranged between 40.8 and 41.3 DIM. All metrics on both training and test sets are shown on table 5.9.

Table 5.9 Metric of all models predicting DIM at conception for both the training and the test set.

Training Set				
		RMSE	R ²	MAE
Linear Regression		51.5	0.030	40.2
Decision Tree		51.1	0.014	40.3
Random Forest		51.0	0.024	39.8
Artificial Neural Networks		116.7	0.005	104.2
MARS		51.6	0.026	40.1
Test Set				
		RMSE	R ²	MAE
Linear Regression		52.3	0.024	41.1
Decision Tree		52.0	0.010	41.0
Random Forest		51.8	0.014	40.8
Artificial Neural Networks		118.7	0.001	106.5
MARS		52.2	0.021	41.3

5.3.2 Inferential Models

5.2.3.1 Study D

The explanatory variables included in the final generalized mixed effects model after backwards elimination were lactation number, calving month and calf mortality.

More specifically, when compared to heifers, cows had decreased odds of insemination success (with odds ratio decreasing with each subsequent lactation), cows that had experienced calf mortality in that same lactation also had decreased odds. The only statistically significant difference when it came to calving month, using January as a baseline, was September having increased odds of insemination success. The results are shown in detail on table 5.10. The Hosmer-Lemeshow goodness of fit test had a p-value of 0.16, indicating no statistical evidence for rejecting the H_0 , meaning that our data appear to be matching the model.

Table 5.10 Odds Ratios with 95% CI for mixed effects logistic regression model of insemination success

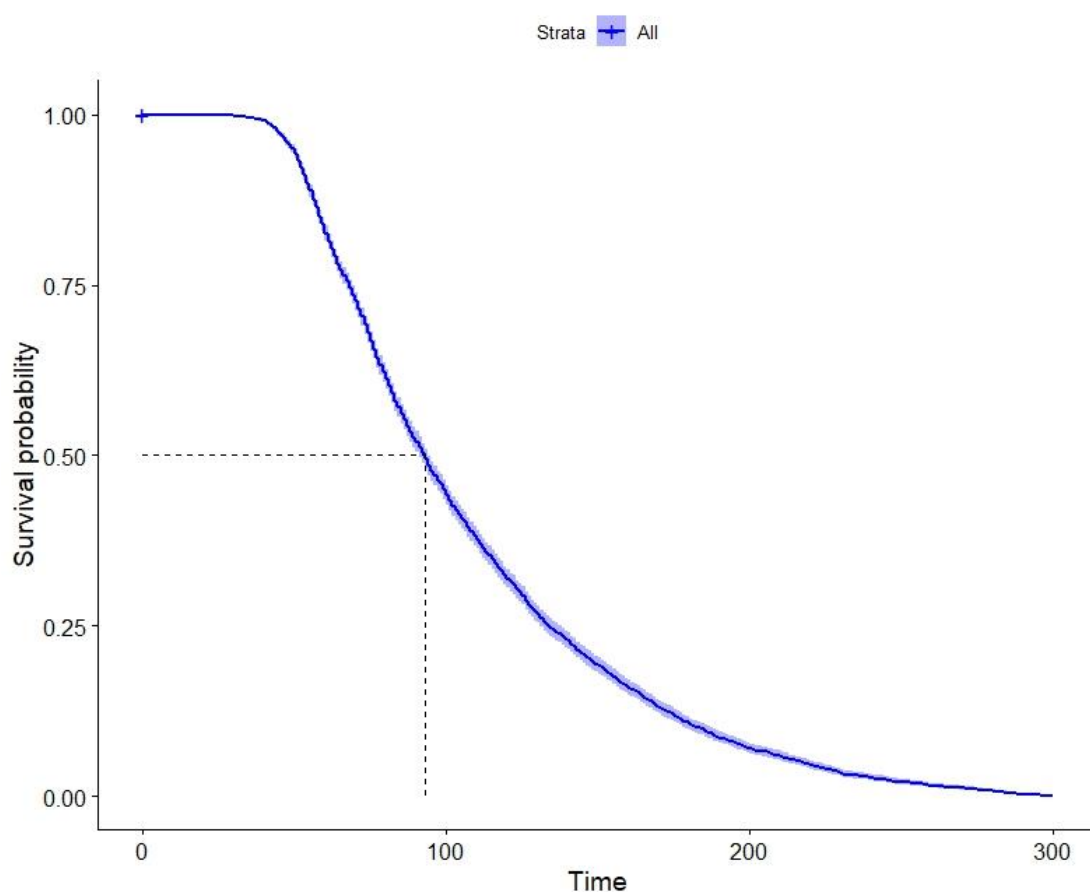
Variable	OR	95% CI	p-value
Lactation 1	Baseline		
Lactation 2	0.810	(0.714, 0.920)	0.001
Lactation 3	0.736	(0.643, 0.842)	>0.001
Lactation 4	0.601	(0.516, 0.701)	>0.001
Lactation >5	0.491	(0.418, 0.577)	>0.001
(Calving month)	Baseline		
January			
February	1.008	(0.804, 1.263)	0.946
March	1.151	(0.924, 1.433)	0.209
April	1.095	(0.871, 1.377)	0.435
May	1.071	(0.850, 1.349)	0.562
June	1.111	(0.889, 1.389)	0.353
July	1.079	(0.872, 1.335)	0.486
August	1.030	(0.829, 1.280)	0.789
September	1.455	(1.165, 1.817)	<0.001
October	1.056	(0.844, 1.322)	0.631
November	0.991	(0.791, 1.242)	0.941

December	1.175	(0.949, 1.455)	0.139
CalfMortality	0.730	(0.546, 0.977)	0.034

5.2.3.2 Study E

The graph below (Figure 5.5) shows the survival probability of a cow becoming pregnant.

Figure 5.5 Kaplan-Meier survival curve of time to pregnancy in cows



The number of observations was 5,516 with a total of 4,500 events. The variables that were included in the final Cox Proportional Hazards model were the lactation number, the calving month, whether the cow was diagnosed with

metritis at that lactation, and the corrected average protein percentage in milk.
(Table 5.11).

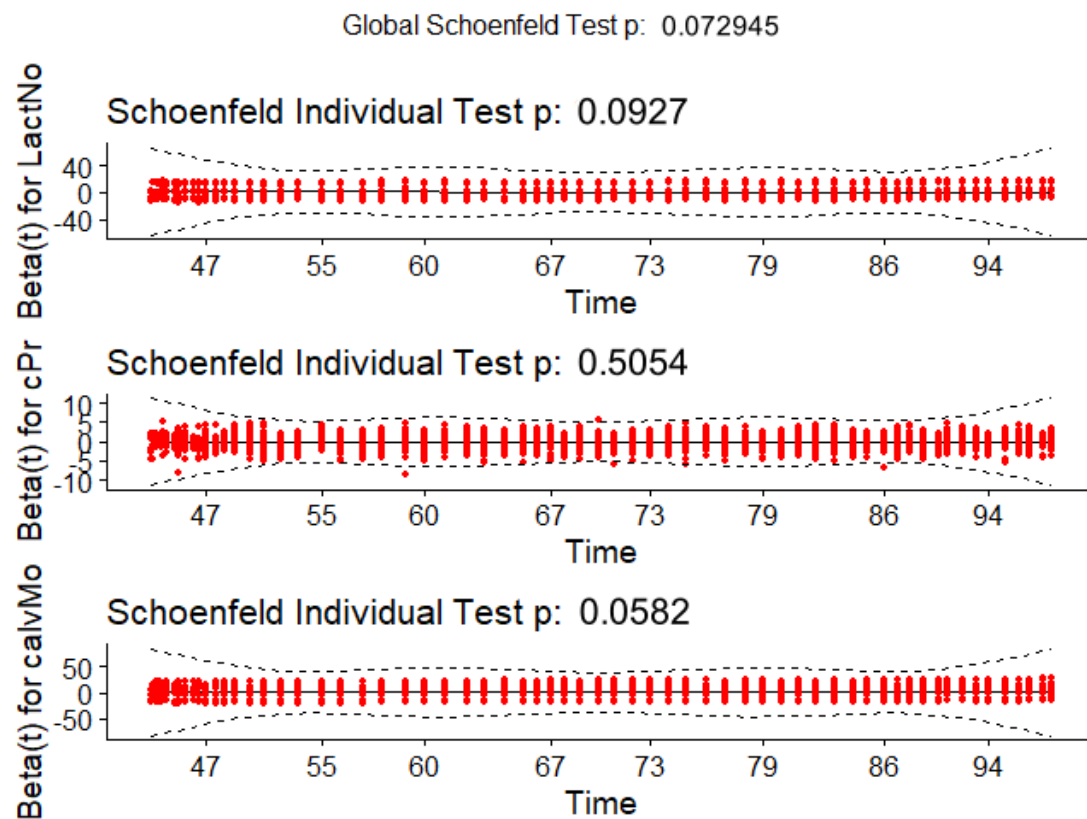
Table 5.11 Hazards Ratios from Cox Proportional Hazards model with outcome time to pregnancy in cows conceiving at <100 DIM.

Variable	Hazard Ratio	95% Confidence Interval	p-value
Metritis	0.84	(0.70, 0.97)	0.01
Corrected Protein %	0.88	(0.84, 0.91)	<0.001
Lactation No 1	(Baseline)		
Lactation No 2	0.90	(0.82, 0.97)	0.01
Lactation No 3	0.87	(0.79, 0.94)	0.002
Lactation No 4	0.86	(0.76, 0.95)	<.003
Lactation >4	0.77	(0.67, 0.86)	<0.001
(Calving Month)	(Baseline)		
January			
February	0.86	(0.68, 0.95)	0.05
March	1.08	(0.90, 1.26)	0.56
April	1.09	(0.91, 1.27)	0.90
May	0.97	(0.77, 1.17)	0.68
June	0.94	(0.76, 1.12)	0.55
July	1.02	(0.84, 1.20)	0.18

August	0.92	(0.74, 1.10)	0.62
September	1.28	(1.14, 1.41)	<0.001
October	1.04	(0.86, 1.22)	0.12
November	0.97	(0.79, 1.15)	0.41
December	1.09	(0.91, 1.27)	0.90

The proportional hazards assumption was tested using a Schoenfeld test, producing a global Schoenfeld test p-value of 0.072945, and individual p-values of 0.0927, 0.5054 and 0.0582 for lactation number, corrected protein percentage and calving month respectively. All values being above the 0.05 cutoff as well as examination of the Schoenfeld residuals graph that highlights that the residuals do not change over time (Figure 5.6) supports the belief that the proportional hazards assumptions is not violated.

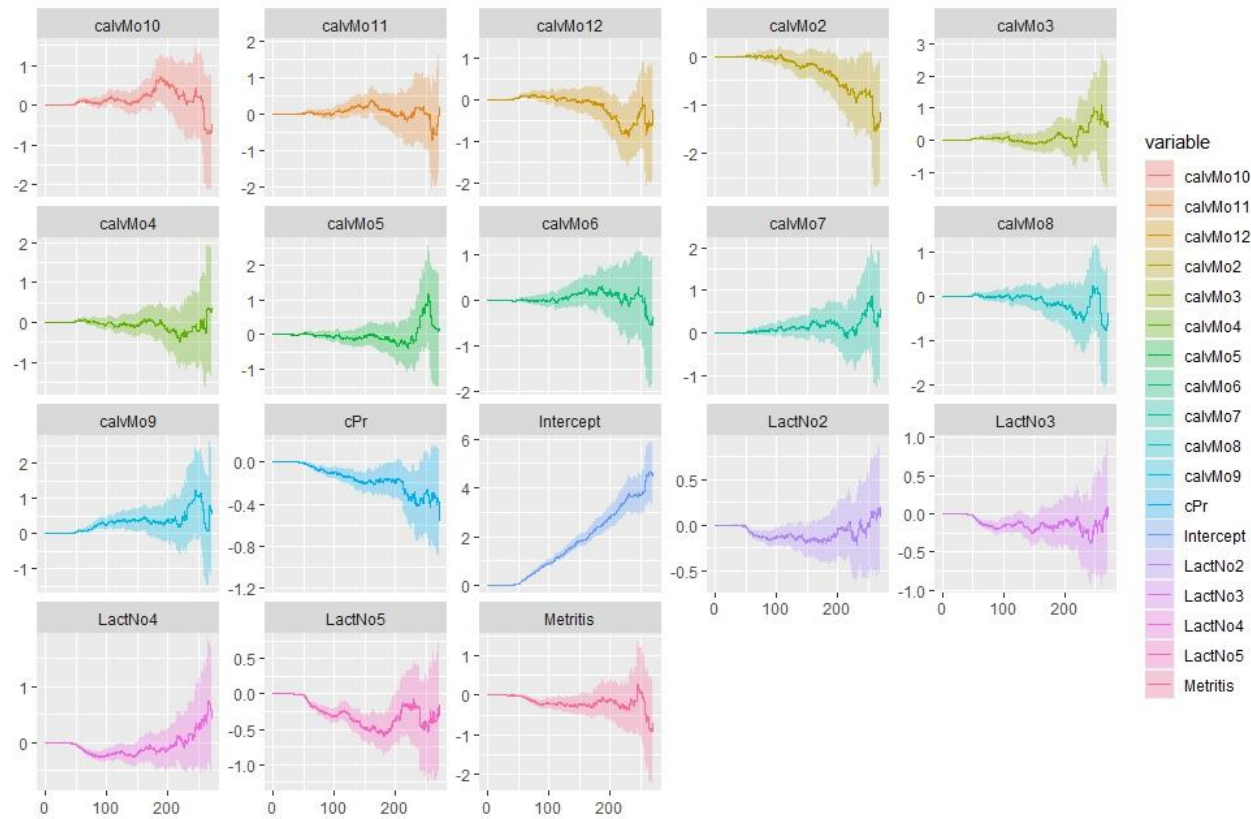
Figure 5.6 Global Schoenfeld test and individual Schoenfeld tests for each independent variable included in the Cox Proportional Hazards Model with outcome time to pregnancy in cows conceiving at <100 DIM.



Further analysis was conducted to look into the slopes of the variables included in the model, to challenge the assumption that they stay constant and see if and how they change throughout time. The slopes and their behaviour over time are graphically depicted in Figure 5.7.

All slopes seem relatively stable, apparently only diverging towards the end of the time period, which is to be expected to a degree since as more cows get pregnant we have a smaller sample size and therefore wider confidence intervals. The most noteworthy changes are in lactation 5 where it seems that if a cow has not become pregnant by day 50 then the chances are starting to decrease greatly.

Figure 5.7 Slopes of Hazards Ratios over time



5.4 Discussion

For study A (predictive models with the outcome of conception for a given insemination) it is evident that the models do not produce adequate predictive value, with the Kappa was consistently low, again with the highest value appearing in the random forest with 0.18. The sensitivity and specificity balance each other, as in models when one increased the other appeared to decrease dramatically. One such example is the random forest model with the highest sensitivity of 82% and the lowest specificity of 27%. The Naïve Bayes model on the other hand behaves differently achieving the highest specificity of 74% and the lowest sensitivity of 35%. The rest of the models follow a similar pattern with the logistic regression and SVM models trying to fit both sensitivity and specificity closer to 50%. The PPV of all models approximated 70% with the NPV at around 40%.

The possibility of a predictive tool that can be used on farm to assess the result of a given insemination has been tackled by research in the previous years. Two papers had already investigated the possibility of predicting insemination success, similarly to Study A (Hempstalk et al., 2015, Shahinfar et al., 2014). However, both these studies lack reporting of some informative metrics, such as the Kappa, balanced accuracy or even specificity and sensitivity, and only present AUROC as a measure of evaluation for their models. AUROC may give an overly optimistic evaluation of a model's performance on imbalanced datasets, as it can be skewed by the model's ability to classify the majority class while overlooking the minority class (Lobo et al., 2008, King et al., 2021, Hancock et al., 2023, Bednarski et al., 2022).

After using a variety of phenotypic and genotypic variables, Shahinfar et al. (2014) produced a random forest model indicating the most predictive variables to be mean conception rate per herd in the past 3 months, the herd-year-month group the insemination took place in, DIM at insemination, times the cow had been inseminated for the current lactation and the stage of lactation at the time of insemination. They also reported a decision tree model, using the C4.5 methods as opposed to the C5.0 used for our project, which identified incidence of ketosis, mastitis, RFM, lameness for primiparous cows, and LDA, mastitis and RFM for multiparous cows. These results seem to align for the greater part with ours, having found that the variables that added the most predictive value in the model were DIM during lactation, incidence of metritis and milk fever, lactation number, service number, and various milk variables. The AUROC reported for the random forest model, which was the highest performing algorithms, ranged from 72.3% and 75.6%, while for the rest of the methods, which included decision trees, Naïve Bayes, Bootstrap Aggregation and Bayesian Network, ranged between 60.0% to 68.0%. This pattern was similar to our results where the random forest model had an overall better performance compared with the rest of the algorithms, which was reflected in both accuracy and AUROC. However, as Sharinfar et al. (2014) reported this trend could be attributed to the random forest method overfitting the data, rather than it producing the most predictive algorithms. Their overall performance in regards of AUROC was higher, but the difference was not overwhelming and it should be noted that they used variables that we did not utilise, mainly the herd-month-year which effectively adds a random effect for herd-month that cannot be utilised in a farm setting to make predictions.

The second study with a similar approach (Hempstalk et al. 2015) and conception rate of 47.3%, also reported AUROC values, the highest one being generated by a logistic regression model at 66.5% and the lowest by a C4.5 decision tree at 49.2%. They also used an independent dataset for further external validation in order to more accurately assess the generalizability of their models and, again, the best performing one was found to be the logistic regression with an AUROC value of 66.5% and standard deviation of 2.5%. The rest of the algorithms included C4.5 decision trees, Naïve Bayes, Bayes Network, SVM, Partial Least Squares, random forest and rotation forest with their AUROC after external validation being between 52.1% and 65.7%. The predictive variables used included information on the lactation number, the number of insemination and DIM when it occurred in the current and the previous lactation, breed, milk production, energy balance, BCS and its changes as well as the day of the week and month of the year. Once again, the results looked fairly similar to our metrics even prior to the external validation. In this case the logistic regression model seemed to outperform the other algorithms, however the AUROC reported was no more than fair for all of them. As mentioned above AUROC was the only metric reported making it difficult to determine if there is actual any predictive value in the models, however in either case the values were rather low to have a clinically significant impact.

When the models of Study A were used to make predictions on aggregated data (study B) all models produced a low R^2 with the MARS model having the highest value of 23.5% and the linear regression the lowest of 7.4%. The reasoning behind aggregating the data here was based on the fact that in many biological situations group-level predictions appear to be easier to make

compared to predictions on the individual level. In this case, across a given group of serves it'd be expected to be able to make reasonable predictions of what percentage would be successful, even using a model that was relatively poor at predicting which of the inseminations would be successful. Eventually, in this study the R^2 improved somewhat, with the value for the best performing increasing from 18.0% to 23.5%. However, as a whole this increase did not appear to meaningfully increase the predictiveness of the models as the variation of the outcome explained continued to be at under a fourth of the total variation.

Study C (where the predictive model was built and tested on data aggregated at herd-month level) had similarly low R^2 values throughout all models. The linear regression was the model that achieved the highest R^2 at just 2.4%. Since in study B it was determined that predictions are not possible using an aggregated, continuous outcome and aggregated variables, the reasoning behind study C was to investigate whether predictions can be made on an individual, continuous outcome without aggregating any variables. However, it is evident that the predictions actually lay far from the actual values rendering the models unsuccessful in making adequate predictions.

Study D was effectively an inferential model with the same outcome and predictors offered as the predictive model in Study A (inseminations as units of data, with conception as the binary outcome). Here, the variables included in the final model were the lactation number, BCS pre- and post-partum, month of previous calving and calf mortality. It is interesting to note that the most important predictive variables included in the model of study A (predicted 305 milk yield, residual daily milk yield and corrected protein percentage in milk)

were not even present in the inferential model. A possible explanation is that since all three of these variables were calculated from the lactation curves that are specific to each herd, their contribution to the predictive model was in fact the herd effect. And since the inferential model had already controlled for herd effect by adding the random effect (herd) in the mixed effects model the milk variables did not have any further contribution.

The survival analysis model (study E) suggested that cows diagnosed with metritis as well as lactation numbers being larger than 1, have a decreased chance of getting pregnant. The calving month also seem to affect the chances with February having decreased chances compared to January, while September having increased chances, again compared to January. Gröhn and Rajala-Schultz (2000) described a similar survival analysis looking into the time to conception using milk variables, parity, disease and calving season as likely variables. Both studies identified lower chances of conception in older cows. They also found lower chances of conception in cows with metritis, retained placenta or ovarian cysts, and while the present study did not find a change in hazard ratio for RFM, there was an evident reduction for cows with metritis. Lastly, they reported a drop in hazard ratio when calving during the spring months (March to May). In this study the drop is apparent in February instead, however if we are taking into consideration how the seasonal temperature may have shifted in the decade that separates the two studies it is possible that they are describing a similar effect. Our study also reported an increase in conception chances when calving in September, which is not reported in Gröhn and Rajala-Schultz (2000), possibly due to the fact that they did not focus on individual months but in seasons.

The failed inseminations in our data compose the 66% of the final dataset as opposed to 34% successful ones, making it somewhat imbalanced though not as much as the dataset used for Chapter 4. Nevertheless, as shown by the Kappa values in contrast with the Accuracy produced the same issue that occurred during the analysis of that dataset reoccured. In fact before using sampling methods, the accuracy of all models lay between 48-63% and Kappa between 0.06 and 0.10, the random forest being the best performing model with its accuracy very closely approaching the incidence of failed inseminations. This is a strong indication that the models are taking advantage of the numerical difference of the two classes to achieve a numerical overall superior accuracy by assigning the label of the majority class in most predictions.

Even after upsampling the majority of the models did not seem to improve in terms of predictiveness. The random forest model, which was already the best performing before upsampling, had the most drastic change in metrics with Kappa rising from 0.10 to 0.18, specificity from 27% to 36%, while the sensitivity, AUROC, balanced accuracy and accuracy remained relatively unchanged. In all models, according to the scale set by Viera and Garrett (2005), Kappa values indicated only slight agreement of predicted and actual values. Furthermore, the balanced accuracy never managed to achieve values much higher than 50%, again reinforcing the suggestion by Brodersen et al. (2010) that balanced accuracy is a metric that more appropriately measures model performance where classes are imbalanced compared to regular accuracy.

As Hempstalk et al. (2015) explained, relatively poor predictive performance in this context is perhaps not surprising considering the very wide variety of

factors that could influence the outcome of an insemination (e.g. individual cow energy balance, semen characteristics and handling, heat stress etc). Interestingly, performance barely improved when the models were used to make predictions on aggregated groups of inseminations. So not only did the models fail to identify which specific inseminations would be specific, but could not even measure the overall impact on success on a herd-quarter-year group. This could point to the fact that there are variables affecting the process which cannot be measured beforehand and therefore cannot be accounted for. These could include herd-season-year group, the capability of the technician, as well as the bull's fertility (Hempstalk et al., 2015). This seems to align with our findings in relation with the inferential models, where the herd/quarter-year group random effect played an important role in explaining the variation of the outcome.

When it comes to the inferential models, the main variables that were found to have a statistically significant association with the outcomes in both the logistic regression and the Cox proportional hazards model were the lactation number, the calving month (and in particular September), BCS pre as well as post partum, predicted milk yield, metritis diagnosis and calf mortality. The reproductive health has been associated with poor insemination success (Shahinfar et al., 2014), which is in accordance with our results for metritis and calf mortality reducing the insemination success odds as well as increasing the time to pregnancy. Similar associations have been found for milk yield and energy balance (Shahinfar et al., 2014) which would explain the association we have found with BCS and predicted milk yield. Lactation number could be associated with numerous other factors, such as cow health which also has a

positive correlation with insemination success (Shahinfar et al.2014). This would explain why primiparous cows, having less health issues, might have a greater chance at getting pregnant. Finally the month could be possibly interpreted as a heat stress proxy, however since only September seemed to differ it seems a bit unlikely.

Our studies had a number of limitations that affected them. Namely, the dataset used was gathered through convenience sampling. While that does not greatly impact the predictive models, since we are not looking for potentially causal associations, however it would mean that in the event of a usable predictive model we would have to assess its generalisability and potentially consider external validation to confirm potential predictiveness. For the inferential models, it similarly means that the results are not necessarily generalisable, at least not to the entirety of dairy cows in the UK. It still has great value to look into the measured and the effects for that specific population of cows as it represents a great deal of the dairy farms a clinician would visit.

All the potential risk factors identified during the inferential analyses were used in the predictive models, so all the statistically significant associations did appear to help with the predictive process somewhat. It is interesting to note that the predictive models did pick up some extra variables, perhaps identifying more subtle or intricate associations, that improved their kappa value. Even still, however, the overall predictiveness of our models remained poor.

It is also interesting to note that the inferential models found a few very strong associations, but while the predictive models included them as well as additional variables they still failed to produce highly accurate predictions. This

is especially important, as before predictive modelling became common practice researchers could potentially draw conclusions about the variation of an outcome based on inferential models instead.

Chapter 6 – Prediction of production Outcomes

6.1 Introduction

Milk production can affect greatly economic losses in dairy cattle, not only by its reduction, but by a decline in protein and lipid composition or a rise in the number of somatic cells (De Amicis et al., 2018). Production normally rises steeply during early lactation, with good health and appropriate feed intake helping ensure a consistent increase (LeBlanc, 2010). It then steadily decreases until the end of the cow's production cycle when the cow is dried off, either at a pre-determined time prior to her next calving, or in some cases where production drops below a threshold set by the herd manager (Martinez Lopez et al., 2019). Evaluating the shape of this "lactation curve" can be a powerful tool in predicting a cow's total milk yield (Martinez Lopez et al., 2019) and assessing the cows' health status (Dudouet, 1982). Functions commonly used to describe the lactation curve are discussed in detail by Martinez Lopez et al. (2019). As summarised, they can be either linear or non-linear and can be all summed up as:

$$Y=\eta(t,\beta)$$

where Y is the milk yield at time t of the lactation and β the unknown parameters of the model that are to be estimated from the data and η is the function that describes their relationship. Among the particular mathematical functions used to calculate the lactation curve include the incomplete gamma (Wood, 1967), the polynomial (Ali and Schaeffer, 1987), the exponential (Wilmink, 1987) and the Legendre polynomial (Kirkpatrick et al., 1994). But the need for more

complex models as the number of available features increased (Murphy et al., 2018) has led researchers to use other methods such as a multivariate regression model using test-day record, month of calving and gene data (Grzesiak et al., 2003), and autoregressive models (Vasconcelos et al., 2004), Macciotta et al., 2002). In addition, models with reduced number of features have been presented for situations where there is a sparsity of data (Græsbøll et al., 2016).

Management factors, including diet, moving cows to a different group and weather conditions, are also shown to influence milk yield variability (LeBlanc, 2010). Frequently, a reduction in milk production precedes the clinical symptoms of a disease (Edwards and Tozer, 2004). Trends in production during early lactation can be used as a source of information for assessing the success of the transition period on the herd-level (Nordlund and Cook, 2004). Milk composition at this stage has been linked with periparturient diseases, through its association with the cow's energy status (Toni et al., 2011). Milk constituents as well as the month of calving have been associated with fertility outcomes, such as the probabilities of the occurrence of pregnancy (Cook and Green, 2016), but other authors have found that milk composition information was unlikely to be usefully predictive for herd conception risk (Hudson and Green, 2018). Lukas et al. (2015) suggested that transition period monitoring using daily milk yield can be valuable for herd managers, allowing them to take action timely and prevent cows from experiencing transition related problems.

Reduced milk production has been linked with disease. In particular, reduced production has been shown in cows with clinical or puerperal metritis (Giuliodori et al., 2012), and 610 kg of lost milk per cow was calculated as a mean measure

for every case of LDA (Grymer et al., 1982). Retained placenta was found to be associated with a 0.8 kg/day loss across the lactation, or 2.5 kg/day loss across the first 100 days in milk (Fourichon et al., 1999). In a systematic review, Fourichon et al. (1999) reported that only 5 out of 13 studies had found production losses after dystocia, while none out of 6 studies reported a loss after a milk fever diagnosis.

The shape of the lactation curve in early lactation has been used as a predictor in various reproductive outcomes, such as the calving to conception interval in dairy herds (Cook and Green, 2016) and insemination outcomes (Hudson and Green, 2018). But the prediction of lactation curves themselves have also been the subject of several studies. ANNs have been used in several studies in order to predict milk yield (Lacroix et al., 1995, Lacroix et al., 1997, Salehi et al., 1998), either in terms of total 305-day yield (Grzesiak et al., 2003, Gorgulu, 2012), daily milk yield (Grzesiak et al., 2006, Torres et al., 2005), 305-day yield in first lactation (Sharma et al., 2006, Sharma et al., 2007, Njubi et al., 2010), or total herd production (Murphy et al., 2014, Sanzogni and Kerr, 2001). Deep learning methods have also been utilised in making milk yield predictions by Liseune et al. (2021), appearing to outperform baseline models.

The lactation curve has been proved critical to the herds' monitoring systems as it is a tool to detect disease such as ketosis and mastitis as an early stage (Grzesiak et al., 2003, Adriaens et al., 2018). This early detection possibility is what makes them so useful for farmers as disease can have a great economic impact in terms of production loss, treatment expenses or animal capital loss (Wilson et al., 2004, Gröhn et al., 2004). One downside is that in order to calculate the lactation curve it is generally necessary to have an initial number

or early milk yields. Therefore, the existence of a model that would be able to predict said lactation curves without the access to such information could be a very important asset.

Several machine learning studies aimed at production outcomes have been published throughout the past two decades. Murphy et al. (2014) for instance, compared 3 different models that focused on predicting total daily milk yield of the herd. Njubi et al. (2010) attempted predictions on next month and first lactation 305-day milk yield of Holstein-Friesian cows in Kenya. In a more recent study Grzesiak et al. (2021) also dealt with primiparous cows, predicting average milk yields with low RMSE and MAE and correlation coefficients of predicted vs actual values ranging between 0.75 and 0.99. Gianola et al. (2011) managed to predict fat, milk and protein yield with some success using genomic data. Grzesiak et al. (2006) focused on daily milk yield producing models with R^2 values ranging from 31% to 79%, while Shahinfar et al. (2012) studied the prediction of breeding values in dairy cattle, including milk yield with a maximum correlation of 0.93. Sefeedpari et al. (2015) focused on forecasting milk yield of 50 target farms in Iran, using energy consumption data, and presenting models with R^2 values ranging between 0.65 and 0.93. Zegler et al. (2020) analysed the prediction of milk production in pastures for each of two months, using variables such as improved legume cover, residual sward height, and non-improved grass cover. Nguyen et al. (2020) launched a small-scale study of 36 cows, aiming to predict daily milk yield by using 35 cows as a training set with the remaining one as a test set, for each one of them, producing 36 models for each method used. The R^2 of their four initial methods averaged at around 70%, showing an increase in their autoregressive models at an average of around

80%. Dallago et al. (2019) explored the prediction of first day milk yield in heifers and provided three different models with less than 4kg MSE, out of which the ANN was considered the best. Fuentes et al. (2020) used feed, weight and weather data as inputs to develop models predicting milk yield, protein and fat content, using daily data from 36 cows gathered over a period of four years, while Ji et al. (2022) explored various production measures, including the prediction of daily milk yield of the next month using data collected by robotic milking systems over a period of five years in a herd of a total of 80 cows, with a mean R^2 of 91.9%. Piwczyński et al. (2020) took advantage of robotic milking systems for data collection as well, on a larger scale study of 37 herds, building a decision tree model predicting monthly milk yield. Gocheva-Ilieva et al. (2022) utilised data from 158 cows throughout 4 farms and identified, farm, udder width, chest width and stature of the cow as important predictors for average 305-day milk yield. Other recent studies have used machine learning methods to predict milk yield. Salamone et al. (2022) have reported a selection of random forest models aiming to predict first day milk yield, with R^2 values of up to 52%. Bovo et al. (2021) also reported models predicting milk yield, based on their median accuracy, using mainly environmental predictors about temperature and heat stress. They built their models using 91 animals from one herd, presenting a relative error of 18% for daily milk yield, which can drop to 2% when they use the total milk yield (of an average 68 test days). To our knowledge, there are no other studies of this size utilizing machine learning to predict predicted 305-milk yield or daily milk residuals through various cow predictors.

Multicollinearity of independent variables is also a significant issue dealt with in this chapter. Multicollinearity is a major issue in predictive modelling, especially in multiple regression analysis, arising when two or more predictor variables are highly correlated. This can hinder accurate estimation of model coefficients, as it inflates their standard errors, making them unreliable and complicating result interpretation (Alin, 2010; Shrestha, 2020; Ayinde & Nwosu, 2021). Such inflation can lead to wider confidence intervals, reduced statistical power, and ultimately diminish the model's predictive accuracy (Shrestha, 2020; Arici, 2023).

Multicollinearity is a concern not only in regression models but also in a range of machine learning algorithms, such as decision trees, random forests, artificial neural networks, support vector machines (SVM), K-nearest neighbours (KNN), and naive Bayes classifiers. Its impact can differ considerably across these techniques, affecting both model performance and interpretability. Tree-based models, such as decision trees and random forests, are generally less sensitive to multicollinearity allowing them to handle correlated predictors effectively, while maintaining robust predictive performance despite high correlations among input features (Abbas et al., 2024). Support vector machines are more sensitive to multicollinearity through the use of kernels, with more complex kernels exacerbating the effects of multicollinearity, leading to overfitting and reduced generalization (Abbas et al., 2024). KNN may also be affected by multicollinearity as it relies on distance metrics that can be distorted by correlated predictors (Singh et al., 2023). The effect of the phenomenon on ANN is less pronounced, however still consequential as the network may struggle to learn the underlying patterns when predictors are highly correlated,

leading to difficulties in training the model and subsequent overfitting (Farrell et al., 2019). For Naive Bayes classifiers, multicollinearity violates the assumption of feature independence that they operate under, leading to biased predictions. The presence of redundant features can cause the classifier to assign equal importance to both relevant and irrelevant features, which may further reduce classification performance (Chen et al., 2021).

One popular method to assess multicollinearity is to calculate the Variance Inflation Factor (VIF). The VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity (Kyriazos and Poga, 2023, Kılıçoğlu and Yerlikaya-Özkurt, 2024). There is no universal consensus as to which VIF cutoff is considered optimal (Vatcheva et al., 2016), however often a VIF value greater than 5 is considered indicative of problematic multicollinearity (Kim, 2019), with others accepting a more lenient cutoff at 10 (Holder and Field, 2019, Mutchler and Anderson, 2010).

Detecting multicollinearity in models that include categorical independent variables as it is in our case can be challenging, but several methods can be employed to assess and address this issue effectively. Categorical variables, when included in regression models, are typically transformed into dummy variables. This transformation can introduce multicollinearity, particularly when the number of categories is high or when categories are correlated with one another. When dealing with categorical variables, it is essential to calculate the VIF for each dummy variable created from the categorical variable. If any dummy variable exhibits a high VIF, it may suggest that the categorical variable is contributing to multicollinearity. It is crucial to omit one category of the dummy variable to avoid a phenomenon which occurs when perfect multicollinearity

arises due to the inclusion of all categories (Dressler et al., 2016). This practice is supported by various studies that emphasize the importance of careful dummy variable management in regression analysis (May et al., 2011, Je & Lee, 2023, Wei et al., 2024). Their findings highlight that when dummy variables are appropriately managed, they can enhance model fit without introducing significant multicollinearity. Moreover, the use of VIF is not limited to traditional regression models. It has been applied in various contexts, including machine learning and econometric models. For instance, in support vector regression, VIF is utilized to establish a multicollinearity threshold for variable selection, ensuring that the model remains robust against multicollinearity (Folli et al., 2020). In conclusion, the incorporation of dummy variables in regression models necessitates careful consideration of multicollinearity, with VIF serving as a critical diagnostic tool. By ensuring that one category of the dummy variable is omitted and monitoring VIF values, researchers can mitigate the risks associated with multicollinearity, thereby enhancing the reliability of their regression analyses.

Another consideration was confounding. Controlling for confounders has been a challenge, especially in biomedical studies (He et al, 2019, Smith and Nichols, 2018, Topol, 2019), as they can interfere with the perceived relationship between input and output variables (Duncan and Northoff, 2013, Jager et al, 2008, Pourhoseingholi et al., 2012). As predictive modelling does not focus on interpreting causation but rather on predictive power, it becomes a concern under certain conditions such as a scanner effect or head motion in neuroimaging or it might affect the generalizability of the model across different contexts as one population might have the confounding effect while another

may not when dealing with predictors such as biomarkers (Spisak, 2022). Therefore, the most pressing issues with confounding in predictive modelling are multicollinearity and overfitting.

6.2 Methods

Four separate lactation curves were fitted for each herd in the dataset, representing expected lactation curve shape for animals in lactation 1, 2, 3 and 4 and above in that herd, using all milk recording test day yields in each herd dataset between 2014 and 2020 and occurring at between 1 and 400 days in milk (DIM). Curves were fitted via the MilkBot equation, a modification of the original Woods curve (Ehrlich, 2010), using the `nsLM` function in the R package “`minpack.lm`” (Timur et al., 2016). For each milk recording test day, the predicted yield was calculated, based on the lactation curve parameters for that herd and parity and the DIM of the animal at that test day. The residual yield (observed minus predicted daily yield for that test day) was calculated for each cow at each test day, representing the absolute difference in daily yield between the individual and her prediction.

The predicted yield divided by observed yield was used to update the “scale” parameter of the herd/parity lactation curve for that particular test day, and the MilkBot formula rearranged to estimate a 305-day yield based on each individual test day yield. For lactations with at least five test day yield records, the mean of the five or more predictions was taken to represent the 305-day yield of that lactation,

Both predictive and inferential models were built using both lactation-level predicted 305-day yield, and test-day-level residual yield as outcomes. Potential predictor variables considered were binary disease occurrence in the relevant lactation (milk fever, LDA, RFM and metritis), existence of lameness both pre- and post-partum, rumen fill, BCS, THI and hock hygiene both pre-

and post-partum, having either twin calves or a dead calf at the beginning of the lactation, and finally the calving month and lactation number.

In addition to the predictive models described above (built using lactation- and test-day-level data) models were also built aggregating data at herd-quarter-year level (i.e. averaging over all lactations/test days in a given herd in a given quarter-year). Predictions were attempted using 3 different methods:

Method A) Predictions at the individual cow level, using cow level models

Method B) Predictions at herd/month level, using cow level models

Method C) Predictions at herd-quarter-year level using herd-quarter-year models

For group C all variables were aggregated on the herd-quarter-year level. Quarter-year was favoured over month groups, in order to avoid groups with very few cows. Furthermore, groups that had less than 10 observations were omitted. Continuous variables were averaged as a mean of the group, while binary ones were converted to a percentage of positive instances in the herd-quarter-year group. Factors with multiple levels, that could not be treated as numerical in order to be averaged (e.g. hock hygiene score), were removed from the analysis. The variables used were mean BCS pre- and post-partum, mean lameness score both pre- and post-partum, mean THI both pre- and post-partum, mean rumen fill score again both pre- and post-partum, mean lactation number, the percentage of LDA, RFM, metritis and milk fever diagnosis and the percentage of twinning and calf mortality of all cows in each herd/trimester group.

A variety of machine learning algorithms appropriate to continuous numeric outcomes were selected: these were linear regression, decision tree, random forest, ANN and MARS. As in previous chapters, 10-fold cross validation was used to fit all initial models. The R^2 , RMSE and MAE were used to evaluate all models' predictive value with an initial emphasis on the R^2 values.

For method B a holdout dataset was used, splitting the train and test data 80% to 20% of the herd-quarter-year groups. The best performing model was chosen to make predictions on the test data and then those predictions were averaged as a mean of each herd/trimester group and compared to the actual means of the groups. The predicted and actual values were plotted against each other and a Pearson's correlated coefficient was calculated, as well as predictive models that used the actual means as an outcome and predicted ones as a predictive variable. The effects of possible multicollinearity were also considered for all models. The Variance Inflation Factor (VIF) described in 6.1 was used to assess the multicollinearity effect of each independent variable. Continuous and binary variables were assessed unmodified, while multi-level categorical variables were split into dummy variables and a VIF value was calculated for each level. The first level of each variable was omitted to avoid instances of perfect multicollinearity (Dressler et al., 2016). In order to calculate the VIF the car package in R was used (Fox and Weisberg, 2019). Multicollinearity was considered an issue when the VIF was over 5 (Kim, 2019). As generally recommended, when VIF values exceed the threshold, the variables were excluded from the model, as high VIF values can lead to unstable estimates and make it difficult to determine the individual effect of each predictor on the dependent variable (Prunier et al., 2015, Ghareeb, 2023, Kroll

and Song, 2013, Xi, 2024). As mentioned in 6.1 confounding is another issue that in predictive modelling can result in multicollinearity and overfitting. Therefore, considering possible confounding, multicollinearity was taken into account through VIF value calculation, as described above, and overfitting was tackled through cross-validation.

6.3 Results

6.3.1 Method A-Individual cow level and models

6.3.1.1 Outcome: Predicted 305-day lactation milk yield

The total number of data points (lactations) used for the models was 7,296. The mean predicted 305 milk yield was at 10,971 kg, with a median of 10,886 kg, a maximum of 19,687 kg and minimum of 3,769 kg.

The variables that were considered for the final model and also included missing data were hock hygiene post-partum with 1,635 missing data points (22.4%), LDA with 6 missing data points (0.08%), THI pre-partum with 5053 (55.4%) and THI post-partum with 8360 (91.7%) missing data points. Due to the volume of missing data, variables THI pre- and post-partum were removed from the analysis to avoid discarding or imputing values for a very high proportion of the data.

The VIF value was calculated for all possible independent variables (Table 6.1). The dummy variables for rumen fill score 5 both pre and post-partum were found to cause perfect multicollinearity so they were excluded from the initial model in order to be able to calculate the VIF values. The rest of the variable levels were included in order to minimise information loss. The variable levels that generated a VIF value higher than the set cutoff of 5 were all 5 levels of the hock hygiene score post-partum variable, rumen fill score post-partum 2, 3 or 4 (so all the variable levels except a score of 1), BCS pre-partum score of 2.5, 2.75, 3, 3.25, 3.5 or higher than 3.5 (so all levels except for a score of 2) and a BCS score post-partum of 2.75 (only level of that variable). Hence those variable levels were excluded from the analysis. Therefore, the final predictive

variables included in the models were lactation number 2, lactation number 3, lactation number 4, lactation number >4, Hock Hygiene Score 1 pre-partum, Hock Hygiene Score 2 pre-partum, Hock Hygiene Score 3 pre-partum, Hock Hygiene Score 4 pre-partum, Hock Hygiene Score 5 pre-partum, Rumen fill score 1 pre-partum, Rumen fill score 2 pre-partum, Rumen fill score 3 pre-partum, Rumen fill score 4 pre-partum, Rumen fill score 1 post-partum, BCS 2 pre-partum, BCS 2 post-partum, BCS 2.5 post-partum, BCS 3 post-partum, BCS 3.25 post-partum, BCS 3.5 post-partum, BCS >3.5 post-partum, twinning, calf mortality, lameness pre-partum, lameness post-partum, calving month February, calving month March, calving month April, calving month May, calving month June, calving month July, calving month August, calving month September, calving month October, calving month November, calving month December, Milk Fever, LDA, RFM and metritis.

Table 6.1 VIF values for all possible predictive variables as calculated when fitting a linear regression model on the predicted 305 milk yield with all variables included

Variable	VIF
Lactation No 2	1.725
Lactation No 3	1.629
Lactation No 4	1.449
Lactation No >4	1.581
Hock Hygiene Score pre-partum 1	1.023
Hock Hygiene Score pre-partum 2	1.026
Hock Hygiene Score pre-partum 3	1.031
Hock Hygiene Score pre-partum 4	1.009
Hock Hygiene Score pre-partum 5	1.006
Hock Hygiene Score post-partum 1	14.020
Hock Hygiene Score post-partum 2	56.840
Hock Hygiene Score post-partum 3	70.375
Hock Hygiene Score post-partum 4	49.849

Hock Hygiene Score post-partum	16.286
5	
Rumen Fill pre-partum 1	1.210
Rumen Fill pre-partum 2	1.934
Rumen Fill pre-partum 3	2.999
Rumen Fill pre-partum 4	2.920
Rumen Fill post-partum 1	4.260
Rumen Fill post-partum 2	14.762
Rumen Fill post-partum 3	18.335
Rumen Fill post-partum 4	12.182
BCS pre-partum 2	1.072
BCS pre-partum 2.5	8.555
BCS pre-partum 2.75	24.364
BCS pre-partum 3	25.238
BCS pre-partum 3.25	27.924
BCS pre-partum 3.5	22.676
BCS pre-partum >3.5	12.875
BCS post-partum 2	1.187
BCS post-partum 2.5	3.741

BCS post-partum 2.75	5.543
BCS post-partum 3	4.599
BCS post-partum 3.25	3.203
BCS post-partum 3.5	1.956
BCS post-partum >3.5	1.374
Twinning	1.080
Calf Mortality	1.046
Lameness pre-partum	1.070
Lameness post-partum	1.063
Calving month-February	1.943
Calving month-March	2.086
Calving month-April	1.853
Calving month-May	2.229
Calving month-June	2.401
Calving month-July	2.633
Calving month-August	2.627
Calving month-September	2.328
Calving month-October	2.126
Calving month-November	1.873

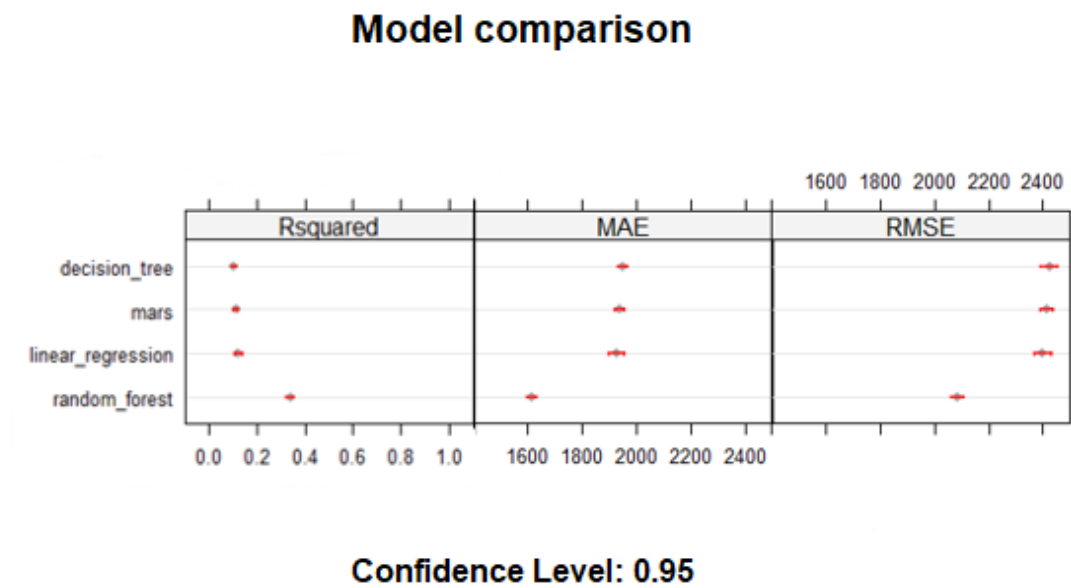
Calving month-December	1.987
Milk Fever	1.090
LDA	1.027
RFM	1.120
Metritis	1.091

The R^2 of all models when applied on the test set ranged from 10.5% (decision tree) to 33.9% (random forest); performance metrics could not be calculated for the ANN model, most likely due to a single value being predicted for all cases (i.e. a lack of variance). RMSE and MAE ranged in similar values for all models with the exception of ANN, which was an outlier for the reasons stated above, and had very high values (11600.62 L and 11313.89 L respectively). For the rest of the models RMSE ranged between 2084.48 L (random forest) and 2423.94 L (decision tree), while MAE ranged between 1614.89 L (random forest) and 1949.63 L (decision tree). Therefore, the overall best model was the random forest explaining around a third of the outcome's variation. All RMSE, R^2 and MAE values for both the training and the test set are shown in Table 6.2. Furthermore, all the values of R^2 , RMSE and MAE values produced when models were applied on the test set are also graphically shown in Figure 6.1.

Table 6.2 R^2 , RMSE and MAE values of all models (excluding ANN) predicting 305-day milk yield, when applied on both the training and the test set.

Training Set			
	RMSE	R^2	MAE
Linear regression	2349.15	0.148	1874.91
Decision Tree	2376.59	0.114	1903.37
Random Forest	1925.70	0.351	1588.77
ANN	11505.12	-	11294.01
MARS	2396.11	0.182	1912.41
Test Set			
	RMSE	R^2	MAE
Linear regression	2398.65	0.124	1925.35
Decision Tree	2423.94	0.105	1949.63
Random Forest	2084.48	0.339	1614.89
ANN	11600.62	-	11313.89
MARS	2412.13	0.115	1936.01

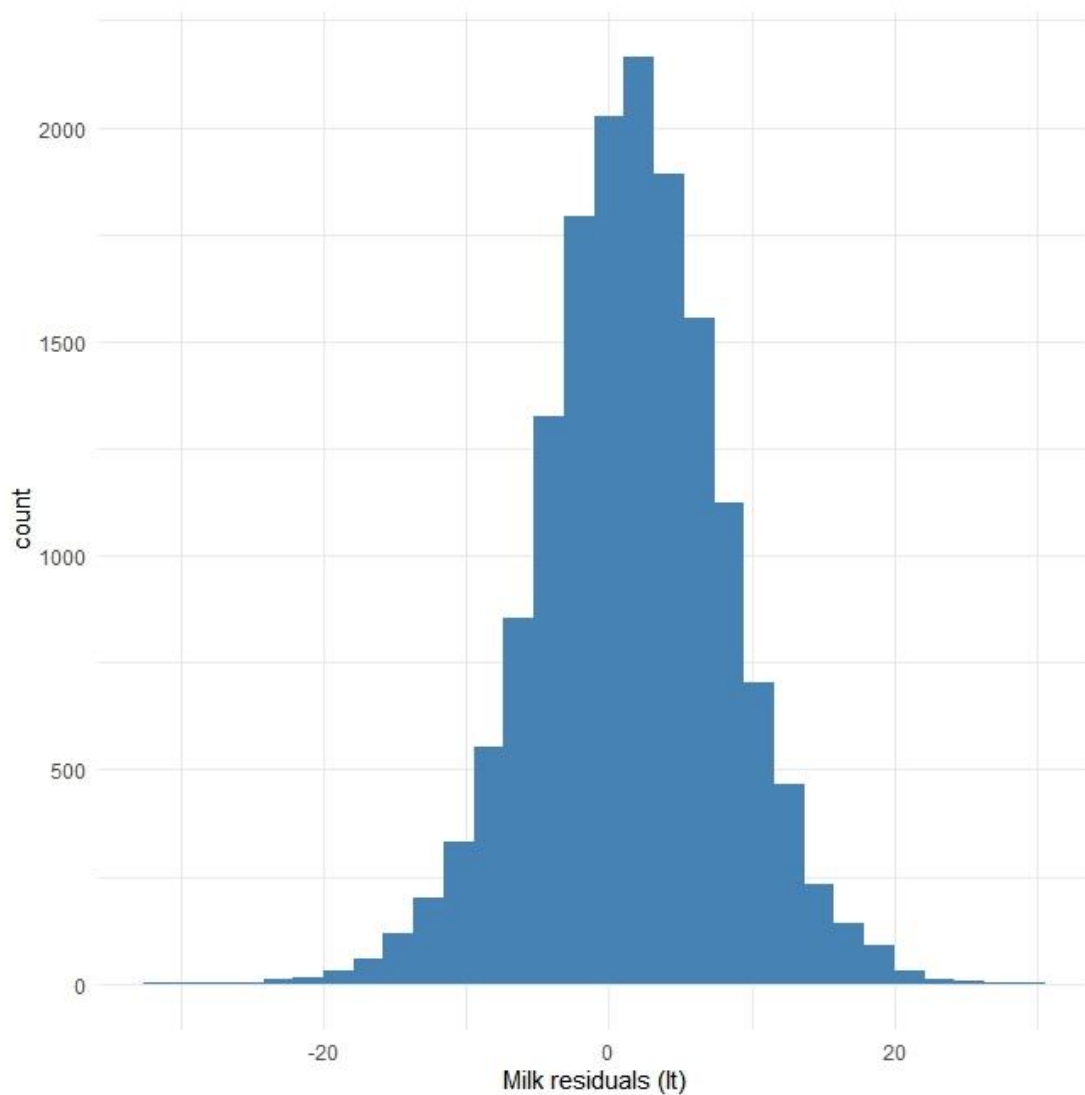
Figure 6.1 R^2 , RMSE and MAE values of all models (excluding ANN) predicting 305-day milk yield, when applied on the test set.



6.3.1.2 Outcome: Residual Daily Milk Yield

The total number of data points used for this part of the analysis after removing the residual milk yield missing data was 15,742. The milk residuals (representing deviation in daily milk yield from what would be predicted by the lactation curve shape for that parity in that herd) had a mean value of 1.56 (median 1.59), with a minimum of -31.28 and a maximum of 29.91 (Figure 6.2)

Figure 6.2 Distribution of milk yield residuals on the final dataset of a total 15,742 data points.



The VIF values were very similar to the ones described in 6.3.1.1 and are shown in Table 6.3. The final predictive values included were the same and included lactation number 2, lactation number 3, lactation number 4, lactation number >4, Hock Hygiene Score 1 pre-partum, Hock Hygiene Score 2 pre-partum, Hock Hygiene Score 3 pre-partum, Hock Hygiene Score 4 pre-partum, Hock Hygiene Score 5 pre-partum, Rumen fill score 1 pre-partum, Rumen fill

score 2 pre-partum, Rumen fill score 3 pre-partum, Rumen fill score 4 pre-partum, Rumen fill score 1 post-partum, BCS 2 pre-partum, BCS 2 post-partum, BCS 2.5 post-partum, BCS 3 post-partum, BCS 3.25 post-partum, BCS 3.5 post-partum, BCS >3.5 post-partum, twinning, calf mortality, lameness pre-partum, lameness post-partum, calving month February, calving month March, calving month April, calving month May, calving month June, calving month July, calving month August, calving month September, calving month October, calving month November, calving month December, Milk Fever, LDA, RFM and metritis. Variables with missing data points were the hock hygiene pre-partum dummy variables with 3,376 (21.4%) and milk fever, LDA, calf mortality and metritis all with 7 missing data points (0.04%).

Table 6.3 VIF values for all possible predictive variables as calculated when fitting a linear regression model on the residual milk yield with all variables included

Variable	VIF
Lactation No 2	1.462
Lactation No 3	1.855
Lactation No 4	1.277
Lactation No >4	1.843
Hock Hygiene Score pre-partum 1	1.010
Hock Hygiene Score pre-partum 2	1.017
Hock Hygiene Score pre-partum 3	1.027
Hock Hygiene Score pre-partum 4	1.012
Hock Hygiene Score pre-partum 5	1.027
Hock Hygiene Score post-partum 1	16.226
Hock Hygiene Score post-partum 2	53.527
Hock Hygiene Score post-partum 3	76.267
Hock Hygiene Score post-partum 4	52.025

Hock Hygiene Score post-partum	14.824
5	
Rumen Fill pre-partum 1	1.173
Rumen Fill pre-partum 2	2.137
Rumen Fill pre-partum 3	3.291
Rumen Fill pre-partum 4	3.182
Rumen Fill post-partum 1	4.637
Rumen Fill post-partum 2	15.845
Rumen Fill post-partum 3	20.573
Rumen Fill post-partum 4	11.457
BCS pre-partum 2	1.038
BCS pre-partum 2.5	9.457
BCS pre-partum 2.75	26.382
BCS pre-partum 3	26.835
BCS pre-partum 3.25	28.358
BCS pre-partum 3.5	23.952
BCS pre-partum >3.5	13.121
BCS post-partum 2	1.255
BCS post-partum 2.5	3.638

BCS post-partum 2.75	5.735
BCS post-partum 3	4.372
BCS post-partum 3.25	3.001
BCS post-partum 3.5	1.724
BCS post-partum >3.5	1.427
Twinning	1.037
Calf Mortality	1.072
Lameness pre-partum	1.037
Lameness post-partum	1.053
Calving month-February	1.725
Calving month-March	2.214
Calving month-April	1.738
Calving month-May	2.173
Calving month-June	2.332
Calving month-July	2.748
Calving month-August	2.979
Calving month-September	2.173
Calving month-October	1.907
Calving month-November	1.453

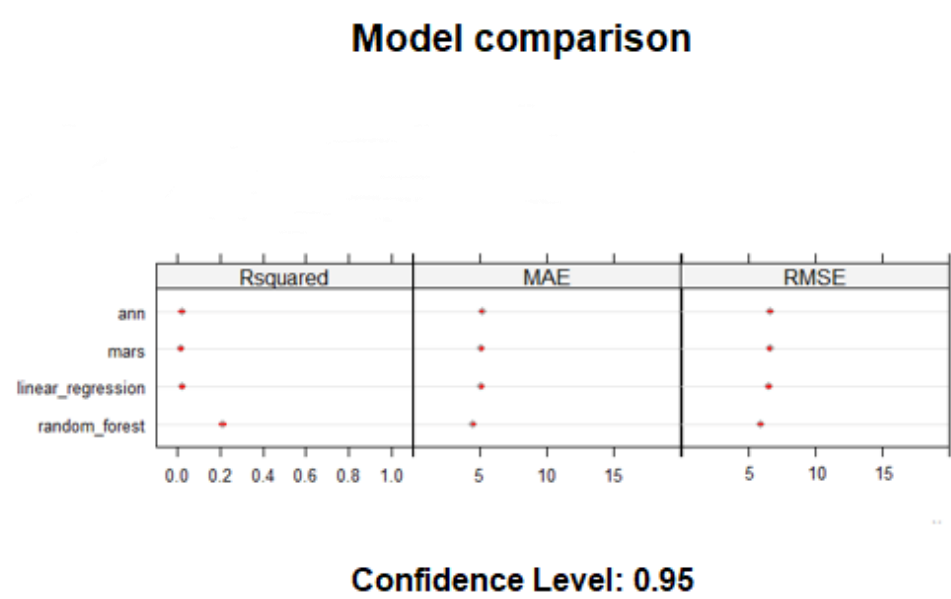
Calving month-December	1.356
Milk Fever	1.084
LDA	1.021
RFM	1.099
Metritis	1.079

R^2 values ranged from 1.7% for the MARS model, to 21.4% for the random forest model on the test set, with all models except for the random forest having a value of less than 3%. R^2 could not be computed for the decision tree model, similar to the ANN model in 6.3.1.1, most likely due to lack of variance. The RMSE ranged from 5.93 (random forest) to 6.66 (decision tree), while the MAE from 4.53 (random forest) to 5.20 (ANN). So overall the best performing model was the random forest, explaining over a fifth of the outcome's variation. The exact metrics of all models as fitted on both the training and the test set are shown in Table 6.4, while the values with their 95% confidence interval are graphically shown in Figure 6.3.

Table 6.4 R^2 , RMSE and MAE values of all models predicting milk yield residuals, when applied on both the training and the test set.

Training Set			
	RMSE	R^2	MAE
Linear regression	6.502	0.039	5.099
Decision Tree	6.616	-	5.110
Random Forest	5.814	0.232	4.468
ANN	6.580	0.034	5.112
MARS	6.554	0.025	5.089
Test Set			
	RMSE	R^2	MAE
Linear regression	6.581	0.025	5.146
Decision Tree	6.664	-	5.197
Random Forest	5.937	0.214	4.531
ANN	6.659	0.021	5.200
MARS	6.605	0.017	5.155

Figure 6.3 R^2 , RMSE and MAE values of all models predicting daily residual milk yield on the test set



6.3.2 Method B– Individual cow models used for predictions at herd-quarter-year level

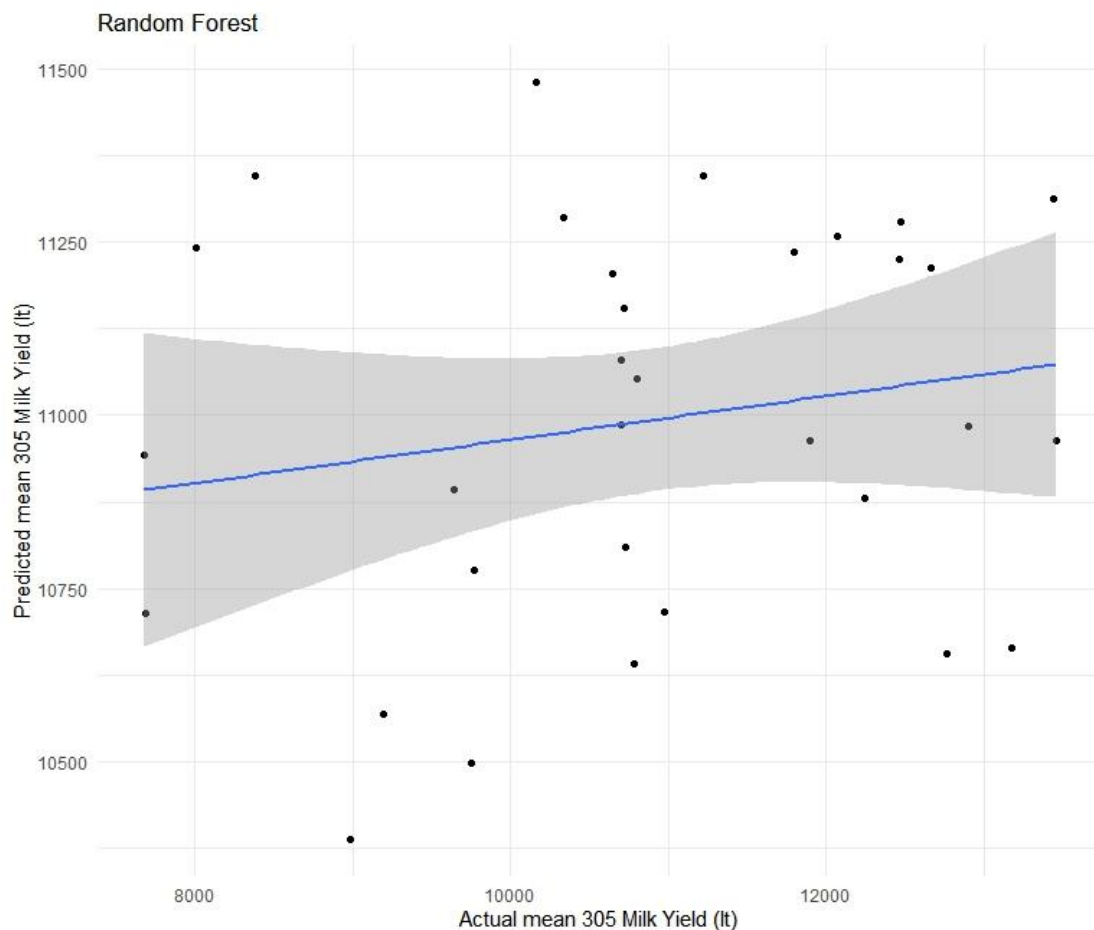
6.3.2.1 Outcome: Predicted 305-day milk yield

The total data points included in these models, after removing the herd-quarter-year groups contributing less than 10 lactation was 6,968. The minimum number of recordings per group was 11, the maximum 129, the mean 34.5 and the median 27. The total number of groups was 202, with 161 of them comprising the train dataset while the remaining 41 the test dataset.

Models were re-trained on the training dataset as described in section 6.3.1.1.1, with the random forest model performing the best and having similar performance characteristics to those described in the previous section (where

the full dataset was used for training). The random forest model was used to generate predictions aggregated at herd-quarter-year level, which were compared with observed values.

Figure 6.4 Scatter plot of predicted mean 305-day milk yield vs observed mean 305-day milk yield per herd-quarter-year group for the random forest model



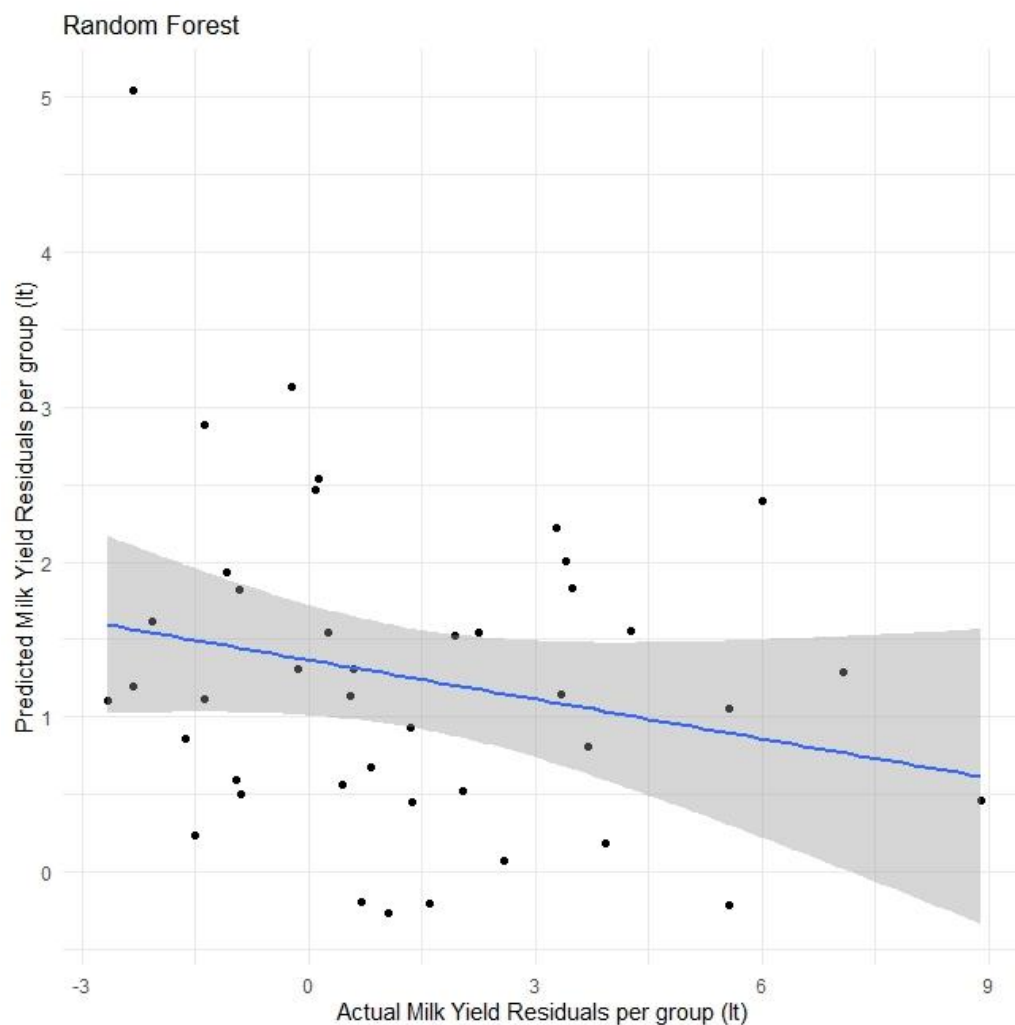
The R^2 describing how much of the actual 305 milk yield's variation is explained by the predictions was found to be just at 2.4%, indicating that only a very small fraction of the variation could be explained by the predictions made by the model.

6.3.2.2 Outcome: Residual milk yield

In total 229 herd/trimester groups were used to fit these versions of the models, after removing the groups with less than 10 observations. Out of those the 183 were used for training the model, while the remaining 46 were used for testing.

Since it was the best performing model, the random forest was chosen for further predictions on the group level. After making the predictions for the individual cows and averaging those predictions to come up with a mean value for the group, predicted values were plotted against the observed mean predicted 305-day milk yield of all the cows in the group (Figure 6.5). The groups used for this testing were a total of 30.

Figure 6.5 Scatter plot of predicted residual milk yield vs observed residual milk yield per herd-quarter-year group for the random forest model



The R^2 describing how much of the actual values' variation is explained by the predictions was found to be at 33.7% meaning a third of the variation of the outcome appears to be explained by the predictions made by initial model.

6.3.3 Method C – Herd-quarter-year level models

6.3.3.1 Outcome: Mean predicted 305-day milk yield

The total number of herd/quarter-year group available for this part of the analysis were 205. In regards to multicollinearity all variables were found to produce low VIF values, all below the 5 cutoff threshold (Table 6.5). The final variables in the analysis were mean BCS both pre and post-partum per group, mean rumen fill both pre and post-partum per group, mean Hock hygiene score both pre and post-partum, mean lactation number of the cows in the group, percentage of cows with milk fever, metritis and/or RFM per group and percentage of cows that had twins at the start of that lactation period, again, per group. Variables with missing data included metritis percentage, RFM percentage, milk fever percentage and twinning percentage all with 3 missing data points (2.4%) and the Hock hygiene pre- and post-partum with 41 missing data points (20%).

Table 6.5 VIF values for all possible predictive variables as calculated when fitting a linear regression model on mean predicted 305-day milk yield per herd-quarter-year group with all independent variables included.

Variable	VIF
Mean BCS pre-partum	2.897
Mean BCS post-partum	2.865
Mean Hock Hygiene score pre-partum	1.096
Mean Hock Hygiene score post-partum	1.191
Mean Rumen Fill score pre-partum	1.728
Mean Rumen Fill Score post-partum	2.004
Mean Lactation No	1.107
Percentage of Milk Fever diagnoses	1.168
Percentage of LDA diagnoses	1.108
Percentage of RFM diagnoses	1.098
Percentage of Metritis diagnoses	1.115
Percentage of Calf Mortality	1.035

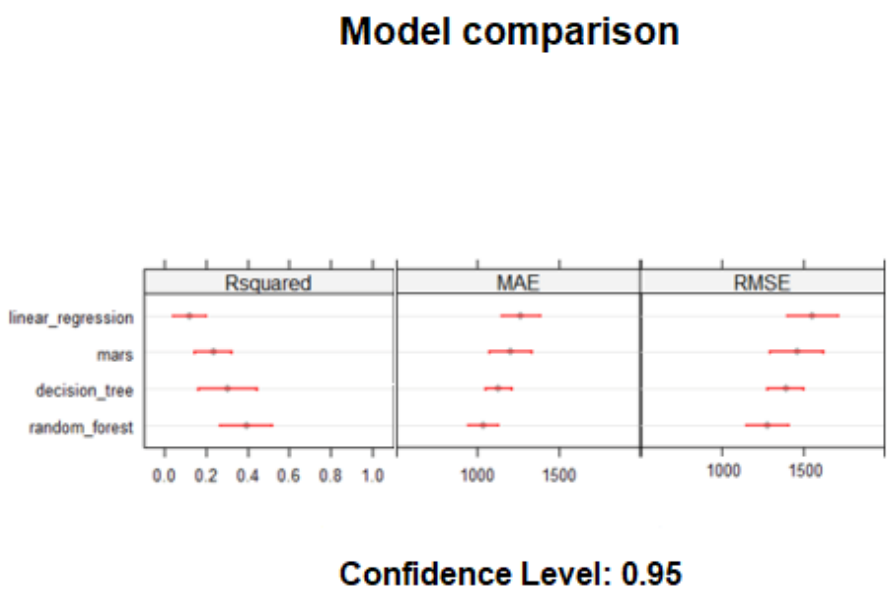
Percentage of Twinning	1.089
------------------------	-------

The R^2 in the final models ranged from 12.0% to 39.4%% for the linear regression and the random forest models respectively. Similarly to 6.3.1.1 ANN could not compute an R^2 value possibly due to lack of variation. Excluding the ANN model, the RMSE ranged from 1280.77 L to 1558.75 L, while MAE ranged from 1035.90 L to 1268.93 L, both for random forest and linear regression respectively. All R^2 , RMSE and MAE values for all models are shown in Table 6.6 and Figure 6.6.

Table 6.6 R^2 , RMSE and MAE values of all models predicting mean 305-day milk yield per herd-quarter-year group, when applied on both the training and the test set.

Training Set			
	RMSE	R^2	MAE
Linear regression	1532.62	0.143	1232.79
Decision Tree	1344.95	0.312	1102.64
Random Forest	1218.60	0.410	1006.26
ANN	11198.49	-	11069.51
MARS	1436.07	0.241	1196.59
Test Set			
	RMSE	R^2	MAE
Linear regression	1558.75	0.120	1268.93
Decision Tree	1392.14	0.303	1131.14
Random Forest	1280.77	0.394	1035.90
ANN	11235.79	-	11119.15
MARS	1461.45	0.234	1205.38

Figure 6.6 R^2 , RMSE and MAE values of all models predicting mean predicted 305-day milk yield per herd-quarter-year group with their 95% confidence intervals, as applied on the test set.



None of the models appeared to be of any considerable predictive value, with the best performing model, (random forest) explaining over a third of the outcome's variation.

6.3.3.2 Outcome: Residual milk yield

The total number of groups and hence data points that were used for this part of the analysis were 229. The means of the residuals ranged from -5.18 to 6.67 with a mean of 0.94 and a median of 0.86. Assessing the presence of multicollinearity, VIF values for all possible predictive variables were below the threshold of 5 (Table 6.7) meaning the possibility of multicollinearity is in fact low. The final predictive variables were mean BCS both pre and post-partum per group, mean rumen fill both pre and post-partum per group, mean Hock

hygiene score both pre and post-partum, mean lactation number of the cows in the group, percentage of cows with milk fever, metritis and/or RFM per group and percentage of cows that had twins at the start of that lactation period, again, per group. Variables with missing data included milk fever percentage, RFM percentage, LDA percentage, metritis percentage, twinning percentage and calf mortality percentage all with 4 missing data points (1.7%) and mean Hock hygiene score both pre- and post-partum with 49 missing data points (21.4%).

Table 6.7 VIF values for all possible predictive variables as calculated when fitting a linear regression model on the mean milk yield residuals per herd-quarter-year group, with all independent variables included.

Variable	VIF
Mean BCS pre-partum	2.629
Mean BCS post-partum	2.515
Mean Hock Hygiene score pre-partum	1.102
Mean Hock Hygiene score post-partum	1.136
Mean Rumen Fill score pre-partum	1.783
Mean Rumen Fill Score post-partum	1.951
Mean Lactation No	1.131
Percentage of Milk Fever diagnoses	1.170
Percentage of LDA diagnoses	1.099
Percentage of RFM diagnoses	1.036
Percentage of Metritis diagnoses	1.198
Percentage of Calf Mortality	1.014

Percentage of Twinning	1.055
------------------------	-------

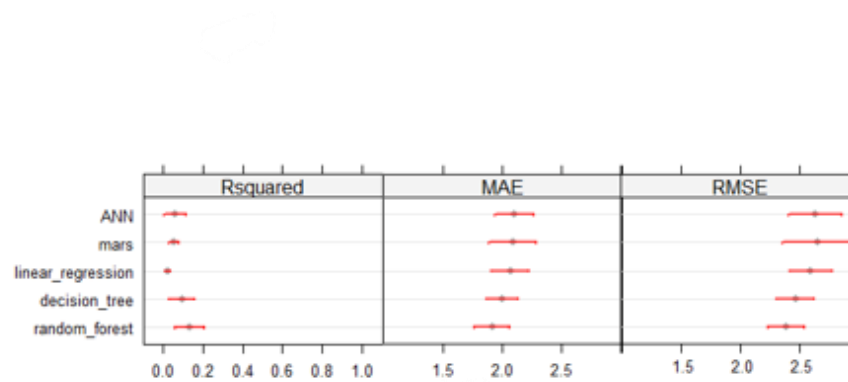
Again, R^2 values on the test set were generally low, from 2.3% (linear regression) to 13.4% (random forest). RMSE and MAE values ranged relatively high, from 2.384 L (random forest) to 2.647 (MARS) for the former, and 1.914 L (random forest) to 2.086 L (MARS) for the latter. All metrics for all models, as produced both on the training and the test set, are shown in Table 6.8 and Figure 6.7.

Table 6.8 R^2 , RMSE and MAE values of all models predicting mean daily residual milk yield per herd-quarter-year group on both the training and the test set

Training Set			
	RMSE	R^2	MAE
Linear regression	2.386	0.035	1.926
Decision Tree	2.529	0.085	2.139
Random Forest	2.302	0.141	1.870
ANN	2.499	0.087	2.007
MARS	2.600	0.060	2.012
Test Set			
	RMSE	R^2	MAE
Linear regression	2.594	0.023	2.066
Decision Tree	2.465	0.095	2.000
Random Forest	2.384	0.134	1.914
ANN	2.630	0.062	2.103
MARS	2.647	0.056	2.086

Figure 6.7 R^2 , RMSE and MAE values of all models predicting mean daily residual milk yield per herd-quarter-year group on the test set

Model comparison



Confidence Level: 0.95

6.4 Discussion

Predictive models for production outcomes showed very varied model performance across the wide range of different machine learning algorithms tested, both of the outcomes under consideration (305-day lactation yield, and deviation from predicted daily yield based on herd- and parity-specific lactation curves), and the different approaches taken to aggregating data for model building and predictions. The best performing model in terms of R^2 value was the random forest model for prediction of residual daily milk yield, explaining just over a third of the observed yield variation; for many of the outcomes, algorithms and aggregation approaches R^2 values were substantially lower. Inferential models found a large number of statistically significant associations between potential predictor variables and production outcomes, but again explained a relatively small proportion of observed variation in milk yields.

For the purposes of this chapter both the predicted 305 milk yield as well as the residual daily milk yield were used as outcomes. One of the major differences is that for the 305-day milk yield, lactation number and herd effects are likely to be major predictors, while with the residual daily milk yield these are effectively accounted for as the expected yield from which the residual is calculated is based on a lactation curve for that herd and lactation number. The models for 305 milk yield did indeed find lactation number to be an important predictor and the inferential model in particular both indicated an association with the outcome and lactation number and showed that herd effect explained a great deal of the outcome's variation. The models for the residual daily milk yield also included lactation number. Similarly to other chapters, none of the models were

likely to be predictive to a practically useful extent. This result is somewhat on a par with Salamone et al. (2022) who used similar random forest models to predict day one milk yields with slightly different cow and herd level variables. The best performing model we reported was the random forest model predicting residual daily milk yield with a R^2 value of 33.5% while Salamone et al. (2022) found an R^2 of 52.0%, which while considerably better is still a moderate value.

In method B the predictive values were similar with the model describing the relationship between averaged predictions and averaged predicted 305 milk yield per herd-quarter-year group having a R^2 value of 2.7%, while the model describing the relationship between averaged predictions and averaged residual daily milk yield having a R^2 value of 33.7%. This means that using our model, inputting individual cow level information, there is little possibility of making reliable conclusions on if the herd will underperform or over perform. The better of the two models in that regard was the residual daily milk yield model, with the aggregated predictions explaining a bit over a third of the aggregated residual daily milk yield values. Variables used for these predictions were health information, such as milk fever, metritis or LDA diagnosis, calf mortality, hock hygiene score pre-partum, rumen fill score both pre- and post-partum, BCS both pre- and post-partum, lameness also both pre- and post-partum, the lactation number and finally the calving month. This differs from existing research, in either the predictive variables used or in the outcome studied. In particular, we studied predictions on the entirety of the 305-milk yield, while Salamone et al. (2022) focused on the day one yields using individual cow information on production and reproduction, as well as herd level

production data. Also, in contrast with Bovo et al. (2021) we used individual cow health data rather than just environmental data.

As described in 6.1 there are a number of studies investigating the possibility of milk yield prediction in dairy cattle, and while their specific outcomes are different, with daily milk yield being amongst the most common, it is still interesting to examine what might cause a decrease in predictive value in our models. One interpretation might be the difference in data size and specifically the lack of variation when it came to herds. In fact, multiple studies only collected data from one herd, as they wanted to focus on robotic system data collection that had not been widely implemented yet (Nguyen et al., 2020, Fuentes et al., 2020, Ji et al., 2022). Murphy et al. (2014) collected data from one herd of 140 cows as well, while Bovo et al. (2021) collected data from 91 cows, once again by one herd. Grzesiak et al. (2006) also had data originating from a single herd and while the total number of daily milk yields was high (>100,000), it only included three lactations of a total 320 cows. In a more recent paper by the same author (Grzesiak et al. 2021) over 900 primiparous cows were included in a dataset used to build ANN models, however once again they were all from the same farm. The lack of farm variation could potentially result in more overfitted models that follow closely the trends and animals of that one specific farm that were perhaps too complex to capture in models where 50 different farms were included, such as ours. Further lack of external validation makes it difficult to assess the generalizability of the one-herd models. Gocheva-Ilieva et al. (2022) gathered data from 4 farms, however the total number of cows did not exceed 158, which could also affect generalizability. In addition, they did in fact identify the farm being an important predictor in their

models, further supporting the theory that a one-herd model would inherently produce better metrics than one with multiple herds. Sefeedpari et al. (2015) did not have this issue since they sampled their data from 50 farms, however their outcome was milk yield of the targeted farm, meaning that 50 was their total sample size. Furthermore, their target population was in Iran, a location with potentially many differences in the dairy industry compared to the UK, while using energy consumption as predictors, again a very different compared to ours. It is perhaps possible that a farm-wide outcome is easier to predict compared to an individual animal one, however even in Method C of our study in which a herd-quarter-year outcome was included the predictiveness was lacking. This could indicate that the energy consumption variables that Sefeedpari et al. (2015) used were of better quality since they were gathered for that purpose and not averaged from existing individual animal-specific variables. Zegler et al. (2020) gathered data from 20 farms including 2 pastures from each farm for 2 separate months, but again the outcome variable was pasture milk yield keeping the total sample size low. Another aspect is that the predictive variables they used included weather and pasture variables, as well as soil characteristics that could be more identifiable in organic farms, which is the kind of farms this particular study investigated. Other studies closer to ours in terms of sample size were Salamone et al. (2022) with 102 herds collected through historical data covering a period of 20 years, Piwczyński et al. (2020) whose data originated from 27 farms, 3,778 cows and 36,005 milk yields, Salamone et al. (2022) presented models predicting first day milk yield of next lactation, with the first day test being anytime from 1 to 60 DIM. In fact, the most important predictor included is the DIM the first day test occurred, with

cumulative 305-day milk yield of previous lactation following as slightly less significant. Since the first couple of months of the lactation however, the lactation curve is so steep, the presence of DIM in the lactation of interest raises a question as to whether the model is ultimately trained to identify its overall general shape that is expected in all cows rather than actually make meaningful predictions for the individual animal. Furthermore, the R^2 of their models ranged from 9% to 52%, not too different from our predicted 305-day milk yield models in Method A (ranging from 10.5% to 33.9%), especially when taking into account that the inclusion of DIM of first day test might artificially decrease variance as well as RMSE and MAE. Piwczyński et al. (2020) demonstrated that the inclusion of several variables in their decision tree model provided a statistically significant variance reduction, as supported by F-test. The most influential variable by a long margin was the milking frequency, where its increase brought a corresponding increase in milk yield. This increase appears plausible as it has been established in some papers (Vijayakumar et al., 2017, Alex et al., 2015), however it has been stressed that it can highly depend on the stage of lactation as well as udder health (Lyons et al., 2014b). It is overall possible that milking frequency is in fact a better-quality predictor than the ones we had available in our study.

Various disease variables such as metritis, LDA and RFM have been associated with a noteworthy drop in milk production (Daetz et al., 2016, Dezfouli et al., 2013, Figueiredo et al., 2021, Fourichon et al., 1999, Giuliadori et al., 2012, Grymer et al., 1982, Lyons et al., 2014a, Ribeiro et al., 2013). LDA was included as a predictor in both our models (methods A and B). This seems to be on par with existing research that have shown that milk yield can decrease

in cows with LDA compared to healthy counterparts (Dezfouli et al., 2013, Lyons et al., 2014a). Metritis, as research has shown its association with milk loss (Daetz et al., 2016, Figueiredo et al., 2021, Giuliadori et al., 2012, Lina et al., 2019, Ribeiro et al., 2013), was also included in both predicted 305-day milk yield models as well as residual milk, along with RFM which is found not only to be associated with milk production losses (Dervishi et al., 2016) but also with the development of metritis (Filho et al., 2012) as well as clinical mastitis (Pinedo and Fleming, 2012), which can exacerbate the existing production issues. Milk fever has also been associated with clinical mastitis (Pinedo and Fleming, 2012), however the overall association between milk fever and milk production appears to be more complex and potentially in some cases positive (Jawor et al., 2012). BCS post-partum was also included in the predictive models, as BCS has a well-established link with production (Kul et al., 2020, Loker et al., 2012, Roche et al., 2009, Rodriguez et al., 2021). Parity has been linked to milk production as well with multiple studies reporting an increase of milk production in subsequent lactations (Hoka et al., 2019, Koc, 2011, Utrera et al., 2013). It is therefore evident that there might have been some actual associations between our predictors and milk yield, however potentially due to the complex nature of the phenomenon they could not be translated to a meaningful predictive model, and the low overall R^2 values suggest that they are not likely to be key drivers of productivity.

There were a number of limitations in our study. Firstly, there were some issues with the quality of our data resulting into not having enough reliable information for some variables, such as the THI values pre and post calving, and ending up dropping them from the analysis. Temperature data have been used before to

predict milk yield by Bovo et al. (2022), however they had used THI information of several days in a row, to assess the effect of potential heat stress; it is doubtful that the single temperature recordings pre- and post-calving per lactation in our data would provide as much information to our models. As with other studies in this project, the nature of the data collection should also be considered. The fact that scoring occasions generally occurred fortnightly for each herd introduces potentially relevant variation depending on when cows are scored relative to calving. For example, a cow scored on the day after calving would be expected to have a lower rumen fill score than if she had been scored 10 days later simply because of the changes in feed intake expected around parturition. Another limitation is the sample size, especially in Method C. While in method A the total data point included in the model were around 7,000 for the predicted 305-day milk yield and around 15,000 for the milk residuals. However, when grouped by herd-quarter-year the number fell to around 200 for both outcomes. Larger sample sizes generally improve the robustness and accuracy of machine learning models. With more data, models can learn more complex patterns and relationships, leading to better generalization to unseen data (Ingathalikar et al., 2021). There is a well-documented issues referred as the curse of dimensionality which points to the phenomenon where the feature space becomes increasingly sparse as the number of dimensions (features) increases. This sparsity makes it difficult for machine learning algorithms to find meaningful patterns unless there is a sufficiently large sample size (Dhiman et al., 2022; Ramezan et al., 2021). As noted by Dhiman et al., (2022) larger sample sizes are necessary when using machine learning methods to mitigate the impact of this curse and improve

model accuracy. Furthermore, smaller sample sizes can lead to overfitting, where the model learns noise in the training data rather than the underlying signal, which can lead to generalization failure when attempting predictions on new data (Infante et al., 2022) and thus poor predictive performance. Takahashi et al. (2020) emphasized that machine learning models typically require larger datasets than traditional statistical methods to achieve robust performance due to the plethora of degrees of freedom that need to be covered. For instance, Collins et al. (2015) reported the recommendation that at least 100 events and 100 non-events need to be included just for the external validation of a predictive logistic regression model. Thus, sample size could have been a potential issue in Method C since it would be difficult to assess the actual predictiveness of our model, especially in the scenario of a highly predictive model, since it could easily be the result of overfitting.

Generalisability is also a potential issue, as our data collection was performed onto farms with similar characteristics (notably from relatively high yielding herds) from within Great Britain. That would mean that predictions might not apply to any farms that do not fit with this set of characteristics. Although there will be some biological characteristics that are consistent across cows within different systems, there are other measures which may have different relevance in different systems. Overall, this study has reinforced the difficulty in accurately predicting milk production outcomes from scoring and other routinely recorded data, while providing insight into some of the factors that are associated with changes in milk yield.

Chapter 7 – General Discussion

7.1 Summary of Results

The aim of our study was to utilise transition period data in order to attempt predictions on the health, reproduction and production of the cow with the use of machine learning. In Chapter 3 the characteristics of the herds participating in our studies was explored in more detail and it became evident that overall we were dealing with farms on the higher end of productivity. In Chapter 4 the models built to attempt predictions on health outcomes proved to be of little value, with the kappa metric not surpassing the expected threshold. When using lactation level models to make predictions on a herd/quarter-year level the averaged predictions on metritis were able to explain over a third (66.9%) of the variation of the averaged outcome. In Chapter 5 the focus was moved to the reproductive performance. Outcomes on both the insemination success and the day of conception were investigated, however predictive value was found to be relatively poor in these models. Inferential modelling did find significant associations for both outcomes, which were seemingly not enough to explain much of these outcomes' variation and make accurate predictions. In Chapter 6 the outcomes of interest were the predicted 305 milk yield and the residual daily milk yield. Not unlike the results of the previous Chapters these models did not produce very high R^2 values, with the best performing ones being the individual lactation models both predicting on an individual and on an aggregated level, which seemed to explain about a third of the outcome's variation.

7.2 Discussion

7.2.1 Predictiveness of Models

The vast majority of the models presented in this work did not reach high levels of predictive value.

Regarding the collective disease outcomes, even when taking steps to increase the sample size of positive instances with up-sampling, the variety of disease that were included in said outcome, may have been each associated with different variables in many complex ways and hence not going towards one clear direction when binned together. It is also important to consider whether the choice of variables was poor or whether important factors were omitted. The individual level models, especially the binary ones, were underwhelming in terms of predictiveness regardless of outcome. The best performing one on an individual lactation level was the residual daily milk yield model, with an R^2 explaining over 30% of the variation of the outcome, which is not likely to be sufficient for practical application.

Out of all the models the best performing overall was the individual lactation model for metritis when predicting on an aggregated level, with the averaged predictions explaining over two thirds of the variation of the aggregated outcome. The improvement on predictions when it comes to metritis may also be apparent on the collective disease status model, where the aggregated predictions of the individual disease model explained almost 45% of the averaged disease outcome, in contrast with the LDA, RFM and milk fever models which all produced R^2 values lower than 20%. So, while predictions on an individual level did not seem possible, that same model managed to produce

predictions on a herd/quarter-year level that could potentially be of some value to farm managers as a possible marker for transition success, on a group level. It is not uncommon with biological outcomes that when failing to make individual level predictions these same models can produce improved results on an aggregated level. This approach, of predicting a probability value for a binary outcome is essentially model calibration, where the evaluation data is separated into groups (usually deciles) and then the model's bias is calculated for each one of the groups (Chen et al., 2022). The improvement is evident in the metritis model and even the collective disease one and could probably be seen in the insemination outcome model, which described over a third of the averaged outcome's variation. This effect is not as evident on the milk outcomes, plausibly since these variables were already on a continuous scale.

In addition to using models built to predict lactation level outcomes to aggregate predictions across groups, we also explored building models using this aggregated dataset. This produces a much smaller dataset (where units of data may, for example, be herd/quarter-years) and a continuous outcome (representing the proportion of lactations affected within that group). Regarding these aggregated models, none seemed to make a significant improvement over the already existing ones. The models for LDA, RFM, as well as milk fever all showed improvement compared to the individual level models making predictions on an aggregated level. However, none of the R^2 values exceeded 40%, indicating low predictiveness. The models for metritis were an exception as the aggregated level model did not surpass the lactation level one. The same applied for the collective disease status with the best aggregated model producing an R^2 value of 32.0% (MARS). The situation reversed for the

insemination success model, while for the two milk outcomes the results were mixed. It is interesting that, overall, each outcome showed a mixed behaviour, since aggregating the predictive variables did not necessarily cause neither an improvement nor a decline in predictive value. So, while it could be argued that aggregating the variables might in fact lead to a loss of information and a drop in predictiveness, in some cases it might be a viable method to improve the model metrics.

Therefore overall, the manipulation of the variables or the change in predictive goal from the individual lactation level to the aggregated one appeared to be working more consistently in binary outcomes with varied results in the continues ones. This was to be expected, since the binary models' results do not have as much margin for error as the continues ones that are just asked to approximate a possible value.

7.2.2 Predictive vs Inferential

For chapter 5, inferential models were attempted to be built for the insemination outcomes, alongside the predictive ones. Despite inferential models showing several statistically significant associations with each outcome, predictive models still underperformed. This is another indication that essential terms, which affect those outcomes, and either were not measured, or perhaps are unmeasurable, were omitted. Hempstalk et al. (2015), when attempting to build machine learning models predicting the conception success to a given cow lactation, suggested this exact thing for the herd-season-year group in

insemination outcomes, along with other factors such as the capability of the technician or the bull's fertility.

This realisation is especially essential since before predictive modelling became mainstream in research, scientists used to rely on inferential models and oftentimes suggest that real life actions on herd health management ought to be taken based on them. While the existing knowledge might still be useful in herd management, it is good to keep that in mind during future research.

7.2.3 Metrics

Throughout the research the subject of metric selection and importance was raised, especially for binary outcomes with imbalanced datasets. Accuracy appeared to be misleading in such cases, even after resampling methods, showcasing that complete reliance on this metric can lead to inappropriate conclusions. Specificity and sensitivity appear to be more robust, however both of them and/or their combination in AUROC should be reported to paint an accurate picture of both classes. Meanwhile, accuracy, specificity and sensitivity appear to be amongst the most popular metrics used to judge a model's predictive performance. Several studies in farm medicine that have used machine learning methods to develop predictive algorithms for classification have reported metrics such as the overall accuracy, (Borchers et al., 2017, Caraviello et al., 2006, Dolecheck et al., 2015, Fenlon et al., 2017, Zaborski et al., 2018, Pastell and Kujala, 2007, Aguias et al., 2012, Chen et al., 2020, Cevik, 2020, Jiménez-Montero et al., 2013, Dolechek et al., 2015, Ebrahimie et al., 2018a, Ebrahimie et al., 2021, Farah et al., 2021, Zaborski et al., 2018, Tamura et al., 2019, Douphrate et al., 2019, Njubi et al. 2010, Sturm et al., 2020, Romadhonny et al., 2019, Rodriguez et al., 2019, Taneja et al.,

2020, Zhao et al., 2020) sensitivity, specificity (Borchers et al., 2017, Caraviello et al., 2006, Dolecheck et al., 2015, Zaborski et al., 2018, Nielen et al., 2015a, Nielen et al., 2015b, Hassan et al., 2009, Sun et al., 2010, Kamphuis et al., 2015, Mammadova et al., 2013, Panchal et al., 2016, Fenlon et al., 2017b, Post et al., 2020, Becker et al., 2021, Lasser et al., 2021, Lardy et al., 2023, Srikok et al., 2020, Volkman et al., 2021, Esener et al., 2021, Sadeghi et al., 2022, Imada et al., 2024, Vergara et al., 2014, Miller et al., 2020, Warner et al., 2020), AUROC (Hempstalk et al., 2015, Shahinfar et al., 2014, Zaborski et al., 2018, Avizheh et al., 2023, Williams et al., 2016, Panchal et al., 2016, Wisnieski et al., 2019, Post et al., 2020, Shahinfar et al., 2021, Imada et al., 2024, Vergara et al., 2014, Merenda et al., 2020, Post et al., 2020, Grzesiak et al., 2010, Keshavarzi et al., 2020, Miller et al., 2020, Warner et al., 2020) or correctly classified instances (CCI) (Shahinfar et al., 2014). When predicting health outcomes in particular, the metrics reported were again accuracy (Ebrahimie et al., 2018, Sharifi et al., 2018), sensitivity and specificity (Kamphuis et al., 2010, Mammadova et al., 2013, Panchal et al., 2016), AUROC (Panchal et al., 2016), success rate (Mammadova et al., 2013) and the diagnostic odds ratio (Panchal et al., 2016). Out of the studies looking at transition period health management Wisnieski et al. (2019) used logistic regression models to predict metabolic stress and reported sensitivity, specificity, AUROC and well as the positive and negative predictive values. Similarly, Vergara et al. (2014) in a study exploring postpartum issues in dairy cows reported the predictive models' AUROC, sensitivity and specificity. In a meta-analysis Shine and Murphy (2021) determined that in 85 studies centred on classification problems, the most frequently utilized evaluation metric was classification accuracy (77%), followed

by recall (66%), specificity (49%), PPV (48%), F1 Score (27%), AUROC (26%), NPV (15%), Cohen's Kappa (12%), false positives (FP) (9%), and false negatives (FN) (6%). It is evident that, especially recently F1-score is more likely to be included in the assessment of the predictive value of their models (Avizheh et al., 2023, Hunter et al., 2021, Keshavarzi et al., 2020, Sturm et al., 2021, de Oliveira et al., 2021, Rodriguez Alvarez et al., 2018, Rodriguez Alvarez et al., 2019, Smith et al., 2016, Vidal et al., 2023, Wang et al., 2020, Carslake et al., 2021, Cantor et al., 2022, Dineva and Atanasova, 2023, Dutta et al., 2015, Ghaffari et al., 2019, Hemalatha et al., 2021, Hyde et al., 2020, Luo et al., 2023, Shafiullah et al., 2019, Sturm et al., 2020, Vázquez-Diosdado et al., 2023, Williams et al., 2019, Wang et al., 2023), however Cohen's Kappa while slowly picking up is still lower on the preference of researchers as a metric of choice (Hassan et al., 2009, Balasso et al., 2021, Hyde et al., 2020, Maciel-Guerra et al., 2021, Esener et al., 2021, Volkmann et al., 2021, Sadeghi et al., 2022, Nagy et al., 2023, Barney et al., 2023, Siachos et al., 2024, Sturm et al., 2020, Imada et al., 2024).

So, it becomes evident that sensitivity, specificity and accuracy up until recent years accuracy were used primarily when reporting predictive models in veterinary medicine. And while the combination of specificity and sensitivity usually helps gain a relatively good understanding of how a model performs in both classes, kappa appears to be more likely to give a definitive picture.

Balanced accuracy could also be more effectively used in imbalanced datasets as suggested by Brodersen et al. (2010). In contrast with regular accuracy that only takes into account the minority class based on how much of the total dataset it consists of, balanced accuracy treats both classes equally and thus

emphasizes lack of predictiveness even in a class that is underrepresented. In most of the predictive models in this study, balanced accuracy was just over 50%, so was little better than would be expected from completely random predictions.

Another metric that could be considered and is used in some of the aforementioned studies is the AUROC. However, as reviewed by Lobo et al. (2008), it is not always an appropriate metric for various reasons, such as the bias of the mean probabilities towards the most frequent class (Hosmer et al. 1980). While it provides a single scalar value that summarizes model performance, it does not account for the distribution of classes in the dataset, which can lead to misleading interpretations in imbalanced scenarios (King et al., 2021, Hancock et al., 2023). In imbalanced datasets, where one class significantly outnumbers the other, the AUROC can give an overly optimistic view of a model's performance. This is because the metric can be inflated by the model's ability to correctly classify the majority class while neglecting the minority class (Bednarski et al., 2022). Hence, it is likely that the kappa value is perhaps the best overall measure of model predictiveness, especially when working with imbalanced data. We should, however, also mention a drawback, that has potentially prevented the wider adoption of kappa as a sole evaluation metric so far, and that is the arbitrary nature of its scale which while enabling the comparison of models with each other, might make their individual interpretation a bit difficult for the end user.

Currently, other metrics dominate the field of veterinary epidemiology as the final selector and judge of predictive model and thus this work supports the

adoption of the alternatives mentioned above especially when the classes are of unequal size. In any case, the use of a good combination of metrics (at least one for each class) is advised.

7.2.4 Limitations of the study

As stated before, one of the major limitations of the study is the convenience sampling that took place in order to gather the dataset. The descriptive statistics presented in Chapter 3 indicate that our herds were almost all following a relatively high input system, with cows calving year-round and likely mostly housed. This could lead towards biased results that reflect the situation in particular types of herds only and not translate in models that could be widely used for all types of dairy cow herds. Dairy farming in the UK is polarising so that farms either pursue an all year round calving system (maximising milk yield, while accepting a higher cost of production per litre) or pursue a block calving system (spring or autumn calving, long grazing, season outdoors, mostly grass based diet, lower yields) where the focus is on minimising costs and accepting level of production will be lower (AHDB, 2017). This could be bypassed in the case of external validation, which unfortunately was not possible during our research. When however, taking into account that the majority of models did not hold significant predictive power this seems like less of a problem. Even so, with the individual lactation metritis model predicting on an aggregated level which showed some predictive potential, it is still valuable to identify this model even it applies for herds similar to those included in our sample. It should also be noted that while it is very likely that the models presented in this thesis would not generalise well to the system of low cost-low productivity cows, the

prevalence of e.g. diseases and subfertility is probably much lower in these systems anyway so predictions would be less important.

Furthermore, the nature of the data collection could potentially have caused some unreliable variables in our study. As stated in previous Chapters, scoring occasions generally occurred fortnightly for each herd which could cause variation in terms of length from day of scoring to the day of calving. This in turn means that scores of rumen-fill for example that were on the calving date would be systematically lower than those from a few days earlier or later, and yet they would be weighted equally in the analysis. Additionally, due to the nature of the data and them being gathered from many different sources and initial datasets, when coming together during the cleaning, there were occasionally variables without any information for the majority of the finalised data points. This certainly led to loss of information that could have potentially improved the predictive power of the models. However, as stated before there are so many complex and immeasurable terms affecting the outcomes that these few variable contributions were likely not that great. It should also be noted that there is likely measurement inaccuracy between observers and occasions, for at least some of the variables, though probably negligible since all the assessors had been trained.

7.2.5 Possible Future Research

This study could be the basis for future research. Most importantly, for the individual level metritis model making predictions on an aggregated level outcome which appeared to be able to explain over two thirds of the outcome's variation, external validation is going to be necessary to determine the predictiveness of the model on a wider context, when including herd with

various characteristics that may not have been present in our thesis. This would ensure the possibility for a widely used assessed tool available for all herd managers to use.

Another angle that studies could take is the use of individual level models for aggregated predictions, at least when it comes to binary outcomes. As seen in our study, binary models such as the one for metritis mentioned above or even the collective disease status showed a significant improvement when converting the results to a percentage and comparing them to the aggregated results of the actual outcome. This is not unheard of in binary outcomes, probably since the approximation of a value allows more flexibility rather than choosing only one of two variables, which also explains why this improvement is not guaranteed in already continuous outcomes. This approach could potentially create value out of algorithms that appeared to be underwhelming at first, like the individual model presented for metritis. A lot of already studied models could be revisited with renewed potential.

Finally, an important consideration for future research aiming to improve the predictive capability of their models is the use of sensor data. Sensor data could have significantly enhanced the quality of our dataset by providing detailed and accurate measurements that are otherwise unattainable. The potential importance of on-animal sensors, due to them making the collection of large amounts of data accessible and therefore enabling the on-farm practical application of predicted models, has been highlighted for the dairy industry (Hudson et al., 2018). Their use is becoming more mainstream especially with the existence of projects such as CowManager, which is a precision

livestock monitoring system, which relies on sensors attached on ear tags that collect data in real-time.

Multiple studies, especially more recent ones, have utilized sensors for their data collection with good results (Benaissa et al., 2019b, Carslake et al., 2021, Chung et al., 2020, Lardy et al., 2023, Post et al., 2020, Post et al., 2021, Sturm et al., 2020, Vázquez-Diosdado et al., 2023). The majority of these studies only include a small sample of cows, especially compared to ours, with only Sturm et al. (2020) achieving a final dataset of 671 cows, Post et al. (2021) including 348 cows and one of 4 datasets utilised by Lardy et al. (2020) reaching 300 cows. Post et al. (2020) collected data from 167 cows, with Vázquez-Diosdado et al. (2023) and Benaissa et al. (2019b) having lower sample sizes (82 and 31 respectively). Finally, Sturm et al. (2020) only included 3 cows in their study. With the increasing adoption of systems like CowManager, future research should prioritize large-scale studies that leverage high-quality sensor data. Such efforts could produce robust predictive models addressing critical outcomes, including disease management, production efficiency, and reproduction in dairy cows.

7.3 Conclusions

This thesis has added to existing knowledge in a number of ways. Firstly, it determined that, at least for our combination of cow level and environmental level variables, out of all disease, reproduction and production outcomes, metritis was the one that had the most potential in terms of accurate predictions. Furthermore, it was showcased that for binary variables in particular, the individual level model results, even if not predictive have the potential to be aggregated to a group level and approximate the group's averaged outcome, providing useful results on that aggregated level. This could indicate that group predicted prevalence of metritis (which in our case was the most predictive model) might be a useful measure for farmers to monitor over time as an overall transition "success" index. In regards with classification models and especially when dealing with imbalanced datasets (which is common in biological outcomes), the importance of reporting the correct metrics was demonstrated. As shown the kappa was among the most useful metrics, being able to capture the difference in predictive performance between the two classes and was proposed to be more widely used for similar situations.

Finally, the overall importance of predictive compared to inferential modelling in terms of making herd-level decisions was emphasised. Inferential modelling is useful in order to look at and understand relationships between predictors and outcomes, while predictive modelling specifically aims at making accurate predictions, without explaining these relationships. Therefore, relying on the former for predictions would not be sensible since the latter are better at that specific task. Again, our models showcased this exact thing with th inferential

models describing strong relationships and yet the predictive ones not producing accurate predictions.

Transition management is really key to successful dairy farming, hence monitoring it, using machine learning methods, could benefit farmers greatly. It is evident from our work that such a thing might be possible in the future, potentially with measures like the group predicted prevalence of metritis over time. However, more work is needed in order to determine and assess such measures.

References

- ABBAS, F., CAI, Z., SHOAIB, M., IQBAL, J., ISMAIL, M., ARIFULLAH, & ALBESHR, M. F. 2024. Machine learning models for water quality prediction: a comprehensive analysis and uncertainty assessment in mirpurkhas, sindh, Pakistan. *Water*, 16(7), 941.
- ABREU, S., 2019. Automated architecture design for deep neural networks, *arXiv*.1908.10714.
- ADHIKARI, M., LONGMAN, R.J., GIAMBELLUCA, T.W., LEE, C.N., HE, Y.,2022. Climate change impacts shifting landscape of the dairy industry in Hawai'i. *Transl Anim Sci*. txac064.
- ADRIAENS, I., HUYBRECHTS, T., AEMOUTS, B., GEERINCKX, K., PIEPERS, S., DE KETELAERE, B., SAEYS, W. 2018. Method for short-term prediction of milk yield at the quarter level to improve udder health monitoring, *J Dairy Sci*, Volume 101, Issue 11, 10327-10336.
- AGGARWAL, C. C., HINNEBURG, A., & KEIM, D. A. 2001. On the surprising behavior of distance metrics in high dimensional space, in Database Theory — ICDT 2001, vol. 1973 of Lecture Notes in Computer Science, Springer, 420-434.
- AGUIAR, G. F., BATISTA, B. L., RODRIGUES, J. L., SILVA, L. R., CAMPIGLIA, A. D., BARBOSA, R. M. & BARBOSA, F., JR. 2012. Determination of trace elements in bovine semen samples by inductively coupled plasma mass spectrometry and data mining techniques for identification of bovine class. *J Dairy Sci*, 95, 7066-73.
- AHDB, 2017, Delivering a more competitive industry through optimal dairy systems,
<https://projectblue.blob.core.windows.net/media/Default/Imported%20Publication%20Docs/Delivering%20a%20more%20competitive%20indus>

- try%20through%20optimal%20dairy%20systems.pdf, Accessed 17 January 2023
- AHDB, 2023a, Body Condition Scoring flow chart, <https://ahdb.org.uk/knowledge-library/body-condition-scoring-flow-chart> Accessed 10 February 2023
- AHDB, 2023b, UK milk productivity: The global context, <https://ahdb.org.uk/news/uk-milk-productivity-the-global-context>. Accessed 14 January 2023
- ALEX, A., COLLIER, J., HADSELL, D., & COLLIER, R. 2015. Milk yield differences between 1x and 4x milking are associated with changes in mammary mitochondrial number and milk protein gene expression, but not mammary cell apoptosis or socs gene expression. *J Dairy Sci*, 98(7), 4439-4448.
- ALIN, A. 2010. Multicollinearity. *WIREs Computational Statistics*, 2(3), 370-374.
- ALIZADEHI, G., VAFAKHAH, M., AZARMSA, A., and TOBARI, M., 2011, Using an artificial neural network to model monthly shoreline variations, 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Deng Feng, China, 2011, pp. 4893-4896
- ALLEN, M. S. & BRADFORD, B. J. 2009. Control of eating by hepatic oxidation of fatty acids. A note of caution. *Appetite*, 53, 272-3; author reply 274-6.
- ALSAAOD, M., ROMER, C., KLEIMANNS, J., HENDRIKSEN, K., ROSE-MEIERHOFER, S., PLUMER, L., BUSCHER, W. 2012. Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. *Appl. Anim. Behav. Sci.* 2012, 142, 134–141
- ALZAHAL, O., MCGILL, H., KLEINBERG, A., HOLLIDAY, J. I., HINDRICHSEN, I. K., DUFFIELD, T. F. & MCBRIDE, B. W. 2014. Use of a direct-fed microbial product as a supplement during the transition period in dairy cattle. *J Dairy Sci*, 97, 7102-14.
- ANDREU-VAZQUEZ, C., GARCIA-ISPIERTO, I., GANAU, S., FRICKE, P. M. & LOPEZ-GATIUS, F. 2012. Effects of twinning on the subsequent

- reproductive performance and productive lifespan of high-producing dairy cows. *Theriogenology*, 78, 2061-70.
- ANGLART, D., HALLEN-SANDGREN, C., EMANUELSON, U., & RONNEGARD, L., 2020. Comparison of methods for predicting cow composite somatic cell counts. *J. Dairy Sci.*, 103, 8433–8442.
- ANSARI-MAHYARI, S., OJALI, M. R., FORUTAN, M., RIASI, A. & BRITO, L. F. 2019. Investigating the genetic architecture of conception and non-return rates in Holstein cattle under heat stress conditions. *Tropical Animal Health and Production*.
- ARICI, Y., OZKAN, M., & KOCABAS, Z. 2023. Effects of multicollinearity on type i error rate and test power of binary logistic regression model: a simulation study. *Medicine Science | International Medical Journal*, 12(4), 1180.
- ATKINSON, O. 2009. Guide to the rumen health visit. *Practice*, 31, 314–325.
- AVIZHEH, M., DADPASAND, M., DEHNAVI, E., & KESHAVARZI, H. 2023. Application of machine-learning algorithms to predict calving difficulty in Holstein dairy cattle. *Animal Production Science*, 63(11), 1095-1104.
- AYINDE, O. O. A. K. and NWOSU, U. I. 2021. Solving multicollinearity problem in linear regression model: the review suggests new idea of partitioning and extraction of the explanatory variables. *Journal of Mathematics and Statistics Studies*, 2(1), 12-20.
- AYRANCI, A., ATAY, S., & YILDIRIM, T. 2021. Speaker accent recognition using mfcc feature extraction and machine learning algorithms. *International Journal of Advances in Engineering and Pure Sciences*, 33, 17-27.
- AZAWI, O. I. 2008. Postpartum uterine infection in cattle. *Anim Reprod Sci*, 105, 187-208.
- BARKER, Z. E., LEACH, K. A., WHAY, H. R., BELL, N. J. & MAIN, D. C. 2010. Assessment of lameness prevalence and associated risk factors in dairy herds in England and Wales. *J Dairy Sci*, 93, 932-41.
- BATES, A. J. Saldias, B. 2019. A comparison of machine learning and logistic regression in modelling the association of body condition score and submission rate, *Preventive Veterinary Medicine*, 171, 104765, 0167-5877.

- BAYKAN, N. & YILMAZ, N. 2011. A mineral classification system with multiple artificial neural network using k-fold cross validation. *Mathematical and Computational Applications*, 16(1), 22-30.
- BEDNARSKI, B., SINGH, A., ZHANG, W., JONES, W., NAEIM, A., & RAMEZANI, R., 2022. Temporal convolutional networks and data rebalancing for clinical length of stay and mortality prediction. *Scientific Reports*, 12(1).
- BEKUMA, A. 2019. Combating negative effect of negative energy balance in dairy cows: comprehensive review. *Approaches in Poultry Dairy & Veterinary Sciences*, 6(2).
- BELL, A. W. 1995. Regulation of organic nutrient metabolism during transition from late pregnancy to early lactation. *J Anim Sci*, 73, 2804-19.
- BENAISSA, S., TUYTTENS, F. A. M., PLETS, D., CATTRYSSSE, H., MARTENS, L., VANDAELE, L., JOSEPH & W., SONCK, B. 2019a. Classification of ingestive-related cow behaviours using RumiWatch halter and neck-mounted accelerometers. *Appl. Anim. Behav. Sci.*, 211, 9–16.
- BENAISSA, S., TUYTTENS, F. A. M., PLETS, D., DE PESSEMIER, T., TROGH, J., TANGHE, E., MARTENS, L., VANDAELE, L., VAN NUFFEL, A., JOSEPH, W. & SONCK, B. 2017. On the use of on-cow accelerometers for the classification of behaviours in dairy barns. *Research in Veterinary Science*. 125, 425–433.
- BENAISSA, S., TUYTTENS, F. A. M., PLETS, D., DE PESSEMIER, T., TROGH, J., TANGHE, E., MARTENS, L., VANDAELE, L., VAN NUFFEL, A., JOSEPH, W., & SONCK, B. 2019b. On the use of on-cow accelerometers for the classification of behaviours in dairy barns. *Res. Vet. Sci.*, 125, 425–433.
- BERRY, D. P., BUCKLEY, F. & DILLON, P. 2007. Body condition score and live-weight effects on milk production in Irish Holstein-Friesian dairy cows. *Animal*, 1, 1351-9.
- BERRY, D. P., BUCKLEY, F., DILLON, P., EVANS, R. D., RATH, M. & VEERKAMP, R. F. 2003. Genetic relationships among body condition score, body weight, milk yield, and fertility in dairy cows. *J Dairy Sci*, 86, 2193-204.

- BERRY, D. P., VEERKAMP, R. F. & DILLON, P. 2006. Phenotypic profiles for body weight, body condition score, energy intake, and energy balance across different parities and concentrate feeding levels. *Livestock Science*, 104, 1-12.
- BERTONI, G., TREVISI, E. & LOMBARDELLI, R. 2009. Some new aspects of nutrition, health conditions and fertility of intensively reared dairy cows. *Italian Journal of Animal Science*, 8, 491-518.
- BICALHO, R. C., CHEONG, S. H., GALVAO, K. N., WARNICK, L. D. & GUARD, C. L. 2007. Effect of twin birth calvings on milk production, reproductive performance, and survival of lactating cows. *J Am Vet Med Assoc*, 231, 1390-7.
- BIFFANI, S., PAUSCH, H., SCHWARZENBACHER, H. & BISCARINI, F. 2017. The effect of mislabeled phenotypic status on the identification of mutation-carriers from SNP genotypes in dairy cattle. *BMC research notes*, 10, 230-230.
- BONFATTI, V. TURNER, S.-A. KUHN-SHERLOCK, B. LUKE, T.D.W. HO, P.N. PHYN, C.V.C. & PRYCE, J.E. 2019. Prediction of blood β -hydroxybutyrate content and occurrence of hyperketonemia in early-lactation, pasture-grazed dairy cows using milk infrared spectra, *Journal of Dairy Science*, 102, 7, 6466-6476.
- BORCHERS, M. R., CHANG, Y. M., PROUDFOOT, K. L., WADSWORTH, B. A., STONE, A. E. & BEWLEY, J. M. 2017. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *J Dairy Sci*, 100, 5664-5674.
- BORDES, A., ERTEKIN, S., WESTON, J., & BOTTOU, L., 2005, Fast kernel classifiers with online and active learning, *Journal of machine learning research*, 6, 1579–1619.
- BORGHART, G. M., O'GRADY, L. E. & SOMERS, J. R. 2021. Prediction of lameness using automatically recorded activity, behavior and production data in post-parturient Irish dairy cows. *Ir. Vet. J.*, 74, 4.
- BOVO, M., AGRUSTI, M., BENNI, S., TORREGGIANI, D., TASSINARI, P. 2021. Random Forest Modelling of Milk Yield of Dairy Cows under Heat Stress Conditions. *Animals (Basel)*. 11(5):1305.

- BRADFORD, B. J., YUAN, K., FARNEY, J. K., MAMEDOVA, L. K. & CARPENTER, A. J. 2015. Invited review: Inflammation during the transition to lactation: New adventures with an old flame. *J Dairy Sci*, 98, 6631-50.
- BRADLEY, A. J., LEACH, K. A., BREEN, J. E., GREEN, L. E. & GREEN, M. J. 2007. Survey of the incidence and aetiology of mastitis on dairy farms in England and Wales. *Veterinary Record*, 160, 253.
- BRAMER, M., 2007, Principles of data mining Springer, vol. 131.
- BRAMLEY, E. COSTA, N.D. FULKERSON, W.J. LEAN, I.J. 2013. Associations between body condition, rumen fill, diarrhoea and lameness and ruminal acidosis in Australian dairy herds. *N. Z. Vet. J.*, 61, 323–329.
- BRAND, W., WELLS, A. T., SMITH, S. L., DENHOLM, S. J., WALL, E., COFFEY, M. P., 2021. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. *J. Dairy Sci.* 104, 4980–4990.
- BREIMAN, L., 1996. Bagging predictors. *Machine Learning* 26(2), 123–140
- BREIMAN, L., 2001. Random forests. *Machine Learning*, 45, 5–32.
- BROUCEK, J., UHRINCAT, M., MIHINA, S., SOCH, M., MREKAJOVA, A. & HANUS, A. 2017. Dairy Cows Produce Less Milk and Modify Their Behaviour during the Transition between Tie-Stall to Free-Stall. *Animals : an open access journal from MDPI*, 7, 16.
- BRUUN, J., ERSBOLL, A. K. & ALBAN, L. 2002. Risk factors for metritis in Danish dairy cows. *Prev Vet Med*, 54, 179-90.
- BURFEIND, O., SEPULVEDA, P., VON KEYSERLINGK, M. A., WEARY, D. M., VEIRA, D. M. & HEUWIESER, W. 2010. Technical note: Evaluation of a scoring system for rumen fill in dairy cows. *J Dairy Sci*, 93, 3635-40.
- BUTLER, S. 2014. Nutritional management to optimize fertility of dairy cows in pasture-based systems. *Animal*, 8, 15-26.
- BUTLER, W. R. & SMITH, R. D. 1989. Interrelationships between energy balance and postpartum reproductive function in dairy cattle. *J Dairy Sci*, 72, 767-83.
- C. BARTLETT, P., KIRK, J. H., WILKE, M. A., KANEENE, J. B. & MATHER, E. C. 1986. Metritis complex in Michigan Holstein-Friesian cattle: incidence,

- descriptive epidemiology and estimated economic impact. *Preventive Veterinary Medicine*, 4, 235-248.
- CABRERA, V.E.,2014. Economics of fertility in high-yielding dairy cows on confined TMR systems. *Animal*. Suppl 1:211-21.
- CABRERA, V. E. and FRICKE, P. 2021. Economics of twin pregnancies in dairy cattle. *Animals*, 11(2), 552.
- CAI, T. Q., WESTON, P. G., LUND, L. A., BRODIE, B., MCKENNA, D. J. & WAGNER, W. C. 1994. Association between neutrophil functions and periparturient disorders in cows. *Am J Vet Res*, 55, 934-43.
- CAIXETA, L. S., HERMAN, J. A., JOHNSON, G. W. & MCART, J. A. A. 2018. Herd-Level Monitoring and Prevention of Displaced Abomasum in Dairy Cattle. *Vet Clin North Am Food Anim Pract*, 34, 83-99.
- CAIXETA, L. S., & OMONTESE, B. O. 2021. Monitoring and Improving the Metabolic Health of Dairy Cows during the Transition Period. *Animals : an open access journal from MDPI*, 11(2), 352.
- CALDERON, D. F. & COOK, N. B. 2011. The effect of lameness on the resting behavior and metabolic status of dairy cattle during the transition period in a freestall-housed dairy herd. *J Dairy Sci*, 94, 2883-2894.
- CAMERON, R. E., DYK, P. B., HERDT, T. H., KANEENE, J. B., MILLER, R., BUCHOLTZ, H. F., LIESMAN, J. S., VANDEHAAR, M. J. & EMERY, R. S. 1998. Dry cow diet, management, and energy balance as risk factors for displaced abomasum in high producing dairy herds. *J Dairy Sci*, 81, 132-9.
- CANTOR, M. C., CASELLA, E., SILVESTRI, S., RENAUD, D. L., & COSTA, J. H. 2022. Using machine learning and behavioral patterns observed by automated feeders and accelerometers for the early indication of clinical bovine respiratory disease status in preweaned dairy calves. *Frontiers in Animal Science*, 3, 852359.
- CARAVIELLO, D. Z., WEIGEL, K. A., CRAVEN, M., GIANOLA, D., COOK, N. B., NORDLUND, K. V., FRICKE, P. M. & WILTBANK, M. C. 2006. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. *J Dairy Sci*, 89, 4703-22.

- CARSLAKE, C., VAZQUEZ-DIOSDADO, J. A. & KALER, J., 2021 Machine learning algorithms to classify and quantify multiple behaviours in dairy calves using a sensor—moving beyond classification in precision livestock. *Sensors*, 21, 88.
- CAUWENBERGHS, G. & POGGIO, T., 2000, Incremental and decremental support vector machine learning, in *Adv. Neural Information Processing Systems (NIPS*2000)*, 13, 409–415.
- CERNEK, P., BOLLIG, N., ANKLAM, K. & DOPFER, D.. Hot topic: Detecting digital dermatitis with computer vision. *J. Dairy Sci.* 2020, 103, 9110–9115.
- CEVIK, K. K. 2020. Deep Learning Based Real-Time Body Condition Score Classification System. *IEEE Access*, 8, 213950–213957.
- CHANDRA, B. & PAUL VERGHESE, P. 2009. Moving towards efficient decision tree construction. *Information Sciences*, 179(8), 1059-1069.
- CHAPWANYA, A., MEADE, K. G., FOLEY, C., NARCIANDI, F., EVANS, A. C., DOHERTY, M. L., CALLANAN, J. J. & O'FARRELLY, C. 2012. The postpartum endometrial inflammatory response: a normal physiological event with potential implications for bovine fertility. *Reprod Fertil Dev*, 24, 1028-39.
- CHELOTTI, J. O., VANRELL, S. R., GALLI, J. R., GIOVANINI, L. L., & RUFINER, H. L., 2018. A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. *Comput. Electron. Agric.*, 145, 83–91.
- CHEN, G., 2018, A Method for the Measurement of Temperature Based on Neural Network PID, *Proceedings of the 2018 3rd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018)}*, 387-390.
- CHEN G. & JI C., 2016, A Method for the Measurement of Temperature Based on Multisensor Data Fusion}, *Proceedings of the 2016 4th International Conference on Advanced Materials and Information Technology Processing (AMITP 2016)}*, 457-460.
- CHEN, H., HU, S., HUA, R., & ZHAO, X. 2021. Improved naive bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, 2021(1).

- CHEN, C., MURPHY, N.R., PARISA, K., SCULLEY, D., UNDERWOOD, T., 2022. *Reliable Machine Learning*, O'Reilly Media, Inc.
- CHUNG, H., LI, J., KIM, Y., VAN OS, J. M. C., BROUNTS, S. H. & CHOI, C. Y. 2020. Using implantable biosensors and wearable scanners to monitor dairy cattle's core body temperature in real-time. *Comput. Electron. Agric.*, 174, 105453.
- CLARKSON, M. J., DOWNHAM, D. Y., FAULL, W. B., HUGHES, J. W., MANSON, F. J., MERRITT, J. B., MURRAY, R. D., RUSSELL, W. B., SUTHERST, J. E. & WARD, W. R. 1996. Incidence and prevalence of lameness in dairy cattle. *Vet Rec*, 138, 563-7.
- COCKBURN M. 2020. Review: Application and Prospective Discussion of Machine Learning for the Management of Dairy Farms. *Animals*. 10(9):1690.
- COFFEY, M. P., SIMM, G., OLDHAM, J. D., HILL, W. G. & BROTHERSTONE, S. 2004. Genotype and diet effects on energy balance in the first three lactations of dairy cows. *J Dairy Sci*, 87, 4318-26.
- COLEMAN, D. A., THAYNE, W. V. & DAILEY, R. A. 1985. Factors affecting reproductive performance of dairy cows. *J Dairy Sci*, 68, 1793-803.
- COLLINS, G. S., OGUNDIMU, E. O., & ALTMAN, D. G. 2015. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine*, 35(2), 214-226.
- CONDE, V. A., SILVA VALENTE, G. D. F., & MINIGHIN, E. C. 2020. Milk fraud by the addition of whey using an artificial neural network. *Cienc. Rural*, 50, 1–8.
- CONTRERAS, L. L., RYAN, C. M. & OVERTON, T. R. 2004. Effects of dry cow grouping strategy and prepartum body condition score on performance and health of transition dairy cows. *J Dairy Sci*, 87, 517-23.
- CONTRERAS, G. A., STRIEDER-BARBOZA, C., SOUZA, J. d., GANDY, J., MAVANGIRA, V., LOCK, A., & SORDILLO, L. M. 2017. Periparturient lipolysis and oxylipid biosynthesis in bovine adipose tissues. *Plos One*, 12(12), e0188621.
- COOK, J. G. & GREEN, M. J. 2016. Use of early lactation milk recording data to predict the calving to conception interval in dairy herds. *J Dairy Sci*, 99, 4699-4706.

- COOK, N. B. 2003. Prevalence of lameness among dairy cattle in Wisconsin as a function of housing type and stall surface. *J Am Vet Med Assoc*, 223, 1324-8.
- COOK, N. B., MENTINK, R. L., BENNETT, T. B. & BURGI, K. 2007. The effect of heat stress and lameness on time budgets of lactating dairy cows. *J Dairy Sci*, 90, 1674-82.
- COOK, N. B., NORDLUND, K. V. & OETZEL, G. R. 2004. Environmental Influences on Claw Horn Lesions Associated with Laminitis and Subacute Ruminant Acidosis in Dairy Cows. *J Dairy Sci*, 87, E36-E46.
- CORREA, M. T., ERB, H. & SCARLETT, J. 1993. Path analysis for seven postpartum disorders of Holstein cows. *J Dairy Sci*, 76, 1305-12.
- CRANINX, M., FIEVEZ, V., VLAEMINCK, B., DE BAETS, B. 2008. Artificial neural network models of the rumen fermentation pattern in dairy cattle. *Comput. Electron. Agric.*, 60, 226–238.
- CRISTIANINI, N., & SHAW-TAYLOR, J. 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.
- CURTIS, C. R., ERB, H. N., SNIFFEN, C. J., SMITH, R. D. & KRONFELD, D. S. 1985. Path analysis of dry period nutrition, postpartum metabolic and reproductive disorders, and mastitis in Holstein cows. *J Dairy Sci*, 68, 2347-60.
- DAETZ, R., CUNHA, F., BITTAR, J. H. J., MAGALHAES, F. d. C., MAEDA, Y., SANTOS, J., & GALVAO, K. N. 2016. Clinical response after chitosan microparticle administration and preliminary assessment of efficacy in preventing metritis in lactating dairy cows. *J Dairy Sci*, 99(11), 8946-8955.
- DALLAGO, G. M., DE FIGUEIREDO, D. M., ANDRADE, P. C. D. R., DOS SANTOS, R. A., LACROIX, R., SANTSCHI, D. E., LEFEBVRE, D. M. 2019. Predicting first test day milk yield of dairy heifers. *Comput. Electron. Agric.*, 166, 105032.
- DANICKE, S., MEYER, U., KERSTEN, S. & FRAHM, J. 2018. Animal models to study the impact of nutrition on the immune system of the transition cow. *Res Vet Sci*, 116, 15-27.

- DANN, H.M., MORIN, D.E., BOLLERO, G.A., MURPHY, M.R., DRACKLEY, J.K. 2005. Prepartum intake, postpartum induction of ketosis, and periparturient disorders affect the metabolic status of dairy cows. *J Dairy Sci.* 3249-64.
- DAROS, R. R., HOTZEL, M. J., BRAN, J. A., LEBLANC, S. J. & VON KEYSERLINGK, M. A. G. 2017. Prevalence and risk factors for transition period diseases in grazing dairy cows in Brazil. *Prev Vet Med*, 145, 16-22.
- DE AMICIS, I., VERONESI, M. C., ROBBE, D., GLORIA, A. & CARLUCCIO, A. 2018. Prevalence, causes, resolution and consequences of bovine dystocia in Italy. *Theriogenology*, 107, 104-108.
- DEFRA, 2022, United Kingdom milk prices and composition of milk: October 2022, <https://www.gov.uk/government/statistics/uk-milk-prices-and-composition-of-milk/united-kingdom-milk-prices-and-composition-of-milk-september-2022> . Accessed 14 January 2023
- DENHOLM, S. J., BRAND, W., MITCHELL, A. P., WELLS, A. T., KRZYZELEWSKI, T., SMITH, S. L., WALL, E. & COFFEY, M. P. 2020. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. *J. Dairy Sci.*, 103, 9355–9367.
- DERVISHI, E., ZHANG, G., HAILEMARIAM, D., DUNN, S. M., & AMETAI, B. N. 2016. Occurrence of retained placenta is preceded by an inflammatory state and alterations of energy metabolism in transition dairy cows. *Journal of Animal Science and Biotechnology*, 7(1).
- DETTMANN, F., WARNER, D., BUITENHUIS, B., KARGO, M., KJELDSSEN, A. M. H., NIELSEN, N. H., LEFEBVRE, D. M., & SANTSCH, D. E. 2020. Fatty acid profiles from routine milk recording as a decision tool for body weight change of dairy cows after calving. *Animals*, 10, 1958.
- DEZFOULI M.M., EFTEKHARI Z., SADEGHIAN S., BAHOUNAR A., JELOUDARI M. 2013. Evaluation of hematological and biochemical profiles in dairy cows with left displacement of the abomasum. *Comp Clin Pathol.*, 22:175-179.

- DHIMAN, P., MA, J., NAVARRO, C. L. A., SPEICH, B., BULLOCK, G. S., DAMEN, J. A., & COLLINS, G. S. 2022. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. *Diagnostic and Prognostic Research*, 6(1).
- DHOBLE, A. S., RYAN, K. T., LAHIRI, P., CHEN, M., PANG, X., CARDOSO, F. C., BHALERAO, K. D. 2019. Cytometric fingerprinting and machine learning (CFML): A novel label-free, objective method for routine mastitis screening. *Comput. Electron. Agric.*, 162, 505–513.
- DÍAZ, C., CARABAÑO, M. J., RAMÓN, M., PÉREZ-GUZMÁN, M. D., MOLINA, A. & SERRADILLA, J. M. 2017. BREEDING AND GENETICS SYMPOSIUM: Breeding for resilience to heat stress effects in dairy ruminants. A comprehensive review¹. *Journal of Animal Science*, 95, 1813-1826.
- DIETTERICH, T.. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, *Machine Learning*, 1–22.
- DINEVA, K., & ATANASOVA, T. 2023. Health Status Classification for Cows Using Machine Learning and Data Management on AWS Cloud. *Animals*, 13(20), 3254.
- DOBSON, H., WALKER, S. L., MORRIS, M. J., ROUTLY, J. E. & SMITH, R. F. 2008. Why is it getting more difficult to successfully artificially inseminate dairy cows? *Animal*, 2, 1104-1111.
- DOHOO, I. R. & WAYNE MARTIN, S. 1984. Disease, production and culling in Holstein-Friesian cows V. Survivorship. *Preventive Veterinary Medicine*, 2, 771-784.
- DOLECHECK, K. & BEWLEY, J. 2018. Animal board invited review: Dairy cow lameness expenditures, losses and total cost. *Animal*, 12, 1462-1474.
- DOLECHECK, K. A., SILVIA, W. J., HEERSCHE, G., JR., CHANG, Y. M., RAY, D. L., STONE, A. E., WADSWORTH, B. A. & BEWLEY, J. M. 2015. Behavioral and physiological changes around estrus events identified using multiple automated monitoring technologies. *J Dairy Sci*, 98, 8723-31.

- DÓREA, J. R. R., ROSA, G. J. M., WELD, K. A. & ARMENTANO, L. E. 2018. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *J Dairy Sci*, 101, 5878-5889.
- DOUPHRATE, D. I., FETHKE, N. B., NONNENMANN, M. W., RODRIGUEZ, A. & DE PORRAS, D. G. R. 2019. Reliability of observational- and machine-based teat hygiene scoring methodologies. *J. Dairy Sci.*, 102, 7494–7502.
- DRACKLEY, J. K. 1999. ADSA Foundation Scholar Award. Biology of dairy cows during the transition period: the final frontier? *J Dairy Sci*, 82, 2259-73.
- DRACKLEY, J. K., OVERTON, T. R. & DOUGLAS, G. N. 2001. Adaptations of Glucose and Long-Chain Fatty Acid Metabolism in Liver of Dairy Cows during the Periparturient Period. *J Dairy Sci*, 84, E100-E112.
- DRESSLER, W. W., BALIEIRO, M. C., ARAUJO, L. F. d., SILVA, W. A., & SANTOS, J. E. d. 2016. The interaction of cultural consonance and a polymorphism in the 2a serotonin receptor in relation to depression in brazil: failure to replicate previous findings. *American Journal of Human Biology*, 28(6), 936-940.
- DUDOUETE. 1982. Courbe de lactation théorique de la chèvre et applications (Theoretical lactation curve of the goat and its applications). *Point Vet.* 14:53–61.
- DUNCAN, N. W. & NORTHOFF, G. 2013. Overview of potential procedural and participant-related confounds for neuroimaging of the resting state. *J. Psychiatry Neurosci.* 38, 84–96.
- DUTTA, R., SMITH, D., RAWNSLEY, R., BISHOP-HURLEY, G., HILLS, J., TIMMS, G. & HENRY, D. 2015. Dynamic cattle behavioural classification using supervised ensemble classifiers. *Comput. Electron. Agric.*, 111, 18–28.
- EBRAHIMIE, E., EBRAHIMI, F., EBRAHIMI, M., TOMLINSON, S. & PETROVSKI, K. R. 2018a. Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. *Computers and Electronics in Agriculture*, 147, 6-11.

- EBRAHIMIE, E., EBRAHIMI, F., EBRAHIMI, M., TOMLINSON, S. & PETROVSKI, K. R. 2018b. A large-scale study of indicators of sub-clinical mastitis in dairy cattle by attribute weighting analysis of milk composition features: highlighting the predictive power of lactose and electrical conductivity. *J Dairy Res*, 85, 193-200.
- EBRAHIMIE, E., MOHAMMADI-DEHCESHMEH, M., LAVEN, R. & PETROVSKI, K. R. 2021. Rule Discovery in Milk Content towards Mastitis Diagnosis: Dealing with Farm Heterogeneity over Multiple Years through Classification Based on Associations. *Animals*, 11, 1638.
- ECKEL, E. F. & AMETAJ, B. N. 2016. Invited review: Role of bacterial endotoxins in the etiopathogenesis of periparturient diseases of transition dairy cows. *J Dairy Sci*, 99, 5967-5990.
- EDMONSON, A. J., LEAN, I., WEAVER, L. D., FARVER, T. B., & WEBSTER, G. L. 1989. A body condition scoring chart for holstein dairy cows. *J Dairy Sci*, 72(1), 68-78.
- EDWARDS, J. L. & TOZER, P. R. 2004. Using activity and milk yield as predictors of fresh cow disorders. *J Dairy Sci*, 87, 524-31.
- EHRLICH, J. L. 2010. Quantifying shape of lactation curves, and benchmark curves for common dairy breeds and parities, *The Bovine Practitioner*, 45(1), 88–95
- EHRET, A., HOCHSTUHL, D., GIANOLA, D. & THALLER, G. 2015. Application of neural networks with back-propagation to genome-enabled prediction of complex traits in Holstein-Friesian and German Fleckvieh cattle. *Genet Sel Evol*, 47, 22.
- ELZHOV, T.V., MULLEN, K.M., SPIESS A.N., BOLKER B. 2016. minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds. R package version 1.2-1. <https://CRAN.R-project.org/package=minpack.lm>
- ERB, H. N., SMITH, R. D., OLTENACU, P. A., GUARD, C. L., HILLMAN, R. B., POWERS, P. A., SMITH, M. C. & WHITE, M. E. 1985. Path model of reproductive disorders and performance, milk fever, mastitis, milk yield, and culling in Holstein cows. *J Dairy Sci*, 68, 3337-49.

- BRADLEY & A. J., DOTTORINI, T. 2018. Discrimination of contagious and environmental strains of *Streptococcus uberis* in dairy herds by means of mass spectrometry and machine-learning. *Sci. Rep.*, 8, 17517.
- ESENER, N., MACIEL-GUERRA, A., GIEBEL, K., LEA, D., GREEN, M. J., BRADLEY, A. J., & DOTTORINI, T. 2021. Mass spectrometry and machine learning for the accurate diagnosis of benzylpenicillin and multidrug resistance of *Staphylococcus aureus* in bovine mastitis. *PLoS computational biology*, 17(6), e1009108.
- ESPEJO, L. A. & ENDRES, M. I. 2007. Herd-level risk factors for lameness in high-producing holstein cows housed in freestall barns. *J Dairy Sci*, 90, 306-14.
- ESPOSITO, G., IRONS, P. C., WEBB, E. C. & CHAPWANYA, A. 2014. Interactions between negative energy balance, metabolic diseases, uterine health and immune response in transition dairy cows. *Anim Reprod Sci*, 144, 60-71.
- FANG K., WU S., ZHU J., XIE B., 2011, A review of random forest methods[J]. *Statistics and Information Forum*, 26(03):32-38.
- FARAH, J. S., CAVALCANTI, R. N., GUIMARAES, J. T., BALTHAZAR, C. F., COIMBRA, P. T., PIMENTEL, T. C., ESMERINO, E. A., DUARTE, M. C. K. H., FREITAS, M. Q., GRANATO, D., et al. 2021. Differential scanning calorimetry coupled with machine learning technique: An effective approach to determine the milk authenticity. *Food Control.*, 121, 107585.
- FARRELL, A. V., WANG, G., RUSH, S. A., MARTIN, J. A., BELANT, J. L., BUTLER, A., & GODWIN, D. 2019. Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. *Ecology and Evolution*, 9(10), 5938-5949.
- FEI, Z., GUAN, C., & GAO, H. 2018. Exponential synchronization of networked chaotic delayed neural network by a hybrid event trigger scheme. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6), 2558-2567.

- FAWAGREH, K., GABER, M., & ELYAN, E., 2014. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 2(1):602–609.
- FENLON, C., O'GRADY, L., BUTLER, S., DOHERTY, M. L. & DUNNION, J. 2017a. The creation and evaluation of a model to simulate the probability of conception in seasonal-calving pasture-based dairy heifers. *Ir Vet J*, 70, 32.
- FENLON, C., O'GRADY, L., MEE, J. F., BUTLER, S. T., DOHERTY, M. L. & DUNNION, J. 2017b. A comparison of 4 predictive models of calving assistance and difficulty in dairy heifers and cows. *J Dairy Sci*, 100, 9746-9758.
- FIGUEIREDO, C., MERENDA, V. R., OLIVEIRA, E. B. d., LIMA, F., CHEBEL, R. C., GALVAO, K. N. & BISINOTTO, R. 2021. Failure of clinical cure in dairy cows treated for metritis is associated with reduced productive and reproductive performance. *J Dairy Sci*, 104(6), 7056-7070.
- FILHO, V. B. S., SCHIAVON, R. S., GASTAL, G. D. A., TIMM, C. D., & LUCIA, T. 2012. Association of the occurrence of some diseases with reproductive performance and milk production of dairy herds in southern brazil. *Revista Brasileira De Zootecnia*, 41(2), 467-471.
- FLEISCHER, P., METZNER, M., BEYERBACH, M., HOEDEMAKER, M. & KLEE, W. 2001. The relationship between milk yield and the incidence of some diseases in dairy cows. *J Dairy Sci*, 84, 2025-35.
- FOLDI, J., KULCSAR, M., PECSI, A., HUYGHE, B., DE SA, C., LOHUIS, J. A., COX, P. & HUSZENICZA, G. 2006. Bacterial complications of postpartum uterine involution in cattle. *Anim Reprod Sci*, 96, 265-81.
- FOLLI, G. S., NASCIMENTO, M. H., PAULO, E. H. d., CUNHA, P. H. P. d., ROMAO, W., & FILGUEIRAS, P. R. 2020. Variable selection in support vector regression using angular search algorithm and variance inflation factor. *Journal of Chemometrics*, 34(12).
- FOURICHON, C., SEEGER, H., BAREILLE, N. & BEAUDEAU, F. 1999. Effects of disease on milk production in the dairy cow: a review. *Prev Vet Med*, 41, 1-35.

- FOURICHON, C., SEEGER, H. & MALHER, X. 2000. Effect of disease on reproduction in the dairy cow: a meta-analysis. *Theriogenology*, 53, 1729-59.
- FOX J., & WEISBERG S., 2019. *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. <https://www.john-fox.ca/Companion/>.
- FRANCESCHINI S, GRELET, C., LEBLOIS, J., GENGLER, N., & SOYEURT H. 2022. Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health? *J Dairy Sci*. 6760-6772.
- FRAZER, G. S. 2005. A rational basis for therapy in the sick postpartum cow. *Vet Clin North Am Food Anim Pract*, 21, 523-68.
- FRIZZARIN, M., GORMLEY, I. C., BERRY, D. P., MURPHY, T. B., CASA, A., LYNCH, A., & MCPARLAND, S. (2021). Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *J Dairy Sci*, 104(7), 7438–7447.
- FROUD, K., BEREฟอร์ด, R., & COGGER, N. (2017). Impact of kiwifruit bacterial canker on productivity of cv. hayward kiwifruit using observational data and multivariable analysis. *Plant Pathology*, 67(3), 671-681.
- FU, Q., SHEN, W., WEI, X., ZHANG, Y., XIN, H., SU, Z., ZHAO, C. 2020. Prediction of the diet energy digestion using kernel extreme learning machine: A case study with Holstein dry cows. *Comput. Electron. Agric.*, 169, 105231.
- FUENTES, S., VIEJO, C. G., CULLEN, B., TONGSON, E., CHAUHAN, S. S., DUNSHEA, F. R., 2020. Artificial intelligence applied to a robotic dairy farm to model milk productivity and quality based on cow data and daily environmental parameters. *Sensors*, 20, 2975.
- GARCIA-ALMANZA, A. L. and TSANG, E. P. K., 2006. Simplifying Decision Trees Learned by Genetic Programming, 2006 *IEEE International Conference on Evolutionary Computation*, 2142-2148.
- GAULY, M., AMMER, S., 2020. Review: Challenges for dairy cow production systems arising from climate changes, *Animal*, Volume 14, Supplement 1, 196-203

- GAUTAM, G., NAKAO, T., YUSUF, M. & KOIKE, K. 2009. Prevalence of endometritis during the postpartum period and its impact on subsequent reproductive performance in two Japanese dairy herds. *Anim Reprod Sci*, 116, 175-87.
- GEISHAUSER, T. 1995. Abomasal displacement in the bovine--a review on character, occurrence, aetiology and pathogenesis. *Zentralbl Veterinarmed A*, 42, 229-51.
- GENEDY, R. A., OGEJO, J. A. 2021. Using machine learning techniques to predict liquid dairy manure temperature during storage. *Comput. Electron. Agric.*, 187, 106234.
- GHAFFARI, M. H., JAHANBEKAM, A., SADRI, H., SCHUH, K., DUSEL, G., PREHN, C., ADAMSKI, J., KOCH, C. & SAUERWEIN, H. 2019. Metabolomics meets machine learning: Longitudinal metabolite profiling in serum of normal versus overconditioned cows and pathway analysis. *J. Dairy Sci.*, 102, 11561–11585.
- GHAREEB, Z. 2023. A new shrinkage method for higher dimensions regression model to remedy of multicollinearity problem. *Periodicals of Engineering and Natural Sciences (Pen)*, 11(3), 18.
- GIANOLA, D., OKUT, H., WEIGEL, K. A. & ROSA, G. J. 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC genetics*, 12, 87-87.
- GLEERUP, K. B., ANDERSEN, P. H., MUNKSGAARD, L. & FORKMAN, B. 2015. Pain evaluation in dairy cattle. *Appl. Anim. Behav. Sci.* 171, 25–32.
- GOICHEVA-ILIEVA S., YORDANOVA A., KULINA H.. 2022. Predicting the 305-Day Milk Yield of Holstein-Friesian Cows Depending on the Conformation Traits and Farm Using Simplified Selective Ensembles. *Mathematics*. 10(8):1254.
- GODDEN, S. M., STEWART, S. C., FETROW, J. F., RAPNICKI, P., CADY, R., WEILAND, W., SPENCER, H. & EICKER, S. 2003. The relationship between herd rbST supplementation and other factors and risk for removal for cows in Minnesota Holstein dairy herds. *Proc. Four-State Nutr. Conf.*, 55-64.

- GOFF, J. P. 2008. The monitoring, prevention, and treatment of milk fever and subclinical hypocalcemia in dairy cows. *Vet J*, 176, 50-7.
- GOFF, J. P. & HORST, R. L. 1997a. Effects of the addition of potassium or sodium, but not calcium, to prepartum rations on milk fever in dairy cows. *J Dairy Sci*, 80, 176-86.
- GOFF, J. P. & HORST, R. L. 1997b. Physiological changes at parturition and their relationship to metabolic disorders. *J Dairy Sci*, 80, 1260-8.
- GOFF, J. P., HORST, R. L., JARDON, P. W., BORELLI, C. & WEDAM, J. 1996. Field trials of an oral calcium propionate paste as an aid to prevent milk fever in periparturient dairy cows. *J Dairy Sci*, 79, 378-83.
- GOHARY, K. & LEBLANC, S. J. 2018. Cost of retained fetal membranes for dairy herds in the United States. *J Am Vet Med Assoc*, 252, 1485-1489.
- GONZALEZ-RECIO, O., WEIGEL, K. A., GIANOLA, D., NAYA, H. & ROSA, G. J. 2010. L2-Boosting algorithm applied to high-dimensional problems in genomic selection. *Genet Res (Camb)*, 92, 227-37.
- GORCZYCA, M. T. & GEBREMEDHIN, K. G. 2020. Ranking of environmental heat stressors for dairy cows using machine learning algorithms. *Comput. Electron. Agric.*, 168, 105124.
- GORGULU, O., 2012. Prediction of 305-day milk yield in Brown Swiss cattle using artificial neural networks, *South African Journal of Animal Science*, 42 (3), 280 – 287
- GRAULET, B., MATTE, J. J., DESROCHERS, A., DOEPEL, L., PALIN, M., & GIRARD, C. (2007). Effects of dietary supplements of folic acid and vitamin b12 on metabolism of dairy cows in early lactation. *J Dairy Sci*, 90(7), 3442-3455.
- GREEN, L. E., HUXLEY, J. N., BANKS, C. & GREEN, M. J. 2014. Temporal associations between low body condition, lameness and milk yield in a UK dairy herd. *Prev Vet Med*, 113, 63-71.
- GRIFFITHS, B. E., GROVE WHITE, D. & OIKONOMOU, G. 2018. A Cross-Sectional Study Into the Prevalence of Dairy Cattle Lameness and Associated Herd-Level Risk Factors in England and Wales. *Frontiers in veterinary science*, 5, 65-65.

- GROHN, Y. T., EICKER, S. W., DUCROCQ, V. & HERTL, J. A. 1998. Effect of diseases on the culling of Holstein dairy cows in New York State. *J Dairy Sci*, 81, 966-78.
- GROHN, Y.T., RAJALA-SCHULTZ, P.J., 2000. Epidemiology of reproductive performance in dairy cows, *Animal Reproduction Science*, Volumes 60–61, 605-614.
- GROHN, Y. T., RAJALA-SCHULTZ, P. J., ALLORE, H. G., DELORENZO, M. A., HERTL, J. A. & GALLIGAN, D. T. 2003. Optimizing replacement of dairy cows: modeling the effects of diseases. *Prev Vet Med*, 61, 27-43.
- GROHN, Y.T. WILSON, D.J. GONZALEZ, R.N. HERTL, J.A. SCHULTE, H. BENNETT, G. SCHUKKENY.H. 2004. Effect of Pathogen-Specific Clinical Mastitis on Milk Yield in Dairy Cows, *J Dairy Sci*, Volume 87, Issue 10, 2004, 3358-3374.
- GROSS, J. J., SCHWARZ, F. J., EDER, K., DORLAND, H. V., & BRUCKMAIER, R. M. (2013). Liver fat content and lipid metabolism in dairy cows during early lactation and during a mid-lactation feed restriction. *J Dairy Sci*, 96(8), 5008-5017.
- GRUMMER, R. R. 1993. Etiology of lipid-related metabolic disorders in periparturient dairy cows. *J Dairy Sci*, 76, 3882-96.
- GRUMMER, R. R. 1995. Impact of changes in organic nutrient metabolism on feeding the transition dairy cow. *J Anim Sci*, 73, 2820-33.
- GRYMER J., WILLEBERG P., HESSELHOLT M. 1982. Milk production and left displaced abomasum: cause and effect relationships. *Nord Vet Med*. 34(11):412-5
- GRZESIAK, W., LACROIX, R., WOJCIK, J., BLASZCZYK, P. 2003. A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records, *Canadian Journal of Animal Science*, 83 (2), 307 – 310
- GRZESIAK, W., BŁASZCZYK, P. & LACROIX, R. 2006. Methods of predicting milk yield in dairy cows—Predictive capabilities of Wood's lactation curve and artificial neural networks (ANNs). *Computers and Electronics in Agriculture*, 54, 69-83.

- GRZEKIAK, W., ZABORSKI, D., SZATKOWSKA, I., & KROLACZYK, K. 2021. Lactation milk yield prediction in primiparous cows on a farm using the seasonal auto-regressive integrated moving average model, nonlinear autoregressive exogenous artificial neural networks and wood's model. *Animal Bioscience*, 34(4), 770-782.
- GIULIODORI M.J., MAGNASCO R.P., BECU-VILLALOBOS D., LACAU-MENGIDO I.M., RISCO C.A., DE LA SOTA R.L., 2013. Metritis in dairy cows: risk factors and reproductive performance. *J Dairy Sci*. 96(6):3621-31.
- GUCKIRAN, K., CANTURK, İ., & OZYILMAZ, L., 2019. Dna microarray gene expression data classification using svm, mlp, and rf with feature selection methods relief and lasso. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 23(1), 126-132.
- GUMEN, A., KESKIN, A., YILMAZBAS-MECITOGLU, G., Karakaya, E., & WILTBANK, M. C. (2011). Dry period management and optimization of post-partum reproductive management in dairy cattle. *Reproduction in Domestic Animals*, 46(s3), 11-17.
- GUMEN, A., RASTANI, R.R., GRUMMER, R.R., WILTBANK, M.C. 2005. Reduced dry periods and varying prepartum diets alter postpartum ovulation and reproductive measures. *J Dairy Sci*. 2401-11.
- HACHENBERG, S., WEINKAUF, C., HISS, S. & SAUERWEIN, H. 2007. Evaluation of classification modes potentially suitable to identify metabolic stress in healthy dairy cows during the peripartal period. *J Anim Sci*, 85, 1923-32.
- HAHN, G. L. 1999. Dynamic responses of cattle to thermal heat loads. *J Anim Sci*, 77 Suppl 2, 10-20.
- HAILEMARIAM, D., MANDAL, R., SALEEM, F., DUNN, S. M., WISHART, D. S., & AMETAJ, B. N. (2014). Identification of predictive biomarkers of disease state in transition dairy cows. *J Dairy Sci*, 97(5), 2680-2693.
- HALADJIAN, J., HAUG, J., NUSKE, S., BRUEGGE, B. 2018. A wearable sensor system for lameness detection in dairy cattle. *Multimodal Technol. Interact.*, 2, 27.

- HANCOCK, J., KHOSHGOFTAAR, T., & JOHNSON, J., 2023. Evaluating classifier performance with highly imbalanced big data. *Journal of Big Data*, 10(1).
- HARRELL, F.E. 2001. Cox Proportional Hazards Regression Model. In: Regression Modeling Strategies. Springer Series in Statistics. Springer, New York, NY.
- HARTONO, H. & ONGKO, E. 2022. Avoiding overfitting dan overlapping in handling class imbalanced using hybrid approach with smoothed bootstrap resampling and feature selection. *Joiv International Journal on Informatics Visualization*, 6(2), 343.
- HASSAN, K. J., SAMARASINGHE, S. & LOPEZ-BENAVIDES, M. G. 2009. Use of neural networks to detect minor and major pathogens that cause bovine mastitis. *J Dairy Sci*, 92, 1493-9.
- HAYIRLI, A., GRUMMER, R. R., NORDHEIM, E. V. & CRUMP, P. M. 2002. Animal and dietary factors affecting feed intake during the prefresh transition period in Holsteins. *J Dairy Sci*, 85, 3430-43.
- HAYKIN, S. 1998. *Neural Networks: A Comprehensive Foundation*, Prentice Hall PTR.
- HE, J., BAXTER, S. L., XU, J., ZHOU, X. & ZHANG, K. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* 25, 30–36.
- HEALD, C. W., KIM, T., SISCHO, W. M., COOPER, J. B. & WOLFGANG, D. R. 2000. A computerized mastitis decision aid using farm-based records: an artificial neural network approach. *J Dairy Sci*, 83, 711-20.
- VAN DER HEIDE, E. M. M., VEERKAMP, R. F., VAN PELT, M. L., KAMPHUIS, C., ATHANASIADIS, I., & DUCRO, B. J. 2019. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. *J Dairy Sci*, 102(10), 9409-9421.
- HEMPEL, S., ADOLPHS, J., LANDWEHR, N., WILLINK, D., JANKE, D., AMON, T., 2020. Supervised machine learning to assess methane emissions of a dairy building with natural ventilation. *Appl. Sci.*, 10, 6938.

- HEMPSTALK, K., MCPARLAND, S. & BERRY, D. P. 2015. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *J Dairy Sci*, 98, 5262-5273.
- HERNANDEZ, B., LOPEZ-VILLALOBOS, N., & VIGNES, M. 2021. Identifying health status in grazing dairy cows from milk mid-infrared spectroscopy by using machine learning methods. *Animals*, 11(8), 2154.
- HERRING, A.D., Beef Cattle, Editor(s): VAN ALFEN, N.K. 2014. Encyclopedia of Agriculture and Food Systems, Academic Press, 1-20.
- HEUER, C., SCHUKKEN, Y. H. & DOBBELAAR, P. 1999. Postpartum body condition score and results from the first test day milk as predictors of disease, fertility, yield, and culling in commercial dairy herds. *J Dairy Sci*, 82, 295-304.
- HIGAKI, S., MIURA, R., SUDA, T., ANDERSSON, L. M., OKADA, H., ZHANG, Y., ITOH, T., MIWAKEICHI, F. & YOSHIOKA, K. 2019. Estrous detection by continuous measurements of vaginal temperature and conductivity with supervised machine learning in cattle. *Theriogenology*, 123, 90-99.
- HO, K. 2017. Effect of non-linearity of a predictor on the shape and magnitude of its receiver-operating-characteristic curve in predicting a binary outcome. *Scientific Reports*, 7(1).
- HOEDEMAKER, M., PRANGE, D. & GUNDELACH, Y. 2009. Body condition change ante- and postpartum, health and reproductive performance in German Holstein cows. *Reprod Domest Anim*, 44, 167-73.
- HOKA, A. I., GICHERU, M. & OTIENO, S. 2019. Effect of cow parity and calf characteristics on milk production and reproduction of friesian dairy cows. JNSR.
- HOLDER, A. M. & FIELD, J. C. 2019. An exploration of factors that relate to the occurrence of multiple brooding in rockfishes (sebastes spp.). *Fishery Bulletin*, 117(3), 56-64.
- HORST, R. L., GOFF, J. P. & REINHARDT, T. A. 1994. Calcium and vitamin D metabolism in the dairy cow. *J Dairy Sci*, 77, 1936-51.
- HOSMER, D.W., HOSMER, T., LEMESHOW, S. 1980. A Goodness-of-Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics*, 10, 1043-1069.

- HUANG, X., HU, Z., WANG, X., YANG, X., ZHANG, J., SHI, D. 2019. An improved single shot multibox detector method applied in body condition score for dairy cows. *Animals*, 9, 470.
- HUDSON, C. 2011. Understanding the factors affecting dairy cow fertility. *Veterinary Record*, 168, 299.
- HUDSON, C., KALER, J. & DOWN, P. 2018. Using big data in cattle practice. *In Practice*, 40, 396.
- HUDSON, C. D., BRADLEY, A. J., BREEN, J. E. & GREEN, M. J. 2012. Associations between udder health and reproductive performance in United Kingdom dairy cows. *J Dairy Sci*, 95, 3683-97.
- HUDSON, C. D. & GREEN, M. J. 2018. Associations between routinely collected Dairy Herd Improvement data and insemination outcome in UK dairy herds. *J Dairy Sci*. 101(12), 11262–11274.
- HUNTER, L. B., BATEN, A., HASKELL, M. J., LANGFORD, F. M., O'CONNOR, C., WEBSTER, J. R., STAFFORD, K. 2021. Machine learning prediction of sleep stages in dairy cows from heart rate and muscle activity measures. *Sci. Rep.*, 11, 10938.
- HUZZEY, J. M., DUFFIELD, T. F., LEBLANC, S. J., VEIRA, D. M., WEARY, D. M. & VON KEYSERLINGK, M. A. 2009. Short communication: Haptoglobin as an early indicator of metritis. *J Dairy Sci*, 92, 621-5.
- HUZZEY, J. M., VEIRA, D. M., WEARY, D. M. & VON KEYSERLINGK, M. A. 2007. Parturition behavior and dry matter intake identify dairy cows at risk for metritis. *J Dairy Sci*, 90, 3220-33.
- HYDE, R. M., DOWN, P. M., BRADLEY, A. J., BREEN, J. E., HUDSON, C., LEACH, K. A., GREEN, M. J. 2020. Automated prediction of mastitis infection patterns in dairy herds using machine learning. *Sci. Rep.*, 10, 4289.
- IIDA, K., & KIYA, H., 2019, IEICE Trans. Inf. Syst., 103-D, 25-32
- IMADA, J., ARANGO-SABODAL, J. C., BAUMAN, C., ROCHE, S., & KELTON, D. (2024). Comparison of Machine Learning Tree-Based Algorithms to Predict Future Paratuberculosis ELISA Results Using Repeat Milk Tests. *Animals : an open access journal from MDPI*, 14(7), 1113

- IMHASLY, S., BIELI, C., NAEGELI, H., NYSTROM, L., RUETTEN, M., & GERSPACH, C. (2015). Blood plasma lipidome profile of dairy cows during the transition period. *BMC Veterinary Research*, 11(1).
- INFANTE, P., JACINTO, G., AFONSO, A., REGO, L., NOGUEIRA, V., QUARESMA, P., & MANUEL, P. 2022. Comparison of statistical and machine-learning models on road traffic accident severity classification. *Computers*, 11(5), 80.
- INGALHALIKAR, M., SHINDE, S., KARMARKAR, A., RAJAN, A., RANGAPRAKASH, D., & DESHPANDE, G. 2021. Functional connectivity-based prediction of autism on site harmonized abide dataset. *IEEE Transactions on Biomedical Engineering*, 68(12), 3628-3637.
- INGVARTSEN, K. L. 2006. Feeding- and management-related diseases in the transition cow: Physiological adaptations around calving and strategies to reduce feeding-related diseases. *Animal Feed Science and Technology*, 126, 175-213.
- JAGER, K. J., ZOCCALI, C., MACLEOD, A. & DEKKER, F. W. 2008. Confounding: what it is and how to deal with it. *Kidney Int.* 73, 256–260.
- JAMALI EMAM GHEISE, N., RIASI, A., ZARE SHAHNEH, A., CELI, P. & GHOREISHI, S. M. 2017. Effect of pre-calving body condition score and previous lactation on BCS change, blood metabolites, oxidative stress and milk production in Holstein dairy cows. *Italian Journal of Animal Science*, 16, 474-483.
- JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. 2014. *An Introduction to Statistical Learning: with Applications in R*, Springer New York.
- JAZAYERI, M. & MOVSHON, J. A. 2006. Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5), 690-696.
- JAWOR, P., HUZZEY, J., LEBLANC, S., & KEYSERLINGK, M. v. 2012. Associations of subclinical hypocalcemia at calving with milk yield, and feeding, drinking, and standing behaviors around parturition in holstein cows. *J Dairy Sci*, 95(3), 1240-1248.
- JE, M. & LEE, J. 2023. Factors associated with smartphone dependence of late school-aged children: a focus on grit and family strengths. *Korean Journal of Health Promotion*, 23(1), 37-42.

- JI, B., BANHAZI, T., GHAHRAMANI, A., BOWTELL, L., WANG, C. & LI, B. 2020. Modelling of heat stress in a robotic dairy farm. Part 2: Identifying the specific thresholds with production factors. *Biosyst. Eng.*, 199, 43–57.
- JI, B., BANHAZI, T., PHILLIPS, C. J., WANG, C., & LI, B. 2022. A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm. *Biosystems Engineering*, 216, 186-197.
- JIMÉNEZ-MONTERO, J. A., GIANOLA, D., WEIGEL, K., ALENDA, R. & GONZÁLEZ-RECIO, O. 2013a. Assets of imputation to ultra-high density for productive and functional traits. *J Dairy Sci*, 96, 6047-6058.
- JIMÉNEZ-MONTERO, J. A., GONZALEZ-RECIO, O. & ALENDA, R. 2013b. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. *J Dairy Sci*, 96, 625-34.
- JING, W., YANG, Y., YUE, X., & ZHAO, X., 2016. A comparison of different regression algorithms for downscaling monthly satellite-based precipitation over north china. *Remote Sensing*, 8(10), 835.
- JOHAM, M., CASTRO, P., UTSCHICK, W., & CASTEDO, L. 2012. Robust precoding with limited feedback design based on precoding mse for mu-miso systems. *IEEE Transactions on Signal Processing*, 60(6), 3101-3111.
- JOHN WALLACE, R., SASSON, G., GARNSWORTHY, P. C., TAPIO, I., GREGSON, E., BANI, P., HUHTANEN, P., BAYAT, A. R., STROZZI, F., BISCARINI, F., et al. 2019. A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. *Sci. Adv.*, 5, eaav8391.
- JOLLYTA, D., GUSRIANTY, G., & SUKRIANTO, D. (2019). Analysis of slow moving goods classification technique: random forest and naïve bayes. *Khazanah Informatika Jurnal Ilmu Komputer Dan Informatika*, 5(2), 134-139.
- KAMPHUIS, C., MOLLENHORST, H., HEESTERBEEK, J. A. & HOGVEEN, H. 2010. Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision-tree induction. *J Dairy Sci*, 93, 3616-27.

- KAMPHUIS, C., PIETERSMA, D., VAN DER TOL, R., WIEDEMANN, M. & HOGEVEEN, H. 2008. 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, 169-181.
- KASIMANICKAM, R., DUFFIELD, T. F., FOSTER, R. A., GARTLEY, C. J., LESLIE, K. E., WALTON, J. S. & JOHNSON, W. H. 2004. Endometrial cytology and ultrasonography for the detection of subclinical endometritis in postpartum dairy cows. *Theriogenology*, 62, 9-23.
- KAWASHIMA, C., KARAKI, C., MUNAKATA, M., MATSUI, M., SHIMIZU, T., MIYAMOTO, A. & KIDA, K. 2016. Association of rumen fill score and energy status during the close-up dry period with conception at first artificial insemination in dairy cows. *Anim Sci J*, 87, 1218-1224.
- KEARNS, M. & VAZIRANI, U. V., 1994, An introduction to computational learning theory. MIT Press, 2.
- KECELI, A. S., CATAL, C., KAYA, A., & TEKINERDOGAN, B. 2020. Development of a recurrent neural networks-based calving prediction model using activity and behavioral data, *Computers and Electronics in Agriculture*, 170, 105285, 0168-1699,
- KELTON, D. F., LISSEMORE, K. D. & MARTIN, R. E. 1998. Recommendations for Recording and Calculating the Incidence of Selected Clinical Diseases of Dairy Cattle. *Journal of Dairy Science*, 81, 2502-2509.
- KEPA, M. and SZYMANSKI, J. 2015. Two stage svm and knn text documents classifier., 279-289.
- KERNBACH, J. and STAARTJES, V. 2021. Foundations of machine learning-based clinical prediction modeling: part ii—generalization and overfitting., *Machine Learning in Clinical Neuroscience*. 15-21.
- KESHAVARZI, H., SADEGHI-SEFIDMAZGI, A., MIRZAEI, A., RAVANIFARD, R., 2020. Machine learning algorithms, bull genetic information, and imbalanced datasets used in abortion incidence prediction models for Iranian Holstein dairy cattle. *Prev. Vet. Med.*, 175, 104869.
- KILICOGLU, Ş. & YERLIKAYA-ÖZKURT, F. 2024. A novel comparison of shrinkage methods based on multi criteria decision making in case of

- multicollinearity. *Journal of Industrial and Management Optimization*, 20(12), 3816-3842.
- KIM, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, 72(6), 558-569
- KIM, T., HEALD, C. W. 1999. Inducing inference rules for the classification of bovine mastitis. *Comput. Electron. Agric.*, 23, 27–42.
- KIMURA, K., GOFF, J. P., KEHRLI, M. E., JR. & REINHARDT, T. A. 2002. Decreased neutrophil function as a cause of retained placenta in dairy cattle. *J Dairy Sci*, 85, 544-50.
- KIMURA, K., REINHARDT, T. A. & GOFF, J. P. 2006. Parturition and hypocalcemia blunts calcium signals in immune cells of dairy cattle. *J Dairy Sci*, 89, 2588-95.
- KING, C., ABRAHAM, J., FRITZ, B., CUI, Z., GALANTER, W., CHEN, Y., & KANNAMPALLIL, T., 2021. Predicting self-intercepted medication ordering errors using machine learning. *Plos One*, 16(7), e0254358.
- KIYAK, E., BIRANT, D., & BIRANT, K., 2021. An improved version of multi-view k-nearest neighbors (mvknn) for multiple view learning. *Turkish Journal of Electrical Engineering & Computer Sciences*, 29(3), 1401-1428.
- KOC, A. 2011. A study of the reproductive performance, milk yield, milk constituents, and somatic cell count of Holstein-Friesian and Montbeliarde cows. *Turk J Vet Anim Sci* 35(5):295-302.
- KOENEN, E. P., VEERKAMP, R. F., DOBBELAAR, P. & DE JONG, G. 2001. Genetic analysis of body condition score of lactating Dutch Holstein and Red-and-White heifers. *J Dairy Sci*, 84, 1265-70.
- KRAMER O., 2013. K-nearest neighbors. In: Dimensionality reduction with unsupervised nearest neighbors. Springer, 13-23.
- KRAWCZEL, P.D., LEE, A.R., 2019. Lying Time and Its Importance to the Dairy Cow: Impact of Stocking Density and Time Budget Stresses, *Veterinary Clinics of North America: Food Animal Practice*, Volume 35, Issue 1, 47-60.
- KRISHNAN, R., G. SIVAKUMAR, & P. BHATTACHARYA 1999, Extracting decision trees from trained neural networks, *Pattern Recognit.*, 32, 1999–2009.

- KROLL, C. & SONG, P. 2013. Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research*, 49(6), 3756-3769.
- KUHN, M. 2008. Caret package. *Journal of Statistical Software*, 28(5)
- KUHN, M. & JOHNSON, K. 2013. *Applied Predictive Modeling*, Springer.
- KUL, E., ŞAHİN, A., UGURLUTEPE, E., & SOYDANER, M. 2020. Association of change in body condition score with milk yield and reproduction traits of holstein cows. *The Journal of Animal and Plant Sciences*, 30(2).
- KYRIAZOS, T. & POGA, M. 2023. Dealing with multicollinearity in factor analysis: the problem, detections, and solutions. *Open Journal of Statistics*, 13(03), 404-424.
- LACROIX R., WADE K.M., KOK R., HAYES J.F. 1995. Prediction of cow performance with a connectionist model, *Transactions of the American Society of Agricultural Engineers*, 38 (5), pp. 1573 - 1579
- LACROIX R., SALEHI F., YANG X.Z., WADE K.M. 1997. Effects of data preprocessing on the performance of artificial neural networks for dairy yield prediction and cow culling classification, *Transactions of the American Society of Agricultural Engineers*, 40 (3), pp. 839 – 846
- LAMP, O., DERNO, M., OTTEN, W., MIELENZ, M., NÜRNBERG, G. & KUHLA, B. 2015. Metabolic Heat Stress Adaption in Transition Cows: Differences in Macronutrient Oxidation between Late-Gestating and Early-Lactating German Holstein Dairy Cows. *PloS one*, 10, e0125264-e0125264.
- LANDIS, J. R. & KOCH, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-74.
- LANIER, P., RODRIGUEZ, M., VERBIEST, S., BRYANT, K., GUAN, T., ZOLOTOR, A., 2020. Preventing infant maltreatment with predictive analytics: applying ethical principles to evidence-based child welfare policy. *J. Fam. Violence* 35(1), 1–13.
- LARDY, R., RUIN, Q., & VEISSIER, I. 2023. Discriminating pathological, reproductive or stress conditions in cows using machine learning on sensor-based activity data. *Computers and Electronics in Agriculture*, 204, 107556.

- LASSER, J., MATZHOLD, C., EGGER-DANNER, C., FUERST-WALTTL, B., STEININGER, F., WITTEK, T., KLIMEK, P., 2021. Integrating diverse data sources to predict disease risk in dairy cattle-a machine learning approach. *J Anim Sci*. 294.
- LEACH, K. A., WHAY, H. R., MAGGS, C. M., BARKER, Z. E., PAUL, E. S., BELL, A. K. & MAIN, D. C. 2010. Working towards a reduction in cattle lameness: 1. Understanding barriers to lameness control on dairy farms. *Res Vet Sci*, 89, 311-7.
- LEAN, I. J., DEGARIS, P. J., MCNEIL, D. M. & BLOCK, E. 2006. Hypocalcemia in dairy cows: meta-analysis and dietary cation anion difference theory revisited. *J Dairy Sci*, 89, 669-84.
- LEAN, I., SHEEDY, D., LEBLANC, S., DUFFIELD, T., SANTOS, J., & GOLDER, H. 2022. Holstein dairy cows lose body condition score and gain body weight with increasing parity in both pasture-based and total mixed ration herds. *JDS Communications*, 3(6), 431-435.
- LEBLANC, S. 2010. Monitoring metabolic health of dairy cattle in the transition period. *J Reprod Dev*, 56 Suppl, S29-35.
- LEBLANC, S. J. 2008. Postpartum uterine disease and dairy herd reproductive performance: a review. *Vet J*, 176, 102-14.
- LEBLANC, S. J., HERDT, T. H., SEYMOUR, W. M., DUFFIELD, T. F. & LESLIE, K. E. 2004. Peripartum serum vitamin E, retinol, and beta-carotene in dairy cattle and their associations with disease. *J Dairy Sci*, 87, 609-19.
- LEBLANC, S. J., LESLIE, K. E. & DUFFIELD, T. F. 2005. Metabolic predictors of displaced abomasum in dairy cattle. *J Dairy Sci*, 88, 159-70.
- LEBLANC, S. J., LISSEMORE, K. D., KELTON, D. F., DUFFIELD, T. F. & LESLIE, K. E. 2006. Major advances in disease prevention in dairy cattle. *J Dairy Sci*, 89, 1267-79.
- LETT, B. M. & KIRKPATRICK, B. (2018). Short communication: heritability of twinning rate in holstein cattle. *J Dairy Sci*, 101(5), 4307-4311.
- LEWIS, G. S. 1997. Uterine health and disorders. *J Dairy Sci*, 80, 984-94.

- LI, S., 2023. Performance comparison of representative methods for few-shot speech gender analysis. *Journal of Physics Conference Series*, 2580(1), 012038.
- LI, Q., LIANG, R., LI, Y., GAO, Y., LI, Q., SUN, D. & LI, J. 2020. Identification of candidate genes for milk production traits by rna sequencing on bovine liver at different lactation stages. *BMC Genetics*, 21(1).
- LI, X. and WANG, J., 2018. Traffic detection of transmission of botnet threat using bp neural network. *Neural Network World*, 28(6), 511-521.
- LIMA, F., VIEIRA-NETO, A., SNODGRASS, J., VRIES, A. D., & SANTOS, J. 2019. Economic comparison of systemic antimicrobial therapies for metritis in dairy cows. *J Dairy Sci*, 102(8), 7345-7358.
- LIN, D., RAJBAHADUR, G., & MING, J. 2021. Towards a consistent interpretation of aiops models. *Acm Transactions on Software Engineering and Methodology*, 31(1), 1-38.
- LISEUNE, A., SALAMONE, M., VAN DEN POEL, D., VAN RANST, B., HOSTENS, M., 2001. Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning, *Computers and Electronics in Agriculture*, Volume 180, 105904.
- LOKER, S., BASTIN, C., MIGLIOR, F., SEWALEM, A., SCHAEFFER, L. R., JAMROZIK, J., ALI, A., & OSBORNE, V. 2012. Genetic and environmental relationships between body condition score and milk production traits in Canadian Holsteins. *J Dairy Sci*, 95(1), 410–419.
- LU, C. Y., MIN, H., GUI, J., ZHU, L. & LEI, Y. K. 2013. Face recognition via weighted sparse representation, *Journal of Visual Communication and Image Representation*, 24, 2, 111–116.
- LUCY, M. C. 2001. Reproductive loss in high-producing dairy cattle: where will it end? *J Dairy Sci*, 84, 1277-93.
- LUKAS, J. M., RENEAU, J. K., WALLACE, R. L. & DE VRIES, A. 2015. A study of methods for evaluating the success of the transition period in early-lactation dairy cows. *J Dairy Sci*, 98, 250-262.

- LUKE, T.D.W. ROCHFORD, S. WALES, W.J. BONFATTI, V. MARETT, L. & PRYCE, J.E. 2019. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra, *J Dairy Sci*, 102, 2, 1747-1760.
- LUO, W., DONG, Q., & FENG, Y. 2023. Risk prediction model of clinical mastitis in lactating dairy cows based on machine learning algorithms. *Preventive Veterinary Medicine*, 221, 106059.
- LYONS, N. A., COOKE, J. S., WILSON, S. B., WINDEN, S. C. L. V., GORDON, P., & WATHES, D. C. 2014a. Relationships between metabolite and igf1 concentrations with fertility and production outcomes following left abomasal displacement. *Veterinary Record*, 174(26), 657-657.
- LYONS, N. A., KERRISK, K. L., & GARCIA, S. C. 2014b. Milking frequency management in pasture-based automatic milking systems: A review. *Livestock Science*, 159, 102–116.
- MACDONALD, K. A., VERKERK, G. A., THORROLD, B. S., PRYCE, J. E., PENNO, J. W., MCNAUGHTON, L. R., BURTON, L. J., LANCASTER, J. A., WILLIAMSON, J. H. & HOLMES, C. W. 2008. A comparison of three strains of holstein-friesian grazed on pasture and managed under different feed allowances. *J Dairy Sci*, 91, 1693-707.
- MACHADO, V. S., CAIXETA, L. S. & BICALHO, R. C. 2011. Use of data collected at cessation of lactation to predict incidence of sole ulcers and white line disease during the subsequent lactation in dairy cows. *Am J Vet Res*, 72, 1338-43.
- MACHADO, V. S., CAIXETA, L. S., MCART, J. A. & BICALHO, R. C. 2010. The effect of claw horn disruption lesions and body condition score at dry-off on survivability, reproductive performance, and milk production in the subsequent lactation. *J Dairy Sci*, 93, 4071-8.
- MACIEL-GUERRA, A., ESENER, N., GIEBEL, K., LEA, D., GREEN, M. J., BRADLEY, A. J., & DOTTORINI, T. (2021). Prediction of *Streptococcus uberis* clinical mastitis treatment success in dairy herds by means of mass spectrometry and machine-learning. *Scientific reports*, 11(1), 7736.
- MAJAJ, N. J., HONG, H., SOLOMON, E. A., & DICARLO, J. J. 2015. Simple learned weighted sums of inferior temporal neuronal firing rates

- accurately predict human core object recognition performance. *The Journal of Neuroscience*, 35(39), 13402-13418.
- MALLARD, B. A., DEKKERS, J. C., IRELAND, M. J., LESLIE, K. E., SHARIF, S., VANKAMPEN, C. L., WAGTER, L. & WILKIE, B. N. 1998. Alteration in immune responsiveness during the peripartum period and its ramification on dairy cow and calf health. *J Dairy Sci*, 81, 585-95.
- MAMMADOVA, N., KESKIN, X. & SMAIL 2013. Application of the Support Vector Machine to Predict Subclinical Mastitis in Dairy Cattle. *The Scientific World Journal*, 2013, 9.
- MARQUARDT, J. P., HORST, R. L. & JORGENSEN, N. A. 1977. Effect of Parity on Dry Matter Intake at Parturition in Dairy Cattle¹. *J Dairy Sci*, 60, 929-934.
- MARTINEZ LOPEZ, I., ORTIZ RODRIGUEZ I. M., RODRIGUEZ TORREBLANCA, C. 2019. A study of lactation curves in dairy cattle using the optimal design of experiments methodology, *Italian Journal of Animal Science*, 18:1, 594-600
- MARTISKAINEN, P., JÄRVINEN, M., SKÖN, J.-P., TIIRIKAINEN, J., KOLEHMAINEN, M. & MONONEN, J. 2009. Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. *Applied Animal Behaviour Science*, 119, 32-38.
- MAY, D. E., CORBIN, A., & HOLLINS, P. D. 2011. Identifying determinants of residential property values in south london. *Review of Economic Perspectives*, 11(1), 3-11.
- MCCARTHY, S., BERRY, D. P., DILLON, P., RATH, M. & HORAN, B. 2007. Influence of Holstein-Friesian Strain and Feed System on Body Weight and Body Condition Score Lactation Profiles. *J Dairy Sci*, 90, 1859-1869.
- MELENDEZ, P., DONOVAN, G. A., RISCO, C. A., LITTELL, R. & GOFF, J. P. 2003. Effect of calcium-energy supplements on calving-related disorders, fertility and milk yield during the transition period in cows fed anionic diets. *Theriogenology*, 60, 843-54.
- MELENDEZ, P. & RISCO, C. A. 2005. Management of transition cows to optimize reproductive efficiency in dairy herds. *Vet Clin North Am Food Anim Pract*, 21, 485-501.

- MELZER, N., WITTENBURG, D., HARTWIG, S., JAKUBOWSKI, S., KESTING, U., WILLMITZER, L., LISEC, J., REINSCH, N. & REPSILBER, D. 2013. Investigating associations between milk metabolite profiles and milk traits of Holstein cows. *J Dairy Sci*, 96, 1521-34.
- MENTA, P., MACHADO, V. S., PINEIRO, J. M., THATCHER, W., SANTOS, J. E. P., & VIEIRA-NETO, A. 2022. Heat stress during the transition period is associated with impaired production, reproduction, and survival in dairy cows. *J Dairy Sci*. 105, 5, 4474-89.
- MERENDA, V. R., RUIZ-MUNOZ, J., ZARE, A., & CHEBEL, R. C. 2021. Predictive models to identify Holstein cows at risk of metritis and clinical cure and reproductive/productive failure following antimicrobial treatment. *Preventive veterinary medicine*, 194, 105431.
- MILIAN-SUAZO, F., ERB, H. N. & SMITH, R. D. 1988. Descriptive epidemiology of culling in dairy cows from 34 herds in New York State. *Preventive Veterinary Medicine*, 6, 243-251.
- MILLER, G. A., MITCHELL, M., BARKER, Z. E., GIEBEL, K., CODLING, E. A., AMORY, J. R., MICHIE, C., DAVISON, C., TACHTATZIS, C., ANDOVONIC, I., 2020. Using animal-mounted sensor technology and machine learning to predict time-to-calving in beef and dairy cows. *Animal*, 14, 1304–1312
- MILTENBURG, C. L., DUFFIELD, T. F., BIENZLE, D., SCHOLTZ, E. L. & LEBLANC, S. J. 2018. The effect of prepartum feeding and lying space on metabolic health and immune function. *J Dairy Sci*, 101, 5294-5306.
- MINNAERT, B., THOEN, B., DAVID, P. A., JOSEPH, W., & STEVENS, N. 2018. Wireless energy transfer by means of inductive coupling for dairy cow health monitoring. *Computers and Electronics in Agriculture*, 152, 101-108.
- MOREIRA, T., FILHO, E., BELL, A., MENESES, R., LEME, F., URIBE, J., RODRIGUES L., & CARVALHO, A. 2018. Metabolic status of crossbreed f1 holstein × gyr dairy cows during the transition period in two different seasons in brazil. *Semina Ciências Agrárias*, 39(6), 2487.
- MOTA, L. F. M., PEGOLO, S., BABA, T., PENAGARICANO, F., MOROTA, G., BITTANTE, G., & CECCHINATO, A. 2021. Evaluating the performance of machine learning methods and variable selection methods for

- predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. *J Dairy Sci*, 104(7), 8107–8121.
- MOUJAHID, A., TANTAOUI, M.E., HINA, M.D., SOUKANE, A., ORTALDA, A., ELKADIMI, A., RAMDANE-CHERIF., 2018. A.: Machine learning techniques in ADAS: a review. 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 235–242.
- MURPHY, M. D., O'MAHONY, M. J., SHALLOO, L., FRENCH, P. & UPTON, J. 2014. Comparison of modelling techniques for milk-production forecasting. *J Dairy Sci*, 97, 3352-63.
- MUTCHLER, M. S. & ANDERSON, S. A. 2010. Therapist personal agency: a model for examining the training context. *Journal of Marital and Family Therapy*, 36(4), 511-525.
- NADIS, 2022a, Part 1 - What does poor fertility cost, <https://www.nadis.org.uk/disease-a-z/cattle/fertility-in-dairy-herds-advanced/part-1-what-does-poor-fertility-cost/> . Accessed 23 November 2022
- NADIS, 2022b, Part 8 - Measuring fertility - Benchmarking your farm, <https://www.nadis.org.uk/disease-a-z/cattle/fertility-in-dairy-herds/part-8-measuring-fertility-benchmarking-your-farm/> . Accessed 23 November 2022
- NAGY, S. Á., KILIM, O., CSABAI, I., GABOR, G., & SOLYMOSI, N. 2023. Impact Evaluation of Score Classes and Annotation Regions in Deep Learning-Based Dairy Cow Body Condition Prediction. *Animals: an open access journal from MDPI*, 13(2), 194.
- NETO, H. A., TAVARES, W. L. F., RIBEIRO, D. C. S. Z., ALVES, R. C. O., FONSECA, L. M., CAMPOS, S. V. A. 2019. On the utilization of deep and ensemble learning to detect milk adulteration. *BioData Min.*, 12, 13.
- NEWSOME, R. F., GREEN, M. J., BELL, N. J., BOLLARD, N. J., MASON, C. S., WHAY, H. R. & HUXLEY, J. N. 2017. A prospective cohort study of digital cushion and corium thickness. Part 2: Does thinning of the digital cushion and corium lead to lameness and claw horn disruption lesions? *J Dairy Sci*, 100, 4759-4771.

- NGUYEN, Q. T., FOUCHEREAU, R., FRENOD, E., GERARD, C., SINCHOLLE, V., 2020. Comparison of forecast models of production of dairy cows combining animal and diet parameters. *Comput. Electron. Agric.*, 170, 105258.
- NIELEN, M., SCHUKKEN, Y. H., BRAND, A., DELUYKER, H. A. & MAATJE, K. 1995a. Detection of subclinical mastitis from on-line milking parlor data. *J Dairy Sci*, 78, 1039-49.
- NIELEN, M., SCHUKKEN, Y. H., BRAND, A., HARING, S. & FERWERDA-VAN ZONNEVELD, R. T. 1995b. Comparison of analysis techniques for on-line detection of clinical mastitis. *J Dairy Sci*, 78, 1050-61.
- NIKOLOSKI, S., MURPHY, P., KOCEV, D., DZEROSKI, S., WALL, D. P. 2019. Using machine learning to estimate herbage production and nutrient uptake on Irish dairy farms. *J. Dairy Sci.*, 102, 10639–10656.
- NIR, O., PARMET, Y., WERNER, D., ADIN, G., HALACHMI, I. 2018. 3D Computer-vision system for automatically estimating heifer height and body mass. *Biosyst. Eng.*, 173, 4–10.
- NJUBI, D. M., WAKHUNGU, J. W. & BADAMANA, M. S. 2010. Use of test-day records to predict first lactation 305-day milk yield using artificial neural network in Kenyan Holstein-Friesian dairy cows. *Trop Anim Health Prod*, 42, 639-44.
- NORDLUND, K., COOK, N. & OETZEL, G. 2006. *Commingleing dairy cows: Pen moves, stocking density, and health*.
- NORDLUND, K. V. & COOK, N. B. 2004. Using herd records to monitor transition cow survival, productivity, and health. *Vet Clin North Am Food Anim Pract*, 20, 627-49.
- NRC 2001. Nutritional requirements of dairy cattle. *Washington, National Academy Press*.
- OEHM, A. W., SPRINGER, A., JORDAN, D., STRUBE, C., KNUBBEN-SCHWEIZER, G., JENSEN, K. C., & ZABLOTSKI, Y. 2022. A machine learning approach using partitioning around medoids clustering and random forest classification to model groups of farms in regard to production parameters and bulk tank milk antibody status of two major internal parasites in dairy cows. *PLoS One*, 17(7), e0271413.

- DE OLIVEIRA, E. B., FERREIRA, F. C., GALVAO, K. N., YOUN, J., TAGKOPOULOS, I., SILVA-DEL-RIO, N., PEREIRA, R. V. V., MACHADO, V. S., & LIMA, F. S. 2021. Integration of statistical inferences and machine learning algorithms for prediction of metritis cure in dairy cows. *J Dairy Sci*, 104(12), 12887–12899.
- OSPINA, P. A., MCART, J. A., OVERTON, T. R., STOKOL, T. & NYDAM, D. V. 2013. Using Nonesterified Fatty Acids and β -Hydroxybutyrate Concentrations During the Transition Period for Herd-Level Monitoring of Increased Risk of Disease and Decreased Reproductive and Milking Performance. *Veterinary Clinics of North America: Food Animal Practice*, 29, 387-412.
- OVERTON, T. R. & WALDRON, M. R. 2004. Nutritional Management of Transition Dairy Cows: Strategies to Optimize Metabolic Health. *J Dairy Sci*, 87, E105-E119.
- PACHERO, V. M., DE SOUSA, R. V., DA SILVA RODRIGUEZ, A. V., DE SOUZA SARDINHA, E. J. & MARTELLO, L. S. 2020. Thermal imaging combined with predictive machine learning based model for the development of thermal stress level classifiers. *Livest. Sci.*, 241, 104244.
- PANCHAL, I., SAWHNEY, I. K., SHARMA, A. K. & DANG, A. K. 2016. Classification of healthy and mastitis Murrah buffaloes by application of neural network models using yield and milk quality parameters. *Computers and Electronics in Agriculture*, 127, 242-248.
- PASTELL, M. E. & KUJALA, M. 2007. A probabilistic neural network model for lameness detection. *J Dairy Sci*, 90, 2283-92.
- PATEL, N.J., JHAVERI, R.H. 2015. Detecting packet dropping nodes using machine learning techniques in mobile ad-hoc network: a survey. 2015 International Conference on Signal Processing and Communication Engineering Systems, 468–472.
- PAWLAK, Z. 2003. Bayes' Theorem — the Rough Set Perspective. In: Inuiguchi, M., Hirano, S., Tsumoto, S. (eds) Rough Set Theory and Granular Computing. Studies in Fuzziness and Soft Computing, vol 125. Springer, Berlin, Heidelberg. 1-12.
- PAZHANIKUMAR, K. and ASWATHI, R. (2020). Performance of naïve bayes, c4.5 and knn using breast cancer, iris and hypothyroid datasets.

- International *Journal of Innovative Technology and Exploring Engineering*, 9(3), 2193-2197.
- PERKINS, K. H., VANDEHAAR, M. J., TEMPELMAN, R. J. & BURTON, J. L. 2001. Negative energy balance does not decrease expression of leukocyte adhesion or antigen-presenting molecules in cattle. *J Dairy Sci*, 84, 421-8.
- PIETERSMA, D., LACROIX, R., LEFEBVRE, D., WADE, K. M., 2003. Induction and evaluation of decision trees for lactation curve analysis. *Comput. Electron. Agric.*, 38, 19–32.
- PINEDO, P. J. & FLEMING, 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. *J Dairy Sci*, 95(12), 7015-7026.
- PIWCZYNSKI, D., SITKOWSKA, B., KOLENDA, M., BRZOZOWSKI, M., AERTS, J., & SCHORK, P. M. 2020. Forecasting the milk yield of cows on farms equipped with automatic milking system with the use of decision trees. *Animal Science Journal*, 91(1).
- POLSKY, L. & VON KEYSERLINGK, M. A. G. 2017. Invited review: Effects of heat stress on dairy cattle welfare. *J Dairy Sci*, 100, 8645-8657.
- POST, C., RIETZ, C., BUSCHER, W., & MULLER, U. 2020. Using sensor data to detect lameness and mastitis treatment events in dairy cows: a comparison of classification models. *Sensors*, 20(14), 3863.
- POST, C., RIETZ, C., BUSCHER, W., & MULLER, U. 2021. The Importance of Low Daily Risk for the Prediction of Treatment Events of Individual Dairy Cows with Sensor Systems. *Sensors (Basel, Switzerland)*, 21(4), 1389.
- POURHOSEINGHOLI, M. A., BAGHESTANI, A. R. & VAHEDI, M. 2012. How to control confounding effects by statistical analysis. *Gastroenterol. Hepatol. Bed Bench* 5, 79–83.
- PRALLE, R.S. WEIGEL, K.W. & WHITE, H.M. 2018. Predicting blood β -hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network, *J Dairy Sci*, 101, 5, 4378-4387,

- PROBO, M., PASCOTTINI, O. B., LEBLANC, S., OPSOMER, G. & HOSTENS, M. 2018. Association between metabolic diseases and the culling risk of high-yielding dairy cows in a transition management facility using survival and decision tree analysis. *J Dairy Sci*, 101, 9419-9429.
- PRUNIER, J., COLYN, M., LEGENDRE, X., NIMON, K., & FLAMAND, M. 2015. multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses. *Molecular Ecology*, 24(2), 263-283.
- PRYCE, J. E., COFFEY, M. P. & SIMM, G. 2001. The relationship between body condition score and reproductive performance. *J Dairy Sci*, 84, 1508-15.
- PRYCE, J. E. & HARRIS, B. L. 2006. Genetics of body condition score in New Zealand dairy cows. *J Dairy Sci*, 89, 4424-32.
- QU, G., ZHANG, H., & HARTRICK, C. T., 2011. Multi-label classification with Bayes' theorem, 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI), 2281-2285.
- OUYANG C.G., CHEN P., 2020, Factor endowment, local industrial sector development and industry choice[J]. *Economic Research*, 55(01):82-98.
- R CORE TEAM 2018. R: A language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- RADOVANOVIC, M. NANOPOULOS, A. & IVANONIC, M. 2010. Hubs in ' space: popular nearest neighbors in high-dimensional data, *Journal of Machine Learning Research (JMLR)*, 11, 2487– 2531.
- RAMEZAN, C. A., WARNER, T. A., MAXWELL, A. E., & PRICE, B. S. 2021. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. *Remote Sensing*, 13(3), 368.
- RAHAMAN, N., BARATIN, A. , ARPIT, D., DRAXLER, F., LIN, M., 2019. Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In International Conference on Machine Learning, 5301–5310.
- RAIKO, T. VALPOLA, H., and LECUN. Y., 2012. Deep learning made easier by linear transformations in perceptrons. In Conference on AI and Statistics (JMLR W&CP), 22, 924–932,

- RANDALL, L. V., THOMAS, H. J., REMNANT, J. G., BOLLARD, N. J. & HUXLEY, J. N. 2019. Lameness prevalence in a random sample of UK dairy herds. *Veterinary Record*, 184, 350.
- REJAB, F., NOUIRA K., & TRABELSI, A., 2014, RTSVM: Real time support vector machines, 2014 Science and Information Conference, 1038-1042
- REN, K., BERNES, G., HETTA, M., KARLSSON, J., 2021. Tracking and analysing social interactions in dairy cattle with real-time locating system and machine learning. *J. Syst. Archit.*, 116, 102139.
- RESHEFF, Y. S., ROTICS, S., HAREL, R., SPIEGEL, O. & NATHAN, R. 2014. AcceleRater: a web application for supervised learning of behavioral modes from acceleration measurements. *Mov Ecol*, 2, 27.
- REUNANEN, J. 2003. Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res* 3, 1371–1382 .
- REYNOLDS, C. K., AIKMAN, P. C., LUPOLI, B., HUMPHRIES, D. J. & BEEVER, D. E. 2003. Splanchnic metabolism of dairy cows during the transition from late gestation through early lactation. *J Dairy Sci*, 86, 1201-17.
- REYNOLDS, C. K., DURST, B., LUPOLI, B., HUMPHRIES, D. J. & BEEVER, D. E. 2004. Visceral tissue mass and rumen volume in dairy cows during the transition from late gestation to early lactation. *J Dairy Sci*, 87, 961-71.
- RIABOFF, L., POGGI, S., MADOUASSE, A., COUVREUR, S., AUBIN, S., BEDERE, N., GOUMAND, E., CHAUVIN, A., PLANTIER, G., 2020. Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data. *Comput. Electron. Agric.*, 169, 105179.
- RIBEIRO, E., LIMA, F., GRECO, L., BISINOTTO, R., MONTEIRO, A., FAVORETO, M. G., & SANTOS, J. 2013. Prevalence of periparturient diseases and effects on fertility of seasonally calving grazing dairy cows supplemented with concentrates. *J Dairy Sci*, 96(9), 5682-5697.
- RIGATTI, S. J., 2017. Random forest. *Journal of Insurance Medicine*, 47(1), 31-39.

- RINGSEIS, R., GESSNER, D. K., & EDER, K. 2014. Molecular insights into the mechanisms of liver-associated diseases in early-lactating dairy cows: hypothetical role of endoplasmic reticulum stress. *Journal of Animal Physiology and Animal Nutrition*, 99(4), 626-645.
- RISCO, C. A., DROST, M., THATCHER, W. W., SAVIO, J. & THATCHER, M. J. 1994. Effects of calving-related disorders on prostaglandin, calcium, ovarian activity and uterine involution in postrartum dairy cows. *Theriogenology*, 42, 183-203.
- ROCHE, J. R. 2003. The incidence and control of hypocalcaemia in pasture-based systems. *Acta Vet Scand Suppl*, 97, 141-4.
- ROCHE, J. R., BERRY, D. P. & KOLVER, E. S. 2006. Holstein-Friesian strain and feed effects on milk production, body weight, and body condition score profiles in grazing dairy cows. *J Dairy Sci*, 89, 3532-43.
- ROCHE, J. R., BERRY, D. P., LEE, J. M., MACDONALD, K. A. & BOSTON, R. C. 2007. Describing the body condition score change between successive calvings: a novel strategy generalizable to diverse cohorts. *J Dairy Sci*, 90, 4378-96.
- ROCHE, J. R., FRIGGENS, N. C., KAY, J. K., FISHER, M. W., STAFFORD, K. J. & BERRY, D. P. 2009. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. *J Dairy Sci*, 92, 5769-801.
- ROCHE, J. R., BURKE, C. R., CROOKENDEN, M. A., HEISER, A., LOOR, J. L., MEIER, S., MITCHELL, M. D., PHYN, C. V. C., TURNER, S. A. 2017. Fertility and the transition dairy cow. *Reproduction, Fertility and Development* 30, 85-100.
- RODRIGUEZ, Z., SHEPLEY, E., FERRO, P. P. C., MORAES, N. L., ANTUNES, A., CRAMER, G. & CAIXETA, L. 2021. Association of body condition score and score change during the late dry period on temporal patterns of beta-hydroxybutyrate concentration and milk yield and composition in early lactation of dairy cows. *Animals*, 11(4), 1054.
- RODRIGUEZ, E., WAISSMAN, J., MAHADEVAN, P., VILLA, C., FLORES, B. L., VILLA, R., 2019, Genome-wide classification of dairy cows using

- decision trees and artificial neural network algorithms. *Genet. Mol. Res.*, 18, gmr18407.
- RODRIGUEZ ALVAREZ, J., ARROGUI, M., MANGUDO, P., TOLOZA, J., JATIP, D., RODRIGUEZ, J. M., TEYSEYRE, A., SANZ, C., ZUNINO, A., MACHADO, C., MATEOS, C., 2018. Body condition estimation on cows from depth images using Convolutional Neural Networks. *Comput. Electron. Agric.*, 155, 12–22.
- RODRIGUEZ ALVAREZ, J. R., ARROQUI, M., MANGUDO, P., TOLOZA, J., JATIP, D., RODRIGUEZ, J. M., TEYSEYRE, A., SANZ, C., ZUNINO, A., MACHADO, C., & MATEOS, C. 2019. Estimating body condition score in dairy cows from depth images using convolutional neural networks, transfer learning and model ensembling techniques. *Agronomy*, 9, 90.
- ROJAS-DUENAS, G., RIBA, J., KAHALERRAS, K., MORENO-EGUILAZ, M., KADECHKAR, A., & GOMEZ-PAU, A. 2020. Black-box modelling of a DC-DC buck converter based on a recurrent neural network. *Institute of Electrical and Electronics Engineers (IEEE)*, 456-461.
- ROLLIN, E., DHUYVETTER, K. C. & OVERTON, M. W. 2015. The cost of clinical mastitis in the first 30 days of lactation: An economic modeling tool. *Prev Vet Med*, 122, 257-64.
- ROMADHONNY, R. A., GUMELAR, A. B., FAHRUDIN, T. M., ADI SETIAWAN, W. P., CAHAYA PUTRA, F. D., NUGROHO, R. D., BUDIANI, J. R. 2019. Estrous Cycle Prediction of Dairy Cows for Planned Artificial Insemination (AI) Using Multiple Logistic Regression. In Proceedings of the 2019 International Seminar on Application for Technology of Information and Communication: Industry 4.0: Retrospect, Prospect, and Challenges, 21–22 157–162.
- SABORÍO-MONTERO, A., VARGAS-LEITÓN, B., ROMERO-ZÚÑIGA, J. J. & SÁNCHEZ, J. M. 2017. Risk factors associated with milk fever occurrence in grazing dairy cattle. *J Dairy Sci*, 100, 9715-9722.
- SADEGHI, H., BRAUN, H. S., PANTI, B., OPSOMER, G., & BOGADO PASCOTINNI, O. 2022. Validation of a deep learning-based image

- analysis system to diagnose subclinical endometritis in dairy cows. *PloS one*, 17(1), e0263409.
- SAED H., IBRAHIM, H., EL-KHODERY, S., YOUSSEF M. A. 2020. Prevalence and potential risk factors of hypocalcaemia in dairy cows during transition period at northern egypt. *Mansoura Veterinary Medical Journal*, 21(1), 21-30.
- SAINANI, K. L. 2014. Explanatory versus predictive modeling. *Pm r*, 6, 841-4.
- SALAMONE, M., ADRIAENS, I., VERVAET A., OPSOMER G., ATASHI H., FIEVEZ V., AERNOUTS B., HOSTENS M. 2022. Prediction of first test day milk yield using historical records in dairy cows. *Animal*. 16(11):100658.
- SALAU, J., HAAS, J. H., JUNGE, W., THALLER, G. 2021. Determination of body parts in holstein friesian cows comparing neural networks and k nearest neighbour classification. *Animals*, 11, 50.
- SALAU, J., KRIETER, J. 2020. Instance segmentation with mask R-CNN applied to loose-housed dairy cows in a multi-camera setting. *Animals*, 10, 2402.
- SALEHI, F. LACROIX, R. WADE, K.M. 1998. Improving dairy yield predictions through combined record classifiers and specialized artificial neural networks, *Computers and Electronics in Agriculture*, Volume 20, Issue 3, 199-213
- SALZER, Y., HONIG, H. H., SHAKED, R., ABELES, E., KLEINJAN-ELAZARY, A., BERGER, K., JACOBY, S., FISHBAIN, B. & KENDLER, S. 2021. Towards on-site automatic detection of noxious events in dairy cows. *Appl. Anim. Behav. Sci.*, 236, 105260
- SANZOGNI, L., KERR, D., 2001. Milk production estimates using feed forward artificial neural networks, *Computers and Electronics in Agriculture*, Volume 32, Issue 1, 21-30
- SCHEFERS, J. M., WEIGEL, K. A., RAWSON, C. L., ZWALD, N. R. & COOK, N. B. 2010. Management practices associated with conception rate and

- service rate of lactating Holstein cows in large, commercial dairy herds. *J Dairy Sci*, 93, 1459-67.
- SCHIRMANN, K., CHAPINAL, N., WEARY, D. M., HEUWIESER, W. & VON KEYSERLINGK, M. A. 2011. Short-term effects of regrouping on behavior of prepartum dairy cows. *J Dairy Sci*, 94, 2312-9.
- SCHLEMMER, A., ZWIRNMANN, H., ZABEL, M., PARLITZ, U. & LUTHER, S., 2014. Evaluation of machine learning methods for the long-term prediction of cardiac diseases, 2014 8th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO), 157-158
- SCHNITZER, M. J. & MEISTER, M. 2003 Multineuronal firing patterns in the signal from eye to brain. *Neuron*, 37(3), 499-511.
- SCHONLAU, M., & YUYAN ZOU, R., 2020. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29.
- SCHWEINZER, V., GUSTERER, E., KANZ, P., KRIEGER, S., SUSS, D., LIDAUER, L., BERGER, A., KICKINGER, F., ÖHLSCHUSTER, M., & AUER W., 2019 Evaluation of an ear-attached accelerometer for detecting estrus events in indoor housed dairy cows. *Theriogenology*, 130, 19–25.
- SEFFEDPARI, P., RAFIEE, S., & AKRAM, A. 2013. Application of artificial neural network to model the energy output of dairy farms in Iran. *Int. J. Energy Technol. Policy*, 9, 82.
- SEFFEDPARI, P., RAFIEE, S., AKRAM, A., CHAU, K. W., & KOMLEH, S. H. P., 2015. Modeling Energy Use in Dairy Cattle Farms by Applying Multi-Layered Adaptive Neuro-Fuzzy Inference System (MLANFIS). *Int. J. Dairy Sci.*, 10, 173–185.
- SEFFEDPARI, P., RAFIEE, S., AKRAM, A., & KOMLEH, S. H. P., 2014. Modeling output energy based on fossil fuels and electricity energy consumption on dairy farms of Iran: Application of adaptive neural-fuzzy inference system technique. *Comput. Electron. Agric.*, 109, 80–85.
- SEPULVEDA-VARAS, P., LOMB, J., VON KEYSERLINGK, M. A. G., HELD, R., BUSTAMANTE, H. & TADICH, N. 2018. Claw horn lesions in mid-lactation primiparous dairy cows under pasture-based systems: Association with behavioral and metabolic changes around calving. *J Dairy Sci*, 101, 9439-9450.

- SHAFIULLAH, A. Z., WERNER, J., KENNEDY, E., LESO, L., O'BRIEN, B., & UMSTATTER, C., 2019. Machine learning based prediction of insufficient herbage allowance with automated feeding behaviour and activity data. *Sensors*, 19, 4479.
- SHAHINFAR, S., MEHRABANI-YEGANEH, H., LUCAS, C., KALHOR, A., KAZEMIAN, M. & WEIGEL, K. A. 2012. Prediction of breeding values for dairy cattle using artificial neural networks and neuro-fuzzy systems. *Comput Math Methods Med*, 2012, 127130.
- 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. *J Dairy Sci*, 97, 731-742.
- SHAHINFAR, S., KHANSEFID, M., HAILE-MARIAM, M., & PRYCE, JE. 2021. Machine learning approaches for the prediction of lameness in dairy cows. *Animal*. 100391.
- SHARIFI, S., PAKDEL, A., EBRAHIMI, M., REECY, J. M., FAZELI FARSANI, S. & EBRAHIMIE, E. 2018. Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. *PLoS One*, 13, e0191227.
- SHARMA, A.K., SHARMA, R.K., & KASANA, H.S.,2006. Empirical comparisons of feed-forward connectionist and conventional regression models for prediction of first lactation 305-day milk yield in Karan Fries dairy cows. *Neural Comput & Applic* 15, 359–365.
- SHARMA, A.K., SHARMA, R.K., & KASANA, H.S.,2007. Prediction of first lactation 305-day milk yield in Karan Fries dairy cattle using ANN modeling, *Applied Soft Computing*, Volume 7, Issue 3, 1112-1120.
- SHAVER, R. D. 1997. Nutritional risk factors in the etiology of left displaced abomasum in dairy cows: a review. *J Dairy Sci*, 80, 2449-53.
- SHELDON, I. M. 2004. The postpartum uterus. *Vet Clin North Am Food Anim Pract*, 20, 569-91.

- SHELDON, I. M., LEWIS, G. S., LEBLANC, S., & GILBERT, R. O. (2006). Defining postpartum uterine disease in cattle. *Theriogenology*, 65(8), 1516–1530.
- SHELDON, I. M., CRONIN, J. G., GOETZE, L., DONOFRIO, G., & SCHUBERTH, H. 2009. Defining postpartum uterine disease and the mechanisms of infection and immunity in the female reproductive tract in cattle¹. *Biology of Reproduction*, 81(6), 1025-1032.
- SHELDON, I. M., LEWIS, G. S., LEBLANC, S. & GILBERT, R. O. 2006a. Defining postpartum uterine disease in cattle. *Theriogenology*, 65, 1516-30.
- SHELDON, I. M., WATHES, D. C. & DOBSON, H. 2006b. The management of bovine reproduction in elite herds. *The Veterinary Journal*, 171, 70-78.
- SHEN, W., CHENG, F., ZHANG, Y., WEI, X., FU, Q., & ZHANG, Y. 2020. Automatic recognition of ingestive-related behaviors of dairy cows based on triaxial acceleration. *Inf. Process. Agric.*, 7, 427–443
- SHINE, P., & MURPHY, M. D. 2022. Over 20 Years of Machine Learning Applications on Dairy Farms: A Comprehensive Mapping Study. *Sensors*, 22, 52.
- SHINE, P., MURPHY, M. D., UPTON, J., & SCULLY, T. 2018a. Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms. *Comput. Electron. Agric.*, 150, 74–87.
- SHINE, P., SCULLY, T., UPTON, J., & MURPHY, M. D., 2018b. Multiple linear regression modelling of on-farm direct water and electricity consumption on pasture based dairy farms. *Comput. Electron. Agric.*, 148, 337–346.
- SHINE, P., SCULLY, T., UPTON, J., & MURPHY, M. D. M. 2019. Annual electricity consumption prediction and future expansion analysis on dairy farms using a support vector machine. *Appl. Energy*, 250, 1110–1119.
- SHMUELI, G. 2010. To Explain or to Predict? *Statist. Sci.*, 25, 289-310.
- SHRESTHA, N. 2020. Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8(2), 39-42.
- SHRESTHA, A., LOUKAS, C., LE KERNEC, J., FIORANELLI, F., BUSIN, V., JONSSON, N., KING, G., TOMLINSON, M., VIORA, L., & VOUTE, L.

2018. Animal lameness detection with radar sensing. *IEEE Geosci. Remote Sens. Lett.*, 15, 1189–1193.
- SIACHOS, N., LENNOX, M., ANAGNOSTOPOULOS, A., GRIFFITHS, B. E., NEARY, J. M., SMITH, R. F., & OIKONOMOU, G. 2024. Development and validation of a fully automated 2-dimensional imaging system generating body condition scores for dairy cows using machine learning. *J Dairy Sci*, 107(4), 2499–2511.
- SILVA DEL RIO, N., STEWART, S., RAPNICKI, P., CHANG, Y. M. & FRICKE, P. M. 2007. An observational analysis of twin births, calf sex ratio, and calf mortality in Holstein dairy cattle. *J Dairy Sci*, 90, 1255-64.
- SILVA, P. R., DRESCH, A. R., MACHADO, K. S., MORAES, J. G., LOBECK-LUCHTERHAND, K., NISHIMURA, T. K., FERREIRA, M. A., ENDRES, M. I. & CHEBEL, R. C. 2014. Prepartum stocking density: effects on metabolic, health, reproductive, and productive responses. *J Dairy Sci*, 97, 5521-32.
- SILVA, P. R., MORAES, J. G., MENDONCA, L. G., SCANAVEZ, A. A., NAKAGAWA, G., BALLOU, M. A., WALCHECK, B., HAINES, D., ENDRES, M. I. & CHEBEL, R. C. 2013. Effects of weekly regrouping of prepartum dairy cows on innate immune response and antibody concentration. *J Dairy Sci*, 96, 7649-57.
- SINGH, B., KUMAR, S., ELANGO VAN, A., VASHT, D., ARYA, S., DUC, N., & CHINNUSAMY, V. 2023. Phenomics based prediction of plant biomass and leaf area in wheat using machine learning approaches. *Frontiers in Plant Science*, 14.
- SLOB, N., CATAL, C., & KASSAHU, A., 2021. Application of machine learning to improve dairy farm management: A systematic literature review, *Preventive Veterinary Medicine*, Volume 187, 105237.
- SMITH, S. M. & NICHOLS, T. E. 2018. Statistical challenges in “big data” human neuroimaging. *Neuron* 97, 263–268.
- SMITH, D., RAHMAN, A., BISHOP-HURLEY, G. J., HILLS, J., SHAHRIAR, S., HENRY, D., & RAWNSLEY, R. 2016. Behavior classification of cows fitted with motion collars: Decomposing multi-class classification into a set of binary problems. *Comput. Electron. Agric.*, 131, 40–50.

- SMOLA, A., SCHOLKOPF, A., & MULLER, R., 1998. The connection between regularization operators and support vector kernels, *Neural Network*, 11,637-649
- SOMASUNDARAM, A. & REDDY, U. S. 2018. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. *Neural Computing and Applications*, 31(S1), 3-14.
- SORDILLO, L. M. 2016. Nutritional strategies to optimize dairy cattle immunity. *J Dairy Sci*, 99(6), 4967-4982.
- SRIKOK, S., PATCHANEE, P., BOONVAVATRA, S., CHUAMMITRI, P. 2020. Potential role of MicroRNA as a diagnostic tool in the detection of bovine mastitis. *Prev. Vet. Med.*, 182, 105101.
- SPISAK, T. 2022. Statistical quantification of confounding bias in machine learning models. *Gigascience*, 11.
- ST-PIERRE, N. R., COBANOV, B. & SCHNITKEY, G. 2003. Economic Losses from Heat Stress by US Livestock Industries¹. *J Dairy Sci*, 86, E52-E77.
- STURM, V., EFROSININ, D., ÖHLSCHUSTER, M., GUSTERER, E., DRILLICH, M., & IWERSEN, M. 2020. Combination of Sensor Data and Health Monitoring for Early Detection of Subclinical Ketosis in Dairy Cows. *Sensors (Basel, Switzerland)*, 20(5), 1484.
- SUN, Z., SAMARASINGHE, S. & JAGO, J. 2010. Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks. *J Dairy Res*, 77, 168-75.
- SURIYASATHAPORN, W., HEUER, C., NOORDHUIZEN-STASSEN, E. N. & SCHUKKEN, Y. H. 2000. Hyperketonemia and the impairment of udder defense: a review. *Vet Res*, 31, 397-412.
- TAMURA, T., OKUBO, Y., DEGUCHI, Y., KOSHIKAWA, S., TAKAHASHI, M., CHIDA, Y. & OKADA, K. 2019. Dairy cattle behavior classifications based on decision tree learning using 3-axis neck-mounted accelerometers. *Anim Sci J*, 90, 589-596.
- TANEJA, M., BYABAZAIRE, J., JALODIA, N., DAVY, A., OLARIU C., & MALONE, P. 2020. Machine learning based fog computing assisted

- data-driven approach for early lameness detection in dairy cattle. *Comput. Electron. Agric.*, 171, 105286.
- TAKAHASHI, Y., UEKI, M., YAMADA, M., TAMIYA, G., MOTOIKE, I. N., SAIGUSA, D., & TOMITA, H. 2020. Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. *Translational Psychiatry*, 10(1).
- TEDDE, A., GRELET, C., HO, P. N., PRYCE, J. E., HAILEMARIAM, D., WANG, Z., PLASTOW, G., GENGLER, N., FROIDMONT, E., DEHARENG, F., BERTOZZI, C., CROWE, M. A. & SOYEURT, H., 2021a. Multiple country approach to improve the test-day prediction of dairy cows' dry matter intake. *Animals*, 11, 1316.
- TEDDE, A., GRELET, C., HO, P. N., PRYCE, J. E., HAILEMARIAM, D., WANG, Z., PLASTOW, G., GENGLER, N., FROIDMONT, E., DEHARENG, F., BERTOZZI, C., CROWE, M. A. & SOYEURT, H., 2021b Validation of Dairy Cow Bodyweight Prediction Using Traits Easily Recorded by Dairy Herd Improvement Organizations and Its Potential Improvement Using Feature Selection Algorithms. *Animals* 2021, 11, 1288.
- THERNEAU, T. 2023. *A Package for Survival Analysis in R*. R package version 3.5-0, <https://CRAN.R-project.org/package=survival>.
- TODDE, G., MURGIA, L., CARIA, M. & PAZZONA, A. 2017. Dairy Energy Prediction (DEP) model: A tool for predicting energy use and related emissions and costs in dairy farms. *Comput. Electron. Agric.*, 135, 216–221.
- TONI, F., VINCENTI, L., GRIGOLETTO, L., RICCI, A. & SCHUKKEN, Y. H. 2011. Early lactation ratio of fat and protein percentage in milk is associated with health, milk production, and survival. *J Dairy Sci*, 94, 1772-83.
- TOPOL, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56.
- TORRES, M., HERVAS, C., & AMADOR, F., 2005. Approximating the sheep milk production curve through the use of artificial neural networks and genetic algorithms, *Computers & Operations Research*, Volume 32, Issue 10, 2653-2670.

- UTRERA, Á. R., CADLERON-ROBLES, R. C., GALAVIZ-RODRIGUEZ, J. R., MURILLO, V. E. V., & LAGUNES-LAGUNES, J. 2013. Effects of breed, calving season and parity on milk yield, body weight and efficiency of dairy cows under subtropical conditions. *International Journal of Animal and Veterinary Advances*, 5(6), 226-232.
- VALVIDIA, V., BARRADO, A., LAAZARO, A., ZUMEL, P., RAGA, C., & FERNANDEZ, C., 2009, Simple Modeling and Identification Procedures for 'Black-Box' Behavioral Modeling of Power Converters Based on Transient Response Analysis, *IEEE Trans. Power Electron.*, 24. 12, 2776–2790.
- VAN DORP, T. E., DEKKERS, J. C., MARTIN, S. W. & NOORDHUIZEN, J. P. 1998. Genetic parameters of health disorders, and relationships with 305-day milk yield and conformation traits of registered Holstein cows. *J Dairy Sci*, 81, 2264-70.
- VAN WINDEN, S. C., BRATTINGA, C. R., MULLER, K. E., NOORDHUIZEN, J. P. & BEYNEN, A. C. 2002. Position of the abomasum in dairy cows during the first six weeks after calving. *Vet Rec*, 151, 446-9.
- VAPNIK, V., 2013. The nature of statistical learning theory. Springer science & business media.
- VARMA, S. & SIMON, 2006. R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* 7, 91.
- VATCHEVA K., LEE, M., MCCORMICK, J. B., & RAHBAR, M. H. 2016. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology: Open Access*, 06(02).
- VÁZQUEZ DIOSDADO, J. A., BARKER, Z. E., HODGES, H. R., AMORY, J. R., CROFT, D. P., BELL, N. J. & CODLING, E. A. 2015. Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. *Animal Biotelemetry*, 3, 15.
- VÁZQUEZ DIOSDADO, J. A., GRUHIER, J., MIGUEL-PACHECO, G. G., GREEN, M., DOTTORINI, T., & KALER, J. 2023. Accurate prediction of calving in dairy cows by applying feature engineering and machine learning. *Preventive Veterinary Medicine*, 219, 106007.

- VERGARA, C. F., DOPFER, D., COOK, N. B., NORDLUND, K. V., MCART, J. A., NYDAM, D. V. & OETZEL, G. R. 2014. Risk factors for postpartum problems in dairy cows: explanatory and predictive modeling. *J Dairy Sci*, 97, 4127-40.
- VIERA, A. J. & GARRETT, J. M. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37, 360-3.
- VIJAYAKUMAR, M., PARK, J. H., KI, K., LIM, D. H., KIM, S. B., PARK, S. M., KIM, T. I. 2017. The effect of lactation number, stage, length, and milking frequency on milk yield in korean holstein dairy cows using automatic milking system. *Asian-Australasian Journal of Animal Sciences*, 30(8), 1093-1098.
- VIJU, K., 2021. Stability of neural networks dependent on time series in anime image recognition. *International Journal of neural network*, 2(3).
- VITALI, A., SEGNALINI, M., BERTOCCHI, L., BERNABUCCI, U., NARDONE, A. & LACETERA, N. 2009. Seasonal pattern of mortality and relationships between mortality and temperature-humidity index in dairy cows. *J Dairy Sci*, 92, 3781-3790.
- VON KEYSERLINGK, M. A., OLENICK, D. & WEARY, D. M. 2008. Acute behavioral effects of regrouping dairy cows. *J Dairy Sci*, 91, 1011-6.
- VOLKMANN N, KULIG B, HOPPE S, STRACKE J, HENSEL O, KEMPER N. 2021. On-farm detection of claw lesions in dairy cows based on acoustic analyses and machine learning. *J Dairy Sci*.104(5):5921-5931.
- WAGNER, N., ANTOINE, V., MIALON, M. M., LARDY, R., SILBERBERG, M., KOKO, J., & VEISSIER, I. 2020. Machine learning to detect behavioural anomalies in dairy cows under subacute ruminal acidosis. *Comput. Electron. Agric.*, 170, 105233.
- WAITMAN, L. R., FISHER, D., & KING, P. 2003. Bootstrapping rule induction. In Proceedings of the IEEE International Conference on Data Mining, Los Alamitos, CA: *IEEE Computer Society*. 677–680.
- WALDRON, M. R., NISHIDA, T., NONNECKE, B. J. & OVERTON, T. R. 2003. Effect of lipopolysaccharide on indices of peripheral and hepatic metabolism in lactating cows. *J Dairy Sci*, 86, 3447-59.
- WALLESER, E., REYES, J. F. M., ANKLAM, K., PRALLE, R. S., WHITE, H. M., UNGER, S., PANNE, N., KAMMER, M., PLATTNER, S., & DOPFER, D.

2023. Novel prediction models for hyperketonemia using bovine milk Fourier-transform infrared spectroscopy. *Preventive veterinary medicine*, 213, 105860.
- WANG, J., BELL, M., LIU, X., & LIU, G. 2020. Machine-Learning Techniques Can Enhance Dairy Cow Estrus Detection Using Location and Acceleration Data. *Animals : an open access journal from MDPI*, 10(7), 1160.
- WANG, H., SHEN, W., ZHANG, Y., GAO, M., ZHANG, Q., A, X., DU, H. & QIU, B. 2023. Diagnosis of dairy cow diseases by knowledge-driven deep learning based on the text reports of illness state, *Computers and Electronics in Agriculture*, Volume 205, 107564, 0168-1699
- WANKHADE, P. R., MANIMARAN, A., KUMARESAN, A., JEYAKUMAR, S., RAMESHA, K. P., SEJIAN, V., RAJENDRAN, D. & VARGHESE, M. R. 2017. Metabolic and immunological changes in transition dairy cows: A review. *Veterinary world*, 10, 1367-1377.
- WARNER, D., VASSEUR, E., LEFEBVRE, D. M., & LACROIX, R. 2020. A machine learning based decision aid for lameness in dairy herds using farm-based records. *Comput. Electron. Agric.*, 169, 105193.
- WARRENS, M. 2010. chance-corrected measures for 2 × 2 tables that coincide with weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 64(2), 355-365.
- WEI, W., DONG, L., YE, J., & XIAO, Z. 2024. Current status and influencing factors of family resilience in families of children with epilepsy: a cross-sectional study. *Frontiers in Psychiatry*, 15.
- WEI, C. & HSU, N., 2008. Derived operating rules for a reservoir operation system: comparison of decision trees, neural decision trees and fuzzy decision trees. *Water Resources Research*, 44(2).
- WEIGEL, K. A., VANRADEN, P. M., NORMAN, H. D. & GROSU, H. 2017. A 100-Year Review: Methods and impact of genetic selection in dairy cattle—From daughter–dam comparisons to deep learning algorithms. *J Dairy Sci*, 100, 10234-10250.
- WEN, J., THIBEAU-SUTRE, E., DIAZ-MELO, M., SAMPER-GONZALEZ, J., ROUTIER, A., BOTTANI, S., DORMONT, D., DURRLEMAN, S., BURGOS, N., & COLLIOT, O., 2020. Alzheimer's Disease Neuroimaging

- Initiative, & Australian Imaging Biomarkers and Lifestyle flagship study of ageing. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63, 101694.
- WILKES, C. O., PENCE, K. J., HURT, A. M., BECVAR, O., KNOWLTON, K. F., MCGILLIARD, M. L. & GWAZDAUSKAS, F. C. 2008. Effect of relocation on locomotion and cleanliness in dairy cows. *J Dairy Res*, 75, 19-23.
- WILLIAMS, M. L., JAMES, W. P., ROSE, M. T., 2019. Variable segmentation and ensemble classifiers for predicting dairy cow behaviour. *Biosyst. Eng.*, 178, 156–167.
- WILLIAMS, M. L., MAC PARTHALÁIN, N., BREWER, P., JAMES, W. P. J. & ROSE, M. T. 2016. A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. *J Dairy Sci*, 99, 2063-2075.
- WILSON, D.J., GONZALEZ, R.N, HERTLJ., SCHULTEH.F., BENNETTG.J., SCHUKKENY.H., & GROHN, Y.T., 2004.Effect of Clinical Mastitis on the Lactation Curve: A Mixed Model Estimation Using Daily Milk Weights, *J Dairy Sci*, Volume 87, Issue 7,2073-2084.
- WISNIESKI, L., NORBY, B., PIERCE, S. J., BECKER, T., GANDY, J. C. & SORDILLO, L. M. 2019. Predictive models for early lactation diseases in transition dairy cattle at dry-off. *Prev Vet Med*, 163, 68-78.
- WITTEK, T., FURLL, M. & CONSTABLE, P. D. 2004. Prevalence of endotoxemia in healthy postparturient dairy cows and cows with abomasal volvulus or left displaced abomasum. *J Vet Intern Med*, 18, 574-80.
- WU, D., FENG, T., NAEHRIG, M., & LAUTER, K., 2016. Privately evaluating decision trees and random forests. *Proceedings on Privacy Enhancing Technologies*, 4, 335-355. XI, W. 2024. Using stepwise regression to address multicollinearity is not appropriate. *International Journal of Surgery*, 110(5), 3122-3123.
- XU. Z. J., 2018. Understanding training and generalization in deep learning by fourier analysis. arXiv preprint arXiv:1808.04295.

- XU, W., VAN KNEGSEL, A. T. M., VERVOORT, J. J. M., BRUCKMAIER, R. M., VAN HOEIJ, R. J., KEMP, B. & SACCENTI, E. 2019. Prediction of metabolic status of dairy cows in early lactation with on-farm cow data and machine learning algorithms. *J. Dairy Sci.*, 102, 10186–10201.
- YAGIS, E., ATNAFU, S.W., GARCIA SECO DE HERRERA, A., MARZI, C., SCHEDA, R., GIANNELLI, M., TESSA, C., CITI, L., & DICIOTTI, S., 2021. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci Rep* 11, 22544.
- YANG, L. & SHAMI, A. 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*, 415, 295-316.
- YANG, H., XIE, X., KADOCH, M. 2022, Machine learning techniques and a case study for intelligent wireless networks. *IEEE Netw.* 34(3), 208–215.
- YANG, Y. & WEBB, G. 2001. Proportional k-interval discretization for naive-bayes classifiers., Machine Learning: ECML 2001. ECML 2001. Lecture Notes in Computer Science(), vol 2167, 564-575.
- YAO, C., SPURLOCK, D. M., ARMENTANO, L. E., PAGE, C. D., JR., VANDEHAAR, M. J., BICKHART, D. M. & WEIGEL, K. A. 2013. Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *J Dairy Sci*, 96, 6716-29.
- YAO, C., ZHU, X., & WEIGEL, K. A. 2016. Semi-supervised learning for genomic prediction of novel traits with small reference populations: An application to residual feed intake in dairy cattle. *Genet. Sel. Evol.*, 48, 84.
- YEOM, S., GIACOMELLI, I., FREDRIKSON, M., & JHA, S. 2018. Privacy risk in machine learning: analyzing the connection to overfitting, 2018 IEEE 31st Computer Security Foundations Symposium, 268-282.
- ZAAIJER, D.; & NOORDHUISEN, J.P.T.M. 2003. A novel scoring system for monitoring the relationship between nutritional efficiency and fertility in dairy cows. *Ir. Vet. J.* 56, 145–151.
- ZABORSKI, D., PROSKURA, W. S. & GRZESIAK, W. 2018. The use of data mining methods for dystocia detection in Polish Holstein-Friesian Black-and-White cattle. *Asian-Australas J Anim Sci*, 31, 1700-1713.

- ZAHRAZADEH, M., RIASI, A., FARHANGFAR, H. & MAHYARI, S. A. 2018. Effects of close-up body condition score and selenium-vitamin E injection on lactation performance, blood metabolites, and oxidative status in high-producing dairy cows. *J Dairy Sci.* 101(11), 10495–10504.
- ZEGLER, C. H., RENZ, M. J., BRINK, G. E., & RUARK, M. D. 2020. Assessing the importance of plant, soil, and management factors affecting potential milk production on organic pastures using regression tree analysis. *Agric. Syst.*, 180, 102776.
- ZHANG, S., & LI, J., 2021. Knn classification with one-step computation. *IEEE Transactions on Knowledge and Data Engineering*, 1-1.
- ZHANG, Y., LI, X., ZHANG, H., ZHAO, Z., PENG, Z., WANG, Z., LIU, G. & LI, X. 2018. Non-Esterified Fatty Acids Over-Activate the TLR2/4-NF-KappaB Signaling Pathway to Increase Inflammatory Cytokine Synthesis in Neutrophils from Ketotic Cows. *Cell Physiol Biochem*, 48, 827-837.
- ZHAO, K., SHELLEY, A. N., LAU, D. L., DOLECHECK, K. A., & BEWLEY, J. M. 2020. Automatic body condition scoring system for dairy cows based on depth-image analysis. *Int. J. Agric. Biol. Eng.*, 13, 45–54.
- ZHOU, X., XU C., WANG, H., XU, W., ZHAO, Z., CHEN, M., JIA, B. & HUANG, B. 2022. The Early Prediction of Common Disorders in Dairy Cows Monitored by Automatic Systems with Machine Learning Algorithms. *Animals*. 12(10):1251.

Appendices

Appendix I

Supplementary to Chapter 2: Sample code for predictive model fitting

```
ctrl <- trainControl(method = "cv", number = 10)
```

or

```
ctrl <- trainControl(method = "cv", number = 10, sampling = "up")
```

when using upsampling

```
train(Outcome~Variable,
```

```
data = data,
```

```
method = method,
```

```
na.action = na.omit, trControl = ctrl, metric = metric)
```

where:

data is the respective dataset used for each analysis,

method the methodology used to fit the model,

and metric was set to “Kappa” for binary models, while left as the default option for continuous outcomes.

Supplementary to Chapter 2: Sample code for Odds Ratios from mixed effects logistic regression model

```
mod <- glmer(Outcome ~ Variable1 +  
  
  ... + VariableN +  
  
  (1| FarmID), data, family = "binomial")  
  
cc <- confint(mod,parm="beta_", method = "Wald")  
  
ctab <- cbind(est=fixef(mod),cc)  
  
rtab <- exp(ctab)  
  
print(rtab,digits=3)
```

where data was the dataset used for each analysis,

Outcome was the outcome variable,

Variable1,..., VariableN the number N explanatory variables used in the analysis

and FarmID each herd identification number

Appendix II

Figure A2. 1Rumen fill distribution based on 28,480 dry cows

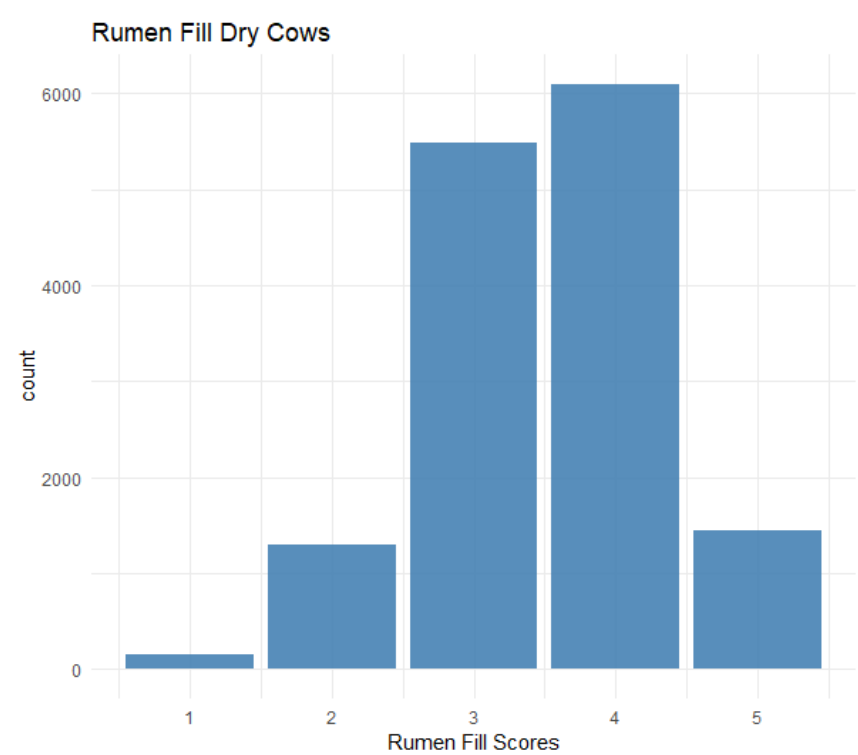


Figure A2.2 Rumen fill distribution based on 43,185 fresh cows

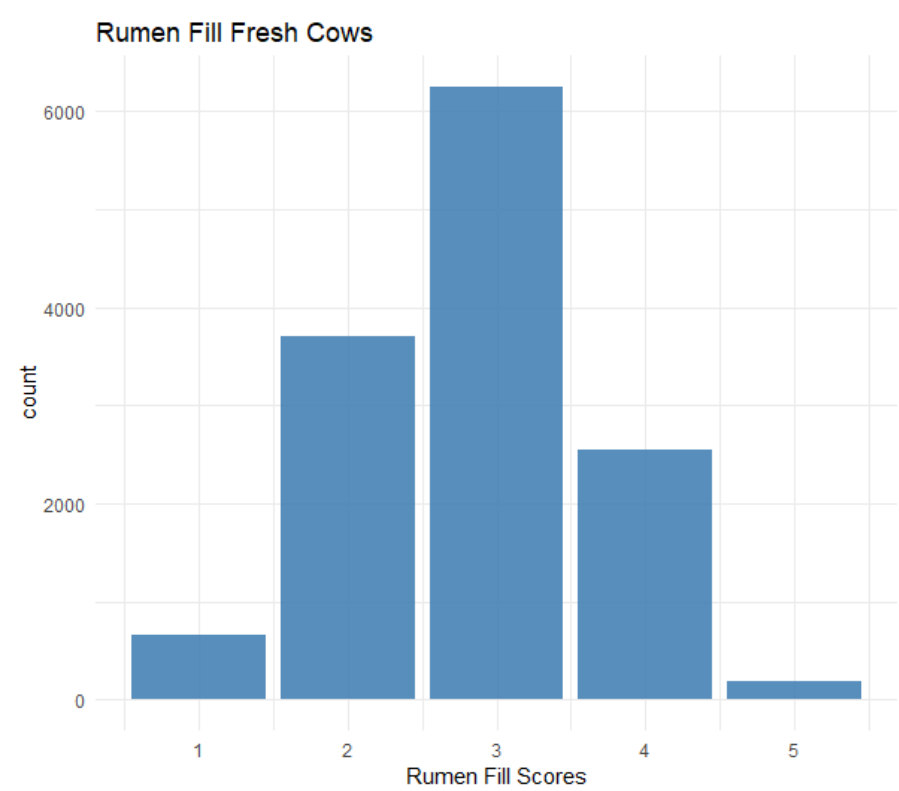


Figure A2.3 Hock Hygiene distribution based on 12,847 lactations

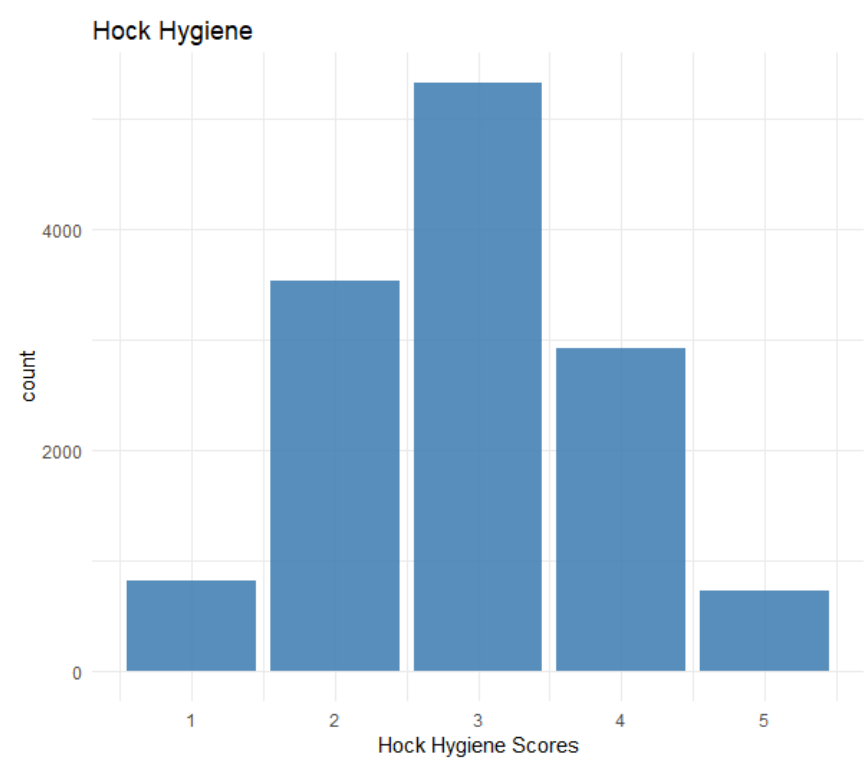


Figure A2.4 Types of pens for both dry and fresh cows, based on 2,787 pens

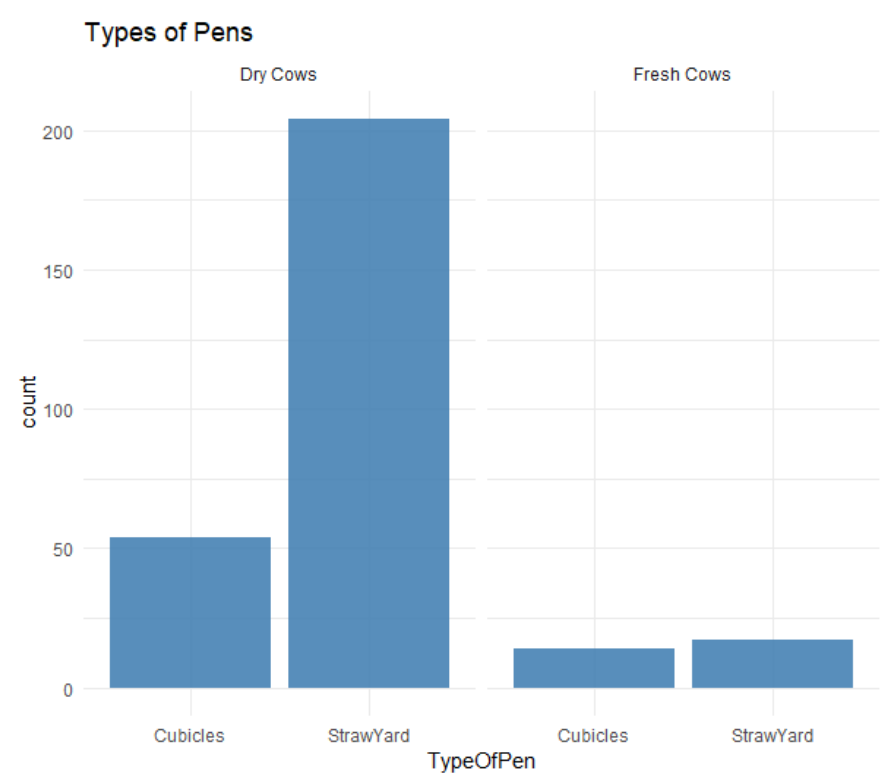


Figure A2.5 Feed Fence space available per cow, based on 2,787 pens

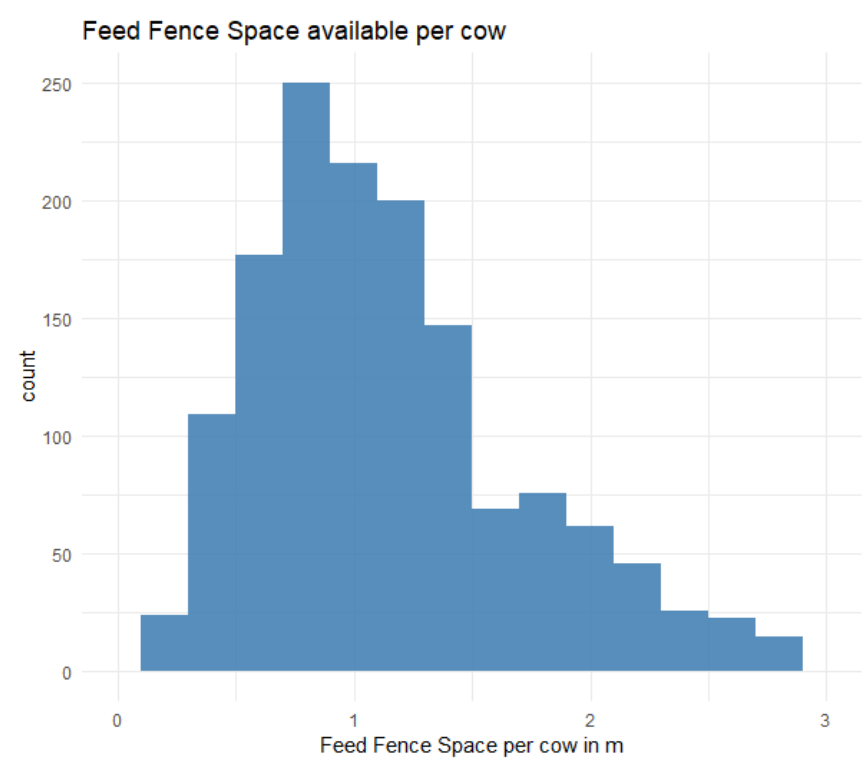


Figure A2.6 Feed Fence space available separately per dry and fresh cows, based on 2,787 pens

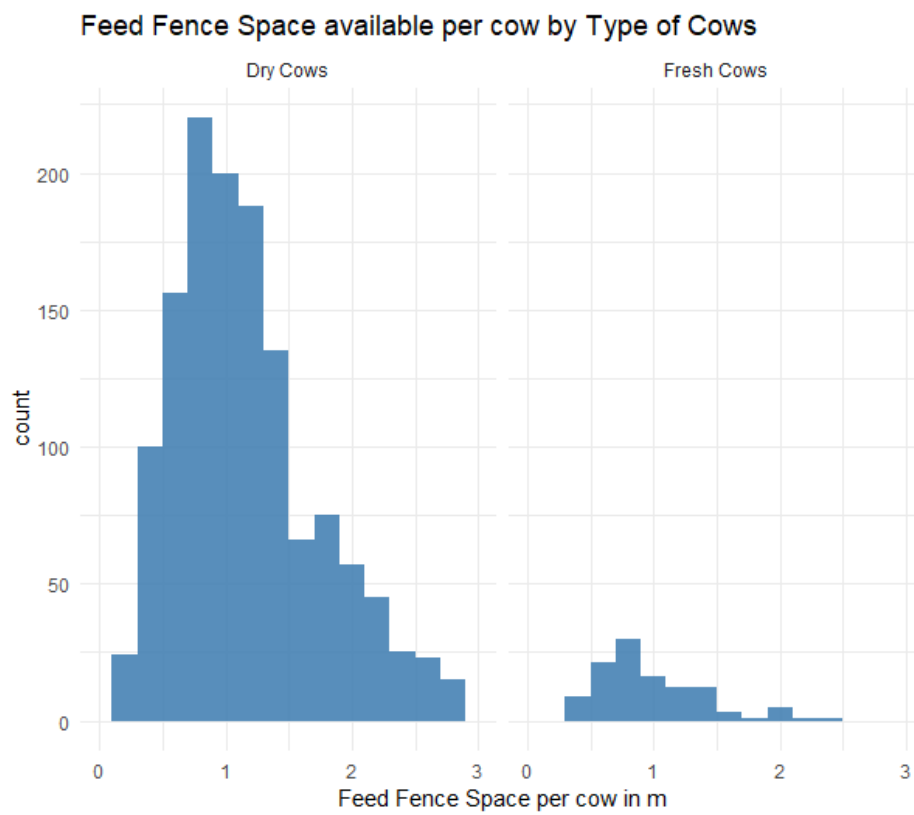


Figure A2.7 Water Trough space available per cow, based on 2,787 pens

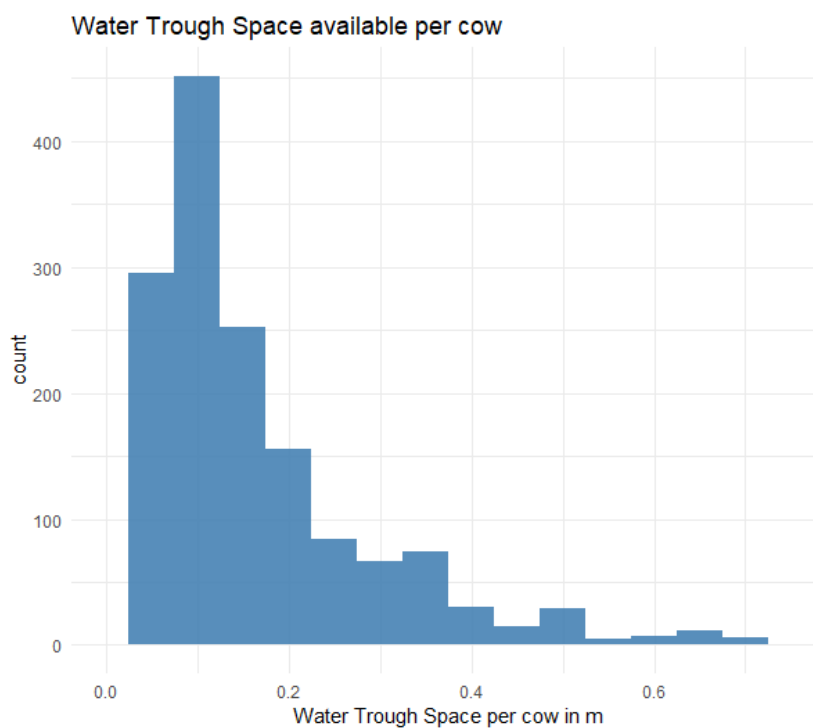


Figure A2.8 Water Trough space available separately per dry and fresh cows, based on 2,787 pens

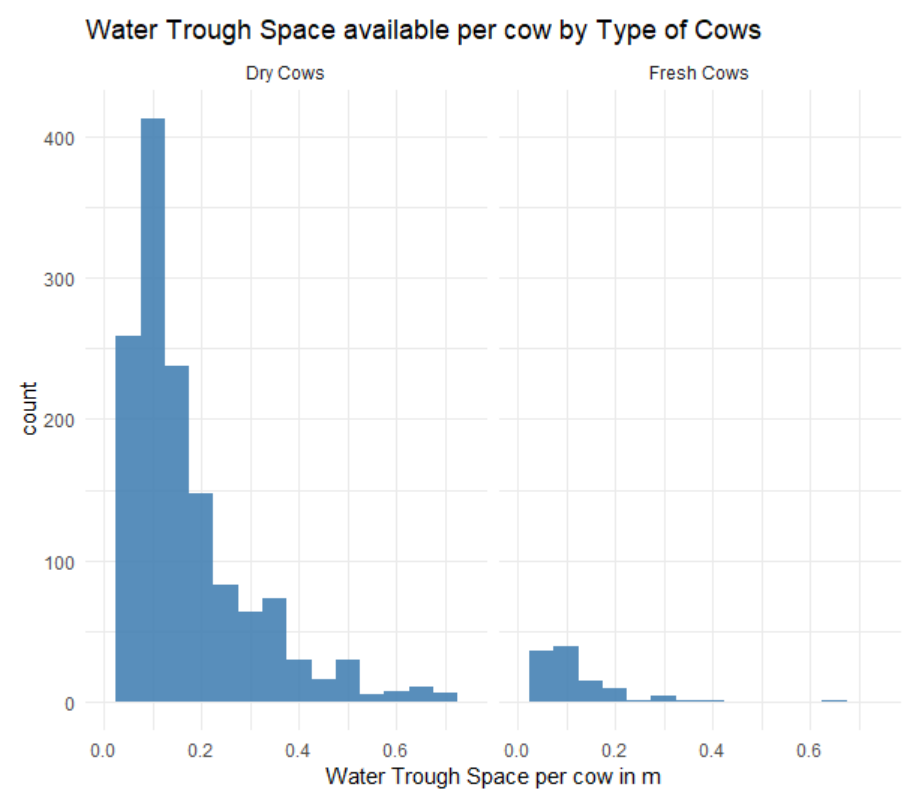


Figure A2.9 Neck Rail Height available based on data on 2,787 pens

