

THESIS

Joseph Oscar Ullomi

2023

Centre for Bioinformatics

University of Veterinary Medicine Budapest



Subclinical Mastitis Detection by Machine Learning

By

Joseph Oscar Ullomi

Supervisor:

Solymosi Norbert Dr. PhD

Budapest, Hungary.

2023

DEDICATION

I dedicate this thesis to God Almighty for the grace of knowledge and courage to work towards its completion. Additionally, I am grateful for the invaluable contributions of my family, university professors, and fellow students throughout my academic journey.

ABSTRACT

The study aimed at developing a robust Machine Learning algorithm that could accurately predict Subclinical Mastitis in dairy cows, which often goes undetected until it is too late. We utilized five Machine Learning models, including Decision Trees, K-nearest neighbours, Logistic Regression, Random Forests, and eXtreme Gradient Boosting, to find the best fit for the work. Data was collected for two years from two large Hungarian dairy farms using two databases: RISKKA and ALPRO. Somatic Cell Count values and Electroconductivity of Milk variables were the key features used for Artificial Neural Network-based classification. By utilizing the features of the Caret Package in the R environment, we filtered out correlated explanatory variables. We estimated the variable importance using a binomial Generalized Linear Model. 1,368 animals were considered for the study, and 7,685 records were available after preprocessing. Modelling was performed in Python with five-fold cross-validation, and Receiver Operating Characteristic curves were collected for all five models. The findings favour the Logistic Regression, which achieved an impressive Area Under Curve value of 0.686, and the Decision Tree had the lowest AUC of 0.547, with K-nearest neighbors of 0.598, XGBoost of 0.637, and Random Forests of 0.667. This study highlights the enormous potential of Machine Learning algorithms to transform mastitis detection and control, providing quick, accurate, and cost-effective solutions. The study's findings offer a promising direction for developing innovative techniques for predicting subclinical mastitis in dairy cows, which could significantly impact the industry. Despite the findings, sole reliance on machine learning algorithms for the detection of mastitis cannot be fully guaranteed, but rather a hybrid approach would yield more accurate and reliable results.

TABLE OF CONTENTS

ABBREVIATIONS.....	- 7 -
LIST OF FIGURES.....	- 8 -
LIST OF TABLES.....	- 9 -
1. CHAPTER I: INTRODUCTION.....	- 10 -
2. CHAPTER 2. LITERATURE REVIEW.....	- 12 -
2.1. Mastitis Overview.....	- 12 -
2.2. Traditional Subclinical Mastitis Detection Methods.....	- 12 -
2.3. Machine Learning Applications in Veterinary Medicine and Other Biological Studies	- 13 -
2.5. Exploring Different Machine Learning Algorithms	- 17 -
2.5.1. Decision Tree (DT) Model	- 17 -
2.5.2. Kernel and Nearest Neighbors (KNN) Model.....	- 17 -
2.5.3. Logistic Regression (LR) Model.....	- 18 -
2.5.4. Random Forest (RF) Model.....	- 18 -
2.5.5. eXtreme Gradient Boosting (XGBoost) Model.....	- 18 -
2.5.6. Artificial Neural Networks (ANNs)	- 19 -
2.6. Machine Learning Programming Languages	- 19 -
2.6.1. The R-environment	- 19 -
2.6.1.1. The Caret package.....	- 20 -
2.6.2. Python	- 20 -
2.7. Model Evaluation Techniques.....	- 20 -
2.7.1. K-Fold Cross-validation	- 20 -
2.7.2. Receiver Operating Characteristic (ROC) and Area Under the ROC Curve (AUC)	- 21 -
2.8. Challenges and Limitations of Predicting Subclinical Mastitis through Machine Learning	- 21 -
2.9. Insights Based on The Literature Review	- 22 -
3. CHAPTER III: MATERIAL METHODS.....	- 23 -
3.1. Dataset Description and Collection.....	- 23 -
3.2. Data Sources.....	- 23 -
3.3. Data Preparation.....	- 23 -
3.4. Target Variable.....	- 23 -

3.5. Feature Selection	- 24 -
3.5. Machine Learning Models Used.....	- 25 -
4. CHAPTER IV: RESULTS	- 25 -
4.5. Performance evaluation of machine learning models	- 25 -
5. CHAPTER V: DISCUSSION	- 26 -
6. CHAPTER VI: CONCLUSION	- 29 -
7. ACKNOWLEDGEMENTS	- 30 -
8. REFERENCES	- 31 -

ABBREVIATIONS

ANN= Artificial Neural Networks

BM=Bovine Mastitis

SM= Subclinical Mastitis

CM= Clinical Mastitis

CMT=California Mastitis Test

SCC=Somatic Cell Count

ECM=Electrical Conductivity of Milk

ML=Machine Learning

DL=Deep Learning

DNN= Deep Neural Networks

CNN=Convolutional Neural Network

IMI= Intramammary Infections

ROC= Receiver Operating Characteristic

AUC=Area Under Curve

XGBoost= Extreme Gradient Model

KNN= Kernel and Nearest Neighbors

LR (logit)= Logistic Regression

RF= Random Forests

DT=Decision Trees

LIST OF FIGURES

Figure 1: Prediction Performance of Models using ROC curves.

LIST OF TABLES

Table 1: Descriptive Table of variables of Electroconductivity of milk.

1. CHAPTER I: INTRODUCTION

Bovine Mastitis (BM) is a significant problem in the dairy industry, impacting productivity and financial stability globally (Hoque M. N. et al., 2019). Pathogens that infiltrate the udder, directly or indirectly, primarily due to shortcomings in milking hygiene procedures, are the leading causes of mastitis (Malcata et al., 2020; Cheng et al., 2020). The distinction between contagious and environmental pathogens, their infection pathways and timing are critical in effectively managing mastitis (Hyde R. M. et al., 2020; Radostis O. M. et al., 1994; Klaas I. C. & Zadok R. N., 2017).

Mastitis can be classified into three main categories based on the course of the disease: sub-clinical, clinical, and chronic mastitis (Cheng W. N. & Han S. G. 2020). *Clinical Mastitis* presents distinct symptoms, including fever, udder swelling, and changes in milk consistency (Henrique et al., 2022). *Subclinical Mastitis* (SM) does not have distinct udder alterations but is characterised by decreased milk production and elevated Somatic Cell Count (SCC) (Hoque et al., 2019; Abebe R. et al., 2016). *Chronic Mastitis* (CM) involves prolonged inflammatory processes, leading to intermittent clinical incidents (Cheng W. N. & Han S. G. 2020).

Mastitis is a complex condition that affects dairy cows and is caused by various factors, including the cow, the farmers' management, and pathogens. Accurate diagnosis and treatment of mastitis are critical to prevent the unnecessary use of antimicrobials, safeguard animal health and welfare, and prolong the animals' lifespan (Hyde R. M. et al., 2020). *Mastitis* is one of the most common ailments in dairy cows and can clinically lead to fatalities by requiring premature culling of non-productive cows (De Vliegher et al., 2018; Bobbo T. et al., 2021; Guet P. et al., 2001; Wang Y. et al., 2022). Therefore, quick, and effective action must be taken to prevent losses from this condition (Wang Y. et al., 2022).

Different bacterial strains cause different types of mastitis infections (TECA) and the prevalence of mastitis varies in different geographical locations (Morales-Ubaldo et al., 2023). Given the urgency to enhance mastitis control and develop new strategies for detection and prevention in response to growing consumer interest in food safety and rational antibiotic use, this study examines how Machine Learning (ML) models can transform its detection and management. These algorithms promise high-precision solutions by integrating expertise from diverse fields (Bobbo T. et al., 2022; Hyde R. et al., 2020), and their application can significantly speed up mastitis detection, prevention, and management, addressing various

causes and consequences affecting dairy production, animal health and global dairy sector outcomes (Bobbo T. et al., 2022; Noor A. et al., 2020).

The objective of this study is to develop an ML algorithm that can predict mastitis using key parameters, such as Electro Conductivity of Milk (ECM), Somatic Cell Count (SCC), and milk yield, collected from two large dairy farms over a specific period through the RISKa and ALPRO databases (CowManager, 2021; Software Informer, 2023). The study explores the potential of ECM as an indicator of SM, which is collected through automated milking systems (Jacobs and Siegford 2012). The main aim is to streamline farmer decision-making, reduce operational costs and ensure ethical and welfare standards for animal health (Cohen E.B. et al., 2022).

The research evaluates five different ML algorithms as part of a PhD thesis titled "Machine Learning Application Developments for Evaluation of Animal Production Processes." The algorithms' potential to improve mastitis control, enhance dairy production, promote animal welfare and achieve better global industry outcomes is being investigated (Jacobs and Siegford, 2012).

The study aims to determine whether machine learning algorithms can provide farmers with the necessary information to make informed decisions on preventive measures (Bobbo T. et al., 2021), save costs through quick assessments (Noor A. et al., 2020) and accurately identify SM by analysing the Electrical Conductivity of milk while considering the ethical and welfare aspects of animal health (Cohen E.B. et al., 2022). The goal is to develop an algorithm to predict subclinical mastitis using data from ECM, SCC, milk yield, and other relevant parameters collected from infected and uninfected milk samples across two large dairy farms.

2. CHAPTER 2. LITERATURE REVIEW

2.1. Mastitis Overview

Mastitis can be identified by changes in the milk's physical, chemical, or bacterial composition, as well as by pathological abnormalities in the glandular tissue of the udder (Shama 2011). Electrical conductivity is one of these modifications—transferring electric current from a solution between two electrodes (Hamilton J. et al., 2017). Moreover, Hamilton (2017) found a high correlation between the SCC rise and the ECM increase.

Among many consequences, *mastitis* can negatively affect the production longevity, well-being, and the welfare of dairy cows, taking away the most intended incomes from farmers through poor-quality milk (Bobbo T. et al., 2021). Other losses may be linked to fatalities that arise due to clinical Mastitis and the culling of non-producing cows at a premature age (Guet P. et al., 2001).

According to research on the genetic resistance of dairy cattle to mastitis by Rupp et al. (2003), they concluded that, for some farms, genetic selection for the udders' susceptibility to mammary pathogens must be included in breeding programs because there was proven high heritability and genetic correlation among phenotypic traits that were related to mastitis such as SCC and clinical cases. Findings also concluded that the mastitis trait depended on a genetic part and physiological and environmental factors, not neglecting infection pressure. A balance was reported between the virulence of pathogenic organisms and udder resistance, as well as the length of time the lactic gland is vulnerable to infection, determining whether mastitis will develop in a cow (Janzekovic M., 2009).

2.2. Traditional Subclinical Mastitis Detection Methods

SM is a common issue on farms, and the California Mastitis Test (CMT) is a widely used method for predicting it. The CMT measures the SCC of milk, which is a reliable indicator of SM. Additionally, farms may use the ECM test to analyse milk conductivity. However, analysing and examining SCC and ECM tests can be laborious and time-consuming (Deshapriya et al., 2011; Hamilton, 2017). For SCC, there is always a significant variation in the number of somatic cells, which should typically be less than 200,000 cells/ml among breeds, animal age, and lactation stage, ruling out the sole dependence on the use of the existing CMT (Deshapriya R. et al., 2011). However, an SCC of 100,000 cells/ml was the most appropriate at

the individual cow level to identify dairy cows' intramammary infection (IMI) in a study conducted in Bangladesh (Sumon SMMR et al., 2020). An SCC of 400000 cells/ml is deemed to be an unquestionable IMI, but the use of SCC is limited because other factors in the body can raise the SCC (Ruegg & Pantoja, 2013).

The ECM has been and continues to be researched for many decades, and most previous publications have deemed it a potential cow-side screening test in devices or in-line, with ECM reported to increase in positive cases. However, differences also exist, depending on whether it was assessed on foremilk or post-milking stripping, directly correlating to the increase in SCC (Fernando R. et al., 1985; Fernando R. et al., 1982; Nielen M. et al., 1992). Typically, the standard ECM of milk falls between 4.0 to 5.5 mS/cm (milli Siemens per centimetre) at 25 °C. During mastitis, the composition of milk changes. Lactose and potassium (K⁺) ions leak into the extracellular fluid. In contrast, sodium (Na⁺) and chloride (Cl⁻) ions, usually found in high concentrations in the extracellular matrix, tend to increase in the milk. This is because the tight junctions, active ion pumping systems, and blood vessels in the inflamed udder tissue are damaged, leading to increased permeability. The increased Na⁺ and Cl⁻ in milk from a mastitic cow helps to maintain osmotic pressure, resulting in an increased conductance of electric current without any change in osmotic pressure (Nielen M. et al., 1992).

2.3. Machine Learning Applications in Veterinary Medicine and Other Biological Studies

Cohen (2022) suggested many artificial intelligence applications in veterinary medicine, namely veterinary imaging and radiation therapy in oncological cases. Proponents have proposed using ML algorithms for effective disease detection strategies (Machuve D. et al., 2022). Their research created a deep Convolutional Neural Network (CNN) model to diagnose poultry diseases by classifying healthy and unhealthy faecal images. In their study, they took several images of poultry faeces that were labelled by experts using an open data kit and then trained baseline CNN models, VGG16, Inception V3, MobileNetV2, and Xception models with farm and laboratory-labelled faecal images with fine-tuning. They concluded that the MobileNetV2 can be used at the farm level to detect poultry diseases early.

In a study that was conducted to make a predictive analysis as a comparative study, up to ten ML algorithms were deployed (Dofadar et al., 2022).

Another study by (Nagy et al., 2023) employed Deep Neural Networks (DNN), in particular Convolutional Neural Networks (CNN), to recover body condition ratings using annotated photos that were first captured by a primary camera and tagged by professionals. The 3 and 12 body condition classes used a pre-trained CNN model for fine adjustment. According to their research, CNNs trained on three BCS classes displayed a more considerable proportion of solid agreement, showing that we need a few training models for accuracy.

As a general observation, in biological studies, computational techniques through ML algorithms were used to predict the correctness of types and functionality of ion channels (Lin H. et al., 2015). Due to positive results in recent publications, ML is another promising field for medical diagnostics (Lee et al., 2020).

According to Obermeyer (2016), algorithms that are developed using ML, in particular Deep Neural Networks (DNN) and expert analysis, have shown tremendous capability in obtaining results by quickly scanning through massive databases and data sets and checking for similarity in data that was obtained and previously trained using models to remember and be able to make predictions on the actual meaning of the data given. ML's potential in biological and biomedical studies has a broad potential for application if massive data are available (Xiang T. W. et al., 2018).

2.4. Innovative Uses of Machine Learning in Detecting Mastitis

ML algorithms can predict the udder health status of cows based on SCC. A study by Bobbo et al. (2021) compared eight ML methods and found that all methods had prediction accuracies above 75%, with Neural Network, Random Forest, and linear methods performing the best. The researchers collected milk samples from a Breeders' Association in Italy. They used a dataset that included information on factors such as milk composition, production, SCC, season of the year, parity, herd test results and other features. The cows were grouped based on the herd-test-date (HTD), and the final dataset used in the study consisted of 18,442 records of 14,064 cows in 791 herds. They performed validation by using 10-fold cross-validation to select the best-performing model. To improve accuracy, they stratified the cross-validation process and repeated it 100 times for each 10-fold cross-validation and a total of one thousand iterations.

Nevertheless, despite the transformation they made to the data, it did not affect the performance of the models. However, only those log-transformed improved performance in the context of

AUC, for instance. They used accuracy, false positive, false negative, total error rates, Cohen's Kappa, and F1 score for model evaluation. They also used the pROC package in the R environment and generated AUC values for performance evaluation. The pROC package is a package in the R environment that is used for displaying and analysing the ROC curves (RDocumentation 2023). Matthew's Correlation Coefficient was also used to measure the classification quality. The resulting prediction by ML was a prevalence of SM at 29%.

In the subsequent study by Bobbo et al. (2022), ML techniques were employed to predict udder health in dairy cows based on SCC. Neural Network (NN) emerged as the most accurate method, identifying key features such as log₁₀SCC, stage of lactation, Differential Somatic Cell Count, protein, and parity. Linear methods (Linear Discriminant Analysis and Generalised Linear Model) and Random Forest (RF) also performed well. The study collected data from buffaloes on commercial farms, including various parameters like milk production, milk composition, SCC, and climatic conditions. They identified SM based on a SCC threshold of 200,000 cells/ml, with a prevalence of 40.3%. The dataset comprised 3,891 records from 1,038 buffaloes in 6 herds. Twenty-seven features were considered, including milk-related traits, climatic parameters, and animal information. The data underwent preprocessing feature selection and was used to build predictive models using Generalized Linear Models (GLM), Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN). Various performance metrics were utilized to assess the models, and the study compared their predictive abilities on both validation and test sets. They reported that including meteorological parameters, such as precipitation, sunshine hours, and soil temperature, in milk production forecast models improved prediction accuracy, with sunshine hours having the most significant effect.

Ebrahimi et al. (2019) conducted a study to predict SM in dairy cows using ECM, milk volume, and various milking variables. The study analyzed a dataset of 297,004 milking samples and applied ML models such as Gradient-Boosted Trees (GBT) and Deep Learning (DL), demonstrating remarkable accuracy and sensitivity in predicting SM. Their conclusion was that, as dairy farms continue to adopt automated data collection, developing and applying such predictive models offer practical solutions to identify infected cows, minimizing economic losses promptly.

Fadul-Pacheco et al. (2021) developed two algorithms for predicting CM in dairy cows, one for short-term daily predictions and the other for identifying first-lactation cows at risk of CM in

the mid-to-long term. The study highlights the importance of integrating algorithms and data sources to manage farm health and tackle CM challenges effectively. The Random Forest algorithm exhibited promising results when considering attributes selected through the Wrapper method, suggesting the potential benefits of combining various data sources and ML techniques in CM prediction for dairy farming.

The research conducted by Pakrashi et al. (2023) sheds light on the challenges dairy farms face due to SM in cows. The study analyzed a large dataset comprising more than 1.3 million milk-day records from seven research farms in Ireland over nine years, encompassing a range of factors such as milk yield, milk composition, milk flow, SCC, and different health and historical indicators. The research utilized ML techniques and constructed predictive models to detect SM up to seven days before its onset. They took advantage of various models during the study and discovered that a Gradient-Boosting Machine, Random Forests, and a Support Vector Machine showed the most potential. The Gradient-Boosting Machine exhibited the highest sensitivity of 69.45% and specificity of 95.64%. The model's efficacy persisted even with reduced data collection frequencies, typical in commercial dairy farming. This research highlighted the potential of ML in the early identification and management of SM in dairy cows.

Among the evaluation metrics, they used sensitivity (SE), AUC and specificity (SP), the metrics used to assess the efficiency of the predictive models trained with different frequencies of milk composition data. The performance was consistently superior when data were collected at a 7-day frequency compared to reduced recording intervals, highlighting the effectiveness of frequent data measurements in early SM detection.

ML techniques have been used to detect mastitis using infrared thermal imaging with promising results (Zhang X. et al., 2022). By utilizing ML algorithms, precisely the Target Detection algorithm to measure the temperature differences between the eyes and the udder and the Enhanced Fusion MobileNetV3 You Only Look Once v3 (EFMYOLOv3) system, it was possible to analyse thermal images of cows and identify early signs of Mastitis by examining the head position.

In an article by Esener (2021), Matrix-Aided Mass Spectrometry using Laser Desorption/Ionization-Time of Flight (MALDI-TOF) together with ML was used to detect BM pathogens by discrimination based on transmission routes used. The findings were suggested to have the potential of classifying BM caused by *Streptococcus uberis*. Another discovery was that there were discernible phenotypic differences.

Hyde et al. (2020) conducted a research study titled "Automated Prediction of Mastitis Infection Patterns in Dairy Herds Using Machine Learning". The objective was to replicate specialist clinical decisions using ML models in R v3.5.1. The RF model achieved an accuracy of 95% in distinguishing between Contagious and Environmental mastitis diagnoses. The model could also accurately replicate herd-level mastitis diagnoses compared to assessments made by specialist veterinary clinicians.

2.5.Exploring Different Machine Learning Algorithms

2.5.1. Decision Tree (DT) Model

(Breiman et al., 1984) made the initial introduction to the application of DTs and made a description about them that they partition input spaces to assign a subsequent output value for every input. There was a relationship between each node in the tree and the input node. DTs are created from a tree root node and end at the leaf nodes from internal nodes, and they usually execute decisions in a top-down system. Their existence may arise from an existing dataset with such observations as records of patients accumulated over time (Quinlan J. R., 1986). Strecht (2019) also stressed that they offer a wide possibility of being combined despite their easiness of interpretation. They look like a tree in representation and do not need much data processing. Simple rules are assigned to the dataset, making decisions called leaves or terminal nodes (Dofadar et al., 2022). Our study used DT as the baseline model from which, after simple classification into 0 and 1 of the variables used, more robust ensembles, the RF and XGboost, were built (Mohammed & Kora, 2023).

2.5.2. Kernel and Nearest Neighbors (KNN) Model

KNN is another well-researched and applied ML model whose main idea is that a given data point class is likelier to be like the nearest classes, thus its application in classification (Carpenter K.A. et al., 2018). The vicinity between these data points, the training data and the new data mean they are similar (Dofadar et al., 2022). K-nearest neighbor (KNN) is a popular algorithm with various applications for pattern recognition. In ML, Jabbar et al. (2013) demonstrated its effectiveness in generating new data points like the initial dataset. This approach involves identifying the k-nearest neighbors of a unique data point and using their attributes to create a new data point. The resulting data points are centred around the original dataset, allowing for the generation of fresh data relevant to the problem at hand.

2.5.3. Logistic Regression (LR) Model

LR is a well-known statistical method and has thus been integrated and used in ML studies for binary classification between two well-understood variables. It uses sinusoidal curves between two groups, differentiating it from linear regression. However, there is a similarity to linear regression in that there is a clear demarcation between two categorical datasets, and the weighted transformation of the logarithmic function is the data points with categories. When utilising the regression function and taking the curve as an axis, human input data will be categorised into one of these categories (Panesar et al., 2019). The LR has a predictive function as a model, can analyse connections between many unrelated variables, and foresees the estimations of related factors ranging between zero and one through the logistic function. Values that are above 0.5 are labelled as one, and vice versa for those that are below 0.5 (Dofadar et al., 2022).

2.5.4. Random Forest (RF) Model

The RF approach is a technique that involves ensemble learning by using randomly sampled feature routes in DTs called vectors. RFs are like a network of DTs; the team looks for or follows a random set of information, but they all use the exact source of information. Consequently, incremental changes in the magnitude of the forest minimize the errors and improve performance accuracy as they tend to vote in favour of the most popular class (Breiman L., 2001). Dieterich (1998) confirmed that another way to lower error rates is by splitting the nodes by random feature selection, thereby enhancing steadiness in node relationships. According to Xiao et al. (2022), the RF model has better simplification capability and can handle data with diverse sizes. They act as a group of DTs, with each tree producing its outcome, and the outcome that appears the most among these trees is voted as the ultimate classification. Hyde (2020) employed well-trained RF algorithms to replicate complex datasets to develop an automated method of predicting mastitis cases, at least on the herd level.

2.5.5. eXtreme Gradient Boosting (XGBoost) Model

The (XGBoost) is an ML algorithm that creates boosted-up trees inside the model and combines multiple types of information across many parameters using individual DTs. In other words, it is a higher version of the Gradient-Boosting Tree algorithm. It also automatically assesses feature importance from a trained analytical model and returns a result for the estimate of each

feature. The XGBoost model thus can handle varying-sized data with an insignificant risk of overfitting. The result of the classification is computed from the outcomes of the vast trees with improved accuracy (Montomoli et al., 2021; Xiao et al., 2022; Dofadar et al., 2022).

2.5.6. Artificial Neural Networks (ANNs)

ANNs have incredibly proven applicable in many studies solving various problems. They are made up of layers of neurons that are interconnected. Practically, they can identify patterns and during training, they develop non-linear models that allow them to make general conclusions and apply them to even patterns they have not encountered before. They can classify large datasets with high accuracy, which improves with datasets of high quality (Baykan and Yılmaz, 2011).

In a recent article titled "Artificial Neural Networks in Animal Product Production" by Nagy et al. (2023), published in the Hungarian Veterinary Journal (Magyar Állatorvosok Lapja) volume 145 number 5, the authors discussed the use of ML in the agricultural industry. With the vast amounts of data generated by large-scale animal production, ML algorithms such as ANNs are gaining popularity due to their increasing use in processing and analyzing this data. The authors noted that ANNs have been the most successful subdivision of ML in agricultural studies.

2.6. Machine Learning Programming Languages

2.6.1. The R-environment

R is a language and programming environment developed as a shared environment that provides packages essential for statistics and visualization techniques; it works to make plots where needed and is available as downloadable software. Its roles include data manipulation, making calculations and graphical displays. It is called an environment because it is a fully prepared system for implementing statistical techniques. R is an environment for the implementation of statistical techniques and through packages, it can be extended easily. Eight packages come with the software and others can be sourced from the Comprehensive R Archive Network (CRAN) (R Core Team; The R Project for Statistical Computing 2023; CRAN Mirrors n.d.).

2.6.1.1.The Caret package.

The Caret package is an acronym for Classification and Regression Training, encompassing many tools employed in the R environment to formulate predictive models using existing models. It contains packages that are used for the modelling and evaluation and feature selection processes. Thus, it played a significant role in the entire process of our methodology of our study.

2.6.2. Python

Python is a versatile all-purpose programming language created by Guido van Rossum to be usable in real-time programming and learning. It is well-designed and a high-level language that uses interpreters to make its applications user-friendly. It is designed in such a way that it can support many forms of programming, like structures and objects. It is also an open source with no limitations in use and application. This program can keep variable names and methods during programming. Within the system, it can work with environmental variables (Sharma et al., 2023). It is a populous language of programming with various scientific libraries that are used for ML (Pedregosa F. et al., 2011).

2.7.Model Evaluation Techniques

2.7.1. K-Fold Cross-validation

In k-fold cross-validation, the dataset is initially randomized to ensure a random distribution of data points. Subsequently, the dataset is divided into k equal-sized groups or 'folds'. During each cross-validation iteration, one of these folds is reserved as the test set, while the remaining k-1 folds collectively form the training dataset. It refers to the process where a particular model for ML is trained on a training dataset and assessed for its performance on the test dataset. This iterative process repeats k times, with each fold being the test set once. Finally, the results obtained from these k iterations are typically averaged to provide a more robust and reliable evaluation of the models' performance (Brownlee, 2023). For our study, this means that k=5, and the process was repeated five times accurately and precisely once on each fold to avoid overfitting (Shanthababu, 2023).

2.7.2. Receiver Operating Characteristic (ROC) and Area Under the ROC Curve (AUC)

A ROC curve is a graph that visualizes, organizes, and compares models' performance which has long been adopted in ML. They are two dimensional in which the y -axis displays the true positive rate while the x -axis shows the false positive rate. Its leading functional architecture is to evaluate a trade-off between these values; both values range between zero and one. A point in the ROC curve is better if the rate of correctly identified positive cases is higher and the rate of incorrectly identified negative cases as positive is lower.

On the other hand, the Area Under the ROC curve is a method used to evaluate the performance of classifiers, and its value is always between 0 and 1. In statistics, the importance of this value of a model means that a randomly selected positive class will be ranked higher than a negatively picked class (Fawcett, 2006).

2.8. Challenges and Limitations of Predicting Subclinical Mastitis through Machine Learning

ML methods have been researched recently for scientific studies, and their application of various methods, like deep DL, is straightforward (De Palva B. et al., 2023). Nevertheless, this success has shortcomings in biological studies, among others, overfitting and underfitting (Xu C. et al., 2019). When ML is exposed to fresh unknown data, it tends to perform in a biased way on new data, which is overfitting and needs simplification and training of the data to be used. Underfitting occurs when an ML model performs poorly on training and new data because it is overly simple (Xu C. et al., 2019). ML algorithms necessitate the availability of a vast amount of data subjected to rigorous quality control procedures to train and assess models effectively. Additionally, the effectiveness of ML techniques can be influenced by the computational expense required, as observed by Yang A. et al. (2020). Acquiring directly helpful information from vast datasets and, in most cases, unstructured cases remain a perpetual challenge in healthcare applications (Miotto R. et al., 2018). Despite these challenges, there is still evidence that ML methods are much better in complex datasets, especially by using DNN (Shamshirband S. et al., 2021).

2.9. Insights Based on The Literature Review

The literature review concludes that the use of ML in mastitis detection is becoming increasingly significant and has the potential to revolutionize dairy farming practices. The review explores ML algorithms for early mastitis diagnosis, including DTs, ANNs, RFs, GLM and many other sophisticated ML methods.

However, it is essential to note that several hurdles must be overcome. Overfitting and underfitting are crucial concerns that require thoroughly developing and validating ML models. Additionally, the quality and size of datasets are pivotal, necessitating the use of substantial, high-quality data for training and evaluating ML models. Furthermore, the computational demands associated with ML may pose practical challenges.

Despite these obstacles, the review highlights the immense potential of ML, particularly DNN, ANN, RF and XGBoost in addressing the complexities of mastitis diagnosis and management. ML can extract valuable insights from vast and intricate datasets, as observed in healthcare applications. The review concludes that continued exploration of ML's capabilities in mastitis detection could lead to improved accuracy, efficiency, and the well-being of dairy cattle and the profitability of dairy farming.

3. CHAPTER III: MATERIAL METHODS

3.1. Dataset Description and Collection

In our work, we performed modelling using a dataset that was created and analysed by Nagy et al. (2023). Moreover, this entails that they provide the data description for my study. According to their data description, SCCs above 200,000 cells/millilitre were considered indicative of Intramammary Infection (IMI), even in the absence of clinical symptoms (Subclinical Mastitis) (Abebe R. et al., 2016). This aligned with standard practices in similar studies, and the dataset also comprised facts about milk samples' electroconductivity.

3.2. Data Sources

Two distinct datasets were used for our Artificial Neural Network (ANN)-based classification. The first dataset was from a farm management system known as RISKKA. In RISKKA, each animal is distinctively identified. Data stored in this system include SCC values obtained from one milking each month in each individually labelled cow. This database had SCC data for 18 dates and 1,368 initial records between 2019-10-28 and 2021-04-27. The second database, ALPRO, contains various automatically measured milk parameters and milk yield data for each cow and milking event.

3.3. Data Preparation

In preparing our data for modelling, the two datasets were merged using a unique identifier of the cows, and the measured data for morning milking were filtered for each individual up to 3 days before the SCC data measurement date.

3.4. Target Variable

To create a binomial field as a dependent variable or, in other words, binary classification, SCC played a fundamental role in our study. Specifically, a value of 1 if SCC was above 200,000 and 0 below 200,000.

3.5. Feature Selection

Using the functions of the caret package (Kuhn M., 2023), we filtered out correlated explanatory variables in the R-environment (R Core Team., 2023). We estimated the variable importance using a binomial generalized linear model. For neural network training, we kept explanatory variables with variable importance values greater than 3, as shown in Table 1.

Variable	Variable description	Variable Importance (VI)
PeakCondLevel	Maximum conductivity of milk during milking (mS)	VI >3 (Important)
AvgCondLevel	Average conductivity of the milk during milking in mS	VI >3 (Important)
RelativeCond	Change in conductivity of the milk.	VI >3 (Important)
Yield	Amount of milk delivered in kg.	VI >3 (Important)
YieldIsLow	Binary expression of production decreases compared to individual production	VI >3 (Important)
Number of actual lactations	Actual lactations	VI >3 (Important)
Days in milk	Days associated with the day of SCC measurement	VI >3 (Important)
PeakCondLevel	Maximum Conductivity of milk measured 1-, 2- and 3-days before SCC measurement (mS)	VI >3 (Important)

Table 1 Descriptive Table of variables of Electroconductivity of milk.

The resulting dataset of this careful feature selection contained 7,685 records.

3.5. Machine Learning Models Used

We used five models to make SCC increase prediction, namely decision tree, k-nearest Neighbors, Logistic Regression, Random Forest and the XGBoost. Modelling was performed in a Python environment with five-fold cross-validation.

4. CHAPTER IV: RESULTS

4.5. Performance evaluation of machine learning models

For the five models applied in this study, ROC curves are collected in **Figure 1**. and they are an illustration of the performance of each model in distinguishing the increase in all the measured variables (indicative of potential Intramammary Infection) and non-increase cases indicating the absence of potential IMI. ROC curves measure the classification efficiency (Bobbo T. et al., 2021).

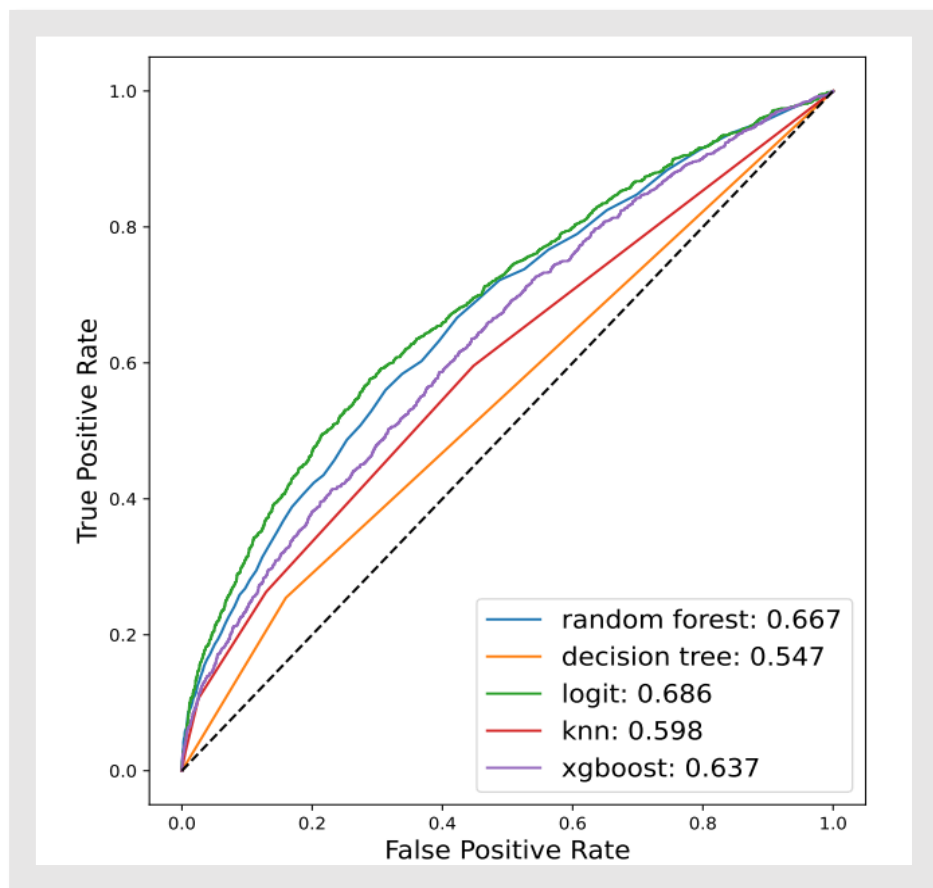


Figure 1. Prediction performance. The ROC curves represent the prediction performance of the five models applied. For each model, the AUC is shown numerically.

Values on the x-axis show the False Positive Rate (specificity) and on the y-axis shows the True Positive Rate (sensitivity). The two are inversely correlated even though adjusting the True Positive Rate does not necessarily mean a decrease in specificity (Nagy et al., 2023).

5. CHAPTER V: DISCUSSION

BM is a critical condition in the dairy industry, necessitating preciseness in its diagnosis, thereby preventing the overuse of antimicrobials, with the potential of ML in biological studies offering hopes of revolutionizing intervention (Hoque M. N. et al., 2019; Hyde R. M. et al., 2020; Xiang T. W. et al., 2018). However, existing techniques have limitations, as Hamilton (2017) reported. In direct contrast, ML-based algorithms provide a solution that is quick, accurate, and cost-saving (Noor A. et al., 2020).

5.1. Comparative analysis of ML models

In this chapter, we embark on a comparative analysis of the performance of five ML predictive models: LR, XGBoost, RF, DT, and KNN, focusing on their AUC values as derived from Figure 1. We will also compare our findings to the existing literature in this study's context and provide the implications and prospects of this study.

Figure 1 is a collective illustration of the AUC values for all five models, and each curve represents the probability of each model making a distinction between elevated SCC as positive to IMI and decreased SCC as unfavourable to IMI cases (Bobbo T. et al., 2021). ROC curves are essential in describing technology and algorithms (Hanley J. A. et al., 1983). They are used to analyse the performance of predictive models. The Logistic Regression model has the highest AUC of 0.686, and the Decision Tree has the lowest AUC of 0.547.

These AUC values are generated by calculating the distance from the x-axis, the false positive rate, regarded as the sensitivity, meaning that the model whose curve is located with the highest distance from the x-axis is the best among the tested models (Fawcett, 2006).

SCC data was numerically easily classifiable with classification models. However, ECM data involved nine explanatory variables of importance after feature selection, necessitating the application of ANN that performs well on complex data. We leveraged the R environment's statistical advantages and ANN's ability to make predictions based on ECM.

The performance of the LR model could relate to its ability to find connections within the dataset through the logistic function to analyse lots of unrelated variables to estimate outcomes (Panesar et al., 2019; Dofadar et al., 2022) and the inherent capability to effectively capture relationships between independent and dependent variables (Niaz et al., 2021).

According to Nagy et al. (2023), the tests used to improve predictive ability must differ in architecture. As you can see, the LR model is so different in terms of architecture from the DT, XGBoost, and the RF, so the use of ensemble modelling through the XGBoost and RF in this study because they are from the simple DTs does not seem to increase the model performance significantly. Even though the ensemble classification has been applied to success in many fields with improved accuracy (Anwar H. et al., 2014), their application to this study was strongly affected by the mediocre performance of DTs with the lowest AUC of 0.547.

5.2. Comparison to Existing Studies and Implications

As earlier presented, more advanced image-based algorithms have been used to detect cases affecting animals more quickly than statistic-based ML algorithms (Nagy S. et al., 2023; Machuwe et al., 2023; Wang et al., 2019). ANN, CNN, Target Detection and EFMYOLOv3 yielded promising results.

In comparing our study to the literature cited in my study, it is unarguably reasonable to say ML studies utilize identical methodologies with fewer modifications, all in the quest to increase the accuracy of predictive models. The logic in many studies is to enhance the quick and correct detection of irregularities in challenging matters in the intended application field. Most involve data collection, preprocessing, choice of algorithms, validation of algorithms and presentation of findings, among other steps.

In the context of mastitis detection, similar methods have been used, contrastingly with more complex datasets (Bobbo T. et al., 2021;) and more features for modelling (Bobbo T. et al., 2022). Moreover, both studies validated their models using ten-fold cross-validation and a thousand iterations for accuracy improvements and SCC data collected in non-invasive ways from automatically measured variables.

According to Norberg et al. (2004), the variation in ECM of milk from an infected quarter may be larger than in ECM of milk from healthy quarters, and a combination of the level and the variation of EC measurements may improve the description of the trait. Regarding this, we performed feature selection for ECM and remained with nine explanatory variables that would

avail real-time data for ANN training. For the validation of models, we used only 5-cross validation. For performance evaluation, we only used ROC curves in this study section. Unlike the many other metrics used in previous studies, AUC values were the basis for comparing models and on which we draw our conclusions.

5.1.Limitations of the study

ML has significantly advanced in recent years, thanks to new techniques and algorithms in diverse scientific domains, including biology (De Palva B. et al., 2023). Nonetheless, these advancements are accompanied by specific challenges. Overfitting occurs when ML models excel in training data but struggle with new, unseen data, necessitating models' simplification and practical training to mitigate this issue (Xu C. et al., 2019). Conversely, underfitting arises when models are excessively simplistic, leading to poor training and new data performance.

ML algorithms require substantial datasets for training and evaluation, and the data's quality significantly impacts their effectiveness, often resulting in considerable computational costs (Yang A. et al., 2020). Extracting valuable insights from extensive and frequently unstructured healthcare datasets remains an enduring challenge (Miotto R. et al., 2018). It is essential to emphasise that AI, including ML, should not be solely relied upon as a diagnostic tool (Nagy et al., 2023).

Despite these hurdles, ML methods, particularly DNN, exhibit remarkable proficiency in managing complex datasets (Shamshirband et al., 2021). In the context of SM detection, exclusive reliance on ECM measurement is intricate due to multifarious factors, such as stress, temperature, milking conditions, genetics, and feed quality, affecting milk's ionic variations (Nielen M. et al., 1992). Additionally, both SCC and ECM measurements lack mastitis-specificity, fluctuating across different breeds, animal ages, and lactation stages. This limitation underscores the challenge of solely relying on conventional methods like the CMT for SM evaluation (Deshapriya R. et al., 2011). Despite suggested SCC thresholds for identifying IMI, its standalone use is restricted due to the influence of various factors (Ruegg & Pantoja, 2013; Sumon SMMR et al., 2020).

Future study directions

In providing future insights for similar studies, I recommend using more inclusive datasets, bearing in mind the potential of ML methods. Automatically measured parameters like (feed intake and exercise activity should be considered for feature selection to be more accurate in ruling out factors correlated to the increase in SCC and ECM. More evaluation metrics should be tried to improve the prediction models' robustness. Additionally, taking into account breed variability of parameters, geographical conditions would improve the generalization of the findings.

6. CHAPTER VI: CONCLUSION

Our study aimed to determine the possibility of developing an ML algorithm that can predict the existence of SM. Our findings indicate that statistical-based ML algorithms can predict the likelihood of SM with reasonable accuracy, as represented by the AUC values. Out of all the models we used, the LR was the most suitable for binary classifications, as it outperformed all other models. Additionally, the use of explanatory variables for ANN-based training related to ECM made the use of ECM more informative and accurately correlated to the increase in SCC. However, ECM still cannot be used as a standalone milk parameter and should be used with other existing methods. Although the CMT is labour-intensive, it still holds its significance for physically examining the udder's health status. Using such predictive potential guarantees high precision.

7. ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the Centre of Bioinformatics at the University of Veterinary Medicine Budapest, especially to Professor Dr Solymosi Norbert, the director, for providing exceptional guidance and support in supplementing this topic. I also want to thank Dr. Nagy Sára Ágnes for her initial description of the methodology and all the other staff members who have helped me in any way during my studies.

Additionally, I am grateful for the incomparable financial assistance provided by The Tempus Public Foundation, which has sponsored my studies through the Stipendium Hungaricum Scholarship.

With the tireless contributions of the people mentioned above, this part of my study was made possible.

8. REFERENCES

- Abebe, R., Hatiya, H., Abera, M. *et al.* Bovine Mastitis: prevalence, risk factors and isolation of *Staphylococcus aureus* in dairy herds at Hawassa milk shed, South Ethiopia. *BMC Vet Res* 12, 270 (2016). <https://doi.org/10.1186/s12917-016-0905-3>
- Anwar, H., Qamar, U., & Qureshi, A. W. M. (2014). Global Optimization Ensemble Model for Classification Methods. *The Scientific World Journal*, 2014, 313164. <https://doi.org/10.1155/2014/313164>
- Baykan, N.A.; Yılmaz, N. A Mineral Classification System with Multiple Artificial Neural Networks Using K-Fold Cross Validation. *Math. Comput. Appl.* **2011**, *16*, 22-30. <https://doi.org/10.3390/mca16010022>
- Bobbo, T., Biffani, S., Taccioli, C., Penasa, M., & Cassandro, M. (123 C.E.). Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. *Scientific Reports* |, *11*, 13642. <https://doi.org/10.1038/s41598-021-93056-4>
- Bobbo, T., Matera, R., Pedota, G., Manunza, A., Cotticelli, A., Neglia, G., & Biffani, S. (2022). Exploiting machine learning methods with monthly routine milk recording data and climatic information to predict subclinical mastitis in Italian Mediterranean buffaloes. *Journal of Dairy Science*. <https://doi.org/10.3168/JDS.2022-22292>
- Breiman, L. (1984). Classification and Regression Trees (1st ed.). Routledge. <https://doi.org/10.1201/9781315139470>
- Breiman, L., 2001. Random forests. *Machine learning*, *45*, pp.5-32.
- Brownlee, J. (2023, October 4). K-Fold Cross-Validation in Machine Learning. Retrieved October 5, 2023, from Machine Learning Mastery website: <https://machinelearningmastery.com/k-fold-cross-validation/>
- Carpenter KA, Huang X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *Current Pharmaceutical Design*. 2018;24(28):3347-3358. <https://doi.org/10.2174/1381612824666180607124038>
- Cheng, W. N., & Han, S. G. (2020). Bovine mastitis: Risk factors, therapeutic strategies, and alternative treatments — A review. *Asian-Australasian Journal of Animal Sciences*, *33*(11), 1699-1713. <https://doi.org/10.5713/ajas.20.0156>
- Cohen, EB, Gordon, IK. First, Do no Harm. Ethical and legal issues of artificial intelligence and machine learning in veterinary radiology and radiation oncology. *Vet Radiol Ultrasound*. 2022; 63(Suppl. 1): 840–850. <https://doi.org/10.1111/vru.13171>
- Comprehensive R Archive Network (CRAN) Mirrors. (n.d.). Retrieved October 10, 2023, from <https://cran.r-project.org/mirrors.html>
- CowManager. (2021). CowManager Creates Interface with RISKÁ Hungary. [online] Available at: <https://www.cowmanager.com/news/cowmanager-creates-interface-with-riska-hungary> (Accessed 23 September 2023)
- De Paiva, B. B. M., Pereira, P. D., de Andrade, C. M. V., Gomes, V. M. R., Souza-Silva, M. V. R., Martins, K. P. M. P., Sales, T. L. S., de Carvalho, R. L. R., Pires, M. C., Ramos, L. E. F., Silva, R. T., de Freitas Martins Vieira, A., Nunes, A. G. S., de Oliveira Jorge, A., de Oliveira

- Maurílio, A., Scotton, A. L. B. A., da Silva, C. T. C. A., Cimini, C. C. R., Ponce, D., ... Marcolino, M. S. (2023). Potential and limitations of machine meta-learning (ensemble) methods for predicting COVID-19 mortality in a large hospital Brazilian dataset. *Scientific Reports 2023 13:1*, 13(1), 1–18. <https://doi.org/10.1038/s41598-023-28579-z>
- de Vlieghe, S., Ohnstad, I., & Piepers, S. (2018). Management and prevention of mastitis: A multifactorial approach with a focus on milking, bedding, and data-management. *Journal of Integrative Agriculture*, 17(6), 1214–1233. [https://doi.org/10.1016/S2095-3119\(17\)61893-8](https://doi.org/10.1016/S2095-3119(17)61893-8)
- Deshapriya, R. M. C., Rahularaj, R., & Ransinghe, R. M. S. B. K. (2011). Relationship of Somatic Cell Count and Mastitis: An Overview. *Asian-Australasian Journal of Animal Sciences*, 66(1), 1. <https://doi.org/10.4038/SLVJ.V66I1.32>
- Dietterich, T. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, *Machine Learning*, 1–22
- Dofadar, D., Dibyo, A., Abdullah, H., Khan, R., Rahman, R., & Ahmed, M. S. (2022). A Comparative Analysis of Lumpy Skin Disease Prediction Through Machine Learning Approaches. Title of the Journal/Conference Proceedings, Volume (Issue), 1-4. <http://dx.doi.org/10.1109/IICAJET55139.2022.9936742>
- Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E., & Petrovski, K. R. (2019). Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models. *Computers in Biology and Medicine*, 114, 103456. <https://doi.org/10.1016/j.combiomed.2019.103456>
- Estes, L., Nakatumba-Nabende, J., Xie, S., & Machuve, D. (n.d.). *Poultry disease diagnostics models using deep learning*. <https://doi.org/10.3389/frai.2022.733345>
- Fadul-Pacheco, L., Delgado, H., & Cabrera, V. E. (2021). Exploring machine learning algorithms for early prediction of clinical mastitis. *International Dairy Journal*, 119, 105051. <https://doi.org/10.1016/j.idairyj.2021.105051>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Fernando RS, Spahr SL, Jaster EH. Comparison of electrical conductivity of milk with other indirect methods for detection of subclinical mastitis. *J Dairy Sci*. 1985 Feb;68(2):449-56. [https://doi.org/10.3168/jds.s0022-0302\(85\)80844-4](https://doi.org/10.3168/jds.s0022-0302(85)80844-4)
- Fernando, R. S., Rindsig, R. B., & Spahr, S. L. (1982). Electrical Conductivity of Milk for Detection of Mastitis. *Journal of Dairy Science*, 65(4), 659–664. [https://doi.org/10.3168/jds.S0022-0302\(82\)82245-5](https://doi.org/10.3168/jds.S0022-0302(82)82245-5)
- Gruet, P., Maincent, P., Berthelot, X., & Kaltsatos, V. (2001). Bovine mastitis and intramammary drug delivery: Review and perspectives. *Advanced Drug Delivery Reviews*, 50(3), 245-259. [https://doi.org/10.1016/S0169-409X\(01\)00160-0](https://doi.org/10.1016/S0169-409X(01)00160-0)
- Hamilton, J., & Solymosi, N. (n.d.). Study of association between the electrical conductivity of milk and subclinical mastitis in dairy cows <http://www.huveta.hu/handle/10832/2162>
- Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148, 839–843. Doi: <https://dx.doi.org/10.1148/radiology.148.3.6878708>

Henrique Orsi, Felipe F. Guimarães, Domingos S. Leite, Simony T. Guerra, Sâmea F. Joaquim, Jose C.F. Pantoja, Rodrigo T. Hernandez, Simone B. Lucheis, Márcio G. Ribeiro, Helio Langoni, Vera L.M. Rall, Characterization of mammary pathogenic Escherichia coli reveals the diversity of Escherichia coli isolates associated with bovine clinical mastitis in Brazil, *Journal of Dairy Science*, Volume 106, Issue 2, 2023, 1403-1413, ISSN 0022-0302, <https://doi.org/10.3168/jds.2022-22126>.

Hoque MN, Istiaq A, Clement RA, Sultana M, Crandall KA, Siddiki AZ, Hossain MA. Metagenomic deep sequencing reveals association of microbiome signature with functional biases in bovine mastitis. *Sci Rep*. 2019 Sep 19;9(1):13536. doi: PMID: 31537825; PMCID: PMC6753130. doi: <https://doi.org/10.1038/s41598-019-49468-4>

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. *Scientific Reports*, 10(1), 1-8. <https://doi.org/10.1038/s41598-020-61126-8>

Implementation of machine learning for the evaluation of mastitis and antimicrobial resistance in dairy cows / Request PDF. (n.d.). Retrieved https://www.researchgate.net/publication/357477281_Implementation_of_machine_learning_for_the_evaluation_of_mastitis_and_antimicrobial_resistance_in_dairy_cows

Jabbar, M.A., Deekshatulu, B.L., & Chandra, P. (2013). Classification of heart disease using K-nearest neighbor and genetic algorithm. *Procedia Technology*, 10, 85-94. <http://dx.doi.org/10.1016/j.protcy.2013.12.340>

Jacobs J.A. & Siegford J.M. Invited review: The impact of automatic milking systems on dairy cow management, behaviour, health, and welfare. *Journal of Dairy Science*. 2012; 95 (22541453): 2227-2247 <https://doi.org/10.3168/jds.2011-4943>

Klaas, I. C., & Zadoks, R. N. 2017. "An Update on Environmental Mastitis: Challenging Perceptions." *Transboundary and Emerging Diseases*, 65, 166-185. Available at: <https://doi.org/10.1111/tbed.12704>

Koeck, A., Miglior, F., Kelton, D. F., & Schenkel, F. S. (2012). Alternative somatic cell count traits to improve mastitis resistance in Canadian Holsteins. *Journal of Dairy Science*, 95(1), 432–439. <https://doi.org/10.3168/JDS.2011-4731>

Kohavi, R., 1995, August. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Kuhn, M. *caret: Classification and regression training* (2023). R package version 6.0-93. <https://CRAN.R-project.org/package=caret>

Lee, C. S., & Lee, A. Y. (2020). Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6), e279–e281. [https://doi.org/10.1016/S2589-7500\(20\)30102-3](https://doi.org/10.1016/S2589-7500(20)30102-3)

Lijetha, J.C., & Umarani, S.G. (2023). A KNN Algorithm Based Predictive Model for Heart Disease Progression. *Recent Advances in Computer Science and Communications*, 16(5), e120522204706. <https://dx.doi.org/10.2174/2666255815666220512223522>

Lin H, Chen W. Briefing in application of machine learning methods in ion channel prediction. *Scientific World Journal*. 2015; 2015:945927. doi: 10.1155/2015/945927. Epub 2015 Apr 16.

PMID: 25961077; PMCID: PMC4415473 [Briefing in application of machine learning methods in ion channel prediction - PubMed \(nih.gov\)](#)

Malcata FB, Pepler PT, O'Reilly EL, Brady N, Eckersall PD, Zadoks RN, Viora L. Point-of-care tests for bovine clinical mastitis: what do we have and what do we need? *J Dairy Res.* 2020 Aug;87(S1):60-66. doi: 10.1017/S002202992000062X. Epub 2020 Jul 30. PMID: 33213589 [Point-of-care tests for bovine clinical mastitis: What do we have and what do we need? - PubMed \(nih.gov\)](#)

Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities, and challenges. *Brief Bioinform.* 2018 Nov 27;19(6):1236-1246. doi: 10.1093/bib/bbx044. PMID: 28481991; PMCID: PMC6455466 [Deep learning for healthcare: review, opportunities, and challenges - PubMed \(nih.gov\)](#)

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2), 757-774. <https://doi.org/10.1016/j.jksuci.2023.01.014>

Montomoli, J., Romeo, L., Moccia, S., Bernardini, M., Migliorelli, L., Berardini, D., Donati, A., Carsetti, A., Bocci, M. G., Wendel Garcia, P. D., Fumeaux, T., Guerci, P., Schüpbach, R. A., Ince, C., Frontoni, E., Hilty, M. P., & Investigators, I. (2021). Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *Journal of Intensive Medicine*, 1(2), 110-116. <https://doi.org/10.1016/j.jointm.2021.09.002>

Morales-Ubaldo, A. L., Rivero-Perez, N., Valladares-Carranza, B., Velázquez-Ordoñez, V., Delgadillo-Ruiz, L., & Zaragoza-Bastida, A. (2023). Bovine mastitis, a worldwide impact disease: Prevalence, antimicrobial resistance, and viable alternative approaches. *Veterinary and Animal Science*, 21, 100306. <https://doi.org/10.1016/j.vas.2023.100306>

Nagy, S. Á.; Csabai, I.; Varga, T.; Póth-Szebenyi, B.; Gábor, G.; Solymosi, N. Neural network-aided milk somatic cell count gain prediction, DOI: 10.21203/rs.3.rs-2865554/v1 (2023). PREPRINT (<https://doi.org/10.21203/rs.3.rs-2865554/v1>).

Nagy, S.Á. et al., 2023. Mesterséges neurális hálózatok az állattermék-előállításban. *MAGYAR ÁLLATORVOSOK LAPJA*, 145(5), pp.309–319 <https://doi.org/10.56385/magyallorv.2023.05.309-319>

Nagy, S.Á.; Kilim, O.; Csabai, I.; Gábor, G.; Solymosi, N. Impact Evaluation of Score Classes and Annotation Regions in Deep Learning-Based Dairy Cow Body Condition Prediction. *Animals* 2023, 13, 194. <https://doi.org/10.3390/ani13020194>

Niaz, R., Zhang, X., Iqbal, N., Almazah, M.M.A., Hussain, T., Hussain, I. (2021). Logistic Regression Analysis for Spatial Patterns of Drought Persistence. *Complexity*, 2021, Article ID 3724919, 13 pages. <https://doi.org/10.1155/2021/3724919>

Nielen, M., Deluyker, H., Schukken, Y. H., & Brand, A. (1992). Electrical Conductivity of Milk: Measurement, Modifiers, and Meta Analysis of Mastitis Detection Performance. *Journal of Dairy Science*, 75(2), 606–614. [https://doi.org/10.3168/jds.S0022-0302\(92\)77798-4](https://doi.org/10.3168/jds.S0022-0302(92)77798-4)

Noor, A., Zhao, Y., Koubaa, A., Wu, L., Khan, R., & Abdalla, F. Y. O. (2020). Automated sheep facial expression classification using deep transfer learning. *Computers and Electronics in Agriculture*, 175. <https://doi.org/10.1016/J.COMPAG.2020.105528>

Norberg, E., Hogeveen, H., Korsgaard, I.R., Friggens, N.C., Sloth, K.H.M.N. and Løvendahl, P., 2004. Electrical conductivity of milk: Ability to predict mastitis status. *Journal of Dairy Science*, 87(4), pp.1099-1107

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, 375(13), 1216. <https://doi.org/10.1056/NEJMp1606181>

Pakrashi, A., Ryan, C., Guéret, C., Berry, D., Corcoran, M., Keane, M., & Mac Namee, B. (2023). Early Detection of Subclinical Mastitis in Dairy Cows: A Machine Learning Approach. *Journal of Dairy Science*, 106(5), 4650-4661. <https://doi.org/10.3168/jds.2022-21095>

Panesar, S. S., D'Souza, R. N., Yeh, F., & Fernandez-Miranda, J. C. (2019). Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database. *World Neurosurgery*: X, 2, 100012. <https://doi.org/10.1016/j.wnsx.2019.100012>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, pp.2825-2830 <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://> .

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at <https://www.R-project.org/>

R Core Team. Title of the webpage. Available at: <https://www.r-project.org/about.html> (Accessed 4 October 2023).

Radostits, O. M., Leslie, K. E. & Fetrow, J. Herd health: food animal production medicine. *Herd Heal. food Anim. Prod. Med.* (1994). <https://vetbooks.ir/herd-health-food-animal-production-medicine-3rd-edition>

RDocumentation (2023) pROC Package Documentation. Available at: <https://www.rdocumentation.org/packages/pROC/versions/1.18.4> (Accessed: October 19, 2023).

Ruegg, P. & Pantoja, J. Understanding and using somatic cell counts to improve milk quality. *Irish Journal of Agricultural and Food Research* 52: 101–117, (2013) (16) (PDF) [Understanding and using somatic cell counts to improve milk quality \(researchgate.net\)](https://www.researchgate.net/publication/312511111_Understanding_and_using_somatic_cell_counts_to_improve_milk_quality)

Rupp, R., & Boichard, D. (2003). Genetics of resistance to mastitis in dairy cattle. *Veterinary Research*, 34(5), 671–688. <https://doi.org/10.1051/VETRES:2003020>

Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T., & Alinejad-Rokny, H. (2021). A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113, 103627. <https://doi.org/10.1016/J.JBI.2020.103627>

Shanthababu. (2023, October 1). Data Analytics in Business. Analytics Vidhya. Retrieved October 7, 2023, from <https://www.analyticsvidhya.com/blog/author/shanthabu/>

Sharma, A., Khan, F., Sharma, D., & Gupta, S. (2023). Python: The Programming Language of the Future. *Publication Volume & Issue*, 6(12), 115-118. <https://ijirt.org/Article?manuscript=149340>

Sharma, N., Singh, N. K., & Bhadwal, M. S. (2011). Relationship of Somatic Cell Count and Mastitis: An Overview. *Asian-Australasian Journal of Animal Sciences*, 24(3), 429–438. <https://doi.org/10.5713/AJAS.2011.10233>

Sharma, N., Singh, N., & Bhadwal, M. (2011). Relationship of Somatic Cell Count and Mastitis: An Overview. *Asian-Australasian Journal of Animal Sciences*, 3(4), 881–906. <https://doi.org/10.3390/DAIRY3040061>

Software Informer. (2023). Alpro Windows. [online] Available at: <https://alpro-windows.software.informer.com> (Accessed 23 September 2023).

Strecht, P., Mendes-Moreira, J., Soares, C. (2019). Generalizing Knowledge in Decentralized Rule-Based Models. In: Monreale, A., *et al.* ECML PKDD 2018 https://doi.org/10.1007/978-3-030-14880-5_3

Sumon SMMR, Parvin MS, Ehsan MA, Islam MT. Relationship between somatic cell counts and subclinical mastitis in lactating dairy cows. *Vet World*. 2020;13(8):1709-1713. doi:10.14202/vetworld.2020.1709-1713 [Relationship between somatic cell counts and subclinical mastitis in lactating dairy cows - PubMed \(nih.gov\)](https://pubmed.ncbi.nlm.nih.gov/34484444/)

TECA. (n.d.). Retrieved February 18, 2023, <https://www.fao.org/teca/en/technologies/10058>

The R Project for Statistical Computing. Retrieved October 10, 2023, from <https://www.r-project.org/about.html>

Wang, Y., Kang, X., He, Z., Feng, Y., & Liu, G. (2022). Accurate detection of dairy cow mastitis with deep learning technology: a new and comprehensive detection method based on <https://doi.org/10.1016/J.ANIMAL.2022.100646>

Xiao, C., Ji, Q., Chen, J., Zhang, F., Li, Y., Fan, J., Hou, X., Yan, F., & Wang, H. (2022). Prediction of soil salinity parameters using machine learning models in an arid region of northwest China. *Computers and Electronics in Agriculture*, 204, 107512. <https://doi.org/10.1016/j.compag.2022.107512>

Xu, C., & Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biology*, 20(1), 1–4. <https://doi.org/10.1186/S13059-019-1689-0/FIGURES/2>

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791 [Gradient-based learning applied to document recognition | IEEE Journals & Magazine | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/726791)

Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020). Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Frontiers in Bioengineering and Biotechnology*, 8, 1032.

Yu XT, Wang L, Zeng T. Revisit of Machine Learning Supported Biological and Biomedical Studies. *Methods Mol Biol*. 2018; 1754:183-204. doi: 10.1007/978-1-4939-7717-8_11. PMID: 29536444. https://doi.org/10.1007/978-1-4939-7717-8_11

Zhang Xudong, Kang Xi, Feng Ningning, Liu Gang, Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector, *Computers and Electronics in Agriculture*, Volume 178,2020,105754, ISSN 0168-1699 <https://doi.org/10.1016/j.compag.2020.105754>



Thesis progress report for veterinary students

Name of student: JOSEPH OSCAR ULLOMI

Neptun code of the student: I6IYI8

Name and title of the supervisor: Norbert Solymosi, assoc. prof.

Department: Centre for Bioinformatics

Thesis title: Subclinical mastitis detection by machine learning

Consultation – 1st semester

Timing				Topic / Remarks of the supervisor	Signature of the supervisor
	year	month	day		
1.	2022	9	9	Literature review	<i>Norbert Solymosi</i>
2.	2022	10	4	Literature review	<i>Norbert Solymosi</i>
3.	2022	11	17	Literature review	<i>Norbert Solymosi</i>
4.	2022	12	6	Literature review	<i>Norbert Solymosi</i>
5.	2023	1	31	Literature review	<i>Norbert Solymosi</i>

Grade achieved at the end of the first semester: 5

Consultation – 2nd semester

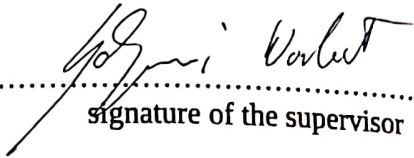
Timing				Topic / Remarks of the supervisor	Signature of the supervisor
	year	month	day		
1.	2023	2	9	Literature review	<i>Norbert Solymosi</i>
2.	2023	3	10	Thesis writing	<i>Norbert Solymosi</i>
3.	2023	4	3	Thesis writing	<i>Norbert Solymosi</i>
4.	2023	4	26	Thesis writing	<i>Norbert Solymosi</i>
5.	2023	5	9	Thesis writing	<i>Norbert Solymosi</i>

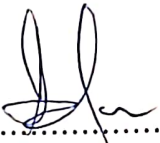
Grade achieved at the end of the second semester: 5

The thesis meets the requirements of the Study and Examination Rules of the University and the Guide to Thesis Writing.



I accept the thesis and found suitable to defence,


signature of the supervisor

Signature of the student: 

Signature of the secretary of the department: NA 

Date of handing the thesis in