UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Ph.D school in Artificial Intelligence – Agrifood and environment

XXXVII cycle

High-tech Farming: Study and Application of Machine Learning Techniques for Improving Production Efficiency in Livestock

Tutor:
Prof. Eng. Nicola Pasquino

Candidate:
Lucia Trapanese
DR995916

Co-Tutor:
DVM Angela Salzano

Academic year 2023/2024

Abstract

The livestock sector faces numerous challenges, including the need to mitigate greenhouse gas emissions and meet the increasing demand for animal-based foods. To address these issues, modernization of breeding techniques is essential for enhancing the overall efficiency of production systems. Additionally, recent trends indicate that consumers are increasingly choosing products based on factors such as raw material quality, animal welfare, and the healthiness of processed foods. Achieving these objectives requires a multi-level approach. Since the 1970s, farmers, veterinarians, and technicians have recognized the benefits of adopting an automated and data-driven approach to improve production and enhance their quality of life. One of the early innovations was the use of individual electronic milk meters for cows, which automatically measured milk yield. The introduction of automatic milking systems in the 2000s further revolutionized the dairy sector, alongside the development of specific devices for recognizing oestrus, animal behaviour and diseases. This progress coincided with significant advancements in engineering, making sensors, devices, and computational units more powerful and affordable. Moreover, the rise of artificial intelligence techniques around 2000 contributed to the widespread dissemination of knowledge in the field. Finally, in 2014, Berckmans defined PLF as the continuous automated real-time monitoring of livestock production, reproduction, health, welfare, and environmental impact (Berckmans, 2014). With the PLF approach, it is now possible to continuously monitor the environment, barn conditions, animal Behaviour, welfare, and production levels.

My Ph.D thesis emerged during this dynamic period, where a multidisciplinary approach has become essential for original research. The collaboration between engineers, data scientists, and veterinarians has been crucial to this endeavor. This thesis aimed to enhance the management of dairy herds by analyzing various aspects of milk production and dairy farming. By focusing on key factors influencing dairy production, the research aimed to improve efficiency and decision-making in the dairy sector through the integration of advanced technologies and data-driven strategies. Despite technological innovations, a significant

amount of data remains underutilized. The techniques utilized in this thesis pertain to the data analysis. Various Machine Learning (ML) techniques were employed not randomly, but out of necessity to manage large volumes of data (from automatic milking systems and herd databases) and extracted valuable insights. ML techniques were chosen for their ability to handle complex, large-scale datasets and identify patterns that can enhance decision-making in the dairy sector, particularly regarding production.

Keywords: Precision Livestock Farming, Dairy, Machine Learning, Milk

Sintesi

Il settore zootecnico sta affrontando numerose sfide, tra cui la necessità di ridurre le emissioni di gas serra e soddisfare la crescente domanda di alimenti di origine animale. Al fine di affrontare tali problematiche, è essenziale modernizzare le tecniche di allevamento, migliorando così l'efficienza complessiva dei sistemi di produzione. Inoltre, le tendenze recenti mostrano che i consumatori scelgono sempre più spesso prodotti in base a fattori come la qualità delle materie prime, il benessere degli animali e la salubrità dei cibi trasformati. Raggiungere questi obiettivi richiede un approccio multidisciplinare. A partire dagli anni '70, allevatori, veterinari e tecnici hanno riconosciuto i benefici dell'adozione di un approccio automatizzato e basato sui dati per migliorare la produzione e la loro qualità della vita. Una delle prime innovazioni è stata l'introduzione dei misuratori elettronici individuali per la misurazione automatica della produzione di latte nelle vacche. L'arrivo dei sistemi di mungitura automatica negli anni 2000 ha ulteriormente rivoluzionato il settore lattiero-caseario, accompagnato dallo sviluppo di dispositivi specifici per il riconoscimento dell'estro, del comportamento e delle malattie degli animali. Questi progressi sono anche il risultato dei significativi sviluppi nel settore ingegneristico, che hanno reso sensori, dispositivi e unità computazionali più potenti ed economici. Inoltre, intorno al 2000, l’ascesa delle tecniche di Intelligenza Artificiale ha contribuito alla diffusione della conoscenza in questo campo. Infine, nel 2014, Berckmans ha formalizzato questo avanzamento tecnologico con il termine "Precision Livestock Farming" (PLF), definendolo come il monitoraggio continuo, automatizzato e in tempo reale della produzione, riproduzione, salute, benessere e impatto ambientale degli animali da allevamento. Grazie all'approccio PLF, è ora possibile monitorare continuamente l'ambiente, le condizioni delle stalle, il comportamento, il benessere e i livelli produttivi degli animali.

La mia tesi di dottorato è nata in questo periodo di grandi trasformazioni, dove è essenziale adottare un approccio multidisciplinare per affrontare i temi discussi. La collaborazione tra

ingegneri, data scientist e veterinari è stata cruciale in questo lavoro. L'obiettivo di questa tesi è stato quello di migliorare la gestione delle mandrie da latte analizzando vari aspetti della produzione di latte e dell'allevamento. Focalizzandosi sui fattori chiave che influenzano la produzione lattiera, la ricerca ha cercato di migliorare l'efficienza e il processo decisionale nel settore lattiero-caseario, integrando tecnologie avanzate e strategie basate sui dati. Nonostante le innovazioni tecnologiche, una quantità significativa di dati rimane ancora inutilizzata. Le tecniche utilizzate in questa tesi si concentrano sull'analisi dei dati ed in particolare su quelle di Machine Learning (ML). Queste ultime sono state impiegate al fine di riuscire a gestire grandi volumi di dati (provenienti dai sistemi di mungitura automatica e dai database delle mandrie) ed estrarre informazioni preziose. Infatti, rispetto alle tecniche statistiche classiche sono in grado di processare dataset complessi e identificare modelli che possono migliorare il processo decisionale nel settore lattiero-caseario, in particolare per quanto riguarda la produzione.

Index

Chapter 1: Introduction to Precision Livestock Farming

Background

The livestock sector is facing many challenges, such as the mitigation of pollutant greenhouse gas emissions and increasing animal-based food demand. Considering these issues, the sector needs modernization for all processes involved in making food process to take control of the global efficiency of production systems. Moreover, according to Peyraud & MacLeo (2020) consumers have increasingly based their product choices on raw materials quality, animal welfare, and the healthiness of processed food. To achieve these objectives, it is necessary to study the livestock sector through a multi-level approach.

Starting from the early of 1900, farmers, veterinaries, and technicians established the milk recording systems especially for improve genetics programs (Weigel et al., 2017). From that moment, recognizing the potential of a data-driven approach, it became more widespread in the livestock sector and was also applied to herd management, leading to significant improvements. One of the first examples was employing individual electronic milk meters for cows to measure each cow's milk yield automatically. Another important contribution was the introduction of the automatic milking system (AMS) in the 2000s which revolutionized the dairy sector (Ozella et al., 2023). At the same time, specific devices for cows were developed and widespread for oestrus and behaviour recognition. This surge happened in combination with some important developments in the engineering field that made sensors, devices, and computational units more powerful and cheaper. Moreover, the increase in Artificial Intelligence techniques across various sectors, which occurred around 2000, contributed to the widespread dissemination of knowledge. Finally, only in the second decade of 2000s this technological epiphany was formalized and described by Berchmans and called Precision Livestock Farming (PLF) (Berckmans, 2014). According to him, PLF is the "management of livestock by continuous automated real-time monitoring of production/reproduction, health and welfare of livestock, and environmental impact". In other words, with the PLF approach, it is possible to monitor continuously the environment, the barn, the animals' behaviour and welfare, and their production. Nowadays, the PLF approach is widespread, and literature is full of papers and practical experience.

PLF workflow

PLF approach includes a plethora of tools, devices, techniques, and net architectures that could work together or alone, depending on the aims of the farmers, veterinaries, and researchers. Fig 1.1 describes a typical PLF workflow.

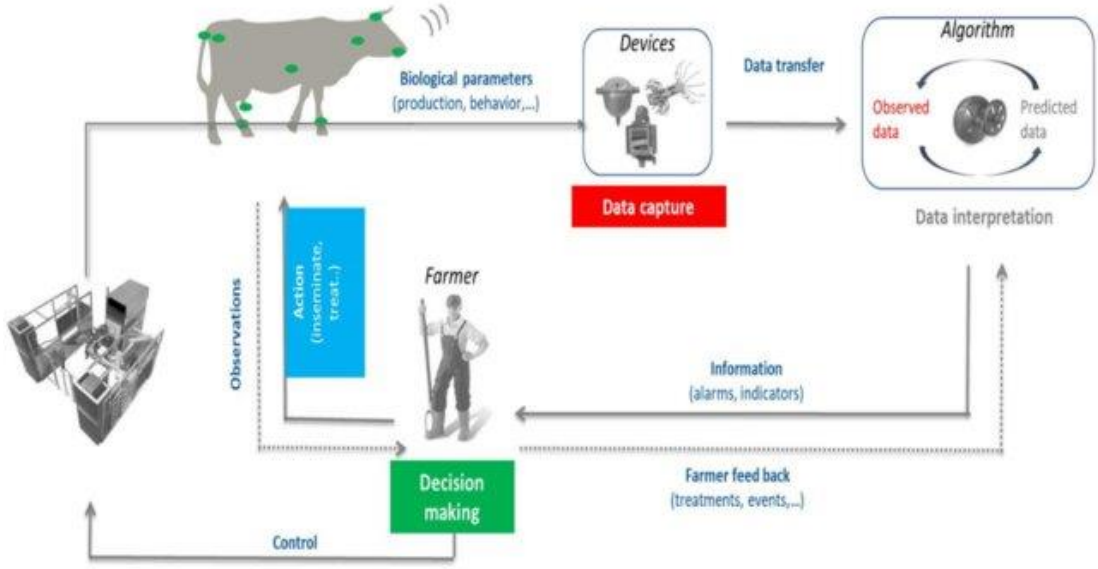

Fig 1.1 *LF is based on different blocks: data capture, data interpretation and decision making all connected in loop (Kleen & Guatteo, 2023)*

Devices such as pedometers, neck collars, ear tags, and weather stations are the primary data sources, collecting information from the environment, barn, and animals representing the first step of the loop. The second step is data analysis, which encompasses all the techniques used to process and interpret the raw data collected by these devices, making it useful and understandable. In this context, machine learning (ML) has been revolutionary, as it enables the management of large datasets and is highly adaptable to non-linear, complex, noisy, and imprecise data, such as dairy production (including aspects like production, reproduction, and animal welfare) (Ray, 2019). The output from the data analysis provides actionable insights that assist farmers in the decision-making process. As shown in Fig 1.1 this system operates in a continuous loop, where the output from one block serves as an input to another, creating a feedback cycle that refines and enhances decision-making over time. In this context, AMS and other advanced devices play a crucial role because they not only improve management efficiency but also collect valuable data to enhance system performance. Moreover, AMS

integrates both sensors and a data analysis block within a single robotic unit. This allows for seamless monitoring of various parameters, such as milk yield, animal health indicators, and udder health (Cogato et al., 2021). Despite the many advantages of PLF, its implementation must face several challenges. Key barriers include the high initial cost of the technology, the still limited technical knowledge among farmers, and ethical concerns regarding the objectification of animals and the potential for further intensification (Schillings et al., 2023). Additionally, issues related to data management, such as integration, interpretation, and privacy present significant obstacles (Maroto Molina et al., 2020).

PLF in the dairy sector

PLF is widespread in various livestock farms because of the versatility and abundance of technologies. However, this thesis is focused on the dairy sector, because of the important role that dairy animals play in animal food-based production. Indeed, Food and Agriculture Organization (FAO) estimated that global milk production in 2023 reached 950 million tons, representing a faster growth rate than that registered in 2022 (*FAO, Dairy Market Review*, 2023). Moreover, as reported by Norton & Berckmans, 2017 milk is the EU's leading agricultural product, representing 15% of the agricultural output by value. Finally, as in other livestock sectors, the challenges of this century have pushed the industry towards a scenario dominated by fewer farms with a larger number of animals in developed countries. In this context, the PLF approach empowers farmers to monitor and manage each animal individually, which would be nearly impossible to achieve without advanced technological systems.

As everyone knows, cows are not a unique dairy species. Buffaloes, goats, and sheep are an important part of the milk market, indeed according to FAOSTAT the sum of milk production of these species in 2022 is the 40% of the total milk produced (*FAOSTAT*, 2024). In particular, the buffalo sector has experienced a steady growth in recent years, especially in Mediterranean countries, Asia, and South America, where buffalo farming plays a significant economic and social role. Despite its importance, ML in buffalo farming remains less developed compared to the bovine sector (Bobbo et al., 2022). Also, for goats and sheep, there are few papers in the literature about the application of PLF compared to the bovine sector. These animals are predominantly raised in developing countries where traditional practices dominate, and the adoption of technology faces several barriers. This is due to the adoption of poorly standardized breeding techniques, which result in challenges in performing systematic

data collection. For example, milking robots are not widely used in this type of breeding, and as a result, only a few data points per lactation can typically be sampled

PLF dairy technologies: data acquisition

The adoption of automatic monitoring systems has been essential to the success of PLF. Before affordable, small, and wearable sensors were introduced, farmers had to record data manually and monitor animals by sight. Processes such as diagnoses and care were typically performed on entire herds or groups of animals, with little focus on individual needs. With modern devices attached to each productive cow and automated systems like milking robots, more detailed information is now available, which simplifies farmers' work and reduces unnecessary costs.

Nowadays, monitoring devices can be classified based on:

- The sensing unit employed: accelerometers, thermometers, and microphones
- The body part where the device is located: legs, neck, or tail
- The health parameters they measure: rumination, lying, walking, hear beat
- Data transmission methods

Fig 1.2 summarizes the wearable sensors used in PLF for the dairy sector.

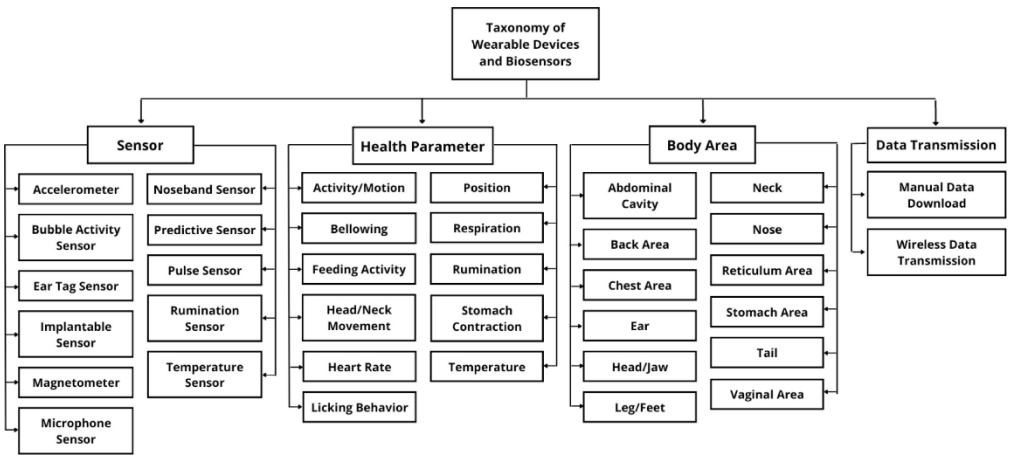

Fig 1.2. *Taxonomy of wearable sensors in PLF. From: Alipio & Villena, 2023*

Accelerometer-based technologies and thermometers are the most commonly used in livestock monitoring (Alipio & Villena, 2023). Devices such as neck and leg collars, pedometers, ear

Fig 1.3 *BMI160 (A), Aficollar (B) and IceQube(C)*

tags, and rumen boluses provide data on cattle behaviour, including rumination, walking, and lying time. These devices typically use accelerometers, which are inertial sensors that measure the vibration or acceleration of movement in a structure. The basic working principle is based on the mass-spring-displacement model, from which the relationship between external acceleration and the linear displacement of the mass element is derived. Changes in acceleration are captured along two axes (2D accelerometers) or three axes (3D accelerometers), per unit of time. Sampling frequencies usually range from 1 to 100 Hz, with full-scale ranges from ± 2 g to ± 16 g. There are now many devices on the market that utilize these technologies, such as IceQube by IceRobotics, SenseHub from Allflex, and AfiCollar by Afimilk and Nedap (Riaboff et al., 2022) . Fig 1.3 shows examples of BMI160 (A), AfiCollar (B), and IceQube (C). It's important to note that devices like the AfiCollar and IceQube are products that incorporate a Microelectromechanical Systems (MEMS) accelerometer, such as the BMI160 provided by BOSCH.

Thermometers, meanwhile, provide information on animals' core body temperature, enabling the monitoring of body temperature. Sometimes, these sensors are embedded in smart boluses and are often combined with other sensors, such as pH monitors, to gather data on rumination and metabolic status (Gasteiner et al., 2015). Some examples of such systems available on the market include smartX and Moonsyst. Additionally, devices located in barns are now widespread for monitoring environmental factors such as temperature, ventilation rates, and

pollutant gases (Leliveld et al., 2024). Finally, also cameras have a high potential to monitor animal behaviours (Fuentes et al., 2023) .

Meanwhile, more complex devices have also transformed the productivity process of the herd improving their efficiency. One of the earliest examples of data acquisition systems in dairy PLF is AMS, also known as milking robots. Introduced for the first time in the early 1990s, this technology led to significant advancements in dairy farming (John et al., 2016). With AMS, cows voluntarily visit the robot for milking, allowing the system to optimize milking intervals based on the lactation stage (Ozella et al., 2023). These robots are equipped with sensors that collect data on cow ID, daily milking frequency, milking gate visits, milk quality (e.g., milk yield per quarter, milk flow, electric conductivity), and the milking process (e.g., milking time, teat-cup attachment, vacuum levels, and the onset of milk letdown). Newer AMS models offer real-time measurements of body condition scores, cow body weight, and feed intake, as automatic feeding systems are now integrated with AMS. The data collected by these systems is processed in the cloud or locally and made available to farmers through reports and statistics, highlighting unusual behaviours. In many dairy developed regions such as Europe and North America, AMS has revolutionized livestock farming by improving both milk quality and quantity and enhancing the overall quality of life for farmers. Due to its advantages, AMS is also being adopted for other dairy species, such as Mediterranean buffaloes in southern Italy (Faugno et al., 2015). Similarly, feeding robots, have enhanced system efficiency by maximizing feed utilization and minimizing waste. These robots ensure that feed is distributed precisely, helping to maintain optimal nutritional levels for each animal.

PLF dairy technologies: Data processing

The adoption of embedded and wearable sensors, biosensors, and digital imaging systems, allows collecting a large amount of data regarding animal conditions and behaviour. To manage this amount of data, a classical statistical approach would not represent a suitable solution because it does not work properly with “big data” as it was designed for few input variables and sample sizes (Bzdok et al., 2018). Moreover, a priori hypothesis and knowledge of the data and observed phenomena are needed to build a useful statistical model (Chen et al., 2022). In this context, ML can successfully overcome these limitations. ML is a branch of Artificial Intelligence applied to studying algorithms for forecasting, inference, and clustering. Among all sub-categories of ML, unsupervised and supervised learning represent the traditional ML techniques that are very common categories of algorithms in livestock

application (Hossain et al., 2022). The major difference between them is the presence of labeled information in the supervised method. The supervised algorithms such as decision trees or neural networks aim to classify or predict a certain value depending on whether the variable is categorical or numerical (Pugliese et al., 2021) while unsupervised techniques such as clustering analysis allow exploratory analysis or dimensionality reduction.

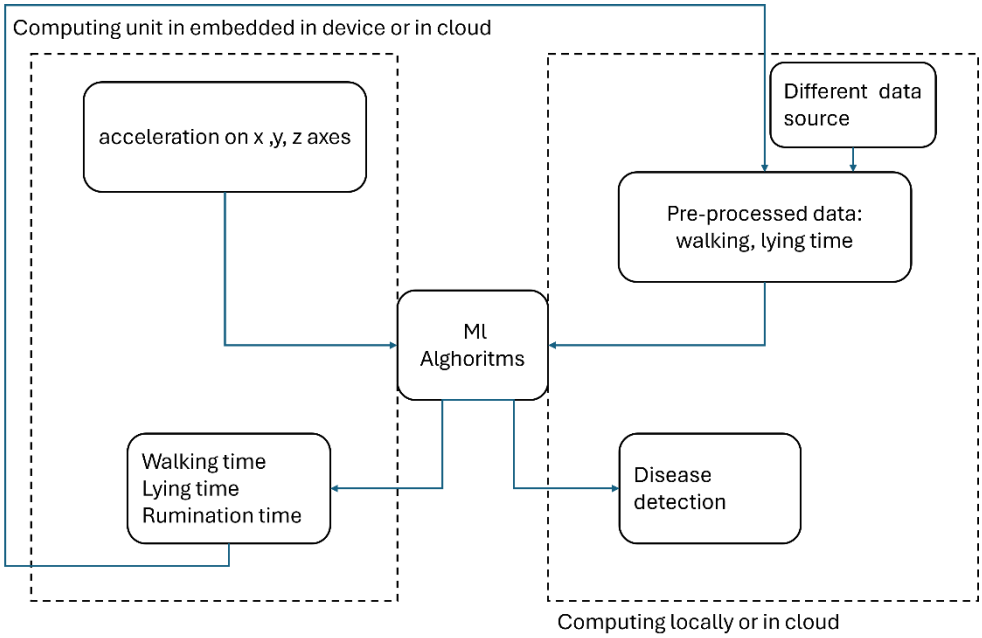

Fig 1.4. *ML algorithms can be employed at various stages: they can process raw data, analyze data from different sources, or make predictions.*

In the context of PLF, ML can be applied to process raw data from sensing units such as accelerometers or post-processed data like walking, lying, or rumination time to timely detect diseases such as lameness. Fig 1.4 represents an example of a possible ML workflow in PLF for accelerometer data and disease detection. Usually, devices such as collars or pedometers are equipped with computing systems in which raw data are processed with ML algorithms. Fig 1.5 shows an example of an acceleration trace of cows.

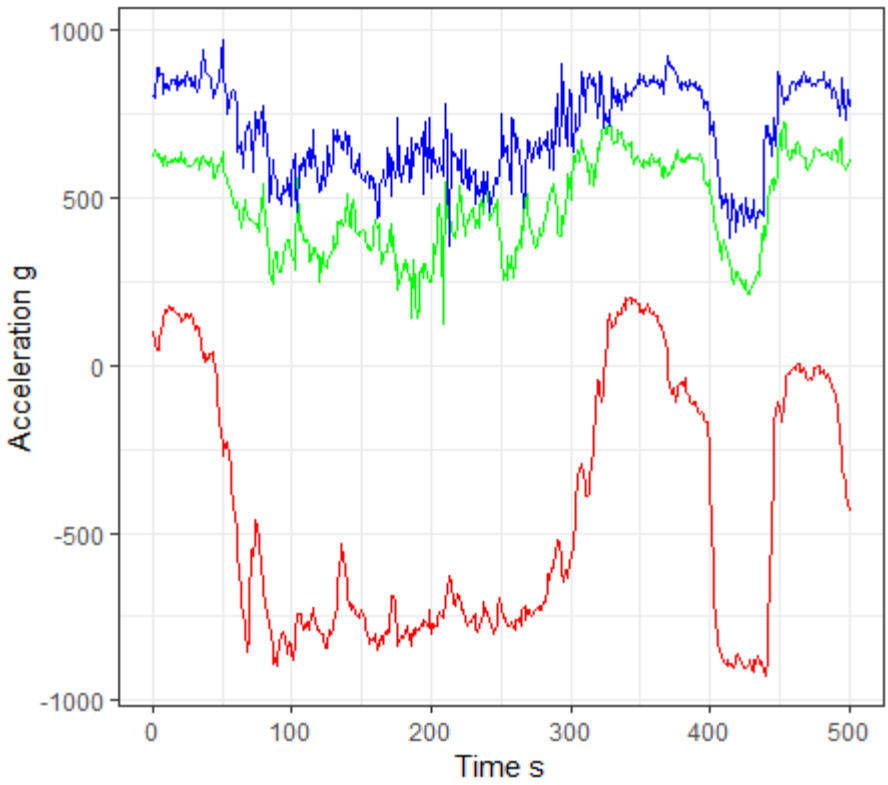

Fig 1.5 *Acceleration traces of the Aficollar over 500 s. The red line is the x-axis, the green y-axis and the blue z-axis.*

For example, Tamura et al. (2019) studied the performance of the decision tree to classify eating, rumination, and lying. They collect the data in 4 different farms (A, B, C, D) by a neck collar. Decision tree learning was trained on the data set of one farm resulting in a precision of 99.2% after cross-validation and 100% in test data sets from other farms (Tamura et al., 2019). The output of these operations is usually the time that the animal spent to perform some behaviours and with other kind of data represent the input for other ML algorithms that can timely recognize diseases such as lameness. In this context, Dervic et al. used data from different sources: environmental data (weather), genetic information, health data (veterinary assessments), data from AMS and sensors, next to information on farm management practices to detect lameness using random forest (RF). They achieve precision scores of up to 93% when predicting lameness (Dervić et al., 2024). In other examples, only data from herd or milk productions were employed to make predictions. Salamone et al. aimed to predict the milk

yield on the first test day, starting with the routine data of the previous lactations with a decision tree based alghoritms employed three different dataset highlighting the importance of days in milk, cumulative milk yield after 305 days and milk yield at the fifth and fourth test days are the features that achieved the highest importance score as fundamental predictor for milk yield on the first test day (Salamone et al., 2022).

The first part of the workflow is typically carried out in the computing system of a device or in the cloud, making it inaccessible to farmers and technicians. In contrast, information on behaviors such as time spent walking, lying, or other activities is more readily available and understandable to farmers. This is one of the reasons why an increasing number of studies in literature rely on this preprocessed data to make predictions. Furthermore, this workflow can be applied to a variety of other devices, such as smart boluses, ear tags, or automated milking systems (AMS), which are equipped with different sensing units

Main objectives

This thesis focuses on enhancing the management of dairy herds by analyzing various aspects of milk production and dairy farming. By focusing on key factors that influence production in dairy animals, this research seeks to improve efficiency and decision-making in the dairy sector by integrating advanced technologies and data-driven approaches. Indeed, despite technological innovations, quite a lot of data are still underutilized. The techniques employed in this thesis belong to the second block of the PLF workflow shown in Fig 1.1: the data analysis. Various ML techniques were employed. The ML was selected for its ability to handle complex, large-scale datasets and uncover patterns that can enhance decision-making in the dairy sector, particularly in areas such as production.

The first aim of the thesis is to perform a literature review on the application of PLF techniques in dairy farms, to understand is potential and limitations. Hence, a systematic review combined with text analysis was carried out to highlight and summarize the key trends in the literature. These methods allowed for a deeper exploration of the papers' structure and enabled clustering based on specific topics, providing a more efficient way to understand the research landscape and identify key areas of focus within PLF applications in the dairy sector. The results of this work are shown in chapter 1. Starting with the results of this systematic review it was possible to focus on one of the eight topics found: ML techniques.

The second aim of the thesis is to describe and report the results of the explorative analysis carried out with unsupervised machine learning in buffalo and goat breeding. These two species were analyzed. In this context, this study aims to fill this gap and provide preliminary results about milk production on minor dairy species. The main results are shown in chapter 2.

The third aim was to identify which lactation models best fit the lactation curves of cows and buffaloes. Specifically, various techniques were employed and compared, including classical methods like exponential equations and more advanced approaches such as deep neural networks, particularly autoencoders. Chapter 4 presents result that explore the potential of using neural networks to predict lactation curves, highlighting the strengths and limitations of these approaches compared to traditional methods.

Chapter 2: Precision Livestock Farming applied to the dairy sector

This chapter is the initial approach to PLF and lays the foundation for the subsequent analyses conducted in this work. It provides the necessary background and understanding of PLF, setting the stage for the more advanced investigations and applications presented in the further chapters. The final aim of this chapter is to explore the evolution of PLF literature over time and identify the key topics within it, aiming to clarify and categorize the various areas of interest related to this field, trying to underscore potentiality and existing knowledge gaps. A comprehensive search on the Scopus bibliometric database was carried out using various welfare-related keywords such as: “precision livestock farming AND dairy”, “sensors AND dairy” and “Machine learning AND dairy”. Based on the findings of this literature review, it is evident that PLF positively impacts the sustainability of livestock production. Addressing one area often influences others: for instance, improving animal health can reduce medical costs, enhance animal welfare, prevent production losses, or even boost production. This, in turn, enhances the environmental, economic, and social sustainability of dairy products. Thus, all these aspects are interrelated, with PLF being a key link between them.

2.1 Text Mining and Topic Analysis for In-Depth Study of precision livestock farming

Over the past century, the global dairy industry has undergone significant changes. At the start of the twentieth century, the shift of the general population from small rural villages to larger cities created a demand for mass-produced and distributed milk products. Since then, advancements in genetics, milking technology, nutrition, and farm management have transformed the industry into what it is today. These changes have increased average milk production per cow, a substantial rise in total annual milk output together with a dramatic reduction in the number of cows (Carpinelli et al., 2019). The PLF workflow involves a highly complex interaction between hardware (such as sensors and devices) and software (including algorithms, and cloud systems). Indeed, literature includes studies from different point of view, such as those focused on disease recognition or assessing the accuracy of tools in providing correct information, as well as technical papers that examine communication protocols or evaluate the performance of different architectural designs. In recent years, systematic reviews have gained significant attention across various disciplines. The availability of online databases (Google Scholar, Scopus and Web of Science), together with the development of bibliometric software and packages, has made it easier to obtain bibliometric datasets from the web. These tools allow for the calculation of statistics and indexes to create a comprehensive overview of a topic's current state, leaving the interpretation of results to the "analyzer" (the review's author). To gain a better understanding of how PLF in the dairy sector has been examined in scientific literature, this paper has employed Text Mining (TM) and Topic Analysis (TA) techniques. These analyses were done to evaluate the most frequent words, their associations, and the hidden relationships in this field.

Text mining is defined as "the knowledge discovery process that aims to identify and analyze useful information within large amounts of textual data, seeking to uncover new structures, patterns, associations, and trends (Nalon et al., 2021). This technology is a knowledge-driven process that utilizes analytical tools to extract meaningful insights from natural language text.

It identifies and uncovers significant patterns within unstructured text data, allowing for the discovery of valuable information from previously unknown textual content. Unlike traditional data mining, which focuses on structured data, text mining specifically targets unstructured data to reveal useful patterns and relationships (Holzinger et al., 2014). These techniques enable the analysis of textual data, extraction of main discussed topics and identification of associations between topics and words. The TM and TA have been applied in various areas, such as animal welfare (Trapanese, Petrocchi Jasinski, et al., 2024) and disease surveillance (Davies et al., 2024). This work delves into this field to understand the evolution of literature over time and identify key topics in this area, clarifying and labeling the different areas of interest for this complex phenomenon. Different algorithms can be used for TA such as Latent Dirichlet allocation (LDA). It is a generative probabilistic model designed for sets of discrete data, such as text corpora. LDA is structured as a three-tier hierarchical Bayesian model. In this model, each item within a collection is represented as a finite mixture over a hidden group of topics. Each topic, in turn, is represented as an infinite mixture over a set of topic probabilities (David M. Blei et al., 2003). In addition, by analyzing the statistical patterns of words in the text, topics that reflect the approximate meaning of the text are abstractly modeled. This abstract representation of the text data in the form of topics allows for dimensionality reduction, simplifying the complexity of the data (Xie et al., 2022).

The aim of the present work is to understand the evolution of literature over time and to identify key topics in this area, clarifying and labeling the different areas of interest also highlighting the strengths and limitations of PLF in the dairy sector.

2.2 Materials and methods

Dataset

A literature search protocol using Scopus® was set up to identify peer-reviewed scientific papers covering the topic of PLF in dairy herds. The search, carried out in April 2024, was refined based on the following criteria:

- Year of publication: from January 1976 to April 2024;
- Scientific area: topics in Veterinary, Agricultural, and Biological Sciences, Computer Science and Engineering;

- Article type: review articles, original articles, book chapters, technical notes and conference papers.

For the systematic review, nine strings were used as Fig 2.1 shows. The string that returned the maximum number of records was “Sensors and dairy” with 2179 papers while the minimum number was gathered from the string “PLF and calves” that returned only 202 papers. Records were collected in an electronic Excel workbook (Microsoft Excel®, v16.0) collecting title, authors, affiliations, abstracts, year of publication, type of record (e.g. article or review) and publication source (i.e. journal name or conference proceedings) for each paper.

After the collection phase, preprocessing was necessary to prepare the dataset for final elaboration. The first step was removing incomplete observations, such as those with missing abstracts (32 papers), authors (94 papers) and with no available abstract in English (190 papers). The second step was to remove duplicates (708 papers), as the same observation could appear in multiple search results. Finally, one researcher screened independently and carefully all the records in the dataset, selecting those eligible for final inclusion. Criteria of inclusion covered all activities and operations performed in cow and buffalo herds according to the definition of PLF furnished by Berckmans (Berckmans, 2014). For example, we included the behaviour analysis, the barn and the environmental monitoring and the comparison of conventional and automatic milking systems, while we excluded records related to different dairy species such as small ruminants, genetic evaluations or operations not performed in the barn such as milk quality analysis. The major results of the systematic scientific literature were schematically summarized in Fig 2.1.
At the end of the cleaning phase, the number of records included was 1794. Once obtained a complete dataset, the information on year, journal and country of publication were employed for the profile and the description of the dataset. The geographical affiliation for each row was determined based on the country of the corresponding author.

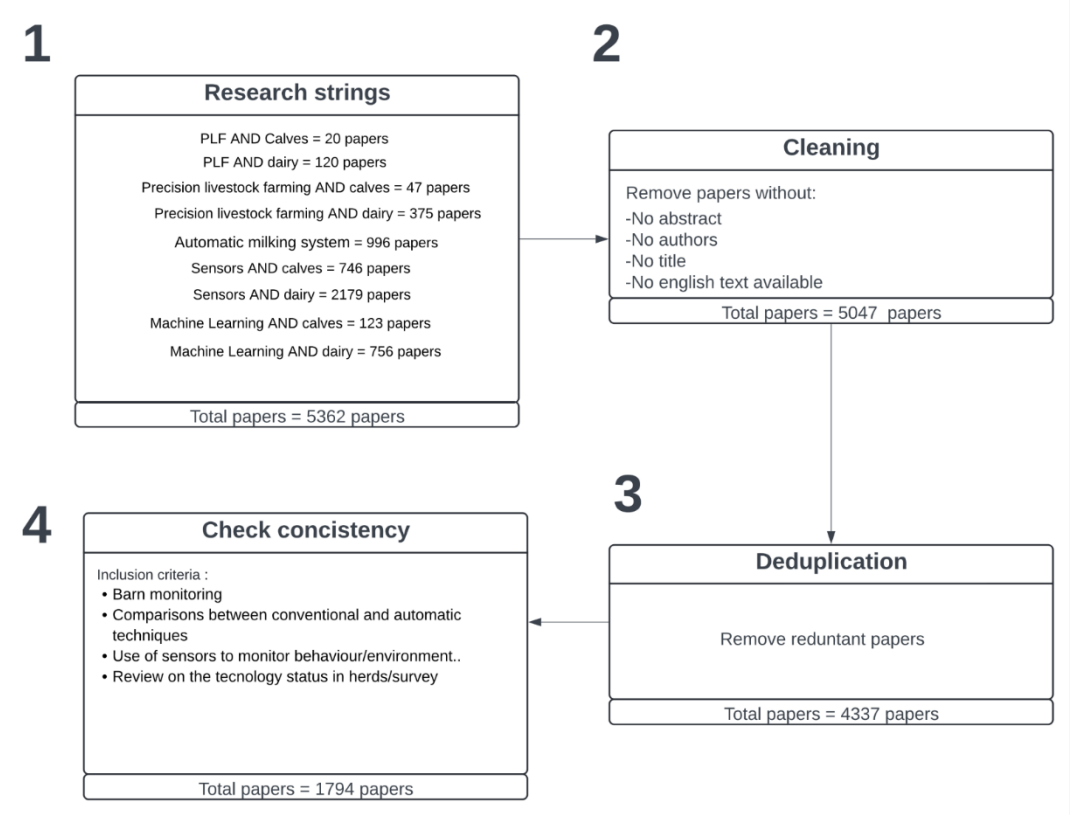

Fig 2.1 *Preprocessing of scientific literature on raw data was conducted to obtain the final dataset. The picture illustrates the details of the exclusion process. Examples of not-relevant papers were studies on calf muscle in human domain or genetic analysis*

Text mining (TM)

The abstracts of collected papers were compiled and organized into a separate Excel spreadsheet containing two columns: 'progressive ID' and 'abstract'. This spreadsheet was then prepared for TM analysis. All subsequent steps were performed in the RStudio environment (RStudio Team, 2022) using a combination of functions from the package's 'tm', 'snowballC', 'ggplot2', 'dplyr' and 'tidyverse'. Following Sebastiani's (2002) preprocessing guidelines, the corpus of records underwent several preprocessing steps. Due to the presence of both American and British English spellings, the authors standardized the corpus using American English exclusively. The preprocessing phases included tokenization, stop-word removal, and stemming-lemmatization. The goal of tokenization was to convert the text into meaningful parts called tokens (Uysal & Gunal, 2014).

Specifically, the tokenization process involved:

- Converting all words to lowercase;
- Replacing escape symbols and fonts (e.g., ‘@,’ ‘/’, ‘*’) with white spaces;
- Excluding certain characters such as punctuation, blanks, and numerical digits.

The removal of stop words aimed to exclude words that were either part of the research string or provided irrelevant information about the document's content, such as common articles and conjunctions. The following words were removed "machine", "learning", "calf", "automatic", "calves”, “sensors", "sensor", "milking", "aim", "group", "%", "/", "vs", "groups", "significative", "observations", "observation", "significant", "significantly", "system", "precision", "livestock", "farming", "PLF", "discussion", "background", "conclusion", "result". The final step in the preprocessing phase involved applying stemming and lemmatization algorithms (Matuszewski, 2023) . Both algorithms aimed to reduce words to their base or root forms, capturing the core meaning by removing suffixes or prefixes. The primary difference between the two is that stemming may produce non-words or partial words, whereas lemmatization returns words based on their dictionary forms (Matuszewski, 2023). To explore the text content deeply, a Document-Term Matrix (DTM) was constructed setting the minimum number of words to include equals three, with documents as rows and terms as columns, showing the frequency of each term in each document (Salton & Buckley, 1988). The Term Frequency–Inverse Document Frequency (TF-IDF) technique was applied to the DTM to assign distinct weights to each term. This weight was determined by the term’s frequency within a specific document and its prevalence across the entire collection of records. This approach ensured that each term's importance was based not only on its frequency within a document but also on its uniqueness and significance within the broader context of the entire dataset. A TF-IDF value threshold greater than or equal to 18 was used to build a histogram of the most frequent words. Additionally, a word cloud representation was made using the R package “wordcloud” setting a weight bigger than 1. The size of each word in the word cloud indicated its TF-IDF value, with larger sizes representing higher values, thus highlighting the most recurrent words. Associations among words with TF-IDF ≥ 18 were also explored. Only correlations ≥ 0.25 were considered. Associations were determined by measuring the frequency of co-occurrence between two words. An association value close to 1 indicated that the two words often appeared together within the documents, while values near 0 indicated limited co-occurrence.

Topic analysis (TA)

The topic analysis aims to explore and understand the document content and to extract meaningful information from unstructured text data. LDA was chosen as the method for performing topic analysis on the DTM. This probabilistic model seeks to uncover latent topics within a collection of documents, based on the assumption that there are K latent topics shared across the corpus (Srivastava & Sahami, 2009). The analysis utilized functions from the R package 'topicmodels', leveraging the Gibbs sampling options (Grün & Hornik, 2011). Since the number of topics is generally unknown, models with several different numbers of topics (6, 7, 8, 9 and 10) were fitted and evaluated. At the end, eight topics were chosen as they provided the most reasonable results. In particular, the authors noted that when choosing six or seven topics, different pieces of information were aggregated into a single topic, whereas with nine or ten topics, some topics did not provide essential information. To better visualize each topic and its most representative words, a bar histogram based on the beta values was created. Beta values represent the probability that words belong to a specific topic. Each topic was named according to the literature, with common consensus among the researchers. As a final step, a thorough investigation into significant trends was carried out. A pivot table was created to capture the frequency of each topic based on the number of papers per year. An exponential function was then applied to model the data, and the coefficient of determination (R^2) was computed. A p-value < 0.05 was considered statistically significant, indicating a meaningful trend in the data over time. Additionally, the most frequent country associated with each topic was identified.

2.3 Results

Descriptive statistics

The literature review identified 1794 peer-reviewed papers. Fig 2.2 A illustrates the temporal trend in the publication of these papers from 1976 to 2024, showing a significant exponential increase over time. Except for 1996, the number of papers published annually from 1976 to 2000 was consistently fewer than ten. However, from 2001 to 2023, there was a notable rise in annual publications, peaking in 2023. Research articles were the most common type of paper, comprising 67.7% (1215/1794) of the total, followed by conference papers (25.4%;

456/1794), reviews (5.3%; 95/1794), book chapters (1.4%; 26/1794), and others (0.2%; including 1 data paper and 2 note papers).

Fig 2.2 B shows the distribution of published papers by journal title, focusing on journals that published at least ten papers on the topic during the period considered. The most representative journals were the Journal of Dairy Science from the American Dairy Science Association, Computers and Electronics in Agriculture from Elsevier, and Animals from MDPI, with 250, 228, and 100 documents, respectively. Fig 2.2 C and D depict the geographical dissemination of the 1794 scientific papers based on the corresponding authors affiliation country. Europe was the most productive continent, accounting for 58.2% (1045/1794) of the papers, followed by Asia (20.2%; 363/1794) and North America (11.6%; 208/1794). A more detailed analysis at the country level revealed that within Europe, the Netherlands and Germany were the most prolific, each producing 156 documents. In the Americas, the United States led with 154 papers. China was the leading country in Asia with 119 papers. Algeria was the most productive African country with six papers. Lastly, Oceania had a substantial contribution, primarily from Australian researchers, who produced 79 papers.

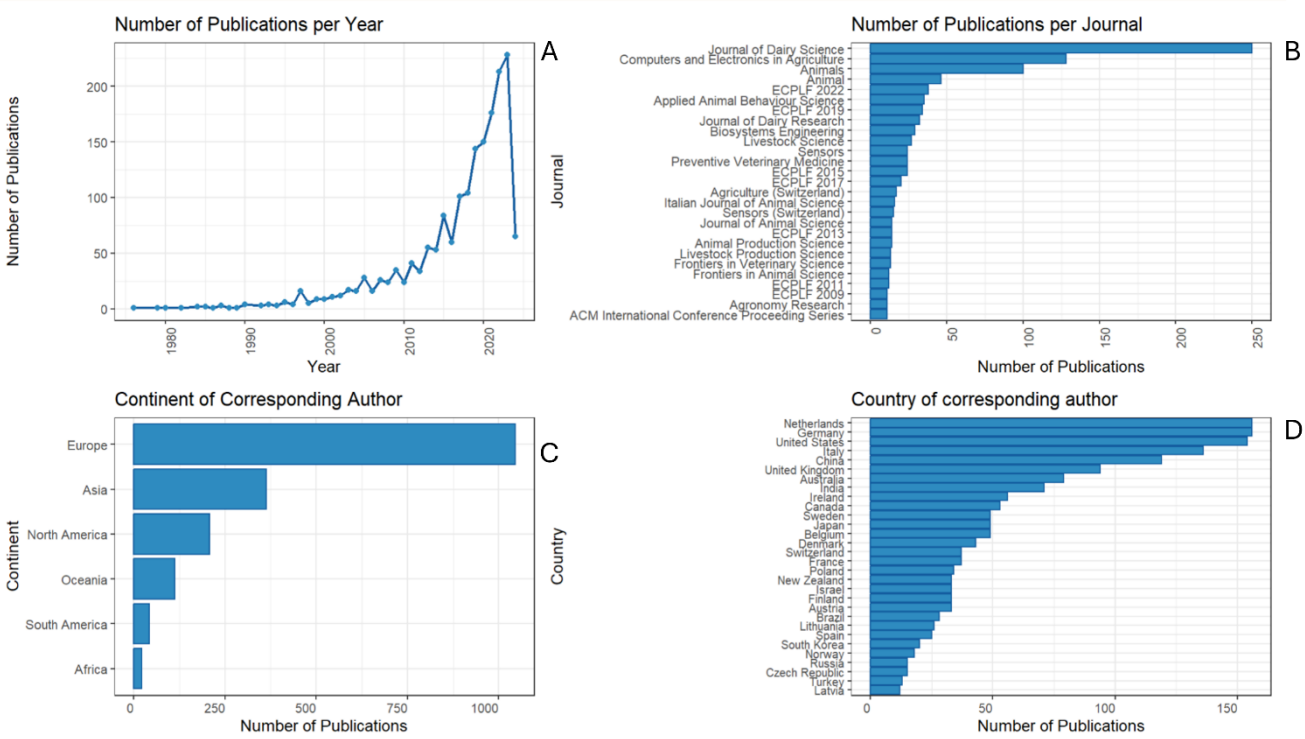

Fig 2.2 *shows the number of publications from 1976 to 2023. The * denoted that the number of papers in 2024 is referred to the period of papers collection that justify the low number. Figure 2B represents the number of papers (> 10 papers) for journals. Figure 2C shows the number of publications per country while Figure 2D the number of publications (> 3) per country of corresponding author.*

Text mining results

After the preprocessing of the data and reduction of sparseness (i.e. exclusion of the 'rare words'), 10003 terms were retained from the selected 1794 records. Fig 2.3 shows the first ten frequent words that showed a weight over 18 (TF-IDF >18). Only milk achieved a TD-IDF greater than 30 while the remaining terms were all below 20. The other words with the highest TF-IDF were "behaviour" (25.63), "model" (24.55), "feed" (22.34), "detect" (22.10), "farm" (21.69), "predict" (19.80), "anim" (19.62), "lame" (18.44) and "time" (18.44).

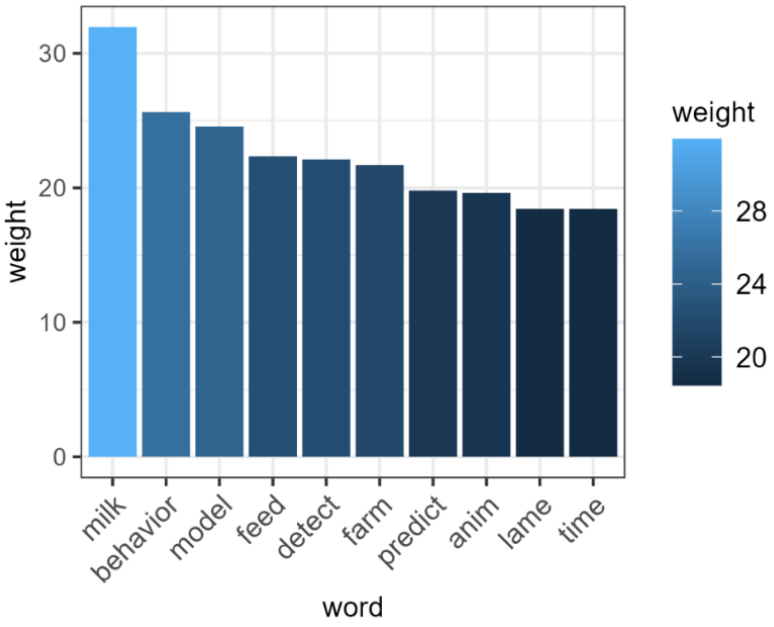

Fig 2.3 *Histogram of the first 10 words with a weight (TF-IDF >18) of 1794 documents selected for inclusion in the study*

A world cloud is shown in Fig 2.4. The size of each word mirrored its height, with the font size proportional to the TF- IDF of each word.

Fig 2.4 *Word cloud with the most frequent words*

Words correlations are presented in Tab 2.1. Only words with TD-IDF ≥ 18 that shown a correlation greater or equal to 0.25 were gathered.

Word	Association
anim	welfar (0.35), health (0.25)
Behaviour	lie (0.34), stand (0.30), acceleromet (0.26)
detect	estrus (0.35), mastiti (0.29)
farm	labor (0.32), cms (0.31),
feed	intak (0.40), ration (0.26)
lame	nonlam (0.55), locomote (0.35), gait (0.33), score (0.33)
milk	yield (0.54), fat (0.33), flow (0.30), protein (0.30), quarter (0.26), co mposit (0.25)
model	regress (0.28)
predict	random (0.27), forest (0.27), accuraci (0.26)
technolog	adopt (0.42), digit (0.26)
time	spent (0.37), day (0.26), eat (0.26)

Tab 2.1 *Correlation for the most frequent words (TF-IDF ≥ 18) and the other words present in the documents analysed.*

Topic analysis results

The topic analysis returned eight distinct topics. Coherent names were assigned based on the first 10 words and articles included in each topic. Fig 2.5 showed a histogram with the top 10 words for each topic. The most representative topic is Topic 7, “Survey”, with 287 records, closely followed by Topic 1, “Animal Welfare and Behaviour”, with 269 documents. A trend analysis was carried out to assess the tendencies of these topics from 1976 to 2023 and reported in Tab 2.2. Topic 6 had publications dating back to 1976, whereas Topics 2 and 5 are more recent, with their first papers published between 1992 and 1995. All topics exhibited a positive trend over time, with R^2 values near 1, indicating that the distribution of publications over the considered period fits an exponential function. European countries were the most productive

in terms of published papers. Specifically, Germany led in Topics 1, 2, and 3, with 27, 31 and 26 papers respectively, while the Netherlands had 36 papers for Topic 5.

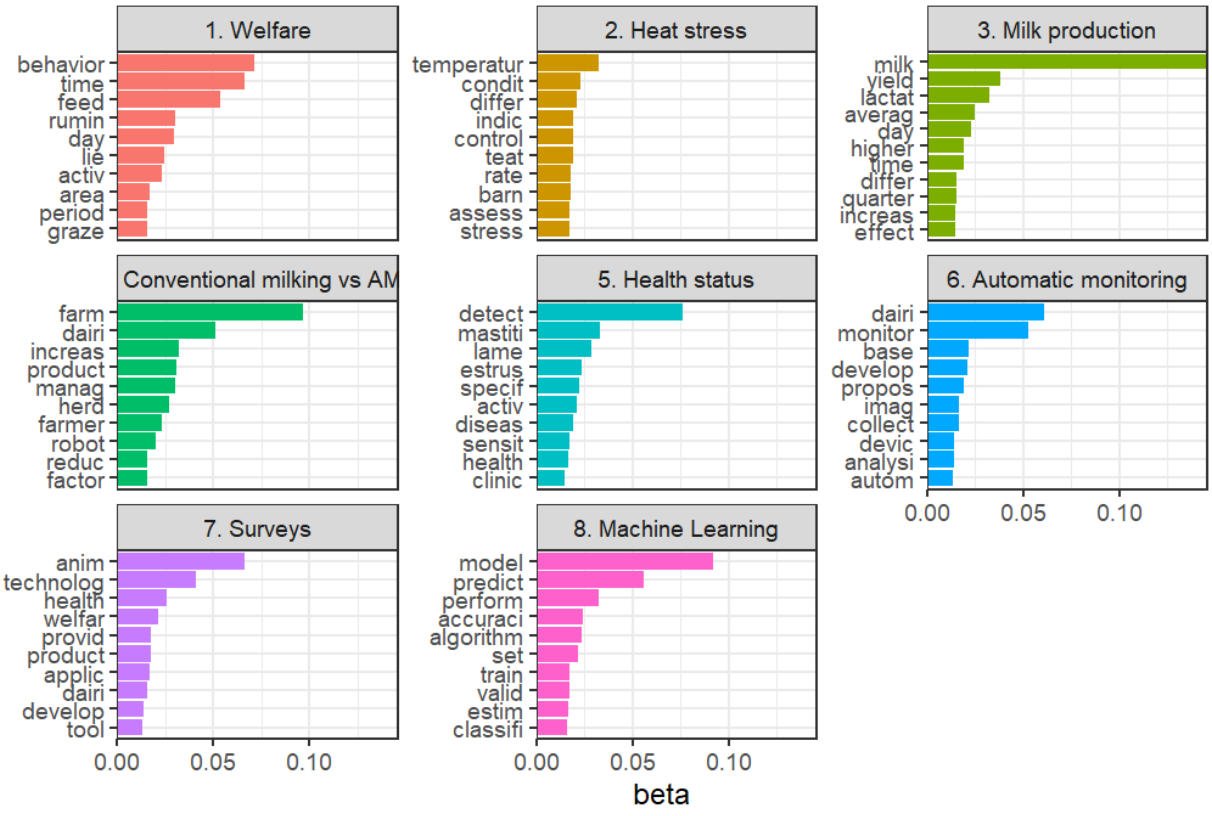

Fig 2.5 *Result of topic analysis. For each topic, the 10 most frequent words were shown.*

Topic	Number and % of publications per topic	R^2 and trend direction	P-value	Year of first publications	Country of most prevalence
1.Animal Welfare and Behaviour	268 (14.99%)	0.68 +	< 0.001	1995	Germany
2.Heat Stress	183 (10.20%)	0.68 +	<0.001	1987	Germany
3.Milk Production	252 (14.05%)	0.80 +	< 0.001	1979	Germany
4.Conventional Milking vs AMS	148 (8.25%)	0.57 +	< 0.001	1984	United States
5.Health Status	194 (10.93%)	0.70+	< 0.001	1992	Netherlands
6.Automatic Monitoring	252 (14.05%)	0.68+	< 0.001	1976	China
7. Surveys	287 (16.00%)	0.63+	< 0.001	1980	India
8.Machine Learning	210 (11.71%)	0.66+	< 0.001	1985	China

Tab 2.2 *Name of each topic, number of publications per topic, R2, p-value, country of most prevalence and year of the first publications. The signs (+ or -) near the R2 denoted a positive or negative trend. AMS = Automatic Milking Systems*

2.4 Discussion

This review aimed to analyse the literature on PL F in the dairy sector in the last 50 years with a TM and TA approach. The number of studies on this topic increased exponentially over the years and our results agreed with other studies on PLF in different farm species (Jiang et al., 2023; Marino et al., 2023). This positive trend is probably due to the growing need to supply more animal-based food, especially for developing countries (Henchion et al., 2017) without increasing the number of animals. Moreover, in recent years, there was also a growing consumer concern about the conditions in which cattle were raised, significantly influencing their food choices.

Starting from 1970s, PLF technology has evolved significantly, with individual electronic milk meters for cows to the creation of sensors to monitor estrus behaviour, rumination activity, and other related studies on breeding and animal management(Jiang et al., 2023). Research in the field of PLF was primarily concentrated in developed countries such as Europe, USA and China. In 2013 the European Union initiated the European Precision Livestock Farming Project (EU-PLF) aimed to translate PLF technology into industrial practice (*CORDIS - EU*, 2024). This project involved the participation of prominent universities such as Wageningen University and Research, Katholieke Universiteit Leuven, and the University of Milan, along with private companies. This could explain why in our work we found out that most of the researchers come from Europe. The USA is one of the countries with the highest degree of modernization in livestock development, especially for the use of milking robots as this study shows. Indeed, most of the papers in topic 4 belong to the USA. This reflects the demand of milking a large number of cows, since all the larger dairy corporations are situated in North America (Du et al., 2022).

From 2000 on, the number of papers on PLF in dairy herd grown quite exponentially over time. Additionally, computational units became more powerful and capable of processing the large amounts of data gathered from these devices. Moreover, the increase in artificial Intelligence techniques across various sectors, that occurred around 2000, also contributed to the widespread dissemination of knowledge (Rashid & Kausik, 2024).

Among the top five journals in terms of publication volume for PLF research, we found "Journal of Dairy Science", "Computers and Electronics in Agriculture", "Animals,"

“Animal”, and "ECPLF 2022", collectively accounting for approximately 30% of the analysed publications. These findings agreed with previous reviews (Jiang et al., 2023; Marino et al., 2023) performed globally on PLF. The journal subject categories belonged to different topics such as animal science, veterinary science, computer science, agricultural engineering and environmental science, highlighting the multidisciplinary nature of PLF.

According to our results, the first three most frequent words with the highest TF-IDF value were “milk”, “behaviour” and “model”. We expected “milk” as the most frequent word, since we aimed to explore the PLF applied to the dairy sector in large ruminants. In 2018, milk ranked third in production tonnage and was the second most valuable agricultural commodity globally, with a total production of 843 billion liters valued at USD 307 billion and in 2023 achieved 950 million tonnes (*FAO, Dairy Market Review*, 2023). The dairy sector is rapidly expanding, with global milk production expected to grow at a rate of 1.7% per year over the next decade. Dairy animals are raised in a variety of production systems, which offer flexibility and increased efficiency in addressing regional constraints but, at the same time, need specific PLF techniques to be applied, according to the need. In the 1794 papers included in our study the word ‘milk’ is often associated with different terms such as: “yield” (0.54), “fat” (0.33), “flow” (0.30), “protein” (0.30), “quarter” (0.26), “composit” (0.25).

The second term was ‘behaviour’, another important aspect to be monitored. This theme is also highlighted by the presence in our analysis of the word ‘animal’ which revealed strong associations with the words “welfare” (0.35) and “health” (0.25). The application of various technologies has been researched in dairy herds to generate effective, efficient, and animal-based indexes of wellbeing. Several devices, such as those for monitoring feeding and drinking Behaviour, diseases, and estrus detection, are integrated and multiple devices per animal were used to produce reliable indicators of animal welfare.

The third word was “model". The data analysis was an important part of PLF techniques because allows the processing of the huge amounts of data gathered by the sensors. The term “model” was often linked to “regress” (0.28): this association referred to the branch of ML techniques, able to predict trends and continues values. Algorithms were mostly used for milk yield composition and prediction or in barn monitoring.

Topic analysis allowed us to organize the research into eight different topics. Topic 1 is about Animal Welfare and Behaviour. There is a growing interest in this topic, initially driven by

ethical considerations and common empathy, and later influenced by economic and political interests that have developed around this issue to meet both public and ethical expectations (Lesimple, 2020). Historically, animal behaviour was documented by humans using focal scan sampling (Allueva Molina et al., 2023). Nowadays, an increasing number of sensors now generate detailed time-series data, often capturing several data points per second. These data are aggregated to minute or hour intervals either by the sensor itself or through software systems processing the sensor output. Aggregation has provided useful information on the time budgets of dairy cows. For example, we now know that a cows' daily time budget is affected by breed, parity, lactation stage and season (Maselyne et al., 2017; Munksgaard et al., 2020). However, cows showed daily variations in behaviour and physiology, so additional meaning is conveyed by determining when an event occurs (Casey & Plaut, 2022). This makes it worthwhile to investigate the 24-h pattern of behaviour, both for management purposes and for disease detection. Among the technologies made available by PLF for welfare monitoring there are devices (collar, limb, ear or even inserted into ruminal boluses) used to identify, in addition to estrus, also alterations in eating in ruminating and resting and activity time, or to identify the presence of diseases (Carpinelli et al., 2019; Ozella et al., 2022; Pahl et al., 2014). Body temperature and heart and respiratory rate can be monitored using specific sensors, such as thermometers and monitors. These devices not only provide information about the animal's health but also help detect stressful situations, often indicated by an increased heart or respiratory rate (Idris et al., 2021). Sound analysis is another technique used to monitor farm animals and to identify signs of illness or other health and well-being issues. Ad hoc programmable microphones or acoustic sensors can detect common symptoms of respiratory diseases, such as coughing and sneezing and analyse them using artificial intelligence algorithms. The characteristics of these sounds are related to the animals' emotional states, providing valuable information for assessing their emotions (Laurijs et al., 2021; Ren et al., 2021). This topic counted very few papers (34 out of 269) on young animals. This result is unsurprising, since animals in their unproductive phase are still less studied than the lactating ones, even from a behavioral point of view.

The second topic was about "Environmental related problems". Even if it started in 1987, it is one of the topics with few papers, probably because half of these were published after 2019. Livestock farming is a significant contributor to global greenhouse gas emissions and faces challenges from climate change, such as heat stress, land and water competition, and food security concerns (Cao et al., 2023; Fournel et al., 2017; Patra et al., 2017). Climate variables

like temperature and water availability can impact livestock production, health, and reproduction, while changes in crop cultivation, influenced by CO2 levels, can affect animal feed quality and methane emissions from ruminants (Cammarano et al., 2019). In this topic we can see two subcategories: one related to housing system and heat stress and another related to environmental problems and climate change. Regarding the first category, the barn environment, where animals spend their lives, plays a crucial role in their health and welfare. Monitoring systems like cameras, thermographic cameras, and accelerometers can help manage ventilation, shade, and temperature control, thereby enhancing animal behaviour and welfare. The barn's design influences internal conditions such as temperature, humidity, and solar radiation, which in turn affect gas concentrations and the temperature-humidity index (THI), a key indicator of animal welfare and health (Mayo et al., 2019). Monitoring these parameters helps assess the extent of heat stress experienced by cows. Fournel et al. (2017) reviewed predictive models for heat loss in livestock systems due to heat stress and suggested that PLF could enhance productivity and animal comfort by improving environmental controls in barns. Mattachini et al. (2019) demonstrated that optimizing barn conditions by efficiently using resources and analysing collected data can lead to improved production outcomes.

Regarding the second category (Wang et al., 2016) categorize the technologies available for measuring animal gas concentrations into three groups: (a) rapidly responding sensors that monitor concentrations continuously over time, such as electrochemical cells, chemiluminescence, fluorescence, etc; (b) cumulative concentration devices that provide only time-averaged values, like denuders, passive samplers, and adsorption bottles; and (c) instantaneous devices that offer snapshot measurements. Mélynda Hassouna et al. (2016) analysed the operating principles, advantages, limitations, and costs of these technologies. Very little data on emissions is available that could be used to characterize the diversity of livestock management systems (eg: type of production, type of livestock buildings, practices, soil and climatic conditions) and take into account of emissions in national genetics evaluations (e.g.: emissions inventories) (Mélynda Hassouna et al., 2016). However, a wide range of measurement methods are now associated with the various sources. Most of these are experimental, difficult to use routinely, and their applicability and reliability must be tested before they can be used on commercial farms. Moreover, the lack of standardized protocols makes it still difficult for these methods to be used in the agricultural sector. Although mechanistic models, such as process-based models, are commonly used for estimating emissions, their application can be compromised by complex environments in manure

storages. Genedy & Ogejo developed a deep-learning approach that combines process-based modelling and recurrent neural networks (RNN) as an alternative method to estimate ammonia loss from dairy manure during storage (Genedy & Ogejo, 2023).

Topics 3 and 4 were about milk production in different management and milking systems. These topics are quite old, starting from 1979 and 1984 respectively for topic 3 and 4. The issues discussed in both topics are mostly. Topic 3 focused more on factors that influence the milking efficiency of dairy cows in a modern production system. It considered aspects such as milking intervals, cow parity (primiparous vs. multiparous), milk production levels, and the effectiveness of technological interventions like automatic milking systems (AMS), emphasizing the impact of these elements on the overall success and efficiency of the milking process. (Ströbel et al., 2013) developed both the hardware and software for a teatcup prototype that monitored the individual quarter control system in a CMS to achieve low teat-end vacuum during low milk flow rates and higher vacuum at high flow rates.

Most of the papers of topic 4 described the advantages and disadvantages of AMS and Conventional milking systems (CMS) in different breeding conditions.(Clark et al., 2016) compared pasture utilization, as well as pre- and post-grazing pasture mass, between AMS and CMS farms managed under the same principles at the same site. The results demonstrated that high levels of pasture utilization can be achieved on pasture-based AMS infrastructure by following established grazing management principles. Rotz et al. (2003) studied the advantages of adopting an AMS from the economic point of view. The results showed that the net return decreased by $110 per cow when the economic lifespan of the AMS was shortened by 3 years, leading to faster depreciation compared to CMS. Hyde et al. (2003) analysed the benefits and the factors that justify the switch from a CMS to AMS. They found out that factors like farm capital structure, and depreciation methods have minimal effect on the investment decision. The key factor influencing the adoption of AMS technology is whether the AMS is expected to have a longer lifespan than the existing milking parlour. The introduction of robotics marked an important advancement in dairy production because of labour cost reduction, good working environment and worker health and less civil construction needs with increased milk yield with little effect on milk quality (Tse et al., 2018).

Topic 5 concerned papers focused on algorithms, devices or both used for the early diagnosis of health-related problems in the dairy sector. The most studied disease in our dataset was mastitis with about 60 papers (31%), the second was lameness with 46 papers (24%), the

reproductive issues such as anoestrus problems achieved 34 (18%) papers while ketosis collected very few papers. Automated diagnosis was also performed on calves, especially for bovine respiratory problems (BRD). Mastitis is one of the most widespread diseases in dairy cows and causes important economic loss and pain (Shen et al., 2015). For this reason, in the past five decades, different solutions were proposed for early recognition before clinical mastitis (CM). The first paper in our database was written in 1992. During the 90's, the first sensors used measured the electric conductibility (EC) of milk as an early detection of CM, due to easy usability and low cost. Nowadays, with the widespread of AMS these kinds of sensors are incorporated in the milking robot (Kunes et al., 2021). With this data, models based on time series analysis (Cavero et al., 2008) or other techniques has been explored (de Mol & Ouweltjes, 2001). For example, Kamphuis et al. (2010) studied the possibility to use a data mining technique to classify cows with CM, achieving good results (specificity 97% and sensitivity 47.4%). At the same time other more sophisticated models, belonging to the ML field, were employed such as neural networks, random forest and support vector machine. Together with the features that described milk yield and quality, also NIR technology in synergy with ML showed a good performance to detect bovine mastitis (Ramirez-Morales et al., 2021). More recently, the widespread adoption of embedded systems has enabled researchers to study the possibility of detecting CM through behaviour analysis. Nielen et al. (1995) assessed the relationship between reticuloruminal temperature, reticuloruminal pH, cow activity, and clinical mastitis showing that temperature, pH and cow activity changed before the diagnosis.

Lameness is a painful disease that causes important economic losses, reduced fertility and decreased slaughter weight and carcass value of culled cows (Flower & Weary, 2009). The first paper on lameness was published in 2003; this trend increased from 2010 until now. This is in line with the spread of commercial monitoring devices. Before the increase in adoption of PLF visual monitoring was used to detect and manage lameness. PLF introduced pressure sensors, accelerometers and video cameras as the core of monitoring systems. Maertens et al. (2011) developed a pressure-sensitive walkway (GaitWise system) that allows a gait analysis by employing an array of pressure sensors and spatiotemporal information. Nowadays, this technology has been surpassed, but it served as a good starting point. Devices based on accelerometers were used for classifying sick cows starting with the analysis of daily behaviours. For example, Van Hertem et al. (2013) explored the possibility of using a neck collar to obtain information on rumination and neck activity. The latter was combined with

feeding behaviour and cow performance to develop an informative logistic regression model (specificity 0.85 and sensitivity 0.89). Other studies explored the use of 2D or 3D cameras. In another approach, a piezoelectric sensor was used to analyse the footfall sound of cows to distinguish lame from non-lame animals (Volkmann et al., 2019). Paper that discussed on ML classification algorithms were widespread in the literature. Indeed, from our result "lame" was associated with "non lame" (0.55). From a data analysis perspective, the terms 'lame' and 'not lame' could be used in the dataset to indicate sick and healthy cows, respectively (Schönberger et al., 2023) . The word "gait" had also a strong association that denoted one of the input used for the lameness detection (Borderas et al., 2008).

Topic 6 collected technical papers that dealt with data acquisition systems, architectures of devices, communications protocols and algorithms to process data. These technologies were employed mostly for health (Arcidiacono et al., 2017), identification of animals (Mustafa et al., 2013) and feed especially for grazing cows (Chelotti et al., 2018).Among all monitoring systems, 3D cameras were the most present; indeed, more than 30% of the papers employed those systems, followed by neck collars and pedometers. Together with 3D cameras the most of algorithms discussed are convolutional neural networks (CNN). 3D-cameras and CNN were mostly employed for the automatically evaluation of body condition score (BCS) (Shigeta et al., 2018; Zhao et al., 2020) and to monitor the daily feed behaviour of cattle (Achour et al., 2020). Finally, regarding the communication protocols, Internet of Things (IoT) (Shi et al., 2023) and Radio Frequency Identification (RFID) were employed in identification tasks (Achour et al., 2022). Communication over long distances in grazing systems has employed Long Range Radio (LoRa) technology(Nyamuryekung'e et al., 2023). Time analysis motion systems and neck collars were the oldest with the first publications in 1976. Cameras and computer vision techniques and GPS were the youngest with the first publications in 2012.

Topic 7 collected reviews, surveys and technical reports on the state of the art, challenges and future prospectives of PLF in dairy production. It also explored the relationship between animals and humans and the perception of dairy producers of PLF technology. Reviews aimed to summarise different issues and topics on the ethical concerns to the introduction of technologies in dairy industries and the "shadow aspects" of the most employed ML algorithms and tasks in dairy (Driessen & Heutinck, 2015). The monitoring systems in livestock could be studied from different points of view. Vannieuwenborg et al. (2017) studied

the economic impact of different configurations of monitoring systems, considering the cost of devices and the communication architecture concluding that with the adoption these systems, it is possible to save money and improve the decision-making process of farms. On the other hand, the introduction of PLF is time-consuming and it can be very difficult for farmers to employ new tools. Social and ethical concerns regarding digitalization (Maroto Molina et al., 2020) and automatic milking (Millar et al., 2002) have been summarized. Simões Filho et al. (2020) reviewed the main limitations of AMS such as high investment costs, changes in milk composition (solids, free fatty acids) and increased risk of ketosis in cows. Other papers of topic 7 discussed the challenges and the difficulty of the farmers adopting PLF since they need to understand AI processes and interpret results. Considering this, Vyas et al.(2022) studied the negative effect that AI could have on the farmers and agriculture workers.

Topic 8 collected papers on ML algorithms and, specifically, collected technical papers on the implementation of different algorithms, the comparison of them, the dataset used, and the best practice of data preprocessing employed. From our analysis, ML techniques were employed especially in health recognition (Post et al., 2020) and prediction of milk yield (Pietersma et al., 2003). Most papers discussed the performance of adult dairy cows. Some reports studied the respiratory disease of calves (Hentzen & Holm, 2024). Another hot topic is the prediction with ML algorithms of the environmental conditions of barn and the emissions of greenhouse gas emissions (Cao et al., 2023; Ross et al., 2023). The most commonly used techniques were decision trees, regression trees, and artificial neural networks. These algorithms can be applied as standalone models or compared to determine the most suitable approach (Trapanese, et al., 2024). Regarding the features used to feed the ML algorithms, data from collars and pedometers were used for behaviour/welfare and disease recognition (Dutta et al., 2015). Features for the milk prediction task included herd and milk characteristics. Additionally, data from environmental sensors, diet, and milk composition were utilized to predict methane levels (Negussie et al., 2022).

This study has several limitations due to the methodology employed. For instance, synonymous terms were used in the initial search on the Scopus, and only the Elsevier Database was utilized, excluding others such as Mendeley, Google Scholar, and grey literature. This approach may have limited the number of papers included in our database. Additionally, the data was restricted to English-language papers with available abstracts.

Consequently, out of 5362 articles, we included only 1794, though we aimed for the most informative ones.

2.5 Conclusion

This review examined the literature on PLF in the dairy sector, identifying key research topics and areas needing more scientific evidence. From the findings of this literature review, it emerges that PLF could have positive effects on animal welfare, the lifestyle of farmers, and production efficiency. Intervening on one aspect involves influencing also other aspects: for example, improving health conditions allows reduces the costs for medicines, improves animal welfare, avoids losses or even increases production and consequently improves the environmental sustainability as well as the economic and the social sustainability of dairy products. Hence, all aspects are interconnected, and PLF represents the most promising connection. Finally, it must be noted that technological progress is constantly evolving and hence PLF is also subject to continuous improvements. As sustainability studies appear to be missing in this sector, Life Cycle Assessment (LCA), Life Cycle Cost (LCC) and Social Life Cycle Assessment (SLCA) should be carried out in the near future to analyse the application of PLF on farm and to identify the aspects on which further investigation and improvements should be introduced and to emphasise those in which the livestock sector is already optimal.

Chapter 3: Application of Machine Learning Algorithms to Dairy Routine Data

As Chapter 2 shows, the use of ML in livestock has the potential to positively impact the management of herds. Starting from these results, this chapter aims to employ unsupervised machine learning to herd data of rural species such as buffaloes and goats. These algorithms are designed to uncover hidden patterns within the data and group similar observations based on specific metrics, providing valuable insights into herd dynamics without labeled data. The results show that, according to the algorithms employed, dairy species can be split into two clusters according to their lactation stage and productivity. These findings lay a basis for further and more complex analysis to classify the most productive animals or to predict complications in dairy animal welfare.

3.1 Application of Cluster analysis Algorithms to Herd Data

Cluster analysis CA belongs to the category of unsupervised ML and aim to identify similar instances within the dataset and group them into distinct categories or clusters (Frades & Matthiesen, 2010). This approach is widely used in livestock sector to denote animals that have the same characteristics. Indeed, Tremblay et al. (2018) developed patterns associated with poor metabolic adaptation syndrome (PMAS) with unsupervised algorithms. In particular, the authors investigated the possibility of using the principal component analysis (PCA) combined with the kmeans algorithm to cluster cows affected by the PMAS. The five resulting clusters were then ascribed to a low, intermediate, or high PMAS class based on their agreement with expected PMAS indicators and characteristics, compared with other clusters. In addition, results revealed that PMAS classes were significantly associated with plasma No Esterified Fatty Acids (NEFA) levels. Moreover, the authors evaluated NEFA values that classified observations into appropriate PMAS classes as separation values. In another work, Franceschini et al. (2022) investigated the possibility of assessing the general health status of dairy cattle with large-scale milk recording data. The results showed four clusters close together and one furthest, more distant, in which animals with a higher somatic cell index were grouped. The hierarchical algorithm showed a good fit for exploring the data. Finally, Rebuli et al. (2023) applied kmeans, hierarchical and fuzzy algorithms to denote the most productivity cows using milk data. Supported by those successful applications, the chapter aim was to use those algorithms in dairy species that were less studied.

Indeed, up to now, all of these studies have been performed on dairy cows. However, despite dairy cows representing the largest portion of the milk market, other species, such as goats or buffaloes, are still important producers of milk and dairy products. In particular, the buffalo sector has grown over growing in the last years, especially in Mediterranean countries, Asia and South America, where buffalo farming is considered an important sector from an

economic and social point of view. Despite this increased interest in the buffalo sector, ML has fewer applications compared to the bovine sector (Bobbo et al., 2022). As well as the buffalo sector, also goats represent an important animal-based food resource for many countries. In the last 20 years milk goats have increased by 55% of the milk production 2020. The most productive continent in 2021 was Asia (59.7%) while Europe produced 15% of the total goat milk. Goat milk represents an important source of protein, calcium, niacin, pantothenic acid, phosphorus, potassium, riboflavin, thiamin, and vitamin A (S. Clark and Mora García, 2017). Moreover, the lower amount of αs1-casein led goat milk more suitable for children and allergenic people compared to cow milk (Agradi et al., 2021).

Due to the role of buffalo and goat milk on the market and in human nutrition it is important drive the new technologies to the buffalo and goat sector to align them with bovine sector. In light of this, a preliminary cluster analysis (CA) of the buffalo and goats herd data is proposed. In this work, an exploratory analysis has been carried out to find possible hidden patterns and natural clusters among buffaloes and goats.

3.2 CA on Buffalo data

Dataset description and preprocessing

The analysis was performed on a dataset containing 21 features for 1182915 observations, 379 farms and 96341 animals. The variables of the dataset concerned routine data such as milk yield, parity order, milk quality and reproductive information. Below, the main steps carried out for the ML analysis are described. The preprocessing phase included data cleaning, features aggregation, standardization, and dimensionality reduction. The missing values analysis showed that some data-type variables recorded the highest share of NaN and then where the columns were removed. Moreover, other numerical-type variables such as Urea and Casein were dropped since the presence of negative and inconsistent values. To have an overview of the buffalo performance the total days in lactation (Days lactation) and total milk yield at the end of lactation (Kg Lactation) were employed in the analysis. For the statistical analysis purpose, the somatic cells count (SCC) was transformed into somatic cell score (SCS) as follows (Ali & Shook, 1980).

$$SCS = log_2\left(\frac{scc}{10^5}\right) + 3$$

Equation 3.1 *Logarithmic transformation of Somatic Cell Count*

The feature aggregation aimed to extract as much information as possible from the dataset by the creation of new variables manipulating the original ones. In this work, calving conception interval (CCI) was added since the dataset did not have direct information on fertility. In addition, the animal age was added to the analysis. After these operations, the presence of outliers and measurement errors was assessed. In particular, the observations outside the 1st-99th percentile range were filtered to obtain consistent values. The clean dataset at this point had 10 numeric features 23261 observations, 282 farms and 8435 animals.

The features were arranged in the following manner:

- Four features described the quality of milk (Fat, Protein, Lactose and SCS);
- Five features described some functional aspects of lactation (Days in milk, Daily milk, Lactation days, Animal age, Kg lactation);
- One feature described information on fertility (CCI).

Fig 3.1 showed the structure of the clean dataset. The dataset contains *n* herds, each herd consists of m_n buffaloes. For each buffalo, there are $p_{m,n}$ observations that include multiple features, measured. To analyse the behaviour of the total phenomena, each observation was considered independent. The clean data were standardized with the z-score methods to obtain more efficient and accurate clustering results (Mohamad & Usman, 2013). The analyses were carried out on the entire dataset and on three subsets to validate the algorithms performance in different conditions. Each subset was made by the observations of three randomly selected herds. In this way, it was assessed if the results were coherent and repeatable regardless of the different breeding techniques and numerosity

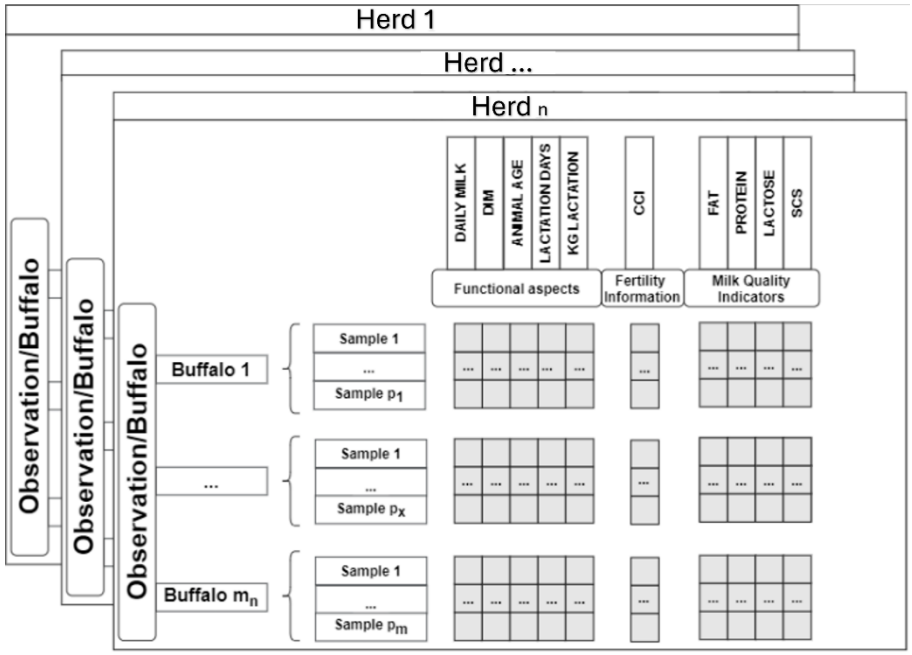

Fig 3.1 *The dataset is made up by n breeding containing m observation. For each breeding, 10 numeric features are available.*

Alghoritms

The first analysis carried out was the Hopkins statistics. It was a statically hypothesis that denotes if a set of data is randomly spread in experimental space or not (Banerjee & Dave, 2004). Then, the Principal Component Analysis (PCA) was performed before the CA to face the high dimensionality problem. PCA is introduced in 1901 by Pearson and the basic idea is to reduce the dimension of a dataset with several interconnected and correlated variables while preserving as much of the variation as possible, which characterizes the studied phenomenon (Greenacre et al., 2022). This technique achieves dimensionality reduction by identifying directions, called principal components, while maximizing the data variation. Starting from the space containing the original variables, a linear orthogonal transformation is performed. Technically, a roto-translation of the original axes is performed to form new axes, which are linear combinations of the original ones and are orthogonal to each other. The original variables are projected into the new Cartesian system, where they are ordered in decreasing variance: thus, the variable with the greatest variance is projected onto the first axis, the second onto the second axis, and so on. Finally, two unsupervised distance-based algorithms: kmeans and hierarchical, were compared in different configurations. The choice of these algorithms

was dictated by their successful applications in similar tasks in the bovine specie (Abreu et al., 2020; Brotzman et al., 2015; Tremblay et al., 2018). K-means is a partitional algorithm because it divides data into K clusters, assigning each observation to the only and one closest cluster. The algorithm is based on the concept of "similarity" between data points, using Euclidean distance as the metric (Ikotun et al., 2023). In detail, to determine which cluster a point should belong to, the distance between the observation and the cluster's centroid is calculated. The centroid represents the mean element of a data set, thus being the most representative element. To assign an observation to a cluster, the distance of that observation to each cluster's centroid is calculated, and it is assigned to the cluster with the nearest centroid. The hierarchical clustering, unlike partition-based clustering, aims to identify a hierarchy of clusters, which can be achieved through an agglomerative or divisive approach (Ran et al., 2023). Graphically, this forms a tree of nested clusters. The goal of this algorithm is to build a binary "merge" tree that allows visualization of the hierarchy of formed clusters. This tree structure is called a dendrogram. A dendrogram can be thought of as a tree where each node represents a cluster merging done by the algorithm, and the leaves are the n single elements to be clustered. The distance at which the cluster merges occur can be visualized on a two-dimensional plane, with the distance displayed on the y-axis. The levels in the dendrogram indicate the cluster hierarchy: clusters closer to the base are more similar, while those higher up are less similar (Ezugwu et al., 2022) . A critical point of unsupervised learning is the choice of an appropriate number of clusters which should be a priori known. Regarding kmeans, the average silhouette and elbow method were employed as suggested by literature (Tremblay et al., 2018). For hierarchical algorithms, the first operation was the choice of an appropriate linkage method. The Ward method was adopted in this work since better results were achieved regarding the cophenetic and agglomerative index comparison (Saraçli et al., 2013). In this case, the number of clusters was chosen as the most frequent suggestion by a set of indexes, in particular, the Hartigan, Scott, Trcovw, Friedman, Ball, and Ch ones (Charrad et al., 2014). After choosing the suitable number of clusters for each algorithm and test, the silhouette method and ρ indexes have been employed to determine the best algorithm. The average silhouette and intracluster correlation coefficient ρ indexes were considered to evaluate which algorithm performed better on the herd's routine data. The Silhouette index is a metric used to evaluate the quality of data partitioning into clusters (Shutaywi & Kachouie, 2021). It assesses the cohesion and separation of clusters obtained from a clustering algorithm.

Results

The results of the Hopkins statistic (0.98) denoted a strong tendency for data clustering. However, this result does not represent the true nature of the data. In particular, the observations occurred close together and were surrounded by isolated noise points. This condition increased the Hopkins statistic despite the absence of natural clusters (Adolfsson et al., 2019). The preliminary results showed that the PCA did not achieve a relevant dimensionality reduction since a low correlation was exhibited between the variables (as shown in the correlation matrix in Fig 3.2). This condition meant that several variables were needed to explain the dataset variance. In particular, according to the cumulative proportion of variance, 8 out of 10 features explained the 92% of variance. The latter result showed the poor performance of the technique, so the PCA algorithm was not employed in the final analysis.

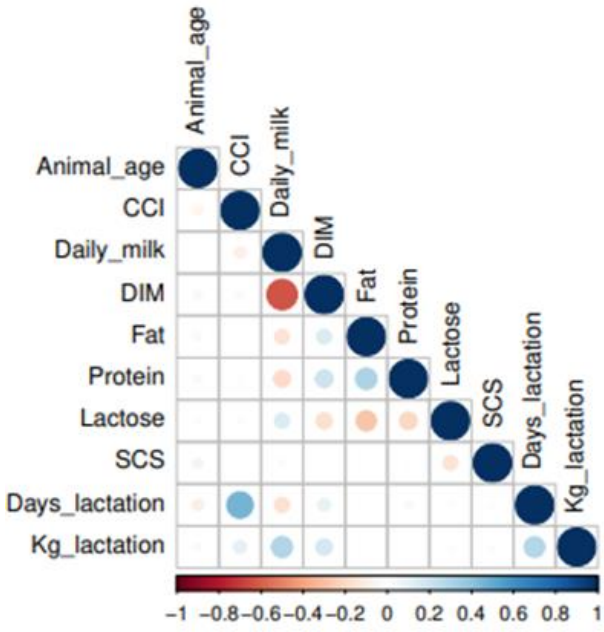

Fig 3.2 *Pearson's correlation matrix. The size of the circle showed a pairwise correlation. The color blue represented a positive correlation, and red a negative one. The functional aspect of milk is correlated with each other as opposed to the other feature categories. CCI: Calving conception interval; DIM: Days in milk; SCS: Somatic cells score.*

The performance results for kmeans and hierarchical algorithms in different conditions are reported in Fig 3.2.

Algorithm	Breeding	Cluster Number	Avg. silhouette	ρ
Kmeans	All	2	0.16	0.20
	First	2	0.20	0.40
	Second	2	0.16	0.35
	Third	2	0.15	0.34
Hierarchical	All	3	0.10	0.29
	First	2	0.11	0.33
	Second	3	0.09	0.30
	Third	7	0.11	0.40

Tab 3.1 *Comparison of kmeans and hierarchical algorithms.*

In general, the results showed low and moderate values of average silhouette and ρ indexes, indicating low separation between the clusters. The values of the ρ coefficient indicated that the true distances between observations are not completely covered by the clustering algorithms. The kmeans performed better than hierarchical, especially in average silhouette. Due to those results, the kmeans was considered the suitable unsupervised algorithm for this work.

Below are shown the kmeans results in terms of cluster content. For all trial, kmeans was performed with the number of clusters set to 2, as the indexes suggested. In Tab 3.2, Tab 3.3, Tab 3.4 and Tab 3.5 summarized each cluster variable's mean and standard deviations among the 2 clusters in all data across different herds. Tab 3.2 shows the results of the first trial carried out on all the data with 8435 animals, 23261 observations and 282 farms. The first cluster had 10690 observations (46%), while the second cluster was the largest with 12571 observations (54%). Tab 3.3 shows the results of the second trial carried out on a herd of 263 animals. This clustering analysis was performed on 887 observations. The first cluster had 398 (44%)

observations, while the second had 489 (66%) observations. Tab 3.4 reported the results of the herd level carried out on a herd of 121 animals. This clustering analysis was performed on 316 observations. The first cluster had 185 (58%), while the second had 131 (42%) observations.

Variable	**Cluster 1 (46% obs.)**	**Cluster 2 (54% obs.)**
Animal age (year)	6.24 ± 2.90	6.54 ± 2.95
CCI (days)	161.47 ± 76.73	159 ± 76.46
Daily Milk (kg)	10.40 ± 1.87	7.61 ± 1.42
DIM (days)	122.03 ± 32.51	210.72 ± 34.06
Fat (%)	8.45 ± 0.91	8.92 ± 0.93
Protein (%)	4.72 ± 0.22	4.86 ± 0.25
Lactose (%)	4.87 ± 0.13	4.80 ± 0.12
SCS (unit)	3.30 ± 1.17	3.40 ± 1.16
Days lactation	299.71 ± 34.30	313.02 ± 36.72
Kg lactation	27649.87 ± 3875.57	27717.40 ± 3857.36

Tab 3.2 *Description of input variables expressed as mean and ± sd for each cluster (all breedings - first trial)*

Variable	**Cluster 1 (44% obs.)**	**Cluster 2 (66% obs.)**
Animal age (year)	6.15± 2.59	5.93 ± 2.60
CCI (days)	138.83 ± 55.29	136.47 ± 52.70
Daily Milk (kg)	11.12 ± 1.74	7.74 ± 1.49
DIM (days)	127.89 ± 32.75	207.25 ± 33.46
Fat (%)	8.30 ± 0.77	9.09 ± 0.79
Protein (%)	4.69 ± 0.19	4.89 ± 0.23
Lactose (%)	4.86 ± 0.12	4.76 ± 0.10
SCS (unit)	3.65 ± 1.41	3.88 ± 1.14
Days lactation	292.29 ± 30.45	291.93 ± 32.27
Kg lactation	28820.50 ± 3913.34	28654.20 ± 3837.52

Tab 3.3 *Description of input variables expressed as mean and ± sd for each cluster (breeding 1 - second trial)*

Variable	Cluster 1 (58% obs.)	Cluster 2 (42% obs.)
Animal age (year)	6.92 ± 2.02	7.27 ± 2.22
CCI (days)	171.55 ± 79.17	173.59 ± 80.20
Daily Milk (kg)	7.42 ± 1.31	9.52 ± 1.74
DIM (days)	227.24 ± 30.21	153.27 ± 41.05
Fat (%)	8.58 ± 0.90	8.16 ± 0.83
Protein (%)	4.87 ± 0.24	4.62 ± 0.17
Lactose (%)	4.79 ± 0.12	4.92 ± 0.11
SCS (unit)	4.51 ± 1.38	5.02 ± 1.26
Days lactation	328.02 ± 37.18	323.17 ± 35.15
Kg lactation	28237.97 ± 3967.25	27928.50 ± 3938.88

Tab 3.4 *Description of input variables expressed as mean and ± sd for each cluster (breeding 2 - third trial)*

Variable	Cluster 1 (45% obs.)	Cluster 2 (55% obs.)
Animal age (year)	5.70 ± 2.64	5.72 ± 2.93
CCI (days)	161.71 ± 67.24	159.12 ± 64.50
Daily Milk (kg)	7.97 ± 1.61	10.92 ± 1.74
DIM (days)	190.31 ± 36.65	117.83 ± 32.80
Fat (%)	9.51 ± 0.79	8.59 ± 0.88
Protein (%)	4.89 ± 0.21	4.67 ± 0.16
Lactose (%)	4.75 ± 0.10	4.88 ± 0.11
SCS (unit)	4.30 ± 1.01	4.05 ± 1.17
Days lactation	311.24 ± 35.76	308.24 ± 35.67
Kg lactation	27184.20 ± 3589.03	27429.47 ± 3722.68

Tab 3.5 *Description of input variables expressed as mean and ± sd for each cluster (breeding 3 - fourth trial)*

Tab 3.5 reports the results of the fourth trial carried out on a herd of 188 animals. This clustering analysis was performed on 562 observations. The first cluster had 56 (45%) observations, while the second had 66 (55%) observations

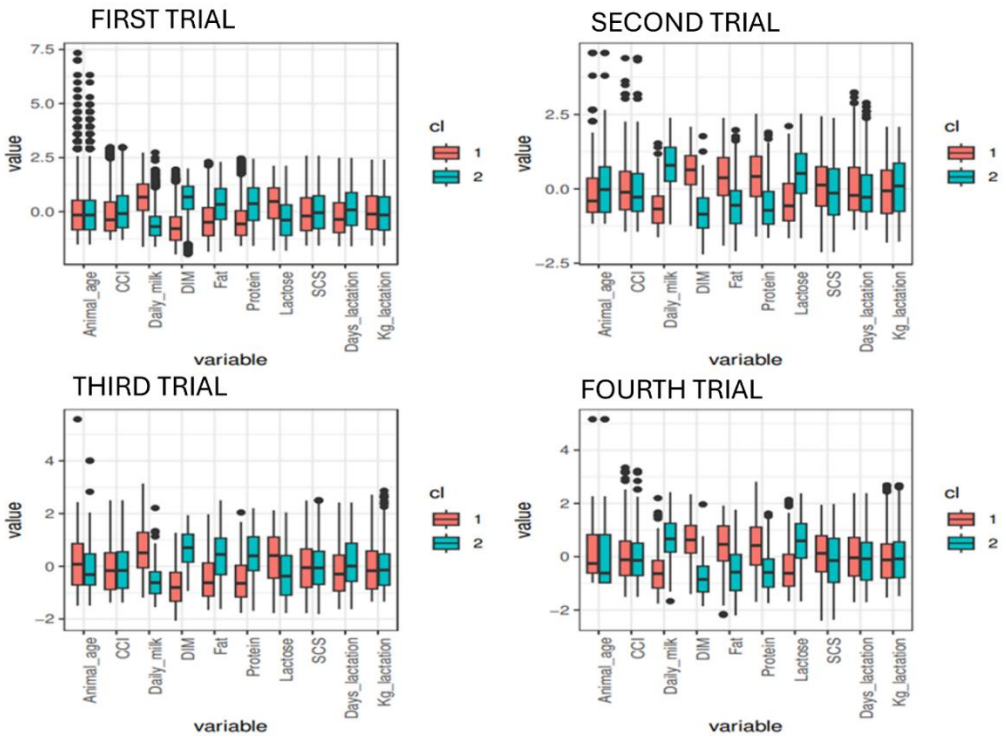

Fig 3.3 *Boxplot of the four different trials. On the x-axis there are the input variables, while on the y-axis, the scaled values that each variable has among two clusters correlated with each other as opposed to the other feature categories. CCI: Calving conception interval; DIM: Days in milk; SCS: Somatic cells score.*

The boxplots in Fig 3.3 shows the distribution of each variable among the two clusters for each trial. Scaled data were employed to clearly visualize the variables. The algorithm returned similar results for each trial. According to Brotzaman et al. (2015) the cluster content exploration was performed. In the first and third trials, cluster 1 grouped all the observations that showed the highest values of daily milk and lactose production. The instances with lower days in milk (DIM), fat and protein values were included in the same cluster. Cluster 1 also grouped the lowest values of CCI, days lactation and SCS. Regarding animal age and kg lactation, cluster 1 grouped the observations with the highest values (this is especially visible for trial three). In contrast, observations that recorded the lowest values of daily milk and lactose content were included in cluster 2. Observations showing the lowest DIM, fat and protein values were also associated with cluster 2. Similarly, the highest values of CCI, days lactation, and SCS were associated with cluster 2, as were the lowest values of animal age and kg lactation. For trials two and four, the assignment of clusters 1 and 2 changes and is reversed. Finally, kmeans, as expected, failed to denote the noise point (outliers).

3.3 CA on Goats data

Dataset and preprocessing

This research was carried out over two years from June 2020 to June 2022 in eight different herds located in province of Potenza and Matera in the south of Italy. The data were provided by regional breeders' association of Basilicata (Potenza, Italy). The original dataset concerned 7836 Test Day (TD) were collected monthly from 1128 goats. Data included general information of the animals (such as ID, parity, number of lactations, breed, and days in milk). Milk yield, fat, protein, lactose, Somatic Cell Count (SCC) were gathered for every lactation. Milk yield, fat, protein, and lactose were predicted using infrared spectroscopy at the milk laboratory of the Breeders Association of Basilicata region (Potenza,Italy) with a MilkoScan FT6000 (Foss Electric A/S, Hillerød, Denmark). To compare the milk yield between the breeds, fat corrected milk at 3.5 % (FCM) was computed according to Currò et al (2019).

$$FCM = Milk(kg) + 0.1046 * Fat(\%)$$

Equation 3.2 *Fat correct milk equation (Currò et al., 2019)*

Somatic cell count (SCC; cells/mL) was determined using Fossomatic FC (Foss Electric, Hillerød, Denmark) and transformed to SCS for statistical purpose through the formula seen in the paragraph 3.2. Finaly, the fat protein ratio was computed as index of negative balance with the formula:

$$FP = Fat/protein$$

Equation 3.3 *Fat protein ratio*

The editing criteria included restriction on breed (Saanen and Alpine), number of test day (TD >1), parity (first to fifth) and removing of all zero and missing values. At the end, the preprocessed dataset counted 6332 observations from 975 goats (n=579 for Saanen and n=396 for Camosciata) reared in eight herds in Southern Italy. The available data contains 19 numeric variables referring to the categories of 1) milk yield, 2) milk quality and 3) milk coagulation.

A graphic summary of the dataset employed is shown in Fig 3.4. The analysis was performed in R version 4.1.2.

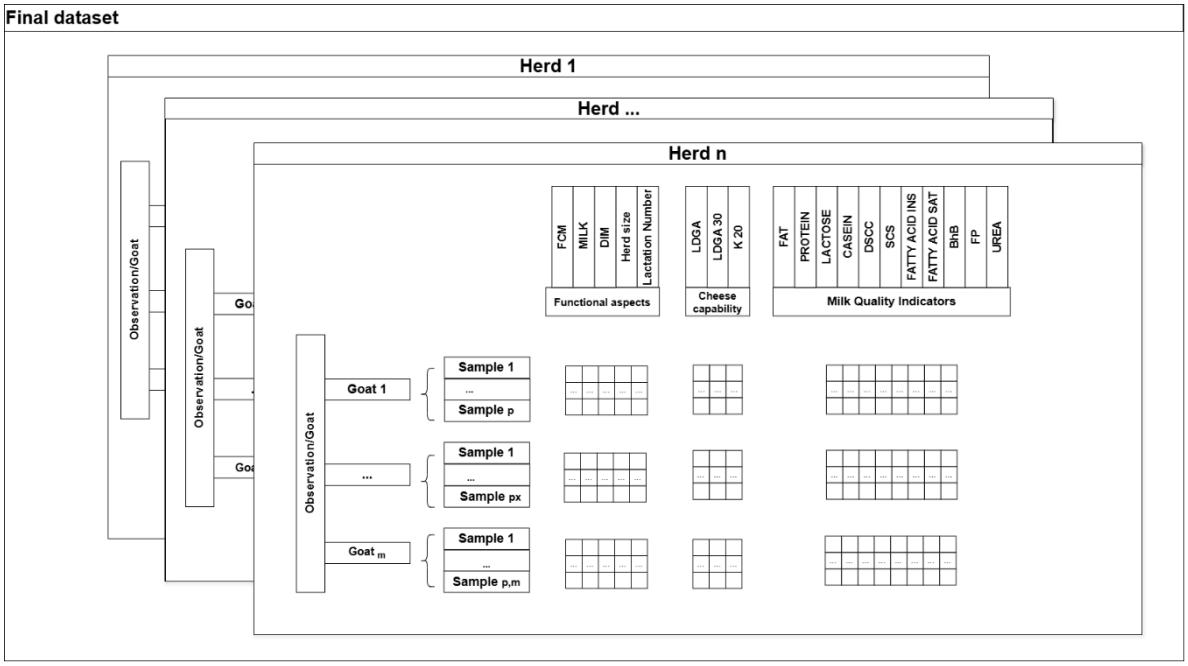

Fig 3.4 *The dataset is made up by n herds containing m observation. For each herd, 19 numeric features are available.*

Algorithms

Similar to buffalo data partitional (Kmeans) and agglomerative algorithms (Hierarchical with the Ward linkage method were employed and compared). Before performing any clustering algorithms, PCA was applied to reduce the dimensionality of the dataset. All techniques were performed on three datasets: 1) both breeds (complete dataset); 2) Saanen subset; 3) Camosciata subset. Then, the silhouette analysis was applied to compare the performance of the two algorithms and to select the best one.

Results

The PCA showed that 94% of the explained variance was captured by ten dimensions for all trials. Tab 3.1 shows the performance of PCA in terms of variance explained by the first two dimensions, total variance expressed and number of dimensions according to the Kaiser rule.

The contribution of each variable for the first two dimensions shows similar results for each trial. Fig 3.5 shows the % contribution of each variable to the first two dimensions for the general dataset. Dimension 1 was mostly correlated with total, unsaturated and saturated fat in milk, while Dimension 2 with lactose and milk coagulation variables.

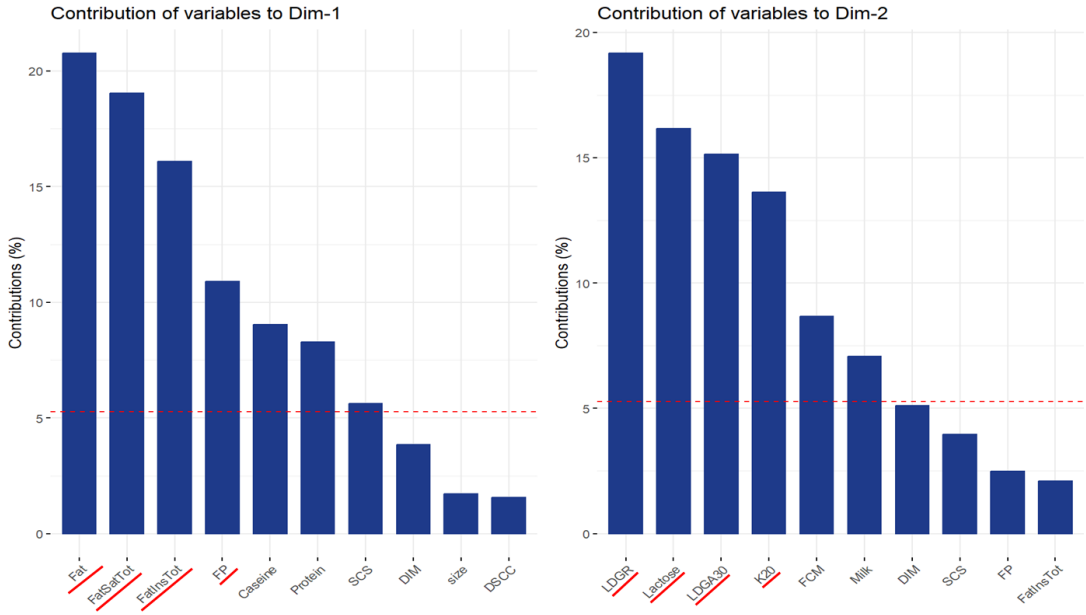

Fig 3.5 *Contribution of each variable to dimensions 1 and 2.*

According to the silhouette analysis, Kmeans performed slightly better than Hierarchical algorithms in each trial. Indeed, for the complete dataset and the Saanen subset, the partitional method gained the highest values (0.14 and 0.17, respectively) with two clusters, while for the Camosciata subset, the silhouette analysis returned the maximum value (0.14) with four clusters. The agglomerative method gained the best values with four clusters (0.12 and 0.11, respectively, for complete and Saanen data). For the Camosciata subset, the Hierarchical algorithm returned two clusters with a silhouette value of 0.11. Tab 3.6 summarizes the results.

Algorithm	Dataset	Cluster Number	Silhouette
Kmeans	All goats	2	0.15
Kmeans	Saanen	2	0.17
Kmeans	Camosciata	4	0.14
Hierarchical	All goats	2	0.12
Hierarchical	Saanen	4	0.11
Hierarchical	Camosciata	2	0.11

Tab 3.6 *Comparison of kmeans and hierarchical algorithms.*

Regarding the cluster content analysis, kmeans on all data set and Camosciata breed returned the similar results: the division happened according to the lactation stage. This result was very close to the result obtained in buffalo data. For Saanen breed kmeans return 4 clusters showing a slightly different behaviour. Indeed, despite animals being split according to the lactation stage, one of the two clusters that grouped the animals at the begging of lactation showed very high fat, SCS content in milk compared to the other groups. Fig 3.6 shows a boxplot of the distribution of DIM, Fat, DSCC (Diffential Somatic Cell Count), SCS , FatInsTot (Insatured Fat Acids) and FatSatTot (Satured Fat Acids) among the four clusters. Scaled data were employed to clearly visualize the variables

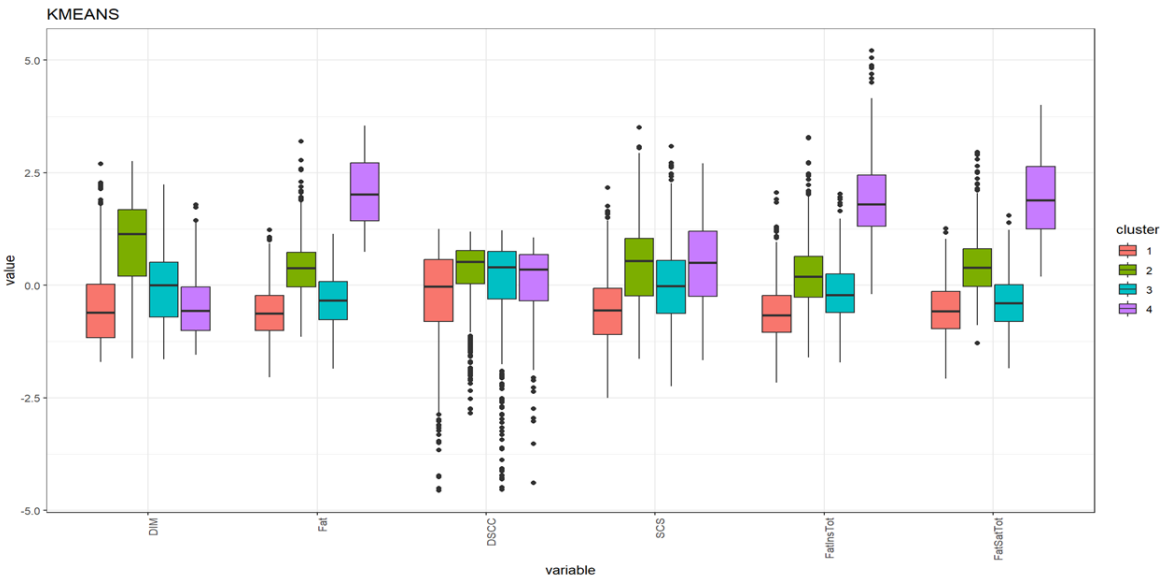

Fig 3.6 *Boxplot of the four different clusters. On the x-axis there are the input variables, while on the y-axis, the scaled values that each variable has among two clusters correlated with each other as opposed to the other feature categories.*

3.4 General discussion and conclusion

This study applied different clustering algorithms to analyze herd data from buffaloes and goats- We aimed to reduce the dimensionality of the dataset with PCA yielded limited benefits due to the high number of variables required to capture the total variance, especially in buffalo data. In the clustering analysis, k-means consistently outperformed hierarchical clustering in terms of silhouette scores, making it the preferred method for both species. For buffaloes, the clustering was first performed across all herds (first trial) and then repeated on three randomly selected individual farms (second, third, and fourth trials) to assess the robustness of the technique under different conditions. A similar approach was followed for goats, where the data was split into three datasets based on breed. The results for buffaloes and goats were comparable. In both species, Principal Component Analysis (PCA) yielded poor results in terms of the number of dimensions needed to explain the total variance. However, in goats, the analysis of the contributions for each dimension was interesting, especially for the Camosciata breed, where milk cheese-making ability and quality made significant contributions. In terms of clustering performance, k-means consistently outperformed hierarchical clustering for both species, as indicated by the higher average silhouette scores. The k-means algorithm produced similar results across all tests, with the behaviour of the input variables remaining consistent. For buffaloes, in the first and third trials, Cluster 1 grouped observations with the lowest days in milk (DIM). In these cases, k-means separated the data into two distinct lactation periods. Cluster 1 that displayed higher daily milk yield compared to Cluster 2, which grouped buffaloes in later lactation stages. These findings align with the typical lactation curve, where milk production is higher in the early months (Catillo et al., 2002). Cluster 1 also exhibited higher lactose levels, consistent with the osmotic regulatory function of lactose (Henao-Velásquez et al., 2014). This trend corresponds with the negative correlation between lactose and somatic cell score (SCS) reported in the literature (Alessio et al., 2021) In the early lactation period, lower SCS values and higher lactose levels were observed. Additionally, the influence of the lactation period on SCS, as reported by Afiani et al. was confirmed by the clustering algorithm (Afiani et al., 2021). Lower fat and protein levels were also noted in Cluster 1, reflecting the typical behaviour of these components during lactation (Silvestre et al., 2009), which was also observed in the third trial. In the second and fourth trials, Cluster 1 grouped observations with higher DIM, fat, protein, and SCS, but lower daily milk and lactose. While the behaviour of the variables remained consistent, the

assignment of clusters changed across the trial s. For goats, the lactation curve followed a similar pattern to that of other dairy animals (Currò,et al., 2019). As with buffaloes, k-means split the goats (and the Saanen breed dataset) into two clusters based on lactation stage. However, in the Saanen dataset, k-means identified four clusters. Two of these clusters grouped animals in early lactation, but one exhibited higher fatty acid content and somatic cell counts (SCC). This finding aligns with the work of Šlyžius et al., who noted that goats with higher SCC in their milk also tended to have higher fatty acid content (Šlyžius et al., 2023). Although k-means produced coherent results for DIM, daily milk, lactose, protein, fat, and SCS, these findings are preliminary, and further investigation is needed to explore the relationships between all variables. Given the strong dependence of milk production on DIM, a future approach could involve splitting animals by lactation stage using the output of k-means and conducting exploratory analysis within each group. A similar analysis could also be performed on herds, calculating aggregate statistics such as mean, minimum, and maximum for each variable, and employing cluster analysis (CA) to identify similarities among herds. Lastly, clustering time series data could be a useful approach. However, in this study, applying this method was not feasible due to the insufficient number of observations available within each lactation.

Finally, the application of ML algorithms like k-means in PLF has shown significant potential for optimizing dairy herd management, particularly in identifying the optimal timing to group buffaloes based on their lactation stage and physiological needs. This study serves as a foundational step toward more sophisticated analyses that incorporate advanced ML methods and improved data quality, ultimately supporting enhanced productivity and welfare in the livestock sector.

Chapter 4: Comparative Assessment of Lactation Models for Evaluating Herd Productivity: Classical model and Machine Learning techniques

Chapter 4 focuses on a detailed study and exploration of lactation curves in buffaloes and cows. Lactation models play a crucial role in evaluating both herd productivity and individual dairy animals' performance. Over time, numerous lactation models have been developed to analyze lactation patterns, offering valuable insights into milk production. In this chapter, a comparative assessment of different lactation models is carried out. Traditional models, such as Wood and Milkbot, have been recognized for their use of exponential equations to describe lactation curves. These models tried to capture the overall shape of lactation curves. In the first part of the chapter, buffalo lactation curves are studied, and the difference between the two models is discussed. However, advancements in technology have led to the development of newer models, including those based on ML. Among these, models that belong to the autoencoder class have gained attention for their ability to handle complex data and provide more accurate predictions. By comparing the performance of traditional models like Wood and Milkbot with newer ML-based models, this chapter aims to evaluate the performance of the new methods and to compare with the classical ones. Through this investigation, we seek to provide valuable insights into the effectiveness of novel prediction models in analyzing lactation dynamics. The second section of the chapter is dedicated to cows. Their lactation curves were analysed using both classical approaches and the autoencoder.

4.1 Analysis of Lactation Curve Models in Dairy Animals

As is extensively known, the production of milk by animals during their lactation varies on a daily basis. Capturing the changes in milk yield during lactation is of primary concern for herd management and genetic selection. The modelling of the phenomena through different mathematical models is usually referred to as a Lactation Curve (LC). A good model should capture the behaviour of milk production changes from the beginning of lactation until the end of the lactation period. Daily milk data can be described through mathematical models that are one of the pivotal tools to fit the lactation data, describe lactation shape and compute LC characteristics (LCC) such as time to peak (days), peak yield (kg/d) or cumulative milk yield (kg/305d). The study of LC shape combined with these indexes can serve to evaluate different aspects of milk production performance at cow and herd level and have many applications in dairy research fields. In feed composition and feeding system, LCC can provide information on how different feeding system strategies affect the production of milk yield (Różańska-Zawieja et al., 2021). For example, studied the effect of dietary energy source and dietary energy level on LCC (Van Hoeij et al., 2017). Additionally, LCC serve as a tool for identifying cows with a specific lactational phenotype (Ehrlich, 2013). For example, it's suggested that cows exhibiting high daily milk yield and long milking intervals are more efficient and thus suited for being milked with an AMS (Masia et al., 2022). Moreover, LCC can characterize perturbations of milk and offer insights into how cows respond to challenges during lactation, such as diseases or changing in the feeding routine (Van Hoeij et al., 2017). For disease detection, LC analysis can significantly contribute to the assessment of both short- and long-term effects of metabolic diseases on milk production (Hostens et al., 2012; Masia et al., 2022; Yamakazi et al., 2009). LC analysis can be used to determine milk production losses due to diseases (Andersen et al., 2011; Steeneveld et al., 2008) or to study the associations between LCC and age at first calving (Atashi & Hostens, 2021; Elahi Torshizi, 2016).

Mathematical approaches vs machine learning approach

Nowadays different mathematical models have been proposed to describe the shape of dairy LC and provide information on milk yield production. Brody, in 1923 was the first to introduced a lactation model employing an exponential function that was able to depict the declining phase of the lactation curve in cows. A year later, the model was improved by incorporating the modelling of the ascending phase (Brody et al., 1923). In 1967, Wood used an incomplete gamma-type function that overcame the limitations of the exponential models in describing the ascending phase of the lactation curve. It consists of three parameters: the scale a (representing the level of production), the ramp b (representing the rising rate of milk to the peak production level) and the declining slope c (Wood, 1967). This classic Wood model is a widely recognized lactation curve model and a starting point for subsequent improvements and innovations. It is often used as a benchmark for evaluating the performance of other models (Radjabalizadeh et al., 2022).

Wood and other models developed (Wilmink, 1987) have been commonly used. While most models failed to describe the shape of the lactation curve beyond 305 days (Dematawewa et al., 2007; VanRaden & Miller, 2006), the MilkBot model is able to adjust to extended lactations (Ehrlich, 2013). The latter is one of the most recently developed mathematical models (Ehrlich, 2013). Parameters a and b had the same interpretation of Wood equation, however, there are the time of maximal creation of productive capacity (offset, c) and the loss of productive capacity (decay, d), which can be easily transformed into a measure of persistency using a conversion index. Having one more parameter compared to Wood, the Milkbot is able to better capture the slope of lactation. Indeed, by incorporating prior information, also offered greater flexibility for accounting for the influence of diseases and management practices, potentially leading to more accurate daily milk yield estimates (Ehrlich, 2010).

Despite all, the LC models also show some weaknesses. The correct adoption of the mathematical models requires good initial parameter values. Those are needed to solve the mathematical fitting. If none are available, the resulting fit can be suboptimal, and in that case, global optimization algorithms such as genetic can be useful. Additionally, these models assume that all lactation curves have the same classical shape: a convex curve increasing at the beginning of the curve and followed by a longer and gradual decline. However, these models neglect animal variability in lactation curve as Fig 4.1 shows.

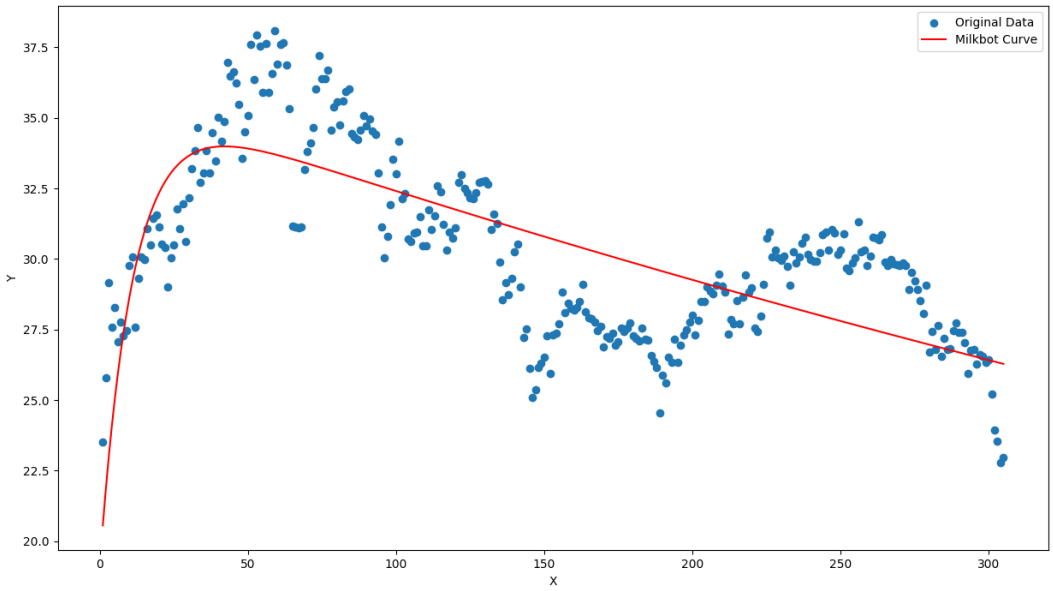

Fig 4.1 *Example of strange lactation curves of cows*

ML techniques have provided new tools to avoid mathematical assumptions. Artificial Neural Networks (ANNs) are a family of ML techniques and are inspired by the structure and function of biological neural networks. They offer a nonlinear approach useful for LC modelling methods and enabling the mapping of complex functional relationships (Bauer & Jagusiak, 2022). Fig 4.2 represents a basically structure of ANN. ANN usually were made up by input signals, interconnected nodes called neurons, and a single output. The neurons are organized into layers (hidden layers) within the network, each neuron receives input signals x_i, for each x_i there is a weight (w_{ij}) that represents the strength of neuron connection, b is the bias; and $f(x)$ is the activation function which is responsible of the nonlinear behaviour of the neuron and give to the ANN the ability of describe complex phenomena. Starting from this structure it is possible to customize the ANN, adding more neurons and hidden layers, changing the activation functions, the learning algorithms and architecture.

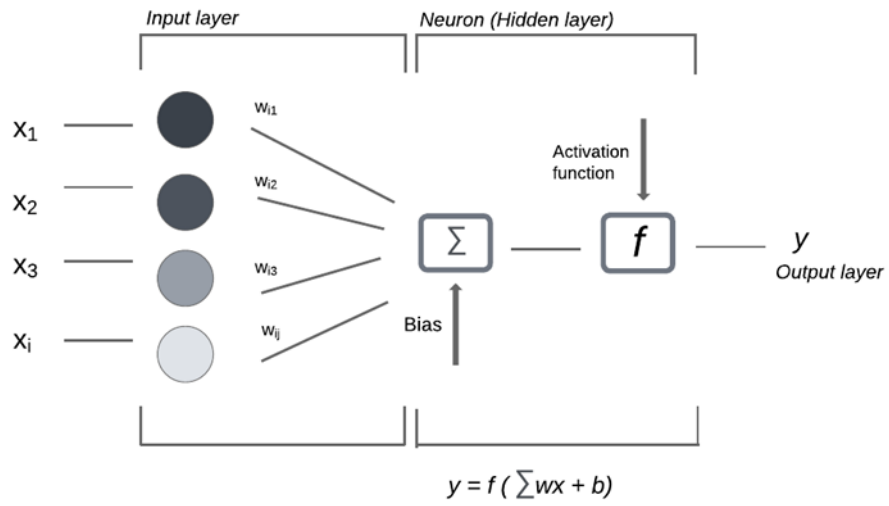

Fig 4.2 *Architecture of one neuron*

In the LC domain, one of the first applications that opened the way to utilization of ANNs was from Lacroix et al. (Lacroix et al., 1995). They compared the performance of the mathematical model and two ANNs to predict the milk 305-d milk, fat and protein production and assessed the individual contribution of the input variables to the prediction process of ANNs. The results showed that both ANNs performed better than the mathematical model, especially at the beginning of lactation. Starting from the promising results obtained by (Lacroix et al., 1995) and helped by the fast progress in the computer science field and the higher data availability, other authors moved toward a ML approach, deeply exploring the capability of ANNs to predict the day milk yield of dairy cows and their LC. Indeed, different ANNs configurations and architectures have been explored. In recent years, (Liseune et al., 2021) proposed an approach using autoencoder that showed very promising results. Autoencoder is an unsupervised learning model that consists in encoder and a decoder. The autoencoder compresses the input in a lower dimensional space and then reconstructs the output from this representation (VanRaden & Miller, 2006). Thanks to this structure, autoencoder is suitable for time series tasks and can predict missing milk yield using all the observations recorded before and after the moment of prediction, without take into account the length of the time interval between these recordings. This means that with autoencoders, it is possible to employ time series that do not have a fixed number of observations and can dynamically update the missing milk yields information along the LC as soon as new information in the corresponding lactation cycle becomes available. Thanks to this flexibility, (2021) explored autoencoders as

an innovative lactation curve model comparing with other ANN in different forecast horizons. The results showed that autoencoders performed better than the other ANN in general and on every task in terms of every evaluation metric (Liseune et al., 2021). Starting from a the description of mathematical and ML-based models, the chapter aims to compare the fitting performance of three lactation models. The initial analysis focused on buffaloes, utilizing only the Wood and Milkbot models. Subsequently, a larger dataset of Holstein cow lactations was analyzed, incorporating the autoencoder model as well.

4.2 Assessment of Mediterranean buffalo lactation curves shape using lactation models

Buffalo data and preprocessing

The analysis was performed on a dataset containing milk yield, calving date, lactation number and days in milk from 33376 animals on 295 buffalo herds from 2013 until 2016 with lactation numbers ranging from 1 to 3. Animals with at least five observations per lactation were chosen to ensure a coherent fitting with at least the number of observations equal to the number of the regression parameters of the model. All data pre-processing was done through R software (version 4.3.2).

Mathematical equations and validity measure

The first mathematical model fitted to the data was the Wood equation (Wood, 1967) with the following formula:

$$Y(t) = at^b e^{-ct}$$

Equation 4.1 *Wood equation*

Where *Y* is the milk production, *t* the days in milk, the magnitude *a*, *b* the time to peak and *c* is the decay. The second model employed was the MilkBot (Ehrlich, 2013). The full MilkBot equation is shown as:

$$\mathrm{Y}\,(t) = a\left(1 - \frac{e^{\frac{c-t}{b}}}{2}\right)e^{-dt}$$

Equation 4.2. *Milkbot equation*

Where *Y* is the milk production, *t* the days in milk, *a* the magnitude, *b* the time to peak *c* the offset and *d* the decay.

The fitting of Wood model was performed with a Python code using the "curve_fit" function from the "scipy" package was used to fit lactation data using Python v3.10. The fitting of Milkbot happened through the API version of the model (1.3), which was available online (Jim

Ehrlich, API Milkbot). Wood and Milkbot equations required the employment of priors for the parameters *a, b,c* and *d*. Prior values were used as initial guesses to search the optimal solution. Initially, Wood and Milkbot models were fitted using priors based on a literature search (Khan et al., 2023; Şeahin et al., 2015). After the first fitting step, mean and standard deviation (sd) of regression parameters from the results were used to fit all lactations for a second time, At the end, the performance of the models was evaluated through the coefficient of determination (R^2).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Equation 4.3 *Coefficient of determination*

where SS_{res} is the sum of the squared residuals and SS_{tot} is the total sum of squares.

Results and discussion

The presented research provides a preliminary approach to a mathematical model of the buffalo LC shape. The results suggest an overall better performance of the Wood equation compared to Milkbot. The results of the fitting of the Wood and Milkbot equations are shown in detail and discussed in this section. Results are shown for lactation number 1, 2 and 3.

Wood analysis

Parity	$\bar{a} \pm \sigma_a$	$\bar{b} \pm \sigma_b$	$\bar{c} \pm \sigma_c$	$\overline{R^2} \pm \sigma_{R^2}$
1	6.1±4.2	0.30±0.30	0.005±0.003	0.72±0.26
2	7.6± 5.1	0.29±0.30	0.006±0.004	0.78±0.22
3	7.9± 5.2	0.30±0.30	0.007±0.004	0.79±0.21

Tab 4.1 *Fitting metrics of Wood model*

The mean values of *a, b, c* are coherent with Khan et al. (2023). We achieved high R^2 values, especially for lactations 2 and 3. Our results suggest that the Wood model achieved a good approximation of real milk yield despite the low sampling rates of time series. On the other

hand, the standard deviation for each parameter and R^2 suggest that data are strongly variable around the mean, probably due to the variability in the number of milk points available for each lactation.

Milkbot analysis

Parity	$\overline{a} \pm \sigma_a$	$\overline{b} \pm \sigma_b$	$\overline{c} \pm \sigma_c$	$\overline{d} \pm \sigma_d$	$\overline{R^2} \pm \sigma_{R^2}$
1	13.5±2.4	30.67±0.06	-0.4992±0.001	0.0015±0.0001	0.58±0.26
2	15.9± 3.2	22.74±0.02	-0.7751±0.001	0.0026±0.0003	0.69±0.22
3	17.1± 3.6	25.07±0.75	0.0039±0.002	0.0029±0.0003	0.69±0.20

Tab 4.2 *Fitting metrics of the Milkbot model.*

Tab 4.2 reported the result for Milkbot equation. No literature was available to compare the results obtained. However, results seem coherent based on dairy cow parameters and their interpretation (Chen et al., 2022). Indeed, the parameter *a* represents the magnitude of milk production. Specifically, the average daily milk production for buffaloes is typically around 10–15 kg. In this case, the mean of *a* values for each lactation were coherent with the generally milk production of buffaloes. Milkbot performed worse than Wood in terms of R^2 for each lactation. This result could be also influenced by the different fitting engines (described in the section of "Mathematical equations and validity measure") of the two models adopted. The R^2 values suggest that Milkbot and Wood equations seem to be a promising technique for evaluating the LC of buffalo considering the very few milk points available during lactation that negatively affect the results. However, since the models are strongly influenced by the choice of the initial priors, more efforts to find suitable values of *a, b, c,* and *d* can improve the model performance. Finally, like in Holstein cows, the first lactations achieved worse result in terms of R^2 compared to the lactations 2+.

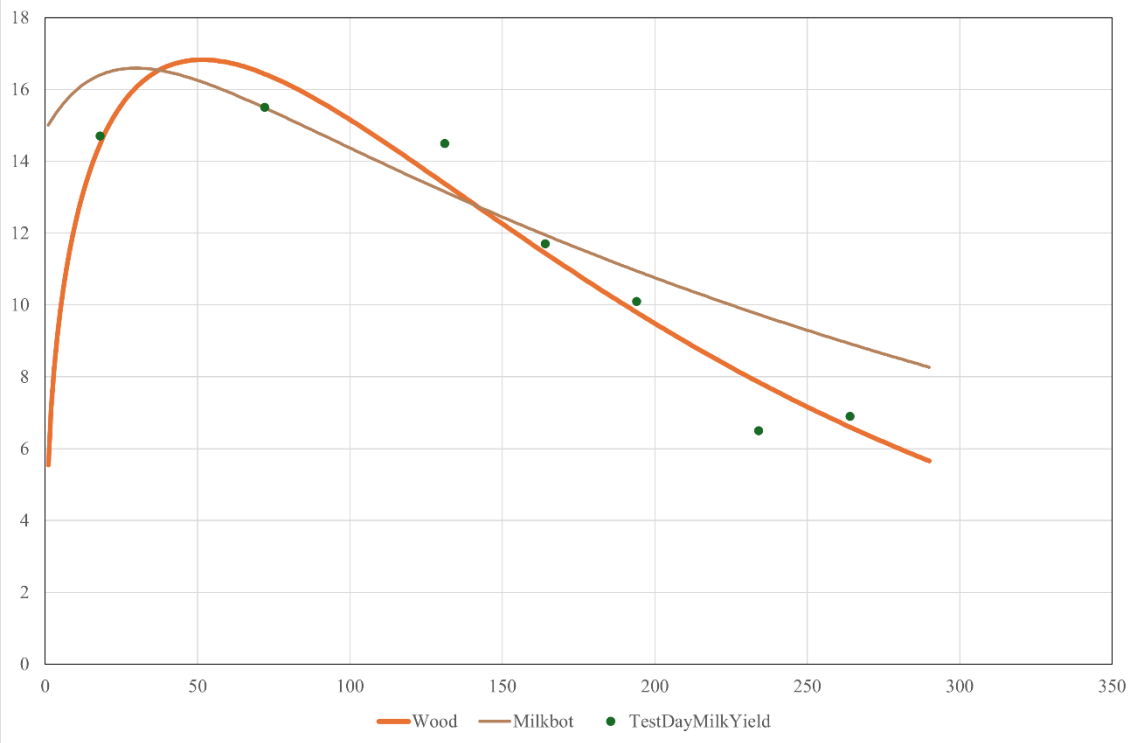

Fig 4.3 *Comparison of the LC fitted with the Wood and Milkbot models*

Fig 4.3 shows a random LC fitted with Wood and Milkbot models. Wood equation provided a closer match to the actual data points, particularly at the beginning and end of the lactation period. At the start of the lactation curve, the MilkBot model struggles to accurately fit the initial data points. This trend is consistent with the boxplot in Fig 4.4. The boxplot illustrates the differences between predicted milk yields from the Wood (orange) and Milkbot (brown) models compared to the true milk yield, within lactation divided into ten periods. Predictions from the Wood model displayed a narrower range, suggesting less variability in its estimations. Conversely, the Milkbot model exhibited a broader range of predicted values, indicating higher variability. Moreover, in the later stages of lactation, the Milkbot model tended to overestimate milk yield, while the Wood model showed a slight underestimation. Those results, were confirmed looking the table of median. This graphic evidence was also supported by the results of Tab 4.1 and Tab 4.2

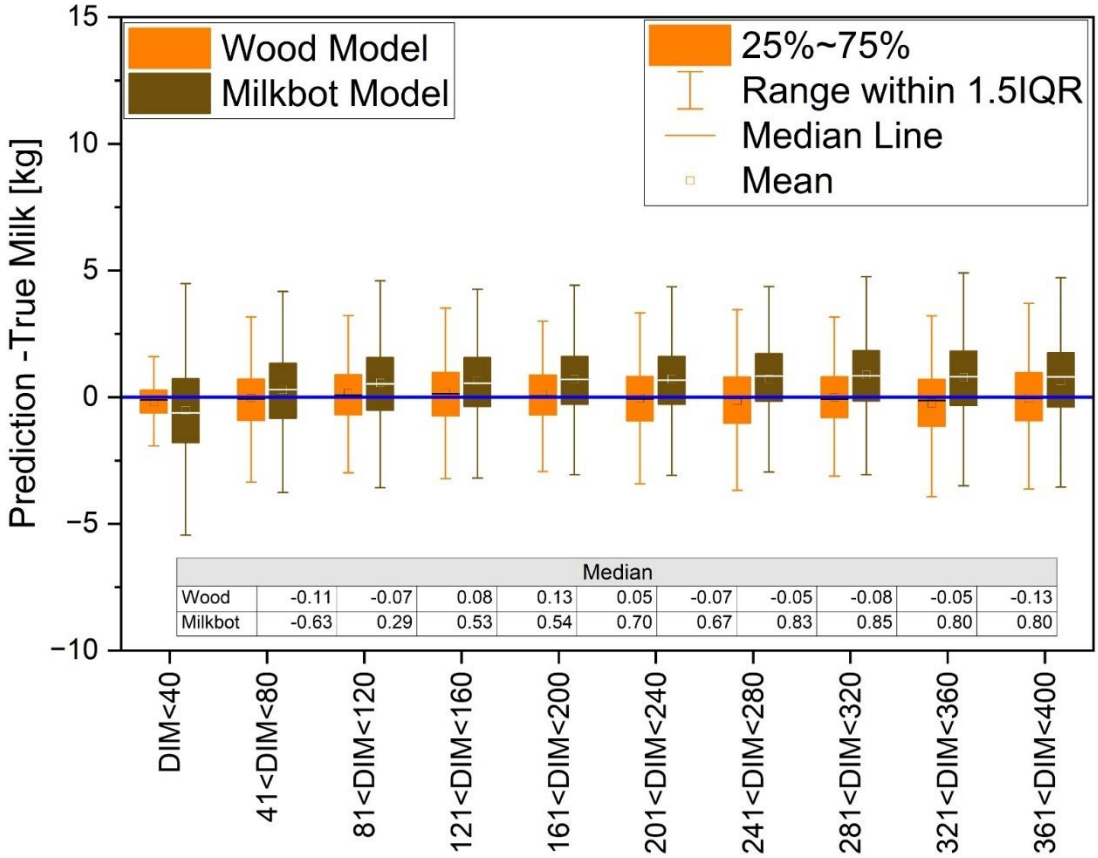

Fig 4.4. *Boxplot of the difference between milk yield predict by the two equations and true milk yield. In the table, there are the median of each box for period. The blue line represents the ideal case when the predicted milk yield equals the true milk yield.*

Discussion

In this study, a comprehensive examination of lactation curves in Mediterranean buffaloes was conducted, emphasizing the effectiveness of mathematical models to estimate milk yield dynamics. This is one of the first analysis applying the Milkbot model to buffalo LC. Utilizing data from 33,376 buffaloes across 295 herds collected between 2013 and 2016, the study applied two mathematical models, the Wood equation and the MilkBot model, to fit lactation data across multiple parities. Initially, parameters for both models were set based on literature priors, while a second fitting phase utilized these results to refine estimates. For Wood equations, priors were found in literature, while for Milkbot different trails were carried out to

find the optimal solutions starting from the priors of cows. The results demonstrated that the Wood equation had a slightly better fit, particularly notable in second and third lactations, where higher R^2 values indicated an accurate representation of actual milk yields. These results were in line with findings from Ghavi Hossein-Zadeh et al. (2016) who compared seven mathematical models for LC fitting and found that the Wood model consistently performed well. Specifically, Ghavi Hossein- Zadeh et al. (2016) highlighted that the Wood model performed better, based on several performance metrics, including the R^2, RMSE, Durbin-Watson statistic, and Akaike information criterion Moreover, the challenge in fitting the LC of buffaloes could lie in its shape, because of the different persistency of plateau phase as the Fig 4.5 shows.

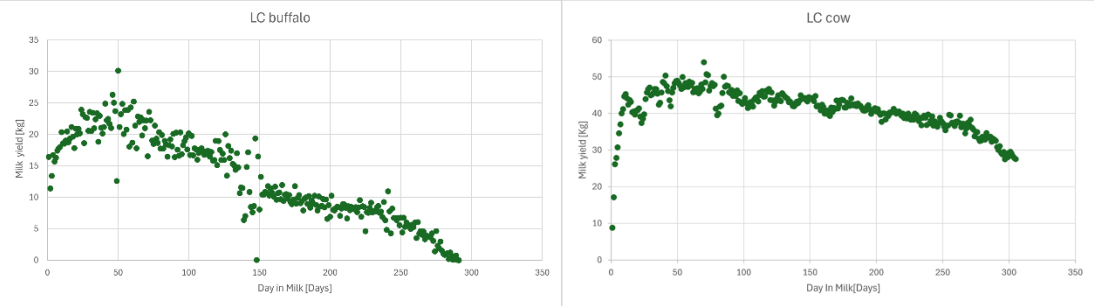

Fig 4.5 *Comparison of a typical buffaloes and cows LC within the same parity order (3).*

Conclusion

This work reported and compared the performance of Wood and Milkbot equations to describe the behaviour of buffalo milk yield. The results suggested that Wood performed better than Milkbot in terms of R^2 in buffalo cows. Moreover, Wood equation achieved better results than Milkbot employing few milk points available. Results are promising, but more efforts are needed to establish more accurate priors for buffalo cows.

4.3 Assessment of cow's lactation curves shape using lactation models: Classical approach vs Deep Learning based model

Cows Data and preprocessing

The dataset employed in this work described the performance of 3498 Holstein cows reared 91 herds in The Netherlands. The dataset was made by 31.693.777 observations. Each observation included general cow information (e.g., Animal ID, Herd ID, days in milk), milk yield (kg) and milk component. Several preprocessing steps were performed. Animals showing one missing value during the first 305 days of lactation were removed. In this way, the dataset had only complete time series for each cow within 305 days. At the end of this process, the final dataset counted 12.661.160 daily milking events of 3399 cows from 86 herds.

Mathematical models and machine learning approach

Three different LC models adopted, and their performances was compared. In particular, two exponential models and an autoencoder were employed. The first mathematical model is the Wood equation (Wood, 1967), while the second model was the MilkBot® (Ehrlich, 2013). Formula of models were explained in section 4.2.

An autoencoder architecture was chosen to apply ML techniques. The structure was made up of three components: an encoder, a bottleneck and a decoder. In this work, one hidden layer for the decoder and a decoder part was employed. The whole dataset was split into training, validation, and test set (80, 10, 10) and then scaled to train the ANN. A sensitivity analysis was carried out to assess the behaviour of the NN with 3,4, and 30 neurons in the NN hidden layers. The matlab function employed to train the autoencoder were "trainAutoencoder". The number of epochs, the activation and loss function, the training alghoritms and number of neurons of layer were set to train the autoencoder. Tab 4.3 summarizes the hyperparameters tested.

Hyperparameter	Setting
N° of epochs	1000
Data	Scaled
Activation function	logsig
Loss function	MSE
Algorithm	Scaled conjugate gradient descent
N° of neurons for hidden layer (n_{HL})	3,4,30

Tab 4.3 *Autoencoder parameters.*

Evaluation metrics

The performance of the models was evaluated through:

- the coefficient of determination (R^2)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Equation 4.3 $Coefficient\ of\ determination$

where SS_{res} is the sum of squared residuals and SS_{tot} is the total sum of squares.

- the Root Mean Squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y\imath})^2}$$

Equation 4.4 *Root Mean Square error formula*

- the percentage of the error on the sum of milk yield at the end of lactation

$$\Delta Milk = \frac{\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \widehat{Yi}}{\sum_{i=1}^{n} Y_i} * 100$$

Equation 4.5 *Error milk formula*

where Y_i is the daily milk yield, $\widehat{Yi}$ is the predict milk yield and *n* the number of samples.

Results and Discussion

In this section the results of the fitting of Wood, Milkbot equations and the autoencoder are shown and discussed. All results are shown for lactation 1 to 5. Tab 4.4 reported the mean of parameters *a,b,c* ± sd and the $\overline{\boldsymbol{R^2}}$, $\overline{\Delta Milk}$ and $\overline{RMSE}$ ± sd.

Parity	$\bar{\boldsymbol{a}} \pm \boldsymbol{\sigma_a}$	$\bar{\boldsymbol{b}} \pm \boldsymbol{\sigma_b}$	$\bar{\boldsymbol{c}} \pm \boldsymbol{\sigma_c}$	$\overline{\boldsymbol{R^2}} \pm \boldsymbol{\sigma_{R^2}}$	$\overline{\Delta Milk} \pm \boldsymbol{\sigma_{\Delta Milk}}$	$\overline{RMSE} \pm \boldsymbol{\sigma_{RMSE}}$
1	13.16±4.22	0.28±0.09	0.003±0.001	0.71±0.16	-0.01±0.06	2.32± 0.6
2	20.19± 6.07	0.25±0.08	0.004±0.001	0.79±0.15	0.0001± 0.09	2.77±0.8
3	21.43± 6.29	0.25±0.09	0.004±0.001	0.81±0.13	0.005±0.11	2.95±0.9
4	21.14±6.39	0.26±0.09	0.004±0.001	0.81±0.13	0.007±0.11	3.03±0.9
5	55.50 ±9.11	23.05±8.92	0.010±0.006	0.79±0.13	0.012±0.15	3.35±0.9

Tab 4.4 *Fitting metrics of the Wood model.*

The mean values of *a, b, c* were in line with Radjabalizadeh et al. (2022). The evaluation metrics denoted a good fit for the true milk yield. In particular, R^2 achieved high values, for lactation 3 and 4. The mean milk error provides information on how the total milk prediction is close to the real total milk yield. In the first lactation, the model predicted a relatively high milk yield compared to the true milk yield. In contrast, in lactation 2 to 5, the model underestimated the true milk yield. The RMSE denoted the highest error in lactation 4 and 5

Parity	$\bar{a} \pm \sigma_a$	$\bar{b} \pm \sigma_b$	$\bar{c} \pm \sigma_c$	$\bar{d} \pm \sigma_d$	$\overline{R^2} \pm \sigma_{R^2}$	$\overline{\Delta Milk} \pm \sigma_{\Delta Milk}$	$\overline{RMSE} \pm \sigma_{RMSE}$
1	38.28 ± 6.37	27.09± 4.35	-0.49±0.001	0.0014±0.0007	0.66± 0.14	-0.30±0.27	2.64± 0.6
2	50.25± 8.56	21.81±3.00	-0.74±0.027	0.0022± 0.0008	0.76± 0.15	-0.26± 0.32	3.05±0.9
3	54.91±8.98	22.56±8.80	0.01±0.006	0.0024±0.0008	0.79±0.13	-0.29±0.30	3.17±1
4	55.64±9.11	22.92±8.72	0.01±0.006	0.0025±0.0008	0.79± 0.13	-0.32±0.31	3.27±1
5	55.50 ±9.11	23.05±8.92	0.01±0.006	0.0025±0.0008	0.79± 0.13	-0.33±0.32	3.35±1

Tab 4.5 *Fitting metrics of the Milkbot model*

The mean values of parameters *a*, *b*, *c*, and *d* were coherent with the work of (Ehrlich, 2013). Milkbot achieved the lowest R^2 in the first lactation. Unlike Wood, the $\overline{\Delta Milk}$ in Milkbot overestimated the true milk yield. Finally, the values of $\overline{\mathrm{RMSE}}$ were higher in Milkbot fitting compared to Wood.

Parity	n_{HL}	$\overline{R^2} \pm \sigma_{R^2}$	$\overline{\Delta Milk} \pm \sigma_{\Delta Milk}$	$\overline{RMSE} \pm \sigma_{RMSE}$
1	3	0.65 ± 0.8	-0.91 ± 4.42	2.41 ± 0.81
	4	0.77 ± 0.38	-0.17 ± 2.36	2.05 ± 0.67
	30	0.90 ± 0.08	-0.01 ± 0.83	1.37 ± 0.42
2	3	0.78 ± 0.41	0.02 ± 2.85	2.76 ± 0.97
	4	0.83 ± 0.22	-0.12 ± 1.73	2.39 ± 0.83
	30	0.93 ± 0.06	-0.04 ± 0.80	1.58 ± 0.54
3	3	0.80 ± 0.26	0.02 ± 2.26	2.97 ± 1.09
	4	0.85 ± 0.16	-0.01 ± 1.44	2.58 ± 0.93
	30	0.94 ± 0.05	-0.07 ± 0.84	1.70 ± 0.66
4	3	0.80 ± 0.21	-0.01 ± 2.16	3.06 ± 1.09
	4	0.85 ± 0.14	-0.05 ± 1.33	2.67 ± 0.92
	30	0.94 ± 0.06	0.08 ± 0.81	1.75 ± 0.6
5	3	0.80 ± 0.19	-0.05 ± 2.24	3.14 ± 1.09
	4	0.85 ± 0.13	-0.07 ± 1.50	2.73 ± 0.93
	30	0.93 ± 0.05	-0.1 ± 0.83	1.80 ± 0.61

Tab 4.6 *Fitting metrics of the Autoencoder models.*

The model with the autoencoder trained with 30 neurons performed better than the models with 3 and 4 neurons for all metrics. Models with 3 and 4 neurons achieved negative R^2, as denoted by the high σ_{R^2}. This trend is evident in the $\overline{\Delta Milk}$ that showed quite high values of $\sigma_{\Delta Milk}$ in models with 3 and 4 neurons.

Comparison of the performance

In Tab 4.7 the comparison of the two exponential models and the autoencoder with 30 neurons is reported. The Wood and Milkbot models showed very similar performances. However, Wood performed relatively better compared to MilkBot for all metrics. The Autoencoder denoted better performances according to both RMSE and R^2 metrics. However, the $\boldsymbol{\sigma}_{\Delta Milk}$ is quite high, denoting a bad prediction of milk yield for some non-conventional LC. Furthermore, since the good performance achieved by the autoencoder, this metric could be a benchmark to assess if the cow production was in line with the expected production. This could be useful to detect health problems such mastitis or stressful situations. Moreover, the capacity for integrating reproduction and health events enhances the model's utility in real-world applications, improving farm management through more realistic milk yield predictions and early detection of deviations from expected production.

Parity	Model	$\overline{R^2} \pm \sigma_{R^2}$	$\overline{\Delta Milk} \pm \sigma_{\Delta Milk}$	$\overline{RMSE} \pm \sigma_{RMSE}$
1	Wood	0.71±0.16	-0.01±0.06	2.32± 0.6
	Milkbot	0.66± 0.14	-0.30±0.27	2.64± 0.6
	AE-30	0.90 ± 0.08	-0.01 ± 0.83	1.37 ± 0.42
2	Wood	0.79±0.15	0.0001± 0.09	2.77±0.8
	Milkbot	0.76± 0.15	-0.26± 0.32	3.05±0.9
	AE-30	0.93 ± 0.06	-0.04 ± 0.80	1.58 ± 0.54
3	Wood	0.81±0.13	0.005±0.11	2.95±0.9
	Milkbot	0.79±0.13	-0.29±0.30	3.17±1
	AE-30	0.94 ± 0.05	-0.07 ± 0.84	1.70 ± 0.66
4	Wood	0.81±0.13	0.007±0.11	3.03±0.9
	Milkbot	0.79± 0.13	-0.32±0.31	3.27±1
	AE-30	0.94 ± 0.06	0.08 ± 0.81	1.75 ± 0.6
5	Wood	0.79±0.13	0.012±0.15	3.35±0.9
	Milkbot	0.79± 0.13	-0.33±0.32	3.35±1
	AE-30	0.93 ± 0.05	-0.1 ± 0.83	1.80 ± 0.61

Tab 4.7 *Fitting metrics comparison among model*

In Fig 4.6 the comparison of the fit of a particular LC is visualized. It can be highlighted that the autoencoder with the 30 neurons also follows the experimental data with smaller fluctuations. The other two models described the general behaviour of milk during lactation. The autoencoder with three neurons achieved performance close to the Wood equation.

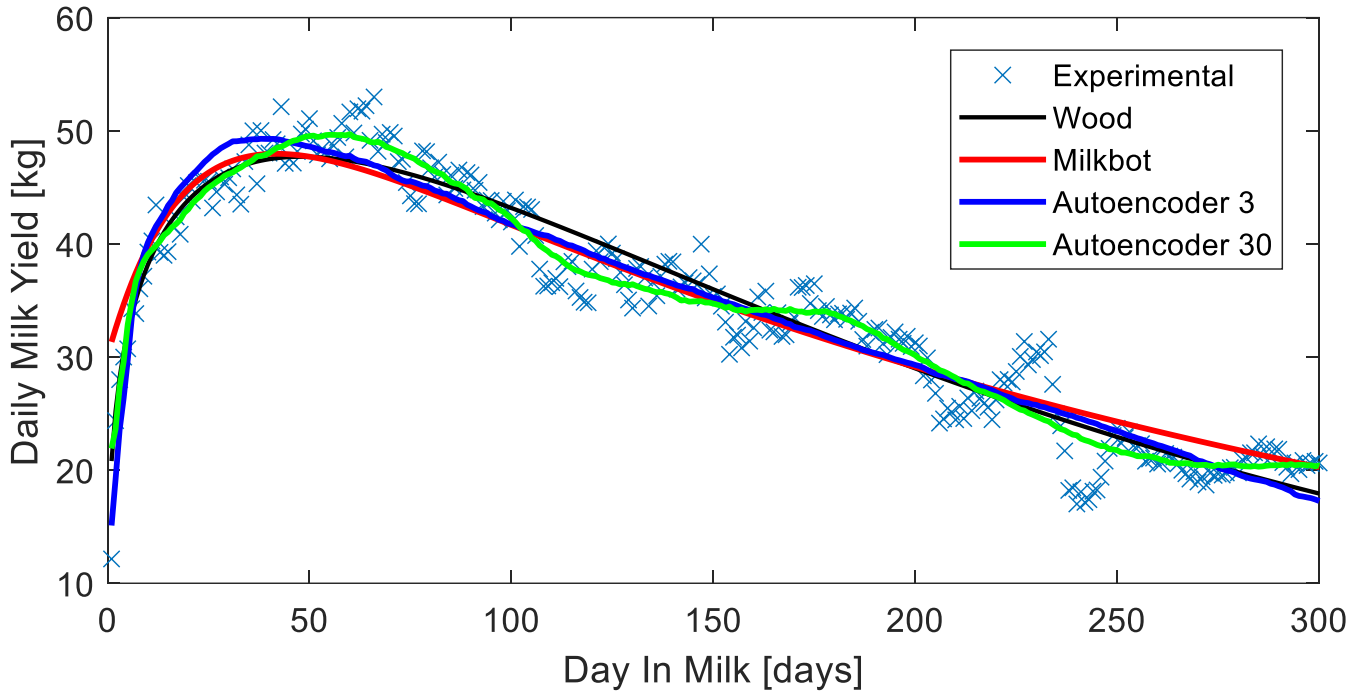

Fig 4.6 *Comparison of the LC fitted with the three models.*

Conclusion

Three different fitting models were tested on LC data of 3498 cows. Generally, all three models were able to the LC. The Wood's model seemed to fitted the milk prediction better. The autoencoder required more parameters (>10 neurons in hidden layer) to offer fitting metrics better to the exponential models. However, their capability to correctly represent the LC makes them an interesting solution for further development and inclusion inside prediction algorithms. This work may pave the way for future studies in which model performances could be used to detect anomalous drops in production or health issues.

Chapter 5: General conclusions

Nowadays, the integration of machine learning (ML) into livestock management is in line with precision livestock farming (PLF) principles, as thoroughly explored in this thesis. By leveraging advanced ML techniques, various aspects of livestock production can be meticulously monitored and analyzed, providing valuable insights that significantly enhance decision-making processes. These techniques enable the continuous and automated real-time monitoring of livestock, encompassing production, reproduction, health, welfare, and environmental impact. One of the primary benefits of ML in PLF was its ability to handle large volumes of data, extracting meaningful patterns and trends that might otherwise go unnoticed. For instance, ML algorithms can analyze data from automated milking system (AMS), herd databases, and environmental sensors to predict milk yield, detect health issues such as mastitis or lameness, and optimize feeding strategies. This data-driven approach allowed for more precise and timely interventions, ultimately leading to improved productivity and sustainability in livestock farming. Moreover, ML techniques facilitate the development of predictive models that can forecast future production trends and identify potential risks. These models can be used to simulate different scenarios, helping farmers make informed decisions about breeding, feeding, and health management. By anticipating issues before they become critical, farmers can take proactive measures to mitigate risks, thereby enhancing the overall resilience of their operations. However, several limitations must be acknowledged. Firstly, the consistency of data is a significant challenge.

The data used in this study was often fragmented and inconsistent. Indeed, despite the availability of viable technologies, data was collected manually, especially in small farms, which led to an inconsistency and an occasional gap in data quality and continuity. This inconsistency can affect the reliability and accuracy of the models. Secondly, the preprocessing of data is extensive and time-consuming. Cleaning, standardizing, and preparing the data for analysis requires substantial effort and can introduce potential biases if not done meticulously. Lastly, the fragmented nature of veterinary data collection poses additional challenges. Data are often collected sporadically and may lack the comprehensive coverage needed for robust analysis, further complicating the modeling process.

This thesis has explored the application of ML techniques to improve production efficiency in livestock, specifically focusing on dairy herds. By integrating advanced technologies and data-

driven approaches, the thesis provides some examples on how ML can effectively handle large datasets, uncover patterns, and improve decision-making processes in the dairy sector. One of the first results presented are from a comprehensive review utilizing ML techniques to analyze the literature on PLF. The increasing number of papers employing, studying, and exploring the PLF approach highlights its growing significance. Given the multidisciplinary nature of PLF, research papers are published across various fields such as veterinary science and engineering. This diversity can sometimes lead to confusion due to the wide range of applications and methodologies. To address this, the thesis employed text mining (TM) and topic analysis (TA) to systematically review and categorize the literature. This approach effectively highlighted the primary trends in PLF, providing a deeper understanding of the sector's challenges and opportunities. By organizing the research into coherent topics, the thesis clarified the various areas of interest and identified key themes, thereby enhancing our comprehension of the current state and future directions of PLF. Subsequently, the cluster analysis (CA) provided valuable insights into milk productivity in buffaloes and goats, uncovering significant characteristics and patterns related to herds, species, and breeds. In goats, for example, different fat percentages were found for the same lactation stage group. The main results indicate that clustering algorithms, particularly k-means, effectively categorized animals based on their lactation stages. This finding was consistent across all trials and species examined. One of the most important results emerged from the analysis of the Camosciata breed, where clusters were differentiated based on the concentrations of fatty acids in the milk. This evidence suggests that k-means was able to identify animals with varying performances in terms of milk quality. Based on these findings, k-means can be a tool in distinguishing animals with different productivity levels and milk quality characteristics. However, further experiments and analyses are necessary to refine these techniques and explore additional factors that may influence clustering outcomes. At the end, the evaluation and study of lactation curve (LC) were carried out. Thes study explored the performance of different lactation models for cows and buffaloes, comparing traditional exponential models with deep learning approaches, such as autoencoders. The analysis revealed both strengths and weaknesses in the methods used, indicating that while classical models remain valuable, advanced ML techniques offer significant potential for accurately predicting lactation curves and enhancing herd productivity. For both buffaloes and cows, the Wood equation provided a better fit than the MilkBot model, particularly for lactations 2 and 3. Despite the close performance between the two models, it is important to note that for buffaloes, very few data points were available for each lactation. Additionally, there is a lack of literature on the

application of the MilkBot model to buffaloes, making it challenging to determine the appropriate priors for initializing the equations. This condition can negatively impact the performance of the MilkBot model. Finally, regarding the autoencoder, despite its superior performance, it requires a higher computational burden. The autoencoder with 3 and 4 neurons performed similarly to the mathematical models, indicating that a more complex architecture with more neurons is necessary to achieve better results. This increased complexity demands more computational resources, which can be a consideration when choosing between traditional models and advanced ML techniques. Nonetheless, the potential benefits of improved accuracy and predictive capabilities offered by the autoencoder justify the additional computational effort. Further optimization and research are needed to balance performance and computational efficiency. Overall, this thesis underscores the importance of a multidisciplinary approach, combining expertise from engineers, data scientists, veterinarians and animal scientists to advance the field of PLF. The findings suggest that ML is a pivotal tool for optimizing livestock management and ensuring the long-term viability of the dairy industry.

Reference

Achour, B., Belkadi, M., Filali, I., Laghrouche, M., & Lahdir, M. (2020). Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on Convolutional Neural Networks (CNN). *Biosystems Engineering*, *198*, 31–49. https://doi.org/10.1016/j.biosystemseng.2020.07.019

Achour, B., Belkadi, M., Saddaoui, R., Filali, I., Aoudjit, R., & Laghrouche, M. (2022). High-accuracy and energy-efficient wearable device for dairy cows' localization and activity detection using low-cost IMU/RFID sensors. *Microsystem Technologies*, *28*(5), 1241–1251. https://doi.org/10.1007/s00542-022-05288-7

Adolfsson, A., Ackerman, M., & Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, *88*, 13–26. https://doi.org/10.1016/j.patcog.2018.10.026

Afiani, F. A., Joezy-Shekalgorabi, S., Amin-Afshar, M., Sadeghi, A.-A., & Jensen, J. (2021). Additive genetic and permanent environmental correlation between different parts of lactation in moderate and cold regions. *Czech Journal of Animal Science*, *66*(4), 112–121. https://doi.org/10.17221/254/2020-CJAS

Agradi, S., Gazzonis, A. L., Curone, G., Faustini, M., Draghi, S., Brecchia, G., Vigo, D., Manfredi, M. T., Zanzani, S. A., Pulinas, L., Sulce, M., Munga, A., Castrica, M., & Menchetti, L. (2021). Lactation Characteristics in Alpine and Nera di Verzasca Goats in Northern Italy: A Statistical Bayesian Approach. *Applied Sciences 2021, Vol. 11, Page 7235*, *11*(16), 7235. https://doi.org/10.3390/APP11167235

Alessio, D. R. M., Velho, J. P., McManus, C. M., Knob, D. A., Vancin, F. R., Antunes, G. V., Busanello, M., De Carli, F., & Neto, A. T. (2021). Lactose and its relationship with other milk constituents, somatic cell count, and total bacterial count. *Livestock Science*, *252*, 104678. https://doi.org/10.1016/j.livsci.2021.104678

Ali, A. K. A., & Shook, G. E. (1980). An Optimum Transformation for Somatic Cell Concentration in Milk. *Journal of Dairy Science*, *63*(3), 487–490. https://doi.org/10.3168/jds.S0022-0302(80)82959-6

Alipio, M., & Villena, M. L. (2023). Intelligent wearable devices and biosensors for monitoring cattle health conditions: A review and classification. *Smart Health*, *27*, 100369. https://doi.org/10.1016/j.smhl.2022.100369

Allueva Molina, Q., Ko, H.-L., Gómez, Y., Manteca, X., & Llonch, P. (2023). Comparative study between scan sampling behavioral observations and an automatic monitoring image system on a commercial fattening pig farm. *Frontiers in Animal Science*, *4*. https://doi.org/10.3389/fanim.2023.1248972

Andersen, F., Østerås, O., Reksen, O., Toft, N., & Gröhn, Y. T. (2011). Associations between the time of conception and the shape of the lactation curve in early lactation in Norwegian dairy cattle. *Acta Veterinaria Scandinavica*, *53*(1), 5. https://doi.org/10.1186/1751-0147-53-5

Arcidiacono, C., Porto, S. M. C., Mancino, M., & Cascone, G. (2017). A threshold-based algorithm for the development of inertial sensor-based systems to perform real-time cow step counting in free-stall barns. *Biosystems Engineering*, *153*, 99–109. https://doi.org/10.1016/j.biosystemseng.2016.11.003

Atashi, H., & Hostens, M. (2021). Genetic parameters for milk urea and its relationship with milk yield and compositions in Holstein dairy cows. *PLOS ONE*, *16*(6), e0253191. https://doi.org/10.1371/journal.pone.0253191

Banerjee, A., & Dave, R. N. (n.d.). Validating clusters using the Hopkins statistic. *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542)*, 149–153. https://doi.org/10.1109/FUZZY.2004.1375706

Berckmans, D. (2014). Precision livestock farming technologies for welfare management in intensive livestock systems. *Revue Scientifique et Technique de l'OIE*, *33*(1), 189–196. https://doi.org/10.20506/rst.33.1.2273

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

Borderas, T. F., Fournier, A., Rushen, J., & de Passillé, A. M. B. (2008). Effect of lameness on dairy cows' visits to automatic milking systems. *Canadian Journal of Animal Science*, *88*(1), 1–8. https://doi.org/10.4141/CJAS07014

Brody, S., Ragsdale, A. C., & Turner, C. W. (1923). THE RATE OF DECLINE OF MILK SECRETION WITH THE ADVANCE OF THE PERIOD OF LACTATION. *Journal of General Physiology*, *5*(4), 441–444. https://doi.org/10.1085/jgp.5.4.441

Brotzman, R. L., Cook, N. B., Nordlund, K., Bennett, T. B., Gomez Rivas, A., & Döpfer, D. (2015). Cluster analysis of Dairy Herd Improvement data to discover trends in performance characteristics in large Upper Midwest dairy herds. *Journal of Dairy Science*, *98*(5), 3059–3070. https://doi.org/10.3168/JDS.2014-8369

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Cammarano, D., Ceccarelli, S., Grando, S., Romagosa, I., Benbelkacem, A., Akar, T., Al-Yassin, A., Pecchioni, N., Francia, E., & Ronga, D. (2019). The impact of climate change on barley yield in the Mediterranean basin. *European Journal of Agronomy*, *106*, 1–11. https://doi.org/10.1016/j.eja.2019.03.002

Cao, M., Yi, Q., Wang, K., Li, J., & Wang, X. (2023). Predicting Ventilation Rate in a Naturally Ventilated Dairy Barn in Wind-Forced Conditions Using Machine Learning Techniques. *Agriculture*, *13*(4), 837. https://doi.org/10.3390/agriculture13040837

Carpinelli, N. A., Rosa, F., Grazziotin, R. C. B., & Osorio, J. S. (2019). Technical note: A novel approach to estimate dry matter intake of lactating dairy cows through multiple on-cow accelerometers. *Journal of Dairy Science*, *102*(12), 11483–11490. https://doi.org/10.3168/jds.2019-16537

Casey, T. M., & Plaut, K. (2022). Circadian clocks and their integration with metabolic and reproductive systems: our current understanding and its application to the management of dairy cows. *Journal of Animal Science*, *100*(10). https://doi.org/10.1093/jas/skac233

Catillo, G., Macciotta, N. P. P., Carretta, A., & Cappio-Borlino, A. (2002). Effects of Age and Calving Season on Lactation Curves of Milk Production Traits in Italian Water Buffaloes. *Journal of Dairy Science*, *85*(5), 1298–1306. https://doi.org/10.3168/jds.S0022-0302(02)74194-5

Cavero, D., Tölle, K.-H., Henze, C., Buxadé, C., & Krieter, J. (2008). Mastitis detection in dairy cows by application of neural networks. *Livestock Science*, *114*(2–3), 280–286. https://doi.org/10.1016/j.livsci.2007.05.012

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *61*(6). https://doi.org/10.18637/jss.v061.i06

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. *Computers and Electronics in Agriculture*, *145*, 83–91. https://doi.org/10.1016/j.compag.2017.12.013

Chen, X., Zheng, H., Wang, H., & Yan, T. (2022). Can machine learning algorithms perform better than multiple linear regression in predicting nitrogen excretion from lactating dairy cows. *Scientific Reports 2022 12:1*, *12*(1), 1–13. https://doi.org/10.1038/s41598-022-16490-y

Chen, Y., Hostens, M., Nielen, M., Ehrlich, J., & Steeneveld, W. (2022). Herd level economic comparison between the shape of the lactation curve and 305 d milk production. *Frontiers in Veterinary Science*, *9*. https://doi.org/10.3389/fvets.2022.997962

Clark, C. E. F., Farina, S. R., Garcia, S. C., Islam, M. R., Kerrisk, K. L., & Fulkerson, W. J. (2016). A comparison of conventional and automatic milking system pasture utilization and pre- and post-grazing pasture mass. *Grass and Forage Science*, *71*(1), 153–159. https://doi.org/10.1111/gfs.12171

Clark, S., & Mora García, M. B. (2017). A 100-Year Review: Advances in goat milk research. *Journal of Dairy Science*, *100*(12), 10026–10044. https://doi.org/10.3168/jds.2017-13287

Cogato, A., Brščić, M., Guo, H., Marinello, F., & Pezzuolo, A. (2021). Challenges and Tendencies of Automatic Milking Systems (AMS): A 20-Years Systematic Review of Literature and Patents. *Animals*, *11*(2), 356. https://doi.org/10.3390/ani11020356

CORDIS - EU. (2024). https://cordis.europa.eu/project/id/311825/it

Currò, S., De Marchi, M., Claps, S., Salzano, A., De Palo, P., Manuelian, C. L., & Neglia, G. (2019). Differences in the Detailed Milk Mineral Composition of Italian Local and Saanen Goat Breeds. *Animals*, *9*(7), 412. https://doi.org/10.3390/ani9070412

Currò, S., Manuelian, C., De Marchi, M., Claps, S., Rufrano, D., & Neglia, G. (2019). Differences in the Detailed Milk Mineral Composition of Italian Local and Saanen Goat Breeds. *Animals*, *9*(7), 412. https://doi.org/10.3390/ani9070412

Currò, S., Manuelian, C. L., De Marchi, M., De Palo, P., Claps, S., Maggiolino, A., Campanile, G., Rufrano, D., Fontana, A., Pedota, G., & Neglia, G. (2019). Autochthonous dairy goat breeds showed better milk quality than Saanen under the same environmental conditions. *Archives Animal Breeding*, *62*(1), 83–89. https://doi.org/10.5194/aab-62-83-2019

David M. Blei, Andrew Y.Ng, & Michael I.Jordan. (2003). *Latent dirichlet allocation* (Vol. 3). Journal of machine learning research.

Davies, H., Nenadic, G., Alfattni, G., Arguello Casteleiro, M., Al Moubayed, N., Farrell, S., Radford, A. D., & Noble, P.-J. M. (2024). Text mining for disease surveillance in veterinary clinical data: part two, training computers to identify features in clinical text. *Frontiers in Veterinary Science*, *11*. https://doi.org/10.3389/fvets.2024.1352726

de Mol, R. M., & Ouweltjes, W. (2001). Detection model for mastitis in cows milked in an automatic milking system. *Preventive Veterinary Medicine*, *49*(1–2), 71–82. https://doi.org/10.1016/S0167-5877(01)00176-3

Dematawewa, C. M. B., Pearson, R. E., & VanRaden, P. M. (2007). Modeling Extended Lactations of Holsteins. *Journal of Dairy Science*, *90*(8), 3924–3936. https://doi.org/10.3168/jds.2006-790

Dervić, E., Matzhold, C., Egger-Danner, C., Steininger, F., & Klimek, P. (2024). Improving Lameness Detection in Cows: A Machine Learning Algorithm Application. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2024-24730

Driessen, C., & Heutinck, L. F. M. (2015). Cows desiring to be milked? Milking robots and the co-evolution of ethics and technology on Dutch dairy farms. *Agriculture and Human Values*, *32*(1), 3–20. https://doi.org/10.1007/s10460-014-9515-5

Du, X., Tejeda, H., Yang, Z., & Lu, L. (2022). A General-Equilibrium Model of Labor-Saving Technology Adoption: Theory and Evidences from Robotic Milking Systems in Idaho. *Sustainability*, *14*(13), 7683. https://doi.org/10.3390/su14137683

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., & Henry, D. (2015). Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture*, *111*, 18–28. https://doi.org/10.1016/j.compag.2014.12.002

Ehrlich, J. L. (2010). Quantifying shape of lactation curves, and benchmark curves for common dairy breeds and parities. *The Bovine Practitioner*, 88–95. https://doi.org/10.21423/bovine-vol45no1p88-95

Ehrlich, J. L. (2013). Quantifying inter-group variability in lactation curve shape and magnitude with the MilkBot ® lactation model. *PeerJ*, *1*, e54. https://doi.org/10.7717/peerj.54

Elahi Torshizi, M. (2016). Effects of season and age at first calving on genetic and phenotypic characteristics of lactation curve parameters in Holstein cows. *Journal of Animal Science and Technology*, *58*(1), 8. https://doi.org/10.1186/s40781-016-0089-1

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743

FAO, Dairy Market Review. (2023). https://openknowledge.fao.org/server/api/core/bitstreams/68f7f25d-b3cb-418e-b04d-5708e5bcea1e/content

FAOSTAT. (2024). https://www.fao.org/statistics/en

Faugno, S., Pindozzi, S., Okello, C., & Sannino, M. (2015). Testing the application of an automatic milking system on buffalo (Bubalus bubalis). *Journal of Agricultural Engineering*, *46*(1), 13–18. https://doi.org/10.4081/JAE.2015.437

Flower, F. C., & Weary, D. M. (2009). Gait assessment in dairy cattle. *Animal*, *3*(1), 87–95. https://doi.org/10.1017/S1751731108003194

Fournel, S., Ouellet, V., & Charbonneau, É. (2017). Practices for Alleviating Heat Stress of Dairy Cows in Humid Continental Climates: A Literature Review. *Animals*, *7*(5), 37. https://doi.org/10.3390/ani7050037

Frades, I., & Matthiesen, R. (2010). *Overview on Techniques in Cluster Analysis* (pp. 81–107). https://doi.org/10.1007/978-1-60327-194-3_5

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? *Journal of Dairy Science*, *105*(8), 6760–6772. https://doi.org/10.3168/JDS.2022-21975

Fuentes, A., Han, S., Nasir, M. F., Park, J., Yoon, S., & Park, D. S. (2023). Multiview Monitoring of Individual Cattle Behavior Based on Action Recognition in Closed Barns Using Deep Learning. *Animals*, *13*(12), 2020. https://doi.org/10.3390/ani13122020

Gasteiner, J., Horn, M., & Steinwidder, A. (2015). Continuous measurement of reticuloruminal pH values in dairy cows during the transition period from barn to pasture feeding using an indwelling wireless data transmitting unit. *Journal of Animal Physiology and Animal Nutrition*, *99*(2), 273–280. https://doi.org/10.1111/jpn.12249

Genedy, R., & Ogejo, J. (2023). Quantifying ammonia lost to the atmosphere during manure storage on a dairy farm as influenced by management and meteorological parameters. *Agriculture, Ecosystems & Environment*, *354*, 108563. https://doi.org/10.1016/j.agee.2023.108563

Ghavi Hossein-Zadeh, N. (2016). Comparison of non-linear models to describe the lactation curves for milk yield and composition in buffaloes (Bubalus bubalis). *Animal*, *10*(2), 248–261. https://doi.org/10.1017/S1751731115001846

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Grün, B., & Hornik, K. (2011). **topicmodels** : An *R* Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13). https://doi.org/10.18637/jss.v040.i13

Henao-Velásquez, A. F., Múnera-Bedoya, O. D., Herrera, A. C., Agudelo-Trujillo, J. H., & Cerón-Muñoz, M. F. (2014). Lactose and milk urea nitrogen: fluctuations during lactation in Holstein cows. *Revista Brasileira de Zootecnia*, *43*(9), 479–484. https://doi.org/10.1590/S1516-35982014000900004

Henchion, M., Hayes, M., Mullen, A., Fenelon, M., & Tiwari, B. (2017). Future Protein Supply and Demand: Strategies and Factors Influencing a Sustainable Equilibrium. *Foods*, *6*(7), 53. https://doi.org/10.3390/foods6070053

Hentzen, A. H. R., & Holm, D. E. (2024). A novel production profile classification system for incoming calves that predicts feedlot growth performance. *Animal Production Science*, *64*(3). https://doi.org/10.1071/AN23395

Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). *Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges* (pp. 271–300). https://doi.org/10.1007/978-3-662-43968-5_16

Hossain, M. E., Kabir, M. A., Zheng, L., Swain, D. L., McGrath, S., & Medway, J. (2022). A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. *Artificial Intelligence in Agriculture*, *6*, 138–155. https://doi.org/10.1016/j.aiia.2022.09.002

Hostens, M., Ehrlich, J., Van Ranst, B., & Opsomer, G. (2012). On-farm evaluation of the effect of metabolic diseases on the shape of the lactation curve in dairy cows through the MilkBot lactation model. *Journal of Dairy Science*, *95*(6), 2988–3007. https://doi.org/10.3168/jds.2011-4791

Hyde, J., Stokes, J. R., & Engel, P. D. (2003). Optimal investment in an automatic milking system: an application of real options. *Agricultural Finance Review*, *63*(1), 75–92. https://doi.org/10.1108/00215010380001142

Idris, M., Uddin, J., Sullivan, M., McNeill, D. M., & Phillips, C. J. C. (2021). Non-Invasive Physiological Indicators of Heat Stress in Cattle. *Animals*, *11*(1), 71. https://doi.org/10.3390/ani11010071

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jiang, B., Tang, W., Cui, L., & Deng, X. (2023). Precision Livestock Farming Research: A Global Scientometric Review. *Animals*, *13*(13), 2096. https://doi.org/10.3390/ani13132096

Jim Ehrlich. (n.d.). *API Milkbot*. Retrieved April 26, 2024, from https://api.milkbot.com/

John, A. J., Clark, C. E. F., Freeman, M. J., Kerrisk, K. L., Garcia, S. C., & Halachmi, I. (2016). Review: Milking robot utilization, a successful precision livestock farming evolution. *Animal*, *10*(9), 1484–1492. https://doi.org/10.1017/S1751731116000495

Kamphuis, C., Mollenhorst, H., Feelders, A., Pietersma, D., & Hogeveen, H. (2010). Decision-tree induction to detect clinical mastitis with automatic milking. *Computers and Electronics in Agriculture*, *70*(1), 60–68. https://doi.org/10.1016/j.compag.2009.08.012

Khan, Z., Pasha, T. N., Bhatti, J. A., Sharif, N. R. M., Sahin, T., Naveed, S., & Tahir, M. N. (2023). Fitting Various Growth Equations to the Daily Milk Yield Data of Nili-Ravi Buffaloes and Cholistani Cows at Intake at Maintenance Levels. *Kafkas Universitesi Veteriner Fakultesi Dergisi*. https://doi.org/10.9775/kvfd.2023.29278

Kleen, J. L., & Guatteo, R. (2023). Precision Livestock Farming: What Does It Contain and What Are the Perspectives? *Animals*, *13*(5), 779. https://doi.org/10.3390/ani13050779

Kunes, R., Bartos, P., Iwasaka, G. K., Lang, A., Hankovec, T., Smutny, L., Cerny, P., Poborska, A., Smetana, P., Kriz, P., & Kernerova, N. (2021). In-Line Technologies for the Analysis of Important Milk Parameters during the Milking Process: A Review. *Agriculture*, *11*(3), 239. https://doi.org/10.3390/agriculture11030239

Laurijs, K. A., Briefer, E. F., Reimert, I., & Webb, L. E. (2021). Vocalisations in farm animals: A step towards positive welfare assessment. *Applied Animal Behaviour Science*, *236*, 105264. https://doi.org/10.1016/j.applanim.2021.105264

Leliveld, L. M. C., Brandolese, C., Grotto, M., Marinucci, A., Fossati, N., Lovarelli, D., Riva, E., & Provolo, G. (2024). Real-time automatic integrated monitoring of barn environment and dairy cattle behaviour: Technical implementation and evaluation on three commercial farms. *Computers and Electronics in Agriculture*, *216*, 108499. https://doi.org/10.1016/j.compag.2023.108499

Lesimple, C. (2020). Indicators of Horse Welfare: State-of-the-Art. *Animals*, *10*(2), 294. https://doi.org/10.3390/ani10020294

Liseune, A., den Poel, D. Van, Hut, P. R., van Eerdenburg, F. J. C. M., & Hostens, M. (2021). Leveraging sequential information from multivariate behavioral sensor data to predict

the moment of calving in dairy cattle using deep learning. *Computers and Electronics in Agriculture*, *191*, 106566. https://doi.org/10.1016/j.compag.2021.106566

Maertens, W., Vangeyte, J., Baert, J., Jantuan, A., Mertens, K. C., De Campeneere, S., Pluk, A., Opsomer, G., Van Weyenberg, S., & Van Nuffel, A. (2011). Development of a real time cow gait tracking and analysing tool to assess lameness using a pressure sensitive walkway: The GAITWISE system. *Biosystems Engineering*, *110*(1), 29–39. https://doi.org/10.1016/j.biosystemseng.2011.06.003

Marino, R., Petrera, F., & Abeni, F. (2023). Scientific Productions on Precision Livestock Farming: An Overview of the Evolution and Current State of Research Based on a Bibliometric Analysis. *Animals*, *13*(14), 2280. https://doi.org/10.3390/ani13142280

Maroto Molina, F., Pérez Marín, C. C., Molina Moreno, L., Agüera Buendía, E. I., & Pérez Marín, D. C. (2020). Welfare Quality ® for dairy cows: towards a sensor-based assessment. *Journal of Dairy Research*, *87*(S1), 28–33. https://doi.org/10.1017/S002202992000045X

Maselyne, J., Pastell, M., Thomsen, P. T., Thorup, V. M., Hänninen, L., Vangeyte, J., Van Nuffel, A., & Munksgaard, L. (2017). Daily lying time, motion index and step frequency in dairy cows change throughout lactation. *Research in Veterinary Science*, *110*, 1–3. https://doi.org/10.1016/j.rvsc.2016.10.003

Masia, F., Molina, G., Vissio, C., Balzarini, M., de la Sota, R. L., & Piccardi, M. (2022). Quantifying the negative impact of clinical diseases on productive and reproductive performance of dairy cows in central Argentina. *Livestock Science*, *259*, 104894. https://doi.org/10.1016/j.livsci.2022.104894

Mattachini, G., Pompe, J., Finzi, A., Tullo, E., Riva, E., & Provolo, G. (2019). Effects of Feeding Frequency on the Lying Behavior of Dairy Cows in a Loose Housing with Automatic Feeding and Milking System. *Animals*, *9*(4), 121. https://doi.org/10.3390/ani9040121

Matuszewski, P. (2023). How to prepare data for the automatic classification of politically related beliefs expressed on Twitter? The consequences of researchers' decisions on the number of coders, the algorithm learning procedure, and the pre-processing steps on the performance of supervised models. *Quality & Quantity*, *57*(1), 301–321. https://doi.org/10.1007/s11135-022-01372-2

Mayo, L. M., Silvia, W. J., Ray, D. L., Jones, B. W., Stone, A. E., Tsai, I. C., Clark, J. D., Bewley, J. M., & Heersche, G. (2019). Automated estrous detection using multiple commercial precision dairy monitoring technologies in synchronized dairy cows. *Journal of Dairy Science*, *102*(3), 2645–2656. https://doi.org/10.3168/jds.2018-14738

Mélynda Hassouna, Thomas Eglin, Pierre Cellier, Vincent Colomb, Jean-Pierre Cohan, Céline Décuq, Monique Delabuis, & Nadége Edouard. (2016). *Measuring emissions from livestock farming: greenhouse gases, ammonia and nitrogen oxides*.

Millar, K. M., Tomkins, S. M., White, R. P., & Mepham, T. B. (2002). Consumer attitudes to the use of two dairy technologies. *British Food Journal*, *104*(1), 31–44. https://doi.org/10.1108/00070700210418721

Mohamad, I. Bin, & Usman, D. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303. https://doi.org/10.19026/rjaset.6.3638

Munksgaard, L., Weisbjerg, M. R., Henriksen, J. C. S., & Løvendahl, P. (2020). Changes to steps, lying, and eating behavior during lactation in Jersey and Holstein cows and the relationship to feed intake, yield, and weight. *Journal of Dairy Science*, *103*(5), 4643–4653. https://doi.org/10.3168/jds.2019-17565

Mustafa, M. Y., Hansen, I., Eilertsen, S. M., Pettersen, E., & Kronen, A. (2013). Matching mother and calf reindeer using wireless sensor networks. *2013 5th International Conference on Computer Science and Information Technology*, 99–105. https://doi.org/10.1109/CSIT.2013.6588765

Nalon, E., Contiero, B., Gottardo, F., & Cozzi, G. (2021). The Welfare of Beef Cattle in the Scientific Literature From 1990 to 2019: A Text Mining Approach. *Frontiers in Veterinary Science*, *7*. https://doi.org/10.3389/fvets.2020.588749

Negussie, E., González-Recio, O., Battagin, M., Bayat, A.-R., Boland, T., de Haas, Y., Garcia-Rodriguez, A., Garnsworthy, P. C., Gengler, N., Kreuzer, M., Kuhla, B., Lassen, J., Peiren, N., Pszczola, M., Schwarm, A., Soyeurt, H., Vanlierde, A., Yan, T., & Biscarini, F. (2022). Integrating heterogeneous across-country data for proxy-based random forest prediction of enteric methane in dairy cattle. *Journal of Dairy Science*, *105*(6), 5124–5140. https://doi.org/10.3168/jds.2021-20158

Nielen, M., Schukken, Y. H., Brand, A., Haring, S., & Ferwerda-Van Zonneveld, R. T. (1995). Comparison of Analysis Techniques for On-Line Detection of Clinical Mastitis. *Journal of Dairy Science*, *78*(5), 1050–1061. https://doi.org/10.3168/jds.S0022-0302(95)76721-2

Norton, T., & Berckmans, D. (2017). Developing precision livestock farming tools for precision dairy farming. *Animal Frontiers*, *7*(1), 18–23. https://doi.org/10.2527/af.2017.0104

Nyamuryekung'e, S., Duff, G., Utsumi, S., Estell, R., McIntosh, M. M., Funk, M., Cox, A., Cao, H., Spiegal, S., Perea, A., & Cibils, A. F. (2023). Real-Time Monitoring of Grazing Cattle Using LORA-WAN Sensors to Improve Precision in Detecting Animal Welfare Implications via Daily Distance Walked Metrics. *Animals*, *13*(16), 2641. https://doi.org/10.3390/ani13162641

Ozella, L., Brotto Rebuli, K., Forte, C., & Giacobini, M. (2023). A Literature Review of Modeling Approaches Applied to Data Collected in Automatic Milking Systems. *Animals*, *13*(12), 1916. https://doi.org/10.3390/ANI13121916/S1

Ozella, L., Price, E., Langford, J., Lewis, K. E., Cattuto, C., & Croft, D. P. (2022). Association networks and social temporal dynamics in ewes and lambs. *Applied Animal Behaviour Science*, *246*, 105515. https://doi.org/10.1016/j.applanim.2021.105515

Pahl, C., Hartung, E., Grothmann, A., Mahlkow-Nerge, K., & Haeussermann, A. (2014). Rumination activity of dairy cows in the 24 hours before and after calving. *Journal of Dairy Science*, *97*(11), 6935–6941. https://doi.org/10.3168/jds.2014-8194

Patra, A., Park, T., Kim, M., & Yu, Z. (2017). Rumen methanogens and mitigation of methane emission by anti-methanogenic compounds and substances. *Journal of Animal Science and Biotechnology*, *8*(1), 13. https://doi.org/10.1186/s40104-017-0145-9

Peyraud, J.-L., & MacLeod, M. (2020). *Future of EU livestock : How to contribute to a sustainable agricultural sector ? : final report*.

Pietersma, D., Lacroix, R., Lefebvre, D., & Wade, K. M. (2003). Induction and evaluation of decision trees for lactation curve analysis. *Computers and Electronics in Agriculture*, *38*(1), 19–32. https://doi.org/10.1016/S0168-1699(02)00105-9

Post, C., Rietz, C., Büscher, W., & Müller, U. (2020). Using Sensor Data to Detect Lameness and Mastitis Treatment Events in Dairy Cows: A Comparison of Classification Models. *Sensors*, *20*(14), 3863. https://doi.org/10.3390/s20143863

Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: global trends, research directions, and regulatory standpoints. *Data Science and Management*, *4*, 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

R. Lacroix, K. M. Wade, R. Kok, & J. F. Hayes. (1995). Prediction of Cow Performance with a Connectionist Model. *Transactions of the ASAE*, *38*(5), 1573–1579. https://doi.org/10.13031/2013.27984

Radjabalizadeh, K., Alijani, S., Gorbani, A., & Farahvash, T. (2022). Estimation of genetic parameters of Wood's lactation curve parameters using Bayesian and REML methods for milk production trait of Holstein dairy cattle. *Journal of Applied Animal Research*, *50*(1), 363–368. https://doi.org/10.1080/09712119.2022.2080211

Ramirez-Morales, I., Aguilar, L., Fernandez-Blanco, E., Rivero, D., Perez, J., Pazos, A., Aguilar, L. ;, Fernandez-Blanco, E. ;, Rivero, D. ;, Perez, J. ;, & Pazos, A. (2021). Detection of Bovine Mastitis in Raw Milk, Using a Low-Cost NIR Spectrometer and k-NN Algorithm. *Applied Sciences 2021, Vol. 11, Page 10751*, *11*(22), 10751. https://doi.org/10.3390/APP112210751

Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, *56*(8), 8219–8264. https://doi.org/10.1007/s10462-022-10366-3

Rashid, A. Bin, & Kausik, M. A. K. (2024). AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Advances*, *7*, 100277. https://doi.org/10.1016/j.hybadv.2024.100277

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 35–39. https://doi.org/10.1109/COMITCON.2019.8862451

Rebuli, K. B., Ozella, L., Vanneschi, L., & Giacobini, M. (2023). Multi-algorithm clustering analysis for characterizing cow productivity on automatic milking systems over lactation

periods. *Computers and Electronics in Agriculture*, *211*, 108002. https://doi.org/10.1016/j.compag.2023.108002

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. *Journal of Systems Architecture*, *116*, 102139. https://doi.org/10.1016/j.sysarc.2021.102139

Riaboff, L., Shalloo, L., Smeaton, A. F., Couvreur, S., Madouasse, A., & Keane, M. T. (2022). Predicting livestock behaviour using accelerometers: A systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. *Computers and Electronics in Agriculture*, *192*, 106610. https://doi.org/10.1016/j.compag.2021.106610

Ross, S., Wang, H., Zheng, H., Yan, T., & Shirali, M. (2023). A Novel Mixed Effects Random Forest Approach for Predicting Dairy Cattle Methane Emissions. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3125–3132. https://doi.org/10.1109/BIBM58861.2023.10385563

Rotz, C. A., Coiner, C. U., & Soder, K. J. (2003). Automatic Milking Systems, Farm Size, and Milk Production. *Journal of Dairy Science*, *86*(12), 4167–4177. https://doi.org/10.3168/jds.S0022-0302(03)74032-6

Różańska-Zawieja, J., Winnicki, S., Zyprych-Walczak, J., Szabelska-Beręsewicz, A., Siatkowski, I., Nowak, W., Stefańska, B., Kujawiak, R., & Sobek, Z. (2021). The Effect of Feeding Management and Culling of Cows on the Lactation Curves and Milk Production of Primiparous Dairy Cows. *Animals*, *11*(7), 1959. https://doi.org/10.3390/ani11071959

RStudio Team. (2022). *R: A Language and Environment for Statistical Computing, Vienna, Austria*. http://www.rstudio.com/

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. https://doi.org/10.1016/j.animal.2022.100658

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, *2013*(1), 203. https://doi.org/10.1186/1029-242X-2013-203

Schillings, J., Bennett, R., & Rose, D. C. (2023). Perceptions of farming stakeholders towards automating dairy cattle mobility and body condition scoring in farm assurance schemes. *Animal*, *17*(5), 100786. https://doi.org/10.1016/j.animal.2023.100786

Schönberger, D., Berthel, R. M., Savary, P., & Bodmer, M. (2023). Analysis of Dairy Cow Behavior during Milking Associated with Lameness. *Dairy*, *4*(4), 554–570. https://doi.org/10.3390/dairy4040038

Şeahin, A., Ulutaş, Z., Arda, Y., Yüksel, A., & Serdar, G. (2015). Lactation Curve and Persistency of Anatolian Buffaloes. *Italian Journal of Animal Science*, *14*(2), 3679. https://doi.org/10.4081/ijas.2015.3679

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Shen, W., Yu, W., Kong, Q., Zhang, Y., Liu, G., & Wang, Q. (2015). Research on Milk Conductivity Real-time Online Monitoring System. *International Journal of Smart Home*, *9*(5), 1–10. https://doi.org/10.14257/ijsh.2015.9.5.01

Shi, Z., Zhang, Z., Jia, Y., Li, J., Wang, X., Qiu, Y., Miao, J., Chang, F., Han, X., & Tang, W. (2023). Internet-of- Things Behavior Monitoring System Based on Wearable Inertial Sensors for Classifying Dairy Cattle Health Using Machine Learning. *2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, 277–282. https://doi.org/10.1109/IICAIET59451.2023.10291766

Shigeta, M., Ike, R., Takemura, H., & Ohwada, H. (2018). Automatic Measurement and Determination of Body Condition Score of Cows Based on 3D Images Using CNN. *Journal of Robotics and Mechatronics*, *30*(2), 206–213. https://doi.org/10.20965/jrm.2018.p0206

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, *23*(6), 759. https://doi.org/10.3390/e23060759

Silvestre, A. M., Martins, A. M., Santos, V. A., Ginja, M. M., & Colaço, J. A. (2009). Lactation curves for milk, fat and protein in dairy cows: A full approach. *Livestock Science*, *122*(2–3), 308–313. https://doi.org/10.1016/j.livsci.2008.09.017

Simões Filho, L. M., Lopes, M. A., Brito, S. C., Rossi, G., Conti, L., & Barbari, M. (2020). Robotic milking of dairy cows: a review. *Semina: Ciências Agrárias*, *41*(6), 2833–2850. https://doi.org/10.5433/1679-0359.2020v41n6p2833

Šlyžius, E., Anskienė, L., Palubinskas, G., Juozaitienė, V., Šlyžienė, B., Juodžentytė, R., & Laučienė, L. (2023). Associations between Somatic Cell Count and Milk Fatty Acid and Amino Acid Profile in Alpine and Saanen Goat Breeds. *Animals*, *13*(6), 965. https://doi.org/10.3390/ani13060965

Srivastava, A. N., & Sahami, M. (Eds.). (2009). *Text Mining*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420059458

Steeneveld, W., Hogeveen, H., Barkema, H. W., van den Broek, J., & Huirne, R. B. M. (2008). The Influence of Cow Factors on the Incidence of Clinical Mastitis in Dairy Cows. *Journal of Dairy Science*, *91*(4), 1391–1402. https://doi.org/10.3168/jds.2007-0705

Ströbel, U., Rose-Meierhöfer, S., Öz, H., & Brunsch, R. (2013). Development of a Control System for the Teat-End Vacuum in Individual Quarter Milking Systems. *Sensors*, *13*(6), 7633–7651. https://doi.org/10.3390/s130607633

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. *Animal Science Journal*, *90*(4), 589–596. https://doi.org/10.1111/ASJ.13184

Trapanese, L., Hostens, M., Salzano, A., & Pasquino, N. (2024). Short review of current limits and challenges of application of machine learning algorithms in the dairy sector. *Acta IMEKO*, *13*(1), 1–7. https://doi.org/10.21014/actaimeko.v13i1.1725

Trapanese, L., Petrocchi Jasinski, F., Bifulco, G., Pasquino, N., Bernabucci, U., & Salzano, A. (2024). Buffalo welfare: a literature review from 1992 to 2023 with a text mining and topic analysis approach. *Italian Journal of Animal Science*, *23*(1), 570–584. https://doi.org/10.1080/1828051X.2024.2333813

Tremblay, M., Kammer, M., Lange, H., Plattner, S., Baumgartner, C., Stegeman, J. A., Duda, J., Mansfeld, R., & Döpfer, D. (2018). Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis. *Journal of Dairy Science*, *101*(8), 7311–7321. https://doi.org/10.3168/JDS.2017-13582

Tse, C., Barkema, H. W., DeVries, T. J., Rushen, J., & Pajor, E. A. (2018). Impact of automatic milking systems on dairy cattle producers' reports of milking labour management, milk production and milk quality. *Animal*, *12*(12), 2649–2656. https://doi.org/10.1017/S1751731118000654

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Van Hertem, T., Maltz, E., Antler, A., Romanini, C. E. B., Viazzi, S., Bahr, C., Schlageter-Tello, A., Lokhorst, C., Berckmans, D., & Halachmi, I. (2013). Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *Journal of Dairy Science*, *96*(7), 4286–4298. https://doi.org/10.3168/jds.2012-6188

van Hoeij, R. J., Dijkstra, J., Bruckmaier, R. M., Gross, J. J., Lam, T. J. G. M., Remmelink, G. J., Kemp, B., & van Knegsel, A. T. M. (2017). The effect of dry period length and postpartum level of concentrate on milk production, energy balance, and plasma metabolites of dairy cows across the dry period and in early lactation. *Journal of Dairy Science*, *100*(7), 5863–5879. https://doi.org/10.3168/jds.2016-11703

Vannieuwenborg, F., Verbrugge, S., & Colle, D. (2017). Designing and evaluating a smart cow monitoring system from a techno-economic perspective. *2017 Internet of Things Business Models, Users, and Networks*, 1–8. https://doi.org/10.1109/CTTE.2017.8260982

VanRaden, P. M., & Miller, R. H. (2006). Effects of Nonadditive Genetic Interactions, Inbreeding, and Recessive Defects on Embryo and Fetal Loss by Seventy Days. *Journal of Dairy Science*, *89*(7), 2716–2721. https://doi.org/10.3168/jds.S0022-0302(06)72347-5

Volkmann, N., Kulig, B., & Kemper, N. (2019). Using the Footfall Sound of Dairy Cows for Detecting Claw Lesions. *Animals*, *9*(3), 78. https://doi.org/10.3390/ani9030078

Vyas, P., Ghosh, S., Roy, M., & Sharma, A. (2022). Machine Learning applications in dairy farm management. In *Advances in Dairy Microbial Products* (pp. 385–396). Elsevier. https://doi.org/10.1016/B978-0-323-85793-2.00006-0

Wang, X., Ndegwa, P. M., Joo, H., Neerackal, G. M., Harrison, J. H., Stöckle, C. O., & Liu, H. (2016). Reliable low-cost devices for monitoring ammonia concentrations and emissions in naturally ventilated dairy barns. *Environmental Pollution*, *208*, 571–579. https://doi.org/10.1016/j.envpol.2015.10.031

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. *Journal of Dairy Science*, *100*(12), 10234–10250. https://doi.org/10.3168/jds.2017-12954

Wilmink, J. B. M. (1987). Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. *Livestock Production Science*, *16*(4), 335–348. https://doi.org/10.1016/0301-6226(87)90003-0

Wood. (1967). Algebraic Model of the Lactation Curve in Cattle. *Nature*, *216*(5111), 164–165. https://doi.org/10.1038/216164a0

Xie, C., Tian, X., Feng, X., Zhang, X., & Ruana, J. (2022). Preference Characteristics on Consumers' Online Consumption of Fresh Agricultural Products under the Outbreak of COVID-19: An Analysis of Online Review Data Based on LDA Model. *Procedia Computer Science*, *207*, 4486–4495. https://doi.org/10.1016/j.procs.2022.09.512

Yamakazi, T., Takeda, H., Nishiura, A., & Togashi, K. (2009). Relationship between the lactation curve and udder disease incidence in different lactation stages in first-lactation Holstein cows. *Animal Science Journal*, *80*(6), 636–643. https://doi.org/10.1111/j.1740-0929.2009.00695.x

Zhao, K., N. Shelley, A., L. Lau, D., A. Dolecheck, K., & M. Bewley, J. (2020). Automatic body condition scoring system for dairy cows based on depth-image analysis. *International Journal of Agricultural and Biological Engineering*, *13*(4), 45–54. https://doi.org/10.25165/j.ijabe.20201304.5655