Mobility classification of cattle with micro-Doppler radar

Linardopoulou Konstantina

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

School of Biodiversity, One Health & Veterinary Medicine
College of Medical, Veterinary and Life Sciences
University of Glasgow

May 2023

Abstract

Lameness in dairy cattle is a welfare concern that negatively impacts animal productivity and farmer profitability. Micro-Doppler radar sensing has been previously suggested as a potential system for automating lameness detection in ruminants. This thesis investigates the refinement of the proposed automated system by analysing and enhancing the repeatability and accuracy of the existing scoring method in cattle mobility scoring, used to provide labels in machine learning. The main aims of the thesis were (1) to quantify the performance of the micro-Doppler radar sensing method for the assessment of mobility, (2) to characterise and validate micro-Doppler radar signatures of dairy cattle with varying degrees of gait impairment, and (3) to develop machine learning algorithms that can infer the mobility status of the animals under test from their radar signatures and support automatic contactless classification.

The first study investigated inter-assessor agreement using a 4-level system and modifications to it, as well as the impact of factors such as mobility scoring experience, confidence in scoring decisions, and video characteristics. The results revealed low levels of agreement between assessors' scores, with kappa values ranging from 0.16 to 0.53. However, after transforming and reducing the mobility scoring system levels, an improvement was observed, with kappa values ranging from 0.2 to 0.67. Subsequently, a longitudinal study was conducted using good-agreement scores as ground truth labels in supervised machine-learning models. However, the accuracy of the algorithmic models was found to be insufficient, ranging from 0.57 to 0.63. To address this issue, different labelling systems and data pre-processing techniques were explored in a cross-sectional study. Nonetheless, the inter-assessor agreement remained challenging, with an average kappa value of 0.37 (SD = 0.16), and high-accuracy algorithmic predictions remained elusive, with an average accuracy of 56.1 (SD =16.58). Finally, the algorithms' performance was tested with high-confidence labels, which consisted of only scores 0 and 3 of the AHDB system. This testing resulted in good classification accuracy (0.82), specificity (0.79), and sensitivity (0.85). This led to the proposal of a new approach to producing labels, testing vantage point changes, and improving the performance of

machine learning models (average accuracy = 0.7 & SD = 0.17, average sensitivity = 0.68 & SD = 0.27, average specificity = 0.75 & SD = 0.17).

The research identified a challenge in creating high-confidence diagnostic labels for supervised machine learning-based algorithms to automate the detection and classification of lameness in dairy cows. As a result, the original goals were partially overridden, with the focus shifted to creating reliable labels that would perform well with radar data and machine learning. This point was considered necessary for smooth system development and process automation. Nevertheless, we managed to quantify the performance of the micro-Doppler radar system, partially develop the supervised machine learning algorithms, compare levels of agreement among multiple assessors, evaluate the assessment tools, assess the mobility evaluation process and gather a valuable data set which can be used as a foundation for subsequent studies. Finally, the thesis suggests changes in the assessment process to improve the prediction accuracy of algorithms based on supervised machine learning with radar data.

Table of Contents

List of Tables

List of Figures

List of Equations

Acknowledgements

I am filled with immense gratitude and joy as I mark the completion of my PhD journey. It has been a remarkable expedition filled with ups and downs, and I can confidently say that I have no regrets. If given the chance, I would gladly undertake this journey once more. It is a chapter of my life that has shaped me profoundly, and I owe it all to the incredible support and friendships I have formed along the way. This is my opportunity to express my heartfelt appreciation to each and every one of you.

First and foremost, I would like to extend my deepest gratitude to my main supervisors, Nicholas Jonsson and Julien Le Kernec. Their unwavering support and encouragement over the past 3.5 years have been invaluable. Without their guidance, the work presented in this thesis would not have been possible. They have always kept their doors open for me, providing guidance and advice in every aspect of my research. Their mentorship has not only expanded my research and writing skills but has also helped me grow as an individual. I am forever grateful to both Nicholas and Julien for their exceptional guidance and belief in my abilities.

I would also like to express my gratitude to the rest of my supervisory team, Lorenzo Viora and Francesco Fioranelli, who served as co-supervisors during this project. Lorenzo's enthusiasm and positivity have been a constant source of inspiration, and his ability to offer solutions to my problems has been invaluable. I have cherished our fieldwork trips and witnessed his exceptional social skills in action. Despite our limited time together, Francesco provided valuable input and suggestions that were always on point and immensely helpful.

I extend my appreciation to Elena Borelli, George King, June Kamonchanok, Zhenghui Li, Yixin Huang, and all the clinicians and farm staff who have assisted me at various stages of this project. Each of you has contributed to different aspects, enriching the outcome of this research.

Finally, I must express my deepest gratitude to my family for their support throughout this journey. Your constant belief in me has been a pillar of strength. And to Dimitris, your patience, presence, support, and ability to uplift and make me feel valued have been a blessing. Thank you for standing by my side and being a steadfast source of support.

To everyone who has played a part, big or small, in this remarkable journey, thank you. Your support, encouragement, and friendship have made all the difference.

Author’s declaration

“I declare that except where explicit reference is made to the contribution of others, this thesis is solely the result of my own work and does not include work presented for another degree at the University of Glasgow or any other institution.”

Konstantina Linardopoulou

May 2023

Definitions/Abbreviations

AD	Anno Domini
AHDB	Agriculture & Horticulture Development Board
AI	Artificial Intelligence
AIC	Akaike information criterion
ANOVA	Analysis of variance
AUC	Area under the curve
CAD	Computer-aided detection
DBi	Decibels relative to isotropic
FFT	Fast Fourier transformation
FMCW	Frequency modulated continuous wave
GLM	Generalized linear model
HD	High definition
ICC	Intraclass correlation coefficient
KNN	K-nearest neighbour
LBP	Local binary pattern
ML	Machine learning
MTI	Moving target indicator
PC	Personal computer
PRF	Pulse repetition frequency
ROC	Receiver operating characteristic
ROI	Region of interest
ROMS	Register Of Mobility Scorers
SD	Standard deviation
SML	Supervised machine learning
SNCR	signal-to-noise-and-clutter ratio
SNR	signal-to-noise ratio
STFT	Short-time Fourier transformation
SVD	Singular value decomposition
SVM	Support vector machine
TOF	Time-of-flight
UK	United Kingdom
US	United States

Chapter 1

General Introduction

Lameness is a common health problem for animals, especially dairy cattle, significantly affecting their productivity, welfare, and longevity. It can affect all cattle breeds and at any age, leading to economic losses because of reduced production (Bicalho et al., 2007) and increased culling rates (Booth et al., 2004). In addition to financial losses, lameness is often associated with pain and discomfort in animals (Coetzee et al., 2017), leading to poor welfare and decreased quality of life. Therefore, timely detection and treatment of lameness are critical for welfare and productivity. Early identification and appropriate treatment of lameness can prevent the condition from worsening and reduce the herd's lameness rate (Nicole, 2007; Whay, 1999). Various methods are available for lameness detection, including visual observation, gait analysis, and automated monitoring systems.

Radar technology has presented great potential for animal health monitoring (Busin et al., 2019a; Manteuffel, 2019; Shrestha et al., 2018; Wang et al., 2020). Using radar allows for non-invasive, real-time detection of various physiological parameters such as respiration (Matsumoto et al., 2022) and movement patterns (Busin et al., 2019a). A micro-Doppler radar system could be particularly useful in situations where direct observation is challenging, such as in a herd with a large number of animals. Radar technology and automation can potentially reduce the need for human intervention and thereby minimise animal stress and disturbance and offer benefits in animal husbandry.

This chapter aims to provide a brief overview of the existing literature on lameness in dairy cattle, the traditional ways of detection, and the benefits and limitations of implementing a radar-based sensing system to automate the process, as well as a brief explanation of the machine learning analysis approaches that will follow in the next chapters. The overall objective was to introduce the basic concepts and establish the rationale for this research project and its aims, which were (1) the quantification of the performance of

the micro-Doppler radar sensing method, (2) the characterisation and validation of the micro-Doppler radar signatures of dairy cattle with varying degrees of gait impairment, and (3) the development of machine learning algorithms that can classify the mobility status of the animals from their radar signatures.

1.1 Lameness

Cattle lameness is a clinical disorder or a sign rather than a disease (Adams, 2014) that primarily affects the limbs, resulting in deviations from normal posture, gait, and locomotion (Ross & Dyson, 2010; Stashak, 2008; Weishaupt, 2008; Wyn–Jones, 1988). It is associated with pain and discomfort and compromises other aspects of the animal, such as reproduction (Melendez et al., 2003) and milk yield (Hernandez et al., 2002). However, it is not a new concept as one of the earliest references describing the care of animals related to their feet and the first attempts at animal shoes is the "Hippiatrica," a collection of Greek and Roman veterinary texts dating back to the 4th century AD (McCabe, 2007). A few centuries later, manuals and books (e.g., Hunting, 1895) were written about the care and health of animals' feet, focusing on horses and means of protecting the hooves. More recently, emphasis has been placed on lameness in cattle, and several studies have been conducted on prevention and treatment (Alawneh et al., 2012; Boelling & Pollott, 1998; Leach et al., 2010b, 2010a; Randall et al., 2015, 2016). The focus on animal welfare and lameness stems from the need for animals to serve for more extended periods, have increased production compared to previous years and because society's views on welfare have shifted significantly over time from one of indifference to one that recognizes the importance of treating animals humanely and protecting their wellbeing (Fraser, 2013).

It is essential at this point to introduce the differences between normal gait and lameness. Normal gait refers to the regular walking pattern of an animal. It is characterised by a balanced, rhythmic movement of the limbs, with each foot landing smoothly on the ground and body weight shifting from one side to the other (Alsaaod et al., 2017; Tijssen et al., 2021). A normal gait is often determined by the individual's conformation and ability to maintain balance

and coordination during locomotion. The impact of conformation on an animal's gait is particularly relevant when considering cases of lameness, as underlying conformational abnormalities may cause deviations from a normal gait (Vermunt & Greenough, 1995). Conformation refers to the physical structure and shape of the body, including the skeletal structure, muscle development, and overall body proportions. Conformational abnormalities can sometimes lead to a deviation from a normal gait, but not all deviations result in lameness. For example, some animals may have a conformational abnormality, such as a long or short limb, that causes them to walk with a slightly different gait than normal. This deviation from normal gait may not necessarily cause lameness, as the animal may still be able to move without pain or discomfort. However, other conformational abnormalities can cause lameness or increase the risk of injury. For example, a cow with a rotated hoof may have an abnormal gait (Anees et al., 2022) that causes excessive stress on its joints or muscles, leading to an increased risk of injury and lameness. Understanding the relationships between normal gait, conformation, and lameness is vital for detecting the causes of diminishing welfare.

1.2 Aetiology and consequences of lameness

Lameness is a multifactorial condition, including production, nutrition, genetics, environment, and management practices. To research and mitigate its impact, navigation through complex interactions is necessary. A better understanding of the underlying mechanisms and their relationship with other health and welfare issues can help develop practical research questions and try to minimise the adverse effects.

1.2.1 Economic impact and prevalence

Lameness can be costly depending on the lesion type and severity. The total cost of lameness should be calculated, considering not only the expenses for the treatment but also the losses because of lameness. Several studies are concerned with calculating expenses due to lameness; some include the whole herd (Davis-Unger et al., 2017), and others focus on the individual level,

suggesting that this will help the breeder make better decisions about the welfare of each animal (Cha et al., 2010b). One study that estimated the cost of a generic lameness case found that average producer costs can range from $76 to $533 per cow (Dolecheck & Bewley, 2018). The calculation of costs included veterinary and pharmaceutical fees; the economic loss for the period during which the cow remains out of production; the reduction in milk production, reproductive capacity, in body weight; and finally, the decrease in the cow's economically prosperous life, as it has been observed that lame animals end up in the slaughterhouse faster than non-lame cows (Booth et al., 2004; Randall et al., 2016; Sogstad et al., 2007). Another study (Cha et al., 2010b) reported the average cost per case of sole ulcer to be $216.07, $132.96 for digital dermatitis and $120.70 for interdigital necrobacillosis. The analysis included costs related to milk yield, fertility and treatment.

Prevalence in dairy cattle varies widely depending on the region, herd management practices such as foot bathing and housing condition, and diagnostic criteria. Studies have reported prevalence rates ranging from 1.2% to 60.5% in different countries (Table 1.1). Other factors affecting lameness prevalence rates include implementing prevention strategies, such as regular hoof trimming, providing proper housing and flooring, and controlling infectious diseases, which can help reduce the prevalence of the disease (Carvalho et al., 2005; Eicher et al., 2013; Refaai et al., 2013).

Table 1.1 Selected examples of lameness prevalence worldwide.

Author & Year	Location	Prevalence
(Kielland et al., 2009)	Norway	60.5% +/- 21.2%
(Von Keyserlingk et al., 2012)	North America (British Columbia, California, North-eastern US)	27.9% +/- 14.1% (British Columbia), 30.8 +/- 15.5% (California), 54.8 +/- 16.7% (Nort-east US)
(Sarjokari et al., 2013)	Finland	21%
(Brenninkmeyer et al., 2013)	Germany & Austria	50%
(Pérez-Cabal & Alenda, 2014)	Spain	13.8%
(Chapinal et al., 2014)	China	31% +/- 12 (range=7–51)
(Solano et al., 2015b)	Canada	21% (range 0 – 69%)
(Foditsch et al., 2016)	New York	14%
(Rashad et al., 2022)	Egypt	0 - 19%
(Fabian et al., 2014)	New Zealand	1.2 – 36%
(Griffiths et al., 2018)	England and Wales	31.6% (range 5.8 to 65.4%)

1.2.2 Welfare

Lameness in dairy cattle has been identified as a significant welfare concern due to the potential pain and discomfort it can cause the animals (Shearer et al. 2013) and the associated limitations on their mobility and access to food and water (Morton & Griffiths, 1985). The access limitations could potentially lead to a decline in body condition and weight loss (Huxley, 2013). Additionally, limited mobility increases the risk of accidents and further injuries (Van Der Tol et al., 2003). The overall behaviour of the cow is also affected, as evidenced by reduced activity levels and increased lying down durations (Ito et al., 2010; Westin et al., 2016). The five freedoms (Table 1.2), which provide

a framework for assessing the welfare of intensively farmed animals, emphasise the importance of ensuring that animals are free from pain, injury, and discomfort and can express their most normal behaviours (Farm Animal Welfare Council, 2009).

Table 1.2 Five freedoms of intensively farmed animals' welfare.

Freedom	
Freedom from hunger or thirst	by ready access to water and a diet to maintain health and vigour.
Freedom from discomfort	by providing an appropriate environment including shelter and a comfortable resting area.
Freedom from pain, injury, or disease	by prevention or rapid diagnosis and treatment.
Freedom to express (most) normal behaviour	by providing sufficient space, proper facilities and appropriate company of the animal's own kind.
Freedom from fear and distress	by ensuring conditions and treatment, which avoid mental suffering.

1.2.3 Milk production

Multiple studies have examined the link between lameness and milk yield in dairy cows, with varying results. While some studies have reported a reduction in milk yield following the diagnosis of lameness (Green et al., 2002; King et al., 2017; Olechnowicz & Jaśkowski, 2010; Warnick et al., 2010), the exact impact remains difficult to estimate accurately. Treatment of lameness has been observed to result in a decrease in milk yield, with reported milk loss ranging from 160 to 550 kg over a lactation period (Green et al., 2002). Other studies have reported losses of up to 424 kg per cow over a 305-day lactation period (Bicalho et al., 2008) and up to a 10% reduction in milk production compared to non-lame cows (Hernandez et al., 2002).

Several studies have investigated the genetic associations between milk production and lameness in dairy cattle. One study by Salleh et al. (2017) found a genetic correlation of 0.27 between 305-day milk yield and lameness. Another study by König et al., (2008) looked at the genetic relationships between four claw disorders and milk yield before and after diagnosis. The results showed a positive relationship between all claw disorders and milk yield, with estimates ranging from 0.08 to 0.44.

1.2.4 Fertility

Lameness has a profound impact on the productivity of dairy cattle, particularly on reproductive performance. It affects animals of all ages and lactation periods, but with a relatively higher incidence of lameness in early lactation and the dry period (Bicalho et al., 2007; Blowey, 2005; Calderón-Amor et al., 2021; Daros et al., 2019). Studies have reported a significant increase in pregnancy loss in lame Jersey cows (11%) compared to healthy (5%) (Omontese et al., 2020). Lame cows have a reduced ability to conceive, with first-service conception rates reported to be as low as 20%, whereas healthy cows range between 40 and 50% (McNally et al., 2014; Melendez et al., 2003; Omontese et al., 2020). Lame cows have also been observed to be mounted less frequently and express fewer signs of oestrus (Walker et al., 2010), and are more prone to delayed cyclicity, lower ovulation rates, and decreased conception rates (Garbarino et al., 2004; Melendez et al., 2018). The duration of lameness has also been found to have a linear relationship with the odds of metritis, with cows being chronically lame during the dry period to be more susceptible (Daros et al., 2020). All these findings underscore the negative impact of lameness on reproductive performance, emphasising the importance of prompt intervention.

1.2.5 Culling rates

Lameness has been associated with increased culling rates depending on the time of lameness detection and the time of culling (Booth et al., 2004; Cramer et al., 2009; Randall et al., 2016; Sogstad et al., 2007). Booth et al. (2004), reported that lameness was associated with a short-term increase in culling rate during early lactation (between 61 and 120 days in milk) and towards the end of lactation. The debilitating effect of lameness on reduced milk yield and fertility likely contributed to this increase. Interdigital necrobacillosis and sole ulcers were also found to have a negative impact on cow survival, with interdigital necrobacillosis having the greatest effect when diagnosed between 61 and 120 DIM and culling occurred during the same period. However, in this study, no associations were found with other lesion types, such as digital dermatitis. Further studies in the literature presented similar results (Dohoo & Martin, 1984; Rajala-Schultz & Gro Èhn, 1999), but there were also studies with mixed (Collick et al., 1989; Milian-Suazo et al., 1989) or no effects (Beaudeau et al., 1994) of lameness on culling rates. The way of analysis, the models used, and the coefficients included possibly account for the differences. Nonetheless, there is no study showing a negative correlation between lameness incidents and culling rates.

1.3 Factors affecting lameness

Numerous biological, environmental, and management-related features have been identified as potential risk factors for lameness in dairy cattle. These factors can interact with each other, making the assessment of their individual and combined effects challenging. Understanding the aetiology of lameness in dairy cattle would benefit its detection, monitoring, and treatment.

1.3.1 Lesions and disorders

Several types of lesions (Table *1.3*) can be responsible when it comes to the complex of disorders that can cause lameness in cattle. These include infectious and non-infectious lesions on any part of the animal's leg. However, lameness is usually associated with lesions in the animals' hooves (Archer et al., 2010), especially in the hind legs (Clarkson et al., 1996a). Most lesion types are located in the hooves because the hooves are the weight-bearing structures that support the entire body weight of the cow (Shearer & Van Amstel, 2001). The hooves comprise a horned outer layer and a sensitive inner layer called the corium (Figure *1.1*), which contains numerous blood vessels and nerves, and any damage or inflammation to the corium or the horn of the hoof can cause lameness and compromise the cow's ability to walk and stand (Amstel & Shearer, 2008). Additionally, hooves are constantly exposed to various environmental factors, such as mud, manure, and rough surfaces, which can lead to injuries, infections, and other hoof problems.

Some of the most common hoof lesions causing lameness in dairy cattle are claw horn lesions, often caused by improper hoof trimming or inadequate housing conditions (Bergsten et al., 1998; Manske et al., 2003) . Two common non-infectious claw horn lesions are sole ulcers and white line disease (Archer et al., 2010). Sole ulcers are typically caused by excessive pressure on the sole, leading to erosion and inflammation (Bonser et al., 2003; Gregory et al., 2006). White line disease in cattle occurs when the inner layer of the hoof wall, becomes separated from the sole of the hoof. This separation creates a space where bacteria and fungi can grow, leading to infection and deterioration of the white line structure (Shearer & van Amstel, 2017b). The white line serves as a weight-bearing region for the animals and is particularly vulnerable to damage or disease due to its location (Figure *1.1*) and the stresses it undergoes. These conditions can also be predisposed by metabolic disorders, such as rumen acidosis and laminitis, along with physiological changes during the transition period (Shearer & van Amstel, 2017a).

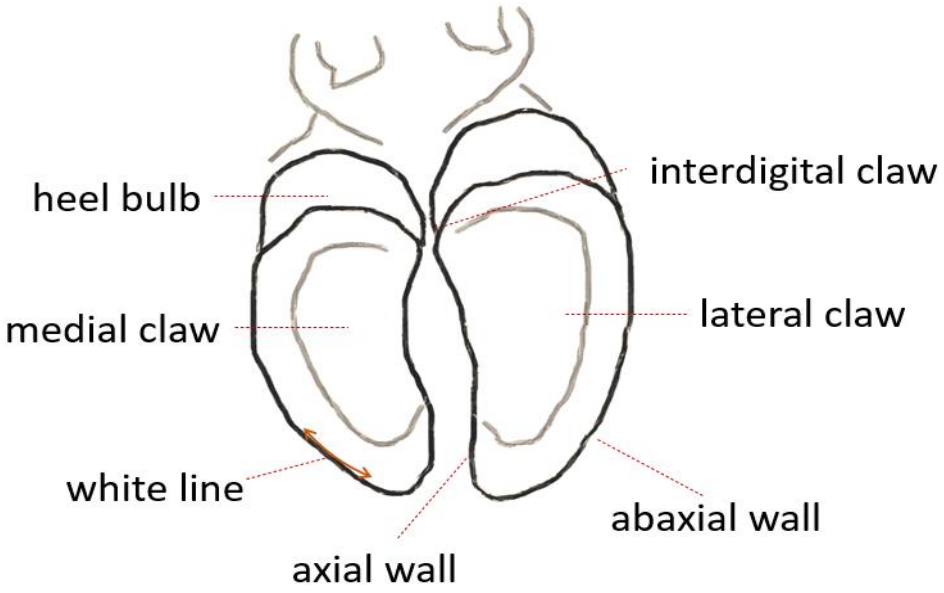


Figure 1.1 Plantar view of hoof anatomy.

Lameness in dairy cattle can be caused by contagious infectious agents too. Among these, digital dermatitis and interdigital necrobacillosis are commonly observed in cattle. Digital dermatitis is a bacterial infection that causes pain and inflammation on the skin of the foot (Palmer & O'Connell, 2015). The disease is primarily seen in the hind feet, particularly between the heel bulbs proximal to the interdigital cleft, and appears as erosive, circumscribed areas of inflammation and epidermal proliferation. The bacterial aetiology of digital dermatitis is complex, and various microorganisms have been detected in the lesions (Beyi et al., 2021; Marcatili et al., 2016; Wilson-Welder et al., 2015). Treponema spp. are consistently observed in digital dermatitis lesions and are present in large numbers at the interface between necrotic and healthy tissue. Interdigital necrobacillosis is another bacterial infection that affects the soft tissues of the foot (Berry, 2001). The condition is caused by various bacteria, including *Dichelobacter nodosus*, *Fusobacterium necrophorum*, and *Bacteroides melaninogenicus*. Wet and muddy conditions are known to predispose cattle to interdigital necrobacillosis (Monrad et al., 1983), as the bacteria responsible for this infection thrive in these environments. Interdigital

necrobacillosis is characterized by severe inflammation and necrosis of the soft tissues of the foot, which can lead to severe lameness (Archer et al., 2010).

Over 90% of lameness-causing injuries may originate in the lower leg (Clarkson et al., 1996b); however, dysfunctions in the animal's spine, nerves and musculoskeletal system can also affect movement and cause lameness (Callan & Garry, 2001; Shearer et al., 2012). A few orthopaedic disorders affecting the upper limb and causing lameness are carpal hygromas, tarsal cellulitis and hip dislocation (Chhatpar et al., 2012; Marchionatti et al., 2014; Nocek, 1997). In these cases, the detection of lameness is immediate with the appearance of the problem since it is in a prominent place and dramatically changes the animal's gait pattern.

Table 1.3 Common hoof lesions in cattle.

Lesion	Description
Sole ulcer	Lesion located typically in the area between the sole and heel bulb at the rear of the pedal bone caused due pressure on the corium - non-infectious (Shearer & van Amstel, 2017b)
Sole haemorrhages, bruises, hoof discolouration	Red and yellow marks present on the sole areas caused due to pressure on the corium, leading to inflammation - non-infectious (Archer, Bell, Huxley, et al., 2010)
Fissures, cracks, separations	Horn defects affecting the wall horn - non-infectious (Clark et al., 2004)
White line disease	The sole separates from the wall of the hoof and foreign materials penetrate and infect the area - non-infectious (Shearer & van Amstel, 2017b)
Under-run sole	A layer of sole, over a layer of keratinised sole caused by interruption of sole horn formation, followed by restoration of horn production - non-infectious (Nocek, 1997)
Digital dermatitis	Skin infection caused by bacteria between the heel bulbs or palmar/plantar pastern area – infectious (Afonso et al., 2021)
Interdigital necrobacillosis (Foul in the foot or Footrot)	The skin between the claws becomes damaged and bacteria infect the soft tissue between the digits causing swelling and necrosis – infectious (Van Metre, 2017)
Interdigital hyperplasia/growth	Mass of tissue between the claws, the degree of lameness depends on the size of the lesion and the presence of infection of the digital tissue (Alsaaod et al., 2023)

1.3.2 Hoof trimming

The hooves of cattle are complex structures that support the animal's weight and absorb the impact of movement on hard surfaces (Blowey, 2015). Hoof trimming in cattle is a critical management practice that helps prevent or detect early lameness incidents (Bergsten et al., 1998), as the trimmer can inspect the claws for signs of lesions and take appropriate action to prevent further damage and promote healing. However, only in recent years has the claw trimming profession been licenced by organisations such as the UK's National Association of Cattle Foot Trimmers (NACFT).

Studies have found associations between frequent hoof trimmings and reduced lameness prevalence (J. A. Hernandez et al., 2007; Manske et al., 2002b; Stoddard & Cramer, 2017). In a survey (Russell et al., 1982) on the incidence of lameness, they reported claw lesions in animals with irregularities in the shape of claws. They established a correlation between claw conformation and lameness occurrences. The findings of another study (Boettcher et al., 1998a) appear to agree that the shape of the claw can contribute to lameness susceptibility. Neichev et al., (1980) confirmed the positive effects in daily milk production when cows with overgrown hooves were trimmed and the reduction in the daily milk yield in the opposite case, where cows were left with overgrown hooves (Diaz & Bodurov, 1986). A negative correlation has also been mentioned between trimming and slipping, with a study (Phillips et al., 2000) proving that recently trimmed cows are less likely to slip. Several more studies (Andersson & Lundström, 1981; Boelling & Pollott, 1998; Manson & Leaver, 1989) have been conducted, and the results reveal the importance of the hoof size and shape in mobility improvement.

Trimming frequency has been the subject of a few studies. Although some have indicated that hoof trimming before dry-off (Thomsen et al., 2019) or during early lactation (Pedersen et al., 2022) may lower the likelihood of lameness, the findings are inconclusive. According to Shearer et al. (2001), hoof trimming once to twice per year is beneficial for the cattle, and in cases of clinical lameness, it might be necessary more frequently. In smaller ruminants such as

sheep and goats, routine hoof trimming is also suggested to be performed twice a year, but not more often if there is no reason to do so, as it is possible to increase the risk of lameness problems in the herd (Wassink et al., 2003). Other studies such as this by Maxwell et al., (2015) have suggested that trimming only lame first-lactation heifers between 50 and 80 days in milk (DIM) resulted in higher milk production relative to non-trimmed heifers. This approach was also deemed to be more cost-effective. Unfortunately, no specific evidence indicates the optimal frequency of claw trimmings. Most of the studies, though, agreed that routine inspection is vital for lameness control (Pedersen et al., 2022; Sadiq et al., 2020; Shearer & Amstel, 2001). The hoof trimming should comply with the farm's management system and the animals' needs (Manske et al., 2003). Some suggested periods, which tend to be the most effective in terms of overall management, according to studies, are the mid-lactation and the dry period because they will have less impact on the productivity and well-being of the animal from a pain and stress perspective (Mason & Offer, 2007; Shearer & Amstel, 2001). This is because if the animals are not accustomed to being handled or restrained, trimming can release stress hormones such as cortisol, which can adversely affect overall health and well-being. Finally, regular inspections do not necessarily mean that claws must be trimmed. The main aim should be balancing the weight-bearing among the feet and claws (Raven, 2003). Attention should be given during the trimming to not remove excessive tissue (over-trim) as it could result in injuries to the animal’s hoof and lameness (Raven, 2003; Reicher, 1985).

1.3.3 Space, Waking area and Bedding

Healthy cows typically spend about 7 to 14 hours in a lying position daily (Cook et al., 2004a; Ito et al., 2009; Jensen et al., 2005), including approximately 4 hours sleeping in small bouts throughout the day and 8 hours drowsing under normal conditions (Kull et al., 2019; Ternman et al., 2012). Thus, their living conditions should provide the necessary comfort and enable normal behaviours. When stalls are uncomfortable, animals tend to decrease their rest time, which may be a contributing factor to lameness (Ceballos et al., 2004). Several studies have been conducted on floor suitability (Flower et al., 2007; Phillips & Morris, 2000; Telezhenko et al., 2009; Telezhenko & Bergsten, 2005)

and its association with lameness occurrence. A few housing options for dairy cows include indoor free-stall or tie-stall barns, bedded pack barns and outdoor pasture-based systems. In free stalls, animals are housed in straw yards where they rest together or in cubicles with individual places. There is also a tie-stall system, an older animal tether, but it is considered to be opposed to animal welfare.

Research has demonstrated the impact of different living conditions on animal health and behaviour. Hernandez-Mendo et al. (2007) and Somers et al. (2003) found that despite a reduction in nutrient intake that resulted in lower milk production, animals that grazed on pasture showed improvements in hoof health and gait. In a study by Frankena et al. (2009), gait disturbance was reduced in straw yards compared to cubicles in free-stall barns. However, other studies (Hultgren, 2002; Webster, 2001) have suggested that cubicles can increase clinical foot disturbances, resulting in smooth but thin heels that exacerbate claw horn lesions over time. Finally, research on tie stalls has highlighted the challenges that animals face in changing positions from lying to standing, as documented by Haley et al. (2000).

Research has shown that the type of flooring on which animals walk can impact their locomotion and health. Phillips & Morris (2001) found that cows on slippery floors take frequent, small steps to maintain speed, while those on floors with more friction take longer steps. Concrete floors have been linked to adverse effects on limb health and physiology. Studies (Cook & Nordlund, 2009; Murray, Russell, et al., 1996; Phillips & Morris, 2000; Rushen et al., 2007; van der Tol et al., 2002) have shown that concrete flooring reduces walking speed, alters limb angles, increases claw exposure, and raises the risk of lameness, leading to incidents of swelling of the carpal joints and negative changes in walking patterns.

In contrast to concrete, rubber mats have emerged as a more suitable flooring option. Studies on rubber flooring have demonstrated its positive effects on claw health and heel erosions (Boyle et al., 2007; Cook & Nordlund, 2009). Furthermore, cows have been observed to exhibit a preference for spending more time lying on soft ground than on concrete (Rushen et al., 2007). Rushen et al. (2007), in their study on the impact of a softer flooring system on leg

injuries, highlighted that cows find it easier to switch between lying and standing postures on softer surfaces than on other types of flooring.

Sand floors and mattresses in cubicles have been suggested as beneficial for lame cows, promoting recovery and reducing hock injuries, according to various studies (Bergsten & Telezhenko, 2005; Cook & Nordlund, 2009; Livesey et al., 2002). Rubber mats, negatively correlated with lameness appearance, can also be a suitable option for cows. Studies (Cook, 2003; Cook et al., 2004b) have shown that cows appeared comfortable on soft rubber bedding, with stride length and movement speed similar to those on pastures, as Jungbluth et al. (2003) observed. However, the preference for a specific type of floor does not necessarily mean that other types are unsuitable for cows, as noted by Cook et al. (2004b).

The type of flooring is not the only factor affecting cow welfare, as the amount of space they have to move around in is also crucial. Research (DeVries et al., 2004; Huzzey et al., 2006) has shown that aggression is linked to pen dimensions, feeding space per animal, and access to outdoor areas. When feeding space is reduced, aggression can increase significantly; for example, reducing manger width by 0.5m per cow doubled fighting between cows (DeVries et al., 2004). Narrow alleys and limited space allowances can also provoke aggression and cause injuries that can lead to lameness (EFSA, 2009). The optimal space for dairy cows depends on factors such as horn status, body size, and herd size, as overcrowding reduces lying times and can increase lameness (Menke et al., 1999; Rowlands et al., 1983). Neck-rail position and pen floor quality also play a significant role in lameness; inadequate floor quality coupled with a high neck-rail position can cause cows to spend more time standing, increasing the risk of lameness. However, when the diagonal between the neck rail and the rear border of the bed is over 1.95m, the risk of lameness is significantly reduced (Mülleder et al., 2004), in agreement with previous research (Murray, Russell, et al., 1996).

1.3.4 Nutrition

Nutrition plays a vital role in the proper functioning of the body, the immune system and the treatment of diseases. Several studies (Cook, 2014; Laven, 2006; Lean et al., 2013; Manson & Leaver, 1988c; Smart, 1985; Westwood et al., 2003) have been conducted over the years on the relationship between diet and lameness.

A diet deficient in essential nutrients, such as protein, vitamins, and minerals, and trace elements can lead to poor bone development, decreased muscle mass, and impaired immune function, all of which can contribute to lameness. For example, copper deficiency has been associated with stiffness and lameness in calves (Smart, 1985), while cows with a history of lameness were found to have elevated levels of copper (Baggott et al., 1988). Zinc and phosphorus imbalances have also been linked to an increased risk of lameness, as these elements have been associated with sensitive limbs, more easily injured, and the hoof growth is abnormal (Smart, 1985; Underwood, 1971). Furthermore, the excessive use of certain chemical elements, such as fluorine and selenium, can interfere with the mineralization process of bones or lead to oxidative damage and cell death in joints, leading to weakened bones and joints that will make the animal more susceptible to lameness (Howell, 1983). Specific vitamins administration, such as biotin, can provide benefits in cows' health and nutritional balance. Incorporating biotin into an animal's diet can aid in preventing lameness by maintaining keratin, as documented by Mülling et al. (2006). Moreover, Hedges et al. (2001) and Pötzsch et al. (2003) have highlighted the significance of biotin in reducing lameness, particularly in cases induced by white line disease.

Another nutritional factor associated with lameness in dairy cattle is subacute ruminal acidosis (SARA). SARA is a condition that occurs when the pH drops to less than 5.8 for more than 330 minutes per day (Zebeli et al., 2008). The severity of SARA can be determined by how frequently this drop in pH occurs (Plaizier et al., 2008). The acidic environment can lead to the breakdown of the rumen epithelial barrier, allowing endotoxins to enter the bloodstream and the small blood vessels in the claws and lead to inflammation, circulatory disturbances, and ischemia (Danscher et al., 2010; Ossent & Lischer, 1998).

This inflammation could possibly damage the hoof and other tissues, causing haemorrhages and bruising of the corium, particularly underneath the flexor tubercle leading to lameness (Greenough, 2007). However this hypothesis has not been proven.

Managing the feeding schedule and ensuring adequate nutrition for dairy cattle can prevent malnourishment and minimize the risk of lameness. Bicalho et al. (2009) observed a positive correlation between body weight and digital cushion thickness, emphasizing that cows with low body weight are more likely to develop lameness. Another study by Donovan et al., (2004) reported that sudden increases in energy intake during sensitive calving periods (i.e., just before or after calving) may also lead to lameness. Several studies have highlighted that increasing the frequency of feedings can lead to animal competition, reduced feed intake time, and increased standing time, ultimately leading to foot injuries and lameness (Huzzey et al., 2006; Olofsson, 1999; Proudfoot et al., 2009) .

1.3.5 Genetics

Genetic selection for desirable traits, such as increased milk yield and meat quality, has inadvertently increased the incidence of lameness (Van Marle-Köster & Visser, 2021). To reduce lameness in their herds, farmers have traditionally chosen indirect traits like conformation as a selection method (McDaniel, 1997). However, more recent research has shown that direct health traits such as foot lesion records are more valuable and practical for genetic selection (Egger-Danner et al., 2015). Studies have suggested that susceptibility to lameness has a substantial genetic component (Boettcher et al., 1998b; Buch et al., 2011; Häggman et al., 2013; Heringstad et al., 2018; Huang & Shanks, 1995b; Koenig et al., 2005; Malchiodi et al., 2017; Ødegård et al., 2013; Onyiro et al., 2008; Sánchez-Molano et al., 2019; van der Spek et al., 2013), with estimated heritability between 0.01 and 0.35. Although the results vary, candidate genes such as the OSR1 gene on BTA-11 that has been linked to conformation traits (Cole et al., 2011), the VWF gene on BTA-5 to foot angle (Kolbehdari et al., 2008), and BTA8 and BTA13 regions have been linked to sole ulcers and white-line disease with the candidate genes involved

in wound healing, bone growth, adipose tissue, and keratinization (Lai et al., 2021). Inflammation response, immune function, and bone growth are also lameness contributors and can be genetically selected for prevalence reduction. Selection can also affect the recovery and healing process after lameness development as some animals might have a better hoof growth rate and quality of the new hoof tissue than others (van der Spek et al., 2013).

Other studies have also concluded that breeds and genetic predisposition contribute to lameness incidence. For example, a study (Huang & Shanks, 1995a) on factors affecting hoof health found that some breeds presented better claw score characteristics than others. In other studies, herds consisting of only Holstein Friesian had an increased risk of lameness (Barker et al., 2010). Brown Swiss cows had the most severe results for corkscrew claws, laminitis and sole ulcers. Guernsey had the worst scores for the white lines and heel erosion. The incidence of digital dermatitis was least favourable in Friesians; last, Jerseys cows had harder feet and were less lame (Chesterton et al., 1989).

Although the genetic basis of lameness is complex, with multiple genes and environmental factors involved, farmers must consider both the prevention and treatment of lameness through genetic selection to ensure the long-term health and productivity of their herds.

1.3.6 Animal, Social and Behavioural Factors

Lameness in cows is associated with certain criteria that determine the well-being of the animal, including the score of limbs and feet, hoof angle, and leg set. According to a study by Pérez-Cabal et al. (2006), cows with better conformation and structure in legs and hooves tend to be more productive and have longer lifespans than other cows. One more factor linked to lameness by Chesterton et al. (1989) is the colour of the claw. They described that animals with light-coloured feet were more likely to suffer from lameness. Other than the claws' colour, an animal's weight can also indicate its susceptibility to lameness. Specifically, according to studies by Boettcher et al. (1998a) and Gudaj et al. (2012), heavier cows are more likely to develop clinical lameness. However, cattle with low body condition scores are also likely to develop lameness, as noted in studies by Randall et al. (2015) and Wells et al. (1993).

Essentially, both underweight and overweight cows are at risk for lameness, highlighting the importance of maintaining appropriate body conditions in cattle to prevent lameness incidents.

Age is also an animal factor contributing to lameness occurrence, according to studies which explain that the risk of a lame animal increases with age (Dembele et al., 2006; Huang et al., 1995; Solano et al., 2015b). As cows age, the cumulative effects of previous injuries and the wear and tear on their joints can possibly lead to lameness. However, some studies suggest that young cows may be at higher risk of developing certain types of lameness. For example, digital dermatitis, a common claw disorder, tends to occur more frequently in younger cows (Manske et al., 2002a; Sogstad et al., 2005). The reasons for this are not entirely clear, but it may be related to the fact that young cows are still growing and developing, and their hooves are more susceptible to trauma and injury. Overall, while age can be a factor in the development of lameness, the specific risk factors can vary depending on the type of lameness and other environmental factors.

The social dynamics within a herd can also play a role in the health of animals' limbs. For example, studies have shown that cows with higher hierarchical status are less likely to experience lameness than those with lower social standing (Olechnowicz & Jaskowski, 2011). Furthermore, cows that develop lameness can become marginalised by healthier cows in the herd, leading to further health issues and discomfort for the affected animal (Galindo & Broom, 2002). This suggests that animal behaviour can contribute to lameness and vice versa. However, providing enough space for animals to engage in their natural behaviours can help mitigate this issue (DeVries et al., 2004; Huzzey et al., 2006).

1.3.7 Footbaths

Lameness studies (Dolecheck & Bewley, 2018; Randhawa et al., 2008) have highlighted that the treatment costs of a lame cow are far more than implementing a simple prevention protocol such as footbaths. Footbaths are a preventative measure commonly used in the dairy industry to reduce the incidence of lameness in cattle. The process involves immersing the cow's

hooves in a solution containing disinfectants or other agents that help control bacteria and fungi growth and thus reduce the infectious lesions such as digital dermatitis (Robcis et al., 2023) and decrease up to 50% other lameness incidents in the herd such as digital skin lesions with regular footbathing (Randhawa et al., 2008). It is a cost-effective and straightforward method for improving the health and welfare of dairy cattle, with benefits such as the reduction of antibiotic treatments, which can have implications for antimicrobial resistance in animals and humans (Schwarz & Chaslus-Dancla, 2001; K. E. Walker et al., 2023).

1.4 Assessment and diagnosis of lameness

1.4.1 Gold standard and methods of lameness detection

The gold standard for cattle lameness detection is a trained and experienced veterinarian conducting a thorough clinical examination of the animal (Desrochers et al., 2001). The veterinarian will visually assess the gait and posture of the cow, as well as palpate the legs and feet to identify any areas of tenderness, swelling, or other abnormalities. In addition to the visual and manual examination, the veterinarian may also use diagnostic tools such as hoof testers, joint flexing, radiographs, or ultrasound to evaluate the animal's condition further. However, due to the large numbers of farm animals, this approach to detecting lameness is not economically beneficial or logistically practical. This is a key reason why other indirect assessment methods have been proposed and have been widely adopted for lameness detection.

The most common method of lameness detection in herds of cattle is visual mobility scoring of the animals (Afonso et al., 2020). This method involves observing the animal’s gait, posture, and behaviour to determine any signs of abnormal mobility. An assessor, usually trained personnel, veterinarian or farmer, watches the animal as it walks or stands, looking for any signs of uneven or abnormal movement, such as limping, favouring a particular leg, uneven weight distribution, or arched back. Assessors may also look for other signs of lameness, such as changes in the animal’s overall posture and mobility, reluctance to move, or visible swelling or inflammation in the affected area.

Finally, a score is given to each animal that meets a category of the severity of impairment or describes specific characteristics that define a lameness case.

1.4.2 Scoring systems

Visual locomotion scoring systems are an indirect method to assess hoof and foot conditions because it relies on the observation of the gait and behaviour rather than direct examination of the feet and legs and are considered subjective as the evaluation depends greatly on the assessor (Van Nuffel et al., 2015). Lameness scoring systems are generally based on numerical scales ranging from 3 to 9 levels (Lorenzini et al., 2017; Manson & Leaver, 1988c; Sprecher et al., 1997; Tranter & Morris, 1991; Wells, Trent, Marsh, & Robinson, 1993; H. Whay, 2002; Winckler & Willen, 2001). These scores indicate the animal's mobility, with lower scores typically denoting healthy mobility and higher scores indicating severe lameness. The scoring involves the assessor making observations of the animal's locomotion and making a classification decision. Different scoring systems may employ varying scales or assign scores based on specific criteria, such as the presence of head nodding, uneven weight bearing, or stiffness in the animal's gait. Despite their differences, mobility scoring systems share some commonalities. For example, they all rely on the subjective observation of the animal's gait and behaviour and are designed to identify the degree of lameness in the animal and aid the diagnosis of the problem. Some systems are more widely adopted than others, potentially because they are simple and easy to understand or because of sector support and industry initiatives, with organizations or associations providing educational resources and training programs (Main et al., 2012). In addition to the original systems for evaluating cow mobility, various modifications have been proposed in studies to make the systems more reliable and repeatable (Haskell et al., 2006; Rajkondawar et al., 2006; Winckler & Willen, 2001). A brief description of a few of the most commonly used mobility scoring systems (Afonso et al., 2020) in research follows.

The locomotion scoring system of Manson & Leaver, (1988b) suggests the scoring to be carried out before the milking on a concrete floor with the animals being scored walking away from the assessor for 5 to 10 meters. The

system has nine levels ranging from 1 to 5 with 0.5 step, and the higher the score, the poorer the cow's locomotion. Animals with a score > 3 are considered clinically lame and should be examined. This system is frequently used in literature, but complexity can arise from its multilevel nature (Channon et al., 2009a). Half of the scores relate to changes in gait and behaviour that precede clinical lameness, intending to detect lameness signs before they become clinically evident. Nonetheless, some definitions can be challenging to comprehend (e.g., "behaviour pattern affected" or "adverse effects on behaviour pattern"), necessitating extensive training. Whay (2002) recommended that an individual with prior knowledge and experience should instruct others to use the system.

The system proposed by Whay et al. (1997) was first introduced for heifer locomotion assessment and consists of 6 levels ranging from 1 to 6; score 1 stands for healthy cow, and score 6 indicates a cow "as lame as possible while upright". It was conducted on a flat concrete floor, with the animal being assessed first while walking away from the assessor and then from the side of the walking animal tracking the hooves' placement and head movement. The system offers a score (score 2 = imperfect locomotion) for cows with an abnormal gait but without apparent lameness signs, indicating distinct concepts in animal movement deviations and disorders, with an abnormal gait not necessarily translating as lameness.

Sprecher et al. (1997) lameness scoring system focused on posture and gait. It has five levels ranging from 1 standing for normal to 5 severely lame with "inability or extreme reluctance to bear weight on one or more of the limbs" Cows with scores of 4 or 5 are recognised as clinically lame. It was the first system to introduce back posture into the assessment criteria, and they linked the observation of the arched-back posture with future reproductive inefficiency.

The Agriculture and Horticulture Development Board (AHDB, 2015) dairy mobility scoring system is the U.K.'s most commonly used system for detecting lameness and assessing cow mobility. It was developed in 2007 and was called the "mobility scoring system" because the term "lameness" was not popular with farmers and was perceived negatively by the industry (Bell & Huxley, 2009). It consisted of 4 levels, detailed in Table 1.4.

Table 1.4 Description of the AHDB mobility scoring system

Category of score	Score	Description
Good mobility	0	Walks with even weight bearing and rhythm on all four feet, with a flat back Long, fluid strides possible
Imperfect mobility	1	Steps uneven (rhythm or weight bearing) or strides shortened; affected limbs or limb not immediately identifiable
Impaired mobility	2	Uneven weight bearing on a limb that is immediately identifiable and/or obviously shortened strides (usually with an arch to the centre of the back)
Severely impaired mobility	3	Unable to walk as fast as a brisk human pace (cannot keep up with the healthy herd) signs of score 2

Grimm & Lorenzini (Lorenzini et al., 2017) is one of the most recently proposed hierarchical locomotion scoring systems. It consists of 3 levels ranging from 1 to 3. Score 1 stands for a sound cow, score 2 indicates an unsound cow and score 3 represents a lame animal with an irregular, uneven and asymmetric gait. It has only a few basic levels that make the assessment straightforward while separating animal mobility into three exclusive and exhaustive classes (Figure *1.2*).

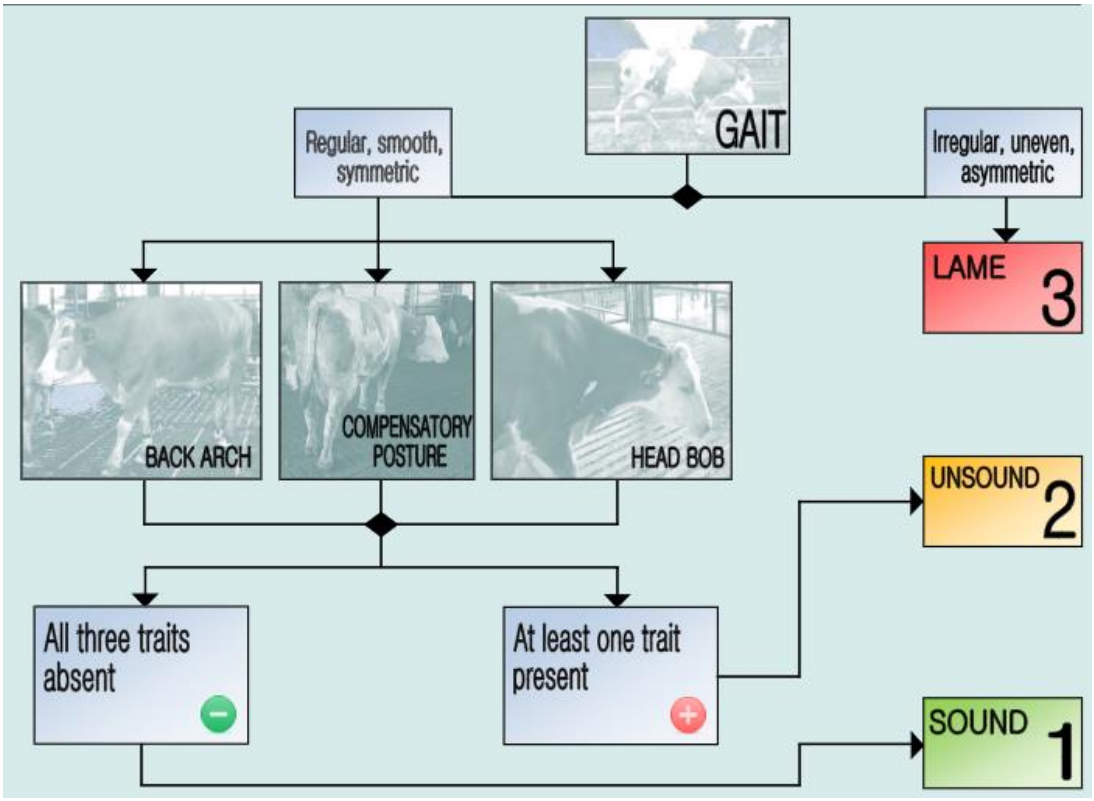


Figure 1.2 Three point locomotion score by Grimm and Lorenzini. From "Using a three point lameness scoring system combined with a clinical examination to increase the reliability of locomotion scoring" by Lorenzini et al., (2017).

1.4.3 Features and characteristics of the mobility scoring systems

Table *1.5* shows some of the most common mobility systems' features. All the features mentioned in the various systems are characteristics that a lame animal can exhibit. However, not all cows express the same characteristics when they are in pain (Reinemann, 2007). Also, lesion severity and location affect the way the cows move differently (H. R. Whay et al., 1997).

Table 1.5 Animal characteristics used in locomotion scoring systems, a short description and a few indicative systems that consider these characteristics.

Characteristics	Explanation	Indicative systems
Arched back	Cow is shifting weight to relieve pressure on one or more limbs	AHDB mobility (2007), Flower and Weary (2006)
Head bob	Indicates that a cow is experiencing pain or discomfort when walking. It can also be a sign of balance issues or a neurological problem	Flower and Weary (2006), Thomsen (2008)
Weight distribution (Reluctance to bear weight)	Uneven weight distribution can be a sign of lameness or other health issues, such as mastitis or a reproductive problem	Manson and Leaver (1988), Sprecher (1997)
Stride length (Short steps)	Shortened stride can be a sign of lameness or pain	AHDB mobility (2007), Sprecher (1997)
Limb placement (Asymmetric gait, limbs' abduction or adduction)	A lame cow may drag its hooves to compensate for lameness problems in an uncoordinated and unbalanced manner, with each limb landing unevenly on the ground	Manson & Leaver (1988), Flower & Weary (2006), AHDB

1.4.4 Pain assessment & limitations of the visual assessment systems

Studies have investigated farmers' and veterinarians' attitudes toward pain in cattle, finding that there is generally a high level of pain recognition but variation in perceptions, with most people not perceiving the same conditions as painful (Huxley & Whay, 2006; Laven et al., 2009; Remnant et al., 2017; Thomsen et al., 2012; Whay & Huxley, 2005). Some methods of pain assessment are more objective than others. For example, measuring variations in heart rate, cortisol levels, and respiratory rate are physiological parameters that can

contribute to a better understanding compared to measures such as vocalisation or facial expressions, that may be objectively quantifiable variables but their association to pain is ambiguous (Adriaense et al., 2020). Some methods considered more thorough, and in case of lameness, examining each individual cow by lifting the legs, trimming, and testing the hooves with either knives or pincers gives a more definitive picture of whether an animal has a lesion in the area being examined compared to solely gait visual assessment. Another reason this type of examination could be more effective in detection is that not all problems manifest with an immediate change in the animal's mobility. Studies have shown that cows will not change their gait despite experiencing discomfort from certain lesions, unless the energy expenditure required for such changes is significant lower than maintaining their normal gait. (Tadich et al., 2010a).

Observer's characteristics influence the mobility scoring outcomes. A study by Polderman et al. (2001) has demonstrated that assessor training can impact scoring and lead to variations in cows' mobility status between assessors. However, other studies in medical fields (Engel et al., 2003; Ford et al., 2000; van Tubergen et al., 2003) have contested these findings and have shown that training does not necessarily influence inter-rater agreement, suggesting that the consistency and production of objective scoring are not affected. Other observer characteristics, such as experience, have been investigated in the evaluation of cow body condition (Kristensen et al., 2006) and found that experienced evaluators are more likely to give scores with greater consistency among themselves (kappa >= 0.86) than non-experienced assessors. This study suggested that if multiple raters with varying experience levels are used, a valid but imprecise estimate of the actual population mean can be obtained. However, in another study, Garcia et al. (2015a) argued about the role of experience in relation to the interrater agreement. Their research found that even inexperienced assessors can achieve high agreement between their scores when evaluating cow mobility via video.

1.4.5 Technologically assisted & automated mobility assessment

The growth in the number of farm animals, the increased demand for dairy products, the lack of time for farmers to monitor all the animals, and the need for objective and valid results are some of the reasons that make it necessary to develop technological systems for automating procedures such as lameness detection. In recent years, various methods based on different principles have been developed for the early detection of lameness. The primary common principle of the developed systems is to classify cattle lameness accurately. Some examples of automated methods that have been proposed for lameness detection are:

- accelerometers that are attached to the animal and measure behavioural characteristics (Beer et al., 2016),
- force–plate system recording the reaction forces of the animals' gait as electric signals that change over time (Rajkondawar et al., 2002)
- balance-platforms in robotic milking systems (Pastell & Kujalaf, 2007)
- cameras and image analysis techniques for gait analysis (Poursaberi et al., 2010)
- pressure sensitive mats estimating the location and time of contact points of the limbs on the mat (Maertens et al., 2012)
- infrared cameras that scan for different heat levels at targeted points on the animal (Alsaaod & Büscher, 2012)
- micro-Doppler radar which operates using electromagnetic waves and observe micro-changes in animal movements (Busin et al., 2019a)
- Video surveillance system which identifies animals with an object-tracking algorithm and uses reference points across frames to assign a mobility score (Anagnostopoulos et al., 2023)

Table 1.6 provides more information on the mentioned automated systems regarding the validation methods, including the number of animals and assessors involved, the scoring system used, and whether hoof lesions were included as references.

Technology-assisted mobility systems have certain constraints that can limit their application or effectiveness. For example, studies that have used a single accelerometer attached to one limb of the animal might not be precise to detect lameness accurately, as altered gait manifests distinctly on the affected limb (Van Nuffel et al., 2013). Therefore, a second accelerometer may be necessary to obtain more indicative measures of lameness. However, incorporating an additional accelerometer can increase the cost of the monitoring system and its environmental impact. Another study by Pastell et al (2007) proposed a 4-balance system for automated lameness detection, but this system was only suitable for farms with milking robots. Other limitations include the impact of weather and ambient conditions on the system's results. For instance, foggy weather can obstruct a camera and prevent it from performing optimally, while infrared cameras may require calibration due to variations in environmental temperature (Alsaaod et al., 2015).

Table 1.6 Validation methods of selected automated systems for lameness detection.

Automated system	Scoring system	Number of animals	Number of assessors	Reference	Foot lesions	Model validation
Accelerometers (Beer et al 2016)	Flower and Weary	12 healthy and 41 lame cows (according to scores)	3 experienced assessors – mean and rounded to the nearest 0.5 point	Video recordings	Recorded but could not be predicted by the automated system	-
Force-plate system (Rajkondawar et al 2002)	Sprecher et al., 1997	11 healthy, 12 unhealthy (according to scores)	3 investigators – the assigned values of two out of the three investigators had to match to constitute a score	Scores (Not mentioned if live or video scoring)	Not recorded	cross-validation method, in which the model was repeatedly reconstructed 23 times by eliminating a single data input at a time (jack–knifing) and then recalculating the probabilities p1, p2, and p3 for the omitted animal.
4-balanced platforms (Pastell & Kujalaf, 2007)	Sprecher et al., 1997	9,942 measurements by 73 cows (72 sound, 17 lame, 18 with pathologies)	Not mentioned	Mobility scoring and clinical examination	Recorded and used in conjunction with the mobility scores	Teaching with 37 cows, validation with 36 cows
Camera and image analysis (Poursaberi et al., 2010)	3-point scale (Van Nuffel et al., 2009)	66 lactating Holstein cows passed	Not mentioned	video	Not recorded	-

Pressure-sensitive mats (Maertens et al., 2012)	Combination of kinematic variables from: Manson and Leaver (1988), Winckler and Willen (2001), Sprecher et al. (1997) systems ----scores using a 3-point scale	58 lame, 58 sound, 58 mild lame cows	One observer	Video images	Not recorded	-
infrared cameras (Alsaaod et al 2012)	Presence or absence of lesions (0vs1) – no severity	626 individual recordings from 24 cows	Not applicable	clinical incidences / recorded lesions	Recorded	-
micro-Doppler radar (Busin et al., 2019)	AHDB 4-level system (binarized-AHDB)	51 cows (31 ‘lame’ and 20 ‘healthy’)	1 observer	On-farm scoring	Not recorded	‘leave-one-out-cross-validation’ method
Video surveillance (Anagnostopoulos et al. 2023)	AHDB 4-level system (and binarized-AHDB)	6,040 cows	2 assessors	On-farm scoring	Recorded	-

1.5 Radar

1.5.1 Radar basic principles and definitions

Radar stands for Radio Detection And Ranging and is a system that requires certain core features to operate; a transmitter that emits waves (electromagnetic-, radio-, microwaves), an antenna that sends the waves towards a target through the air, an antenna that picks up the reflected waves from the target, a receiver and a processor that displays the received information (Skolnik, 2001) as in *Figure 1.3*.

The basic principle of radar is to transmit a radio wave from the radar antenna, which then travels through space until it encounters an object in its path. The wave is reflected back to the radar antenna, and the time it takes for the wave to travel to the object and back can be used to determine the distance to the object (Folger, 2014).

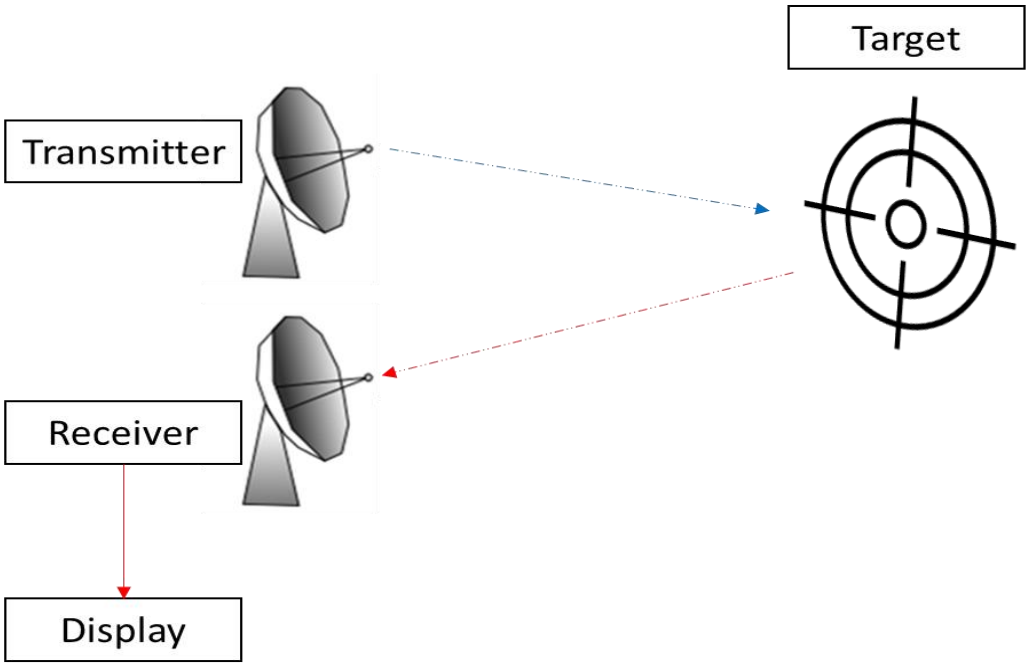


Figure 1.3 Radar system operation. The transmitter antenna sends out electromagnetic waves towards a target; the waves are then reflected back towards the receiver antenna displaying information about the target.

1.5.2 Carrier wave, radio wave & frequency

A carrier wave is a continuous, typically sine wave that is used as a reference signal to transmit information and has a fixed frequency. The information to be transmitted is modulated by varying some aspect of the wave, such as its amplitude, frequency, or phase (Kingsley & Quegan, 1999). In a radar system, the carrier wave is used to transmit the radar pulse, which is the basic signal used to determine the range and velocity of a target by analysing the time delay and frequency shift of the reflected pulses.

Carrier frequency is the specific frequency at which the carrier wave oscillates and is the number of emitted waves per second of a carrier wave, measured in Hertz. For example, if a radar has 50 carrier waves per second, the carrier frequency equals 50Hz. The frequency of the carrier wave (Equation 1.1) determines the wavelength (λ), which is the distance between two consecutive points on the wave that have the same phase. The amplitude is the distance between the origin and the crest or trough of a wave (Figure 1.4).

Radio waves are a type of electromagnetic radiation travelling through space at the speed of light. The frequencies of a radio wave range from 3 Hz to 3·1012 Hz, and the wavelength is between 0.1 mm to 100 000 km.

Equation 1.1 The carrier frequency formula

$$\lambda = c/f$$

where λ is the wavelength, c is the speed of light, and f is the frequency.

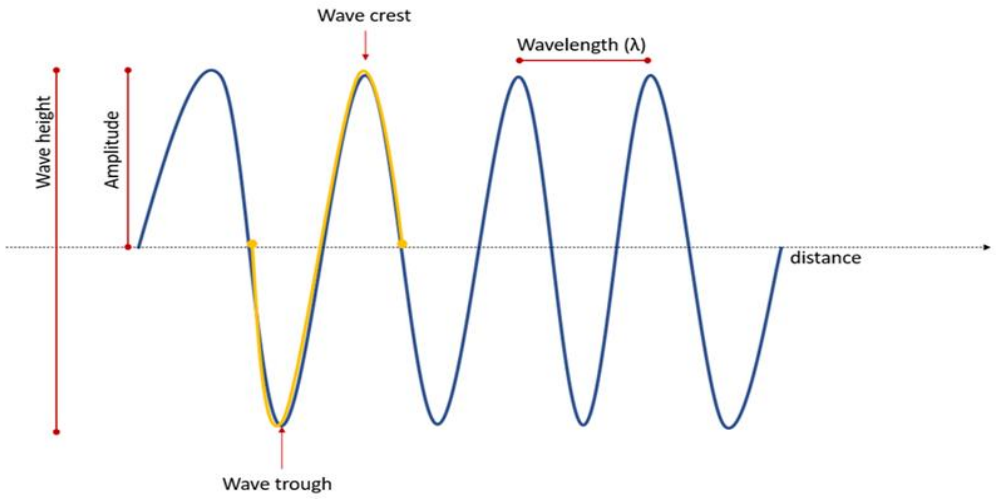


Figure 1.4 A sinewave. The yellow colour represents a wavelength or a wave cycle.

1.5.3 Pulse width & pulse repetition frequency

Pulse width refers to the duration of a radar signal transmitted by the radar system measured in units of time and is determined by the duration of the modulating signal that is used to generate the radar pulse (Melvin & Scheer, 2010). It determines the range resolution of the system, which means the ability of the radar to distinguish between targets that are located at different ranges from the radar antenna. Range resolution is inversely proportional to the pulse width of the signal, with shorter pulse widths resulting in higher range resolution but lower signal power.

Pulse repetition frequency (PRF) is the number of short bursts of electromagnetic energy transmitted by a radar system per second. The PRF determines the rate at which the radar transmits pulses, and the operating parameters affect the range resolution, target detection, and maximum range of the system (Hlawatsch & Auger, 2008). PRF is related to pulse width through the concept of range ambiguity (Equation 1.3), which occurs when the radar pulse is transmitted at a high PRF.

1.5.4 Frequency Modulated Continuous Wave (FMCW)

A traditional radar emits short pulses of radio waves and then awaits reflections from targets (Kingsley & Quegan, 1999). In contrast, a Frequency Modulated Continuous Wave (FMCW) radar transmits a continuous wave signal known as chirps, which is frequency modulated, resulting in a continuously changing carrier wave frequency over time (Atayants et al., 2014). As a result, an FMCW radar can attain high-range resolution, making it capable of distinguishing between objects that are in close proximity using the Doppler effect (Chen, 2003). This is because the Doppler effect causes a change in the frequency of the reflected signal if the target is moving relative to the radar.

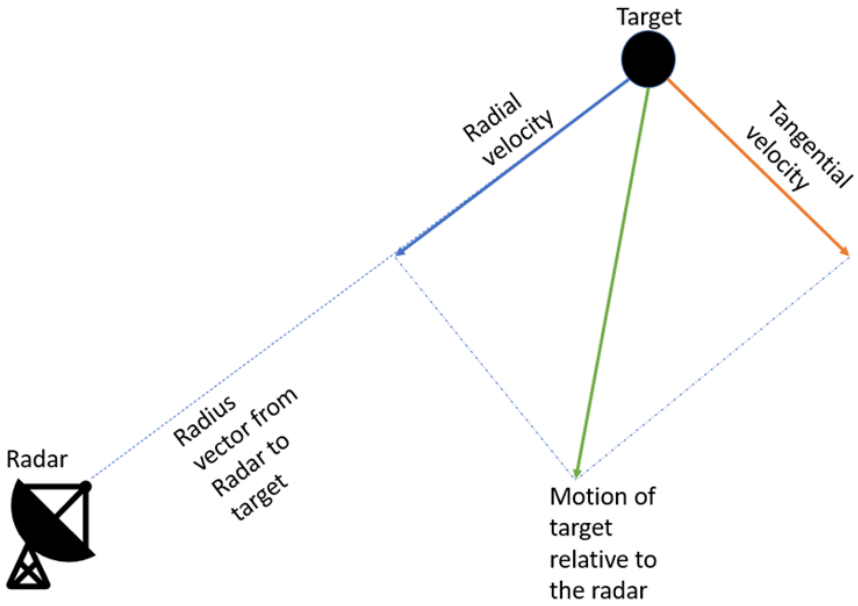


Figure 1.5 The motion of a target relative to the radar. The motion toward or away from the radar is called radial velocity. Motion perpendicular to the direction of the radar is called tangential velocity. The combination of the two motions is the target's velocity.

The velocity of a target relative to the radar system, called radial velocity (Figure 1.5), can be determined from the Doppler shift of the reflected signal. The Doppler shift is the difference between the frequency of the transmitted

pulse and the frequency of the received pulse, and it is proportional to the radial velocity of the target (Equation 1.2).

Equation 1.2 Doppler ambiguity calculation based on the pulse repetition frequency (PRF) shift

$$f_d = \pm \frac{PRF}{2}$$

Where f_d is the maximum unambiguous Doppler frequency.

1.5.5 Doppler & micro-Doppler effect

Doppler effect refers to the change in frequency of the electromagnetic wave reflected by a moving object. The frequency shift is caused by the relative motion between the radar and the target, resulting in a change in the wavelength of the reflected signal (Evans & McDicken, 2000). If the target is moving toward the radar, the frequency of the reflected signal will be higher than the transmitted frequency and vice-versa. Analysing and comparing the transmitted and reflected signals' shifts in frequency can determine the velocity of objects relative to the radar.

Micro-Doppler is the analysis of Doppler shifts caused by the motion of the targets within the main target and is used to identify specific motion characteristics, such as rotations or vibrations (Chen, 2008). Targets performing movements or activities with micromotions have a unique and distinct micro-Doppler signature that can be used to perform detection and classification (Chen et al., 2014a). Particularly, micro-Doppler can identify and accurately classify individual components of an object/target, such as the movements of a person's or animal's limbs while doing a movement or activity (Fioranelli et al., 2015; Shrestha et al., 2017).

1.5.6 Radar radiation beam pattern

The radiation pattern of a radar beam describes the directional properties of the electromagnetic waves emitted by the antenna. It is a graphical representation of the intensity of electromagnetic waves in different directions, depending on the type of antenna used in the system (Balanis, 2016). For example, a parabolic dish reflector antenna (Rudge & Adatia, 1978) will have a different radiation pattern compared to a directional Yagi antenna (Alhalabi & Rebeiz, 2009), as in Figure 1.6. The radiation pattern of a radar beam plays a crucial role in determining the performance of the radar system, as the gain depends on the angle of arrival (Balanis, 2016; Huang & Boyle, 2008). It affects the coverage area, resolution, and sensitivity of the radar and its ability to detect and track targets accurately.

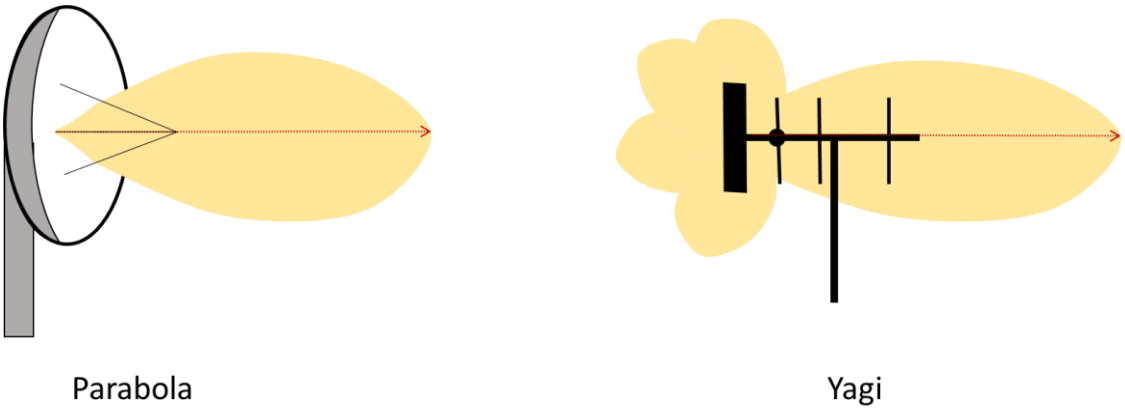


Figure 1.6 Parabolic and Yagi antennas schematic difference of their radiation beam pattern.

The beam pattern of the antenna and the properties of the target can be used in the radar equation (Equation 1.3) to determine the minimum detectable signal power and the maximum detection range of the radar system. The radar equation takes into account various other factors, such as the transmitted power and the gain and the aperture of the antenna, and Equation 1.3 describes the attenuation of electromagnetic waves in free space: the shorter the wavelength, the greater the attenuation in free space.

Equation 1.3 Radar equation

$$P_{Rx} = P_{Tx} \frac{G_{Tx} * G_{Rx} * \lambda^2 * \sigma_0}{(4\pi)^3 * {R_{Tx}}^2 * {R_{Rx}}^2 * L * N}$$

Where,

P_{Rx} and P_{Tx} are the R_x (receiver) and T_x (transmitter) Power,
G_{Tx} and G_{Rx} are the T_x and R_x antenna gain
λ is the signal wavelength
σ_0 is the target radar cross-section
R_{Tx} and R_{Rx} are the T_x – to-target and target-to – R_x distances
L is the systems losses
and N the noise power

1.6 Limiting factors

A limiting factor is signal noise, which is the result of all the electronic accessories that form the radar and other external sources (Kolawole, 2002). Signal noise in radar systems is unpredictable and can result from various factors such as thermal noise (Lange & Hammer, 1978) or flicker noise (Nguyen et al., 2007). The signal-to-noise ratio (SNR) measures the strength of the signal about the noise (Proakis & Salehi, 2002). In radar, a high SNR is desirable as it improves the detection of weak signals. However, achieving a high SNR can be challenging (Equation 1.3). To address this, digital signal processing techniques can be used to filter out unwanted noise and enhance the radar signals, leading to better detection performance (Mahafza, 2016).

In addition to signal noise, clutter is another factor that can compromise the quality of radar signals. Clutter refers to unwanted echoes produced by external sources like rain or ground, which can interfere with radar outputs (Toomay & Hannen, 2004). There are several ways to mitigate clutter in radar systems, such as the moving target indicator (MTI) technique, which can help to separate valuable signals from unwanted clutter by comparing the received

signal from one pulse to the previous pulse. Constant signals are filtered out, and the signals that change from pulse to pulse are amplified (Mahafza, 2016). Alternatively, the employment of micro-Doppler analysis as a technique can be used to differentiate between useful and non-valuable signals (Chen et al., 2014a). This approach can be efficient in situations where clutter is present and can help improve radar readings' accuracy.

The signal-to-noise ratio (SNR) and the signal-to-noise-and-clutter ratio (SNCR) are essential measures that influence the overall performance of radar systems. These metrics are closely related to the radar equation (Equation 1.3) used to determine the maximum range of a radar (Skolnik, 2001). To account for the impact of noise and clutter, the radar equation includes a noise term (Equation 1.3) that can be adjusted to reflect the presence of clutter in the environment. Modifying this term can improve the accuracy and reliability of range detection. The SNR and SNCR are particularly crucial for radar systems operating in challenging environments such as farms where clutter and noise are prevalent because of the presence of various sources of interference such as metallic objects, machinery and other moving animals. By using advanced signal processing techniques and high-quality electronic components, these metrics can be optimised and improve the overall performance of radar systems.

In addition, to signal noise and clutter, the radar receiver's noise, which is critical to the radar design parameters, can also impact the performance of radar systems (Skolnik, 2001). A radar loss budget can be used to mitigate the impact of the radar receiver's noise. A radar loss budget is a technique used to identify and minimize the different sources of loss, i.e. hardware and propagation losses, within a radar system (Kolawole, 2002). It is a comprehensive analysis of the losses from the transmitter to the receiver and provides a detailed breakdown of each component's contribution to the total loss. A radar loss budget aims to optimize the radar's design parameters to ensure optimal performance. The loss budget analysis includes several factors, such as the radar's antenna gain, the transmission line loss, the receiver noise figure, and other losses due to atmospheric absorption. Each of these factors contributes to the overall loss of the radar system, which impacts its performance. Using a loss budget analysis, the sources of loss can be identified, and steps to minimize their impact can be taken. This process can involve

selecting appropriate components, such as high-gain antennas and low-noise amplifiers (Pandey & Singh, 2015), and optimizing the radar's operating parameters, such as the frequency and pulse width (Skolnik, 2001)

1.7 Radar applications

Radars are used for many reasons, and they have broad applications. Nearly every industry uses some form of radar. Radar was first invented to avoid ships colliding which each other with the Telemobiloscope invented by Hulsmeyer (Kendal, 2011). It was quickly taken over during the war for military applications and still is a valuable resource for field operations with smarter multifunction radars, but they are also used in agriculture and other areas.

A few examples of the fields that radars are used in the present are

- Biological research – e.g., tracking birds and insects. (Gauthreaux, 2003; Reynolds et al., 1997)
- Air traffic control, aircraft landing (Li & Bar-Shalom, 1993; Soumekh, 1996)
- Weather-storm forecasting (Baron, 1998; Viswanathan et al., 1997)
- Military (Olsen & Asen, 2017; Pitkethly, 1992)
- Geology, ground analysis, and surface topography (Galagedara et al., 2003; Mellett, 1995)
- Speed radar-traffic radar (Muñoz-Ferreras et al., 2008 ; Teed et al., 1993)
- Biological radar (detects human body movements such as the heart) (Lv et al., 2015)
- Movement detection (Li & Lin, 2008 ; Schleicher et al., 2013 ; Tupin & Couse, 2016)
- Environmental monitoring (Albright, 2004; Koo et al., 2012)
- Terrain Mapping (Graham, 1974; Madsen et al., 1993)
- Ocean Mapping (Hasselmann & Hasselmann, 1991)

- Agriculture, crop classification, and fish (Hedgepeth et al., 1999 ; Mattia et al., 2003 ; Ulaby et al., 1982)

1.8 Radar for animal monitoring

In animal studies, the use of radar technology for animal tracking and behaviour analysis has been investigated, with a focus on applications in wildlife ecology and conservation. A few examples are the studies which investigated the use of weather radars to track the movements of bats (Pennisi, 2011) and quantify the density of migratory birds (Buler & Diehl, 2009). The researchers found that radar technology was highly effective at detecting and tracking individuals and could provide insights into their behaviour and activity patterns. Other examples include animal identification, positioning and tracking changes in the locomotor behaviour of farm animals such as sheep, cows, and horses (Shrestha et al., 2018).

In particular, using radar for diagnostic purposes is a relatively recent proposal (Busin et al., 2019a; Shrestha et al., 2017, 2018), motivated by research such as that of Fioranelli et al. (2015), which concerns the detection of human movements behind walls, saw the possibility of using the same technology in animals. The publication by Shrestha et al. (2018) was a proof of concept for detecting deviations from the typical locomotion of animals (horses, cows and sheep). The study by Busin et al. (2019), aimed to take the micro-Doppler radar sensing method and quantify its performance for lameness detection in dairy cows and sheep on farms. In both studies, they manually extracted features from the collected radar data. Then, an expert veterinarian classified the animals into lameness categories using the 4-level AHDB dairy mobility system, and they used the dichotomised scores (non-lame = scores 0,1 and lame =scores 2,3) as labels to train the machine learning algorithms. In both studies, the authors used a supervised machine learning framework that could classify the animals as either lame or non-lame based on their radar micro-Doppler signatures following the process described in Figure *1.7*. They converted the raw radar data into the range-time domain and then to range-Doppler data through two consecutive fast Fourier transformations. Then, they summed the range dimensions to produce spectrograms used for feature extraction. They

used a “leave-one-out-cross-validation” approach to train and test the classification algorithm, simulating the scenario where the algorithm is presented with an animal with an unknown mobility status on the farm. In the second paper (Busin 2019), they also tested the impact of different parameters, such as different classification algorithms (SVM and KNN classifiers) and the selection of features on the method’s performance. Both studies achieved more than 83% accuracy for cows’ mobility classification (Figure *1.8*), demonstrating radar technology’s potential as a promising tool for automated lameness detection.

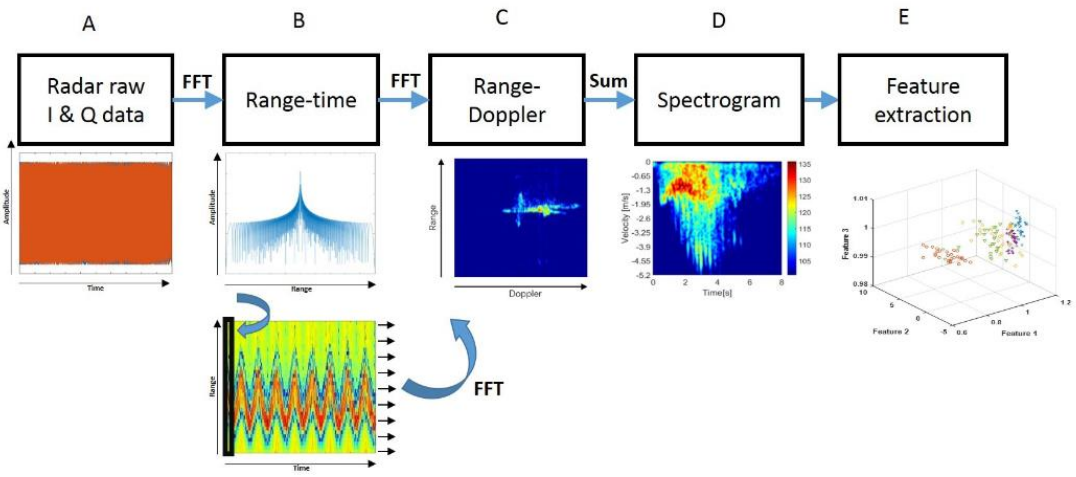


Figure 1.7 Radar signal processing chain followed in previous studies. Figure from “Evaluation of lameness detection using radar sensing in ruminants” by Busin V. et al., 2019, Veterinary Record 185(18), p572. (DOI:10.1136/vr.105407)

A

Accuracy [%]	Predicted Healthy	Predicted Lame
True Healthy	70%	30%
True Lame	8.6%	91.4%

B

	Naïve Bayesian algorithm		
	5 seconds	3 seconds	1.5 seconds
Dairy cows			
True positives	71.6	79.3	82.5
False positives	28.4	20.7	17.5
True negatives	72	86.4	73.9
False negatives	28	13.6	26.1
Sensitivity*	0.72	0.85	0.76
Specificity†	0.72	0.81	0.81
Accuracy‡	0.72	0.83	0.78

Figure 1.8 The classification accuracy results of the two studies. Above (A) are the results from the study of Shrestha et al. 2018 using an SVM classification model, and below (B) are the results from Busin et al. 2019, where they used naïve-Bayes algorithmic classifier model and different time segment durations. Adapted from "Animal Lameness Detection With Radar Sensing"by Shrestha et al. 2018 and "Evaluation of lameness detection using radar sensing in ruminants" by Busin V. et al., 2019

1.9 Advantages and limitations of the proposed radar system

The radar system proposed for lameness detection in cows offers several advantages over other methods. Firstly, it is non-intrusive, meaning it can detect lameness without physically interfering with the animal, reducing the potential for discomfort or distress. In contrast, methods like accelerometers require sensors to be attached to the animal, which can cause discomfort or affect natural movement patterns. Wearable devices also add to the farm's expenses as they need to be increased with every animal increase. Secondly, the radar system can detect changes in an animal's gait pattern from a distance of more than 5 meters away from the animal and its limbs. This is not always possible with computer vision, providing more flexibility in monitoring animal movement. Lastly, the system is not affected by lighting conditions or weather

factors such as rain or fog, making it more robust, reliable, and ideal for farm environments.

Like most developing systems, the radar system has some limitations so far. The previous studies established a proof of concept and evaluated the radar as a tool and the machine learning process for the analysis. However, both studies used only a small number of animals which may not represent the overall population; thus, the results may not be generalisable to other farms and new unseen cases. This is because the finite number of animals used may have been categorised across all levels of lameness classification according to the AHDB mobility system but may not have had all the different lesion types causing lameness present. Reducing the classification levels from four to two may have simplified the data and made it more manageable for analysis. However, the intermediate levels and some classification details were lost, making it more challenging to distinguish between different degrees of lameness. Another limitation concerns the ground truth on which the algorithms were trained. Both studies relied on one expert veterinarian to classify the animals into lameness categories, which can be subjective and may vary between experts. As previously discussed, visual lameness assessments are subjective, but it is widely used in practice and is considered a gold standard upon which machine learning is based on being trained. Other analysis techniques would be good to be considered in the future. These main limitations need further research to validate the radar system for automated lameness detection on farms.

1.10 Machine learning

Machine learning (ML) is a field of artificial intelligence that involves the development of algorithms and models that can learn patterns and make predictions or decisions from data without being explicitly programmed (Jung, 2022). Machine learning aims to enable computers to automatically learn from data, identify patterns, and make accurate predictions or decisions.

There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the

algorithm is trained on a labelled dataset, including input and corresponding output labels (Kotsiantis, 2007). The algorithm learns to recognise patterns in the input data associated with the output labels and can use this knowledge to classify new, unseen data. Unsupervised learning involves training an algorithm on an unlabelled dataset and allowing it to identify patterns and relationships (Albalate & Minker, 2013). Finally, reinforcement learning involves training an algorithm to make decisions based on rewards and punishments received from its environment (Kaelbling et al., 1996).

Machine learning has a wide range of applications, including natural language processing (Sebastiani, 2002), image (Litjens et al., 2017) and speech recognition (Campbell et al., 2006) fraud detection (Ajdani & Ghaffary, 2021), autonomous vehicles (M. Chen et al., 2017), and predictive maintenance (Susto et al., 2015). In recent decades, machine learning has become increasingly important in fields such as healthcare (Jiang et al., 2017), finance (Fischer & Krauss, 2018), and marketing (Kim et al., 2001), where large amounts of data are generated, and there is a need for accurate predictions and decision-making.

1.10.1 Supervised ML

Supervised machine learning algorithms are a type of artificial intelligence that can be trained to classify data based on input-output pairs and make predictions or decisions about new, unseen data by learning from labelled data (Alpaydin, 2014b, 2014a). For example, in the context of lameness detection using radar technology, supervised ML algorithms can be trained to classify radar data into lame and non-lame categories based on known examples of each, as in the study by Busin et al. (2019).

A labelled dataset of radar data is needed to train a supervised ML algorithm for lameness detection. This dataset should include radar data collected from both lame and non-lame animals. The dataset is then split into two parts: training and testing sets. The training set is used to train the ML algorithm to recognise patterns in the radar data that are indicative of lameness, while the testing set is used to evaluate the accuracy of the trained algorithm on new, unseen data.

1.11 Classification process and algorithms

Supervised machine learning classification involves training an algorithm to recognise patterns in input data and make predictions about the corresponding output labels. The classification process typically involves the following steps (Alpaydin, 2014a).

The first step is to prepare the dataset for algorithm training. This includes collecting and cleaning the data, pre-processing if needed and splitting it into training and testing sets, typically in a 80/20 split, where 80% of the data is used for training and the remaining 20% is used for testing (examples Kagiyama et al., 2020; Ratzinger et al., 2018; Tran et al., 2019). The splitting process can be performed later depending on the validation and analysis programs. For example, in Matlab's application for classification learning (*MATLAB R2022b*, 2022), there is the option for n-fold cross-validation, where the data are split by the program n-times for the analysis. Then the relevant features or variables from the input data that will be used to train the algorithm are selected. This is important to ensure that the algorithm can learn the most relevant patterns to the problem being solved. The next step is to choose the appropriate algorithm or model for the classification task. This depends on the nature of the problem being solved and the characteristics of the data. The algorithm is trained using labelled training data. During training, the classifier learns to recognise the patterns in the input data associated with the corresponding output labels. And then, the performance of the trained model is evaluated using the testing data. This helps to determine the model's accuracy in making predictions on new, unseen data. Depending on the outcomes, the algorithm's parameters can be adjusted to improve its performance and repeat the classification testing and validation process.

There are several algorithms used in supervised machine learning classification (Burkov, 2019), including:

- Logistic Regression: A model that predicts the probability of a binary output based on the input features.
- Decision Trees: A model that makes decisions based on a series of if-then rules based on the input features.

- Random Forest: A model that uses multiple decision trees to make predictions.
- Support Vector Machines (SVM): A model that finds the optimal decision boundary between different data classes.
- Neural Networks: A model that consists of layers of interconnected nodes that learn complex patterns in the input data.
- K-Nearest Neighbours (KNN): A model that makes predictions based on the input features of the k-nearest neighbours in the training data. It relies on the assumption that similar data points in the feature space have similar outcomes. The parameter k represents the number of considered neighbours.

The choice of algorithm depends on the nature of the problem being solved and the characteristics of the data. Each algorithm has its strengths and weaknesses, and the performance of the algorithm can be improved by adjusting the parameters and tuning the model.

1.11.1 Features and feature extraction

A feature refers to a measurable aspect of a data point (Salau & Jain, 2019), and feature extraction is selecting, transforming, and combining raw data into a set of meaningful features to train a machine learning model. In supervised ML, where the models are trained on labelled data, the features are linked to the corresponding output labels. The algorithm identifies patterns based on the paired features-labels and updates the internal parameters to optimise the difference between the predicted and true labels (Goodfellow et al., 2016).

To extract features from micro-Doppler radar data, one can use signal processing techniques to extract relevant features from the raw data. Some standard feature extraction techniques used in studies (Khalid et al., 2014) include time-frequency analysis like fast Fourier transformation and principal component analysis. These techniques can extract features such as the radar signal's frequency content, amplitude modulation, and time-frequency distribution. Once the features have been extracted, they can be used to train a model to classify different types of targets based on their micro-Doppler radar signature.

1.11.2 Classifiers' performance

The performance of a classifier can be impacted by several factors, such as class imbalance (Japkowicz & Stephen, 2002), missing values (Rahman et al., 2013) and overlaps in the classes (Alejo et al., 2013). Balanced, well-defined and complete datasets are important in supervised machine learning because they allow the algorithmic model to learn from all classes or categories equally, preventing bias towards the majority class (Batista et al., 2004). For example, in a binary classification problem, if the dataset is imbalanced and one class has significantly more observations than the other class, the machine learning model can become biased towards the majority class and may perform poorly on the minority class. This can result in poor generalisation performance and inaccurate new, unseen data predictions. Likewise, in multi-class classification problems, imbalanced datasets can lead to similar issues, meaning some classes may be underrepresented and not given enough attention during model training, resulting in poor performance in those classes. However, not all machine learning algorithms require balanced datasets. Some algorithms, like decision trees (Pérez et al., 2005) or Naive Bayes (Yang et al., 2013), can handle imbalanced datasets well, while others, like neural networks, may require balanced datasets or appropriate modifications to handle imbalanced datasets effectively (Ren et al., 2020).

Other factors that may affect classifier performance include the presence of noise and outliers in the data (Johnson & Khoshgoftaar, 2022) which can introduce irrelevant or misleading information. Another factor is the choice of features used for training (Li et al., 2017) which should be relevant and informative and allow for discrimination between the different classes.

Once the factors affecting classifier performance have been considered, it is essential to evaluate the performance of classifiers to determine their effectiveness and suitability for a given task. Depending on the task, several metrics can be used to assess the performance (Alpaydin, 2014a). Accuracy is one of the most common metrics. It measures the proportion of correct predictions made by the classifier out of all the predictions made. However, accuracy can be misleading in cases where the dataset is imbalanced, or the cost of misclassifying different classes is not the same. Another metric is

precision which measures the proportion of true positives (TP) out of all the positive predictions made by the classifier. Then, recall measures the proportion of predicted true positives (TP) from all the actual positives in the dataset. It is a good metric when the cost of false negatives (FN) is high. The F1-score is the harmonic mean of precision and recall and is a commonly used metric. Finally, the Receiver Operating Characteristic – Area Under Curve (ROC-AUC) is a metric that measures the performance of a classifier at different classes. Again, it is a good metric to use when the dataset is imbalanced or when the misclassification cost is not the same for all the classes (Fawcett, 2006).

In this project, accuracy and confusion matrices will be primarily presented, although in some cases, F1-scores (true positives, true negatives, false positives, false negatives and ROC-AUC curves) are also presented. Accuracy and confusion matrices were used to evaluate classifier performance as they provide a straightforward and intuitive way to interpret the results. The accuracy metric offers a simple way to measure a classifier's performance. At the same time, confusion matrices provide a more detailed view of the classifier's performance by showing how well it can distinguish between different classes. In addition, confusion matrices allow for the calculation of additional performance metrics such as precision, recall, and F1 score, which can provide a more nuanced understanding of the classifier's performance for each class. Overall, the two metrics were used to be easily interpretable for anyone, regardless of their background in machine learning.

1.12 Labels

A label is a predefined tag or identifier assigned to a data point indicating that it belongs to a particular class or category (Fieguth, 2022). It is a requirement for supervised machine learning classification, serving as the ground truth. For example, suppose we would like a model to recognise images containing different cow breeds. In that case, we should train it by providing images with all the breeds we wish to be identified and a representative label for each image. The more balanced the dataset, i.e., having a similar number of pictures containing each breed, and the more representative the label, i.e.,

“black and white Holstein cow” instead of “Holstein”, the more valid and open to generalisation the results and the model will be. Labels are one of the most fundamental elements in supervised machine learning and the assignment process is essentially the behaviour (decision-making) which we want to be replicated.

In medical fields such as radiology, experts typically interpret images/data and describe the findings providing information such as the presence or absence of specific features, the location and size of abnormalities, and the severity or stage of the condition (Zhang & Sejdić, 2019). While manual annotation by experts is considered the gold standard for creating labels, it can be time-consuming and labour-intensive. To address this challenge, researchers have proposed alternative labelling methods, such as semi-automated approaches that use computer-aided detection (CAD) algorithms and natural language processing (NLP) techniques to extract labels from reports (Jun et al., 2018; Martín-Herrero, 2007; Pesce et al., 2019). Machine learning models fed with representative labels can process information in raw image data and automatically learn relevant patterns for prediction tasks like recognising, localising, and segmenting visual objects (Janiesch et al., 2021). Then, new unlabelled data is introduced during the test, and the algorithm looks for similarities in features and patterns and classifies them into a category. The importance of labels is significant, and they determine the results that the algorithm will produce. There have been cases in the past where unbalanced and poorly characterised data have had negative consequences (Grother et al., 2019; Mehrabi et al., 2019).

Finally, while the terms “ground truth” and “gold standard” are sometimes used interchangeably to address labels, they have different meanings in the context of machine learning. Ground truth refers to the labels we provide to the algorithm as a reference, while the gold standard represents the diagnostic method with the highest accuracy (Cardoso et al., 2014).

1.12.1 Decision-making process

There are a few steps in the decision-making process that someone does to reach a decision (Lunenburg, 2010; Schoenfeld, 2011). The first step is to

define what kind of decision needs to be made. Then one has to gather relevant information to make the decision. During the information gathering, all possible alternatives are identified. The decision-maker then weighs the evidence and chooses one of the pre-defined options.

A practical example is the process of cattle mobility scoring. The assessor first defines the system used for scoring and what should look for in the animal's mobility to evaluate. In this case, the general attitude is that the assessor should look for deviations from normal cow mobility. Then the assessor, after the visual examination (step: information gathering), considers all alternatives – which score corresponds to the animal's condition under assessment and why. And finally, the assessor weighs the evidence (mobility vs scores) and decides whether the animal is lame or not and which score to assign.

In theory, the described process might seem straightforward. However, in several steps through the process, many things deviate among individuals and are considered challenging (Hammond et al., 1998). For example, during the visual examination, an experienced assessor might see something that an inexperienced person will not notice, as research has shown that the process followed in visual examination and decision-making can differ according to experience (T. Donovan & Litchfield, 2013; Jaarsma et al., 2014). In addition, bias and perceptual adaptations could also affect the assessor's decision (Witthoft et al., 2018). An example of perceptual adaptation is that a cow will be assessed as mildly lame when the assessor is not adapted, but the same cow might be reported as sound following prolonged viewing of severely lame cows. The availability of time during the decision-making process has been reported to affect the decision outcome too (Kahneman, 2011). Not everyone can process the same amount of information in a limited time. And the amount of information available could affect the decision. Prior research in the behavioural sciences has reported that humans tend to shorten the time of decision making usually following heuristic approaches as explained in the book "Thinking, fast and slow" (Kahneman, 2011). For example, the decision-makers follow a strategy where they compare the alternatives, and when one is eliminated, they don't consider it an option anymore (Russo & Dosher, 1983)

A final and essential consideration of the decision-making process is the consequences of the call (Semmel, 1979). For example, the assessor in the

beforementioned example of cow mobility scoring decides to assign a score that characterises the cow as lame. Then, some actions need to be taken to address the mobility problem. Either try to cure the animal by administering drugs, performing hoof trimming or surgery, or sending the animal to the slaughterhouse. The example can be generalised in all cases. A decision leads to action, and the assessor should be aware and responsible for the decisions. The same thought could make the decision-maker lenient or strict in their choices depending on their level of commitment (Juliusson et al., 2005), which affects the outcome.

The decision-making process might be challenging to explain in detail. However, it feels natural and relatively easy in everyday life. And in most cases, a small degree of variability is acceptable, as it allows us to consider and explore options, we had not previously perceived (Kahneman, 2011). However, when the goal is to implement a complex decision-making process in a machine for decision support or diagnostic reasons, it is necessary to be as precise as possible so that the outcomes are not negatively impacted or uncertain (Jordan & Mitchell, 2015).

1.13 Measurement of agreement

Quantifying inter-assessor agreement is important in many fields, including healthcare and diagnosis (Dunn, 2004; Mulsant et al., 2002; Szklo et al., 2019). The purpose of calculating agreement is to evaluate the degree to which multiple assessors or measurements are consistent and reliable or to identify sources of variation.

There are several methods for the calculation of agreement, and the choice depends on factors such as the type of data being analysed, the number of assessors, the levels of the measurement system and the questions being addressed.

Some of the commonly used measures of agreement are detailed below:

1.13.1 Cohen's kappa

Cohen's kappa (Cohen, 1960) is used to assess the level of agreement between two or more assessors. It takes into account the possibility of agreement occurring by chance and provides a more accurate measure of agreement than simply looking at the proportion of agreement (percentage agreement). The coefficient ranges from -1 to 1, where a score of 1 indicates perfect agreement, 0 indicates agreement due to chance, and -1 indicates perfect disagreement. While there is no universally accepted method for interpreting kappa scores, one commonly used approach is the method proposed by Landis & Koch, (1977). Values of 0.81 or higher indicate almost perfect agreement, values between 0.61 and 0.8 indicate substantial agreement, values between 0.41 to 0.6 indicate moderate agreement, values between 0.21 to 0.4 indicate fair agreement, and values below 0.2 indicate poor agreement. Cohen's kappa is calculated by comparing the observed agreement between assessors with the agreement expected by chance. The formula for calculating Cohen's kappa is:

Equation 1.4 Cohen's kappa formula

$$Kappa = (Po - Pe) / (1 - Pe)$$

Where

Po is the proportion of observed agreement among assessors, and
Pe is the proportion of agreement expected by chance and they are calculated as follows.

Equation 1.5 Calculation formula of the proportion of observed agreement among raters

$$P_o = \frac{sum\ of\ pairwise\ agreement\ in\ all\ categories}{sum\ of\ pairwise\ agreement + disagreement\ in\ all\ categories}$$

Equation 1.6 Calculation formula of the proportion of chance agreement

$$P_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

Where,

k = categories
N = observations
n_{ki} = number of times rater i predicted category k

Cohen's kappa is commonly used in fields such as medicine (Kraemer et al., 2002; McGinn et al., 2004) and psychology (Herjanic & Reich, 1997; Ruan et al., 2008; Vanasse et al., 2012), where multiple assessors are involved in the evaluation of the same subject or phenomenon. It provides a useful tool for assessing the reliability of ratings and can be used to identify sources of variation, but it has also been noticed to present some limitations. The kappa paradox, also known as the prevalence paradox, occurs when there is a significant disparity between the prevalence of a condition being measured and the prevalence of agreement among assessors (Feinstein & Cicchetti, 1990). This means that the prevalence of the measured condition can impact the interpretation of kappa scores, and assessors who produce similar marginal distributions need to have a higher agreement rate to achieve the same kappa value compared to assessors who produce different marginal distributions (Brennan & Prediger, 1981). This can result in low kappa scores, even when the assessors agree substantially in percentage calculations, which can be a limitation in situations where certain classes are more important or clinically relevant than others and may lead to an underestimation of the level of agreement in these cases. One other consideration when interpreting kappa scores is that the measure treats all classes equally distinct even in ordinal classification, meaning there is no greater agreement between a 1 and a 4 than between a 3 and a 4.

1.13.2 Fleiss' kappa

Fleiss' kappa (Fleiss, 1971) is an extension of Cohen's kappa and is used to assess the level of agreement between three or more assessors. The formula for calculating Fleiss' kappa is:

Equation 1.7 Fleiss' kappa formula

$$Kappa = (P - Pe) / (1 - Pe)$$

Where

P is the proportion of the averaged observed agreement among raters,

and Pe is the proportion of agreement expected by chance alone, and they are calculated as in Equation 1.5 and Equation *1.6* respectively.

As an extension of Cohen's kappa, Fleiss' kappa shares the same interpretation of the scores and limitations (McHugh, 2012).

1.13.3 Interclass correlation coefficient (ICC)

The ICC (Gwet, 2008) is based on the analysis of variance (ANOVA) and takes into account the variation between and within assessors. It is calculated by dividing the inter-assessor variance by the total variance (which includes both the inter- and intra-assessor variances), resulting in a value between 0 and 1.

The ICC has three extensions:

ICC(1) measures the consistency of ratings from a single assessor over multiple assessments. It is calculated by dividing the between-subject variance by the total variance.

Equation 1.8 ICC1 formula

$$ICC1 = (MSR - MSE) / (MSR + (k - 1) * MSE + k * (CR - C))$$

Where:

MSR is the mean square for the assessors
MSE is the mean square error (or residual mean square)
k is the number of assessors (or measurement methods)
CR is the mean of the variances of the k assessor
C is the variance of the true scores

ICC(2) measures the consistency of ratings among multiple assessors who rate the same subjects or phenomena. It is calculated by dividing the between-subject variance by the total variance.

Equation 1.9 ICC2 formula

$$ICC2 = (MSR - MSW) / (MSR + (k - 1)\ MSW + k\ (CR - C) / n)$$

Where:

MSR is the mean square for the raters
MSW is the mean square error (or residual mean square)
k is the number of raters (or measurement methods)
CR is the mean of the variances of the k raters
C is the variance of the true scores
n is the number of subjects

ICC(3) measures the agreement among multiple raters who rate the same subjects or phenomena, regardless of the rater or subject. It is calculated by dividing the between-subject and residual variance by the total variance.

Equation 1.10 ICC3 formula

$$ICC3 = (MSR - MSW) / (MSR + (k - 1)\ MSW)$$

Where:

MSR is the mean square for the raters
MSW is the mean square error (or residual mean square)
k is the number of raters (or measurement methods)

ICC values range from 0 to 1, with higher values indicating better agreement. The interpretation can also be based on Landis & Koch (1977), with values above 0.6 to be considered substantial.

1.13.4 Pearson correlation coefficient

The Pearson correlation coefficient measures the strength and direction of a linear relationship between two variables. It is commonly used to measure agreement between two sets of continuous data. It is calculated as the covariance ratio between the two variables to the product of their standard deviations. The resulting value, called the correlation coefficient I, ranges from -1 to +1, where -1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no correlation.

Equation 1.11 Correlation coefficient formula

$$r = \frac{(n\Sigma x - \Sigma x\Sigma y)}{sqrt((n\Sigma x^2 - (\Sigma x)^2 * (n\Sigma y)^2 - (\Sigma y^2))}$$

Where:

r is the Pearson correlation coefficient
n is the number of paired data points
Σxy is the sum of the products of the x and y values
Σx is the sum of the x values
Σy is the sum of the y values
Σx2 is the sum of the squares of the x values
Σy2 is the sum of the squares of the y values

When using the Pearson correlation coefficient to measure agreement between two data sets, the variables should be calculated on the same scale and have a linear relationship (Mukaka, 2012). If the association is not linear, other measures of agreement, such as the intraclass correlation coefficient, may be more appropriate (Mcgraw & Wong, 1996).

1.13.5 Kendall's tau coefficient

Kendall's tau coefficient, also known as Kendall's rank correlation coefficient, is a statistical measure of the correlation between two variables. It is commonly used to measure the degree of association between two rankings or ordered lists. Kendall's tau coefficient is a non-parametric measure, which means that it does not assume any specific distribution for the variables being measured. It ranges from -1 to 1, where a value of -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

Equation 1.12 Kendall's tau formula

$$\tau = \frac{(2 * Number\ of\ concordant\ pair - 2 * Number\ of\ discordant\ pairs)}{(n * (n - 1))}$$

Where:

τ is Kendall's tau coefficient

Number of concordant pairs is the number of pairs of items that have the same order in both rankings

Number of discordant pairs is the number of pairs of items that have opposite orders in the two rankings

n is the total number of items being ranked

Kendall's tau coefficient is calculated by counting the number of concordant and discordant pairs in the two rankings being compared. A pair of items is considered concordant if they have the same order in both rankings (i.e., they

are both ranked higher or lower in both rankings) and discordant if they have opposite orders in the two rankings (i.e., one is ranked higher in one ranking and lower in the other). Kendall's tau coefficient is then calculated as the difference between the number of concordant pairs and the number of discordant pairs, divided by the total number of pairs.

1.14 Notes on terminology used in the thesis

The fields of machine learning (ML), statistics, and computer science all involve using mathematical and computational tools to analyse data and make predictions (Fieguth, 2022). However, these fields often use different terminology to describe similar concepts, which can create confusion and make it challenging to communicate across disciplines. For example, the terms "binarise" (Merriam-Webster, n.d.-b) and "dichotomise" (Merriam-Webster, n.d.-c) are used to describe converting a continuous variable into a binary variable with only two possible values. However, the term "dichotomise" is more commonly used in statistics, while in machine learning and computer science, the term "binarise" is often preferred (examples:Jung et al., 2007; Murphy, 2012). Similarly, the term "regression" is commonly used in statistics to refer to a method for modelling the relationship between variables (Fox, 1997), while in machine learning, the term "classification" is often used to describe a similar process of assigning objects to different classes (i.e., Alpaydin, 2014a). To effectively communicate across disciplines, it is essential to be aware of these differences in terminology and to clarify any misunderstandings that may arise. Terms corresponding to the ML field will be used in this thesis.

Different terms are often used in the fields of psychology, medicine, and other areas that involve assessments or evaluations, to refer to the individuals who perform these tasks. These terms include "rater," "assessor," "evaluator," and others. While these terms are often used interchangeably, they can have slightly different connotations.

The term "rater" generally refers to an individual responsible for assigning scores or ratings to a particular set of objects, such as patients or tests (Merriam-Webster, n.d.-e). This term is often used in reliability studies, where the goal is to assess the consistency of ratings among multiple raters.

The term "assessor" typically refers to an individual responsible for evaluating a particular attribute or quality of an object or individual (Merriam-Webster, n.d.-a), such as intelligence, skill, or physical health. This term is often used in the context of diagnostic assessments or performance evaluations.

The term "evaluator" is a more general term referring to individuals who perform assessments or evaluations (Merriam-Webster, n.d.-d) in various contexts. For example, this term is often used in the context of program evaluations, where the goal is to assess the effectiveness of a particular intervention or program.

While these terms can have slightly different meanings, they are often used interchangeably in practice. Regardless of the specific terminology used, assessments and evaluations generally aim to provide accurate and reliable information about the attributes or qualities of the objects or individuals being evaluated. In this thesis, the term “assessors” has been chosen and used throughout.

1.15 Conclusion and aims of the thesis

After reviewing the literature, several aspects related to the automation of cattle lameness detection and classification require further investigation, particularly regarding the transition from visual to technologically assisted decision-making. Therefore, the overarching objectives of this project are to quantify the performance of the micro-Doppler radar sensing method, characterize and validate the micro-Doppler radar signatures of dairy cattle with varying degrees of gait impairment, and develop machine learning algorithms capable of inferring the mobility status of the animals being tested from their signatures. This research aims to support the automatic, contactless

classification of cattle mobility status, improving the efficiency and accuracy of lameness detection in the dairy industry.

The following 5 chapters of the document explain how the goals mentioned earlier were achieved. Each chapter is divided based on the chronological order the study was carried out and covers the related literature, methods used, results, and discussion.

Chapter 2

Inter-assessor agreement of mobility classifications based on the AHDB scoring system

2.1 Introduction

The work presented in this chapter is part of a programme of work to develop an automated mobility classification system. After promising initial results with a micro-Doppler radar detection system using on-farm classification by a single assessor (Busin et al., 2019; Shrestha et al., 2018), we aimed to find a robust approach to labelling for the large numbers of observations needed for an effective automated system. It was recognised that using a single assessor for all labels could result in an idiosyncratic system if the same assessor was used throughout the study, or excessive variation if different assessors were used at different time points. These errors correspond to bias and noise, respectively, using the definitions of Kahneman et al., (2021) and applied to pathological diagnosis by Böer-Auer et al. (2022). It was expected that combining the scores of multiple assessors would improve the rigour of classification, but the best way of combining scores and the optimal number of assessors was not known. We initially set out with the hypothesis that it would be possible to aggregate scores from multiple assessors using the UK dairy industry standard AHDB 4-level scoring system (https://ahdb.org.uk/knowledge-library/mobility-scoring-how-to-score-your-cows; accessed January 31, 2023), and our primary aim was to confirm that inter-assessor agreement was adequate. However, unexpectedly low levels of inter-assessor agreement results led to additional studies to provide a more detailed investigation of some of the factors we could address to improve the quality of labels for our system. Therefore, this chapter reports on several linked studies that aimed to investigate inter-assessor agreement and the decision-making process for mobility classification.

Research has shown the potential adverse effects of lameness on the animal. Reproductive capacity, physical condition, and milk production are all negatively affected (Archer et al., 2010; Bicalho et al., 2007; Booth et al.,

2004; Green et al., 2002; Melendez et al., 2003; Mitev et al., 2011; Morris et al., 2011; Sogstad et al., 2007; S. L. Walker et al., 2010). Poor animal welfare because of lameness also affects the farmer as the financial burden can range from $2 to $982 per cow per year (Cha et al., 2010a; Dolecheck & Bewley, 2018). Lameness prevalence (scores 2,3 of a 4-level system) in the UK has been reported to range from 0–79.2% in the past few years, but the figures should be taken with caution as farmers tend to underestimate mild cases of lameness in their herds (Leach et al., 2010a; Šárová et al., 2011a).

In clinical practice, the diagnosis of lameness in cattle is recognised as being less effective than in horses because of the relative difficulty of conducting progressive functional anatomical examinations such as flexion tests, compression tests and nerve blocks (Desrochers et al., 2001). Nonetheless, the definitive diagnosis of lameness in individual cattle is based on detailed history taking and careful individual clinical examination, sometimes with the degree of mobility impairment being scored on one of several possible ordinal categorical systems (see Afonso et al., 2020 for the most common mobility classification systems in the UK). Translating such ordinal scoring systems to applications for large numbers of cows, potentially herds, such that each animal can be scored without being removed for detailed individual clinical examination is more challenging but is necessary for machine learning applications and current benchmarking schemes to ensure high welfare on farms. For example, the Tesco Sustainable Dairy Group require their suppliers to ensure that lameness prevalence in their herds remains below 20% Anon, 2023; available at: https://www.tescoplc.com/sustainability/planet/farming-agriculture/tesco-sustainable-dairy-group/, accessed January 31, 2023). This requirement is based on systems such as the Agriculture and Horticulture Development Board (AHDB) 4-level mobility scoring system, which is supported by the Register of Mobility Scorers (RoMS). The AHDB 4-level mobility system is the most widely used mobility assessment method for dairy cows in the UK dairy industry (Afonso et al., 2020), with RoMS being the regulatory body for anyone wishing to be trained and accredited on this system. The AHDB system consists of 4 levels with distinct scores ranging from 0 to 4 and will be fully described in the following materials and method section. A RoMS accredited scorer is trained by evaluating 1000 cows via video and passing a 20-video evaluation test, which must be repeated annually at membership renewal. In

this way, an assessor is calibrated, and the process is theoretically standardized, that is, maintaining a scoring consistency over time.

In recent years, there have been attempts to introduce technologies onto the farm routine to automate the mobility assessment process. Examples are the use of 2- or 3- dimensional cameras for video images analysis (Bahr et al., 2008; Gu et al., 2018; Kang et al., 2021), pressure-sensitive walkways measuring the ground reaction forces or balance (Maertens et al., 2011), and accelerometers worn by the animal as pedometers (leg-mounted) or as collars (neck-mounted) (Beer et al., 2016; M. Pastell et al., 2009; Shepard et al., 2010; Weigele et al., 2018). Colleagues at the University of Glasgow have recently demonstrated good initial results using micro-Doppler radar (Busin et al., 2019a; Shrestha et al., 2018a), which is not affected by weather and lighting conditions and does not require wearable devices on the animal, thus avoiding potential animal discomfort and reducing the carbon footprint and a cost that increases with each increase in animal numbers.

Automated systems are generally based on supervised machine learning, which depends on annotated data (labelled data) that have been classified into distinct categories to produce an output (Cunningham et al., 2008). For a machine to learn to distinguish classes, there must be an accurate correspondence between labels and data, and each of the classes should be adequately represented in the training data and ideally classified according to a gold standard. Labelling is provided by mobility scores, which fall short of the gold standard of detailed individual animal clinical examination as they are likely subject to bias and noise, and wide variation among assessors' scores have been observed previously in lameness studies (Channon et al., 2009b; Schlageter-Tello et al., 2015b), rendering the label acquisition process challenging.

Variation arises from multiple factors, including the experience and confidence of the observers, and the training they have received. A study by Kristensen et al. (2006) has shown that for a cattle mobility assessment, experienced evaluators agree with each other in classification more than less experienced evaluators. However, another study by Garcia et al. (2015) reported no great differences between experienced and inexperienced observers for lameness scoring. Pre-assessment training has also been shown to be an important factor

in some studies (Polderman et al., 2001) but not in others (Engel et al., 2003; Ford et al., 2000; van Tubergen et al., 2003). Another feature that could affect the decision-making and, consequently, the agreement among observers is whether there is discussion among them during the evaluation process (Brenninkmeyer et al., 2007). Finally, the confidence one might have in a decision has been linked to the familiarity of an event or a process, which acts as a heuristic or cue to assist a judgement (Fitzsimmons et al., 2020). However, confidence in a decision does not imply decision accuracy (Grimaldi et al., 2015; Sen & Boe, 1991).

One approach to evaluating the validity of labels is to compare the output to a gold standard – in a dairy cattle mobility assessment study, the best reference would be a thorough physical examination of the animal's mobility by a veterinarian. Since this is a logistically challenging way to evaluate a large herd, another more accessible way is to calculate intra- and inter-assessor agreement of their mobility evaluation scores, with the assumption that higher levels of agreement suggest a greater likelihood of accurate evaluations. The inter-assessor agreement can be calculated by comparing assessors' scores under the same conditions for the same animals, which according to Cohen (1960), can provide insight into the consistency of the assessment process. The intra-assessor agreement is calculated by comparing assessors' scores for the same animals at different time points under the same conditions (Gwet, 2008). A common way to measure agreement is by calculating percentage agreement ((number of agreement scores/ the total number of scores) * 100). However, this method does not consider the agreement that may occur by chance, thus potentially overestimating the agreement between assessors, as it will produce higher agreement values. Cohen's kappa is a measure of agreement, which accounts for imbalances in class distribution and considers the chance agreement between assessors (Cohen, 1960; McHugh, 2012). It presents some limitations; for example, agreement results are affected by the number of classes into which the assessors should classify the data. That is, in a variable that takes a discrete value (0 or 1 / YES-NO), the agreement is expected to be higher compared to a variable that would be classified into classes with more levels (e.g., 9-point mobility system by Manson & Leaver, 1988). The prevalence and the frequency of the assessor's choices to assign a score to a specific level can also affect the kappa values. This phenomenon is called the

first kappa paradox and should be considered when interpreting the results or setting a threshold. It is present when the examined subjects tend to be classified to one of the possible outcomes, either due to the nature of the outcome itself and its high prevalence or because at least one of the assessors tends to assign more frequently to one specific outcome (Zec et al., 2017). Nevertheless, kappa statistics are widely used in research, and the main advantage is that high kappa values correspond to highly repeatable and accurate agreement strength.

This study aimed to evaluate and quantify inter-assessors' agreement of mobility classifications based on the AHDB scoring system, with the overarching aim to produce reliable labels for machine learning systems. The objectives were to test the agreement of multiple assessors' scores and examine other factors, such as assessors' confidence in decision-making and the role of experience, which may contribute to the improvement of labels and, therefore, to the automation of lameness detection.

2.2 Materials and Methods

This project adhered to the ethical guidelines established by the University of Glasgow and received local ethical approval (Ethics licence EA06 19), despite not involving any procedures regulated under the Animals (Scientific Procedures) Act 1986.

Data collection

The study started with on-farm, real-time assessments and progressed to remote video assessments, as a means of dealing with Covid-19 constraints and a desire to obtain more flexible expert contributions. For this reason, the study consists of assessments carried out with different media (live and camera), and the participating assessors differed.

2.2.1 Farm visits, participating animals and evaluations

Two farms in central Scotland milking exclusively Holstein-Friesian cows were selected for herd evaluation (Farm A, Farm B). A total of 3 visits were made - 2 visits to Farm A and 1 to Farm B (Table *2.1*). On the first visit to Farm A, 49 milking dairy cows were assessed on-site using the AHDB mobility classification system, and videos were recorded. During the second visit to farm A, 52 milking dairy cows were assessed on-site only. The visit to Farm B was conducted solely for video recording of 69 cows.

All the milking cows of Farm A on each visit day were included in the study, and the first 69 milked animals of Farm B were used, to reduce the level of interference in the farm operations. All cows presented were scored, without exception.

Table 2.1 Sites, type of each assessment, the total number of animals assessed, and the dates of each farm visit.

Site	Type of assessment	Number of animals	Date
Farm A	On-farm + video	49	13-02-2020
Farm A	On-farm	52	04-03-2020
Farm B	Video	69	01-09-2020

2.2.2 Video recordings

A rugged camera (Kodak PlaySport Zx5 Full HD 1080P) was used to capture video of the cows walking through a concrete-floored race after exiting the milking parlour. A temporary race (6.9 m length x 1.65 m average width) was created for the study on Farm A, with steel fencing panels, extending beyond the permanent exit passage from the dairy, which had solid walls on either side and in which cattle could not be seen from the side (Figure 2.1 A). The panels were installed five days before each visit to accustom the animals. The race running through the centre of one of the buildings on Farm B was a permanent installation, used by the animals routinely. For flexibility and a wider field of vision, a person operated the camera from 1 – 5 m on the side of the race rather than using a fixed camera. The video recordings were processed with 'mp4compress/mute-video' (FileConverto Network, 2020) and 'video online cutter' (123apps LLC, 2020.) to remove sound and to obtain clips of every cow, with no overlap with another animal.

Figure 2.1 Figure A shows the permanent race with solid walls on the farm. Figures B and C are consecutive snapshots of the temporarily constructed race we used to capture videos for assessing the animals. The vantage-point of the person who recorded the animals is the same vantage-point from which the live assessment took place.

2.2.3 On-site assessments

Prior to the visit, training was provided by scoring online videos of representative of each class, publicly offered by the AHDB (i.e., https://www.youtube.com/watch?v=rFj72vqUwlU, accessed February 2020). During visits to Farm A, two experienced veterinarians and three veterinary students (collectively, the "assessors" in this study) were provided with the AHDB template mobility score form (https://ahdb.org.uk/knowledge-library/dairy-mobility-scoresheet, accessed February 2020) and scored all the cows individually following the guidelines of the Agriculture and Horticulture Development Board (AHDB) system. In each on-site assessment, the assessors scored the animals simultaneously and independently, standing one meter apart from each other, without any discussion or sharing of scores.

2.2.4 Video assessments

The edited videos were saved into a PowerPoint document with detailed instructions and a description of the system the assessors should use to rate the videos from Farm A and B. Each slide showed one cow moving along the race, and a table for the assessors to complete:

(1) the score for the cow
(2) the assessor's confidence in the score
(3) a second score in cases where they responded negatively to the preceding question about their confidence
(4) how many times they watched the video before making a decision
(5) any comments on why they could not provide a score.

Figure 2.2 A shows the slide with the instructions included in the shared PowerPoint file, and Figure 2.2 B provides an example of the slides with the video and the table the assessors were asked to complete.

The assessors involved in the on-farm and video scoring performed the assessments at least one month apart to avoid bias, such as recall of the animals and previously assigned scores.

A

Instructions

- The videos will run while not in presentation mode. Leave it in this mode so that you can type your responses into each of the slides.
- In the following slides you will watch 49 videos of cows walking towards the barn.
- Please look at the videos, as often as necessary, to provide a mobility score for each cow.
- Type your mobility score next to "Score" and below that, type the number of times you viewed the video to be confident of the score.
- State the confidence you have in your score (Confident/Uncertain).
- State the next most likely score you would give if you answered that you are uncertain about the score.
- You are also asked to state how many times you watched each video until you make your final decision on the score.
- For each cow, if there is a reason why you cannot provide a score that you are confident in, please state it in the appropriate comment box.
- You should evaluate the movement of cows according to the scoring system described in detail below.

B

Cow 3	
Score (0 to 3)	
Certainty? (Y/N)	
2nd score if uncertain	
Times video was played	
Comments	

Figure 2.2 (A) Instructions of the shared PowerPoint file for the video assessments, and (B) a selected representative example of one slide with the table the assessors were asked to fill.

2.2.5 Participating assessors

In total, 13 assessors provided scores, with between 5 and 8 assessors providing scores for each site and mode (Table 2.2). The same assessors were not retained for all study elements due to logistic challenges, including the Covid-19 pandemic.

Table 2.2 Number of assessors participating in each evaluation, their experience and occupation. Experienced assessors have performed mobility scores on farms and are active veterinarians, hoof trimmers, and registered mobility scorers. Vet students were familiar with the mobility scoring process and lameness identification in theory but had performed fewer than five assessments in practice.

Assessors	Occupation	Experience in mobility scoring	Assessments			
			Farm A on-site 1	Farm A on-site 2	Farm A Video	Farm B Video
1	Bovine Veterinarian	Experienced	✓	✓	✓	✓
2	Bovine Veterinarian	Experienced	✓	✓	✓	✓
3	Bovine Veterinarian	Experienced	✓			
4	Vet. student	Inexperienced	✓			
5	Vet. student	Inexperienced	✓	✓		
6	Vet. student	Inexperienced	✓	✓		
7	Vet. student	Inexperienced		✓		
8	Bovine Veterinarian	Experienced			✓	✓
9	Bovine Veterinarian	Experienced (Registered mobility scorer)			✓	✓
10	Bovine Veterinarian	Experienced			✓	✓
11	Bovine Veterinarian	Experienced			✓	✓
12	Bovine Hoof trimmer	Experience (NACFT * member, Category 1 qualifications)				✓
13	PhD Student in Animal Sciences	Experienced (Registered mobility scorer)			✓	✓

**National Association of Cattle Hoof Trimmers in the UK (https://nacft.co.uk/)*

2.2.6 Mobility scoring system

Mobility scoring was performed using the AHDB system (AHDB, 2020., https://ahdb.org.uk/), the most widely accepted system in the United Kingdom farm sector. It is a 4-scale system from 0 to 3, where 0 represents a sound cow and three a severely lame cow. The presentation and interpretation of the system follow in Table *2.3*.

Table 2.3 Presentation of the four-level AHDB - Dairy mobility scoring system, which was used for the initial video scoring of the cows

Category of score	**Score**	**Description of cow behaviour**
Good mobility	0	Walks with even weight bearing and rhythm on all four feet, with a flat back. Long, fluid strides possible.
Imperfect mobility	1	Steps uneven (rhythm or weight bearing) or strides shortened; affected limb or limbs not immediately identifiable
Impaired mobility	2	Uneven weight bearing on a limb that is immediately identifiable and/or obviously shortened strides (usually with an arch to the centre of the back)
Severely impaired mobility	3	Unable to walk as fast as a brisk human pace (cannot keep up with the healthy herd). Lame leg easy to identify – limping; may barely stand on lame leg/s; back arched when standing and walking. Very lame.

Scoring system transformations

We transformed the four-level AHDB system for statistical analysis into

(1) a binarised-AHDB system by collapsing from a 4 to a 2-level score, with 0 and 1 becoming 0, and 2 and 3 becoming 1

(2) a convergent-AHDB system was created from those observations for which the assessors provided more than one score each (ie were not 100% confident – Farm A second scores 113 / total scores 315).

   a. Where all assessors had the same first choice score (very rare), that was the convergent score.

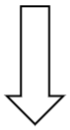
   b. The modal first choice score was selected as the convergent score if it was also nominated as the second choice by the other assessors.

   c. Where there was no unique modal first choice score, the second choice scores were considered for all assessors, and the modal first and second choice score was chosen as the convergent score.

   d. Where there was no clear modal score, but two scores were equally common, it was randomly allocated to the higher or lower score – the first allocation being by coin-flip and subsequently by alternation.

An example of the convergent system is presented in Figure 2.3.

AHDB scores	Assessor 1		Assessor 2		Assessor 3	
	Score A	Score B	Score A	Score B	Score A	Score B
Cow 1	1	2	2	1	2	3
Cow2	1	0	0	1	2	1



Convergent scores	Assessor 1	Assessor 2	Assessor 3
Cow 1	2	2	2
Cow 2	1	1	1

Figure 2.3 Generation process of the Convergent scoring set out using the second given scores in case of uncertainty.

Data handling and statistical analysis

All collected data were manually transferred from paper in the case of on-farm assessment or from PowerPoint files for video assessment into excel spreadsheets. Files were stored on a PC hard drive with backups to an external hard drive and the University of Glasgow's OneDrive cloud. All files used in R program analysis were converted from .xlsx files to .csv files. The files were restricted access, and only the project supervisors and I could access the saved files.

All statistical analysis was performed in R (R Core Team, 2020) using the 'IRR' (Gamer et al., 2019) and 'ggplot2' (Wickham, 2016) packages.

All the available data were used. Percentage agreement and kappa statistics were obtained for agreement among, between and within the assessors. Data from the video evaluation, including the assessor's confidence, their comments and the number of video views, were used in regressions to quantify their relationships with the scores and with the level of agreement.

2.2.7 Inter-assessor agreement

We calculated inter-assessor agreement between and among assessors using the scores from the AHDB-4-level, the convergent-AHDB, and the binary transformed scores of these two systems (binary-AHDB and binary-convergent-AHDB) for the on-site and video assessments. We analysed the agreement using kappa statistics ('kappa2' function for pairwise comparisons and 'kappam.fleiss' function for comparison of multiple assessors).

2.2.8 Intra-assessor agreement for Farm A Live vs Video assessment

We were able to calculate intra-assessor agreement after having the same animals evaluated by the same two assessors (Assessors 1 and 2) on the farm and after one month via video. We again used kappa statistics and percentage agreement to calculate their agreement.

For the calculations, we enabled zero tolerance, meaning that assessors must have given the same score for an agreement to be reached. The interpretation of the kappa statistics was based on Landis & Koch (1977)-Table *2.4*.

Table 2.4 Landis and Koch (1977) kappa indices interpretation.

Kappa	Strength of agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

2.2.9 Generalised linear models (GLM)

We used a generalized linear model to quantify the associations between score (AHDB, binarized AHDB, Convergent-AHDB, binarised-convergent-AHDB) as the dependent variable and the comments (comment = 1, no comment =0), the confidence (yes=1, no =0), and the number of video views before each decision (numeric) as potential explanatory variables.

$$Y = \beta_0 + \beta_1 X$$

Where,

Y = the dependent categorical variable – the scores

β_0 = intercept

β_1 = the coefficient of the variable

X = the variable (confidence (Yes/No) or number of views or comments (Yes/No))

Models with the lowest Akaike information criterion (AIC) values were selected. Bonferroni adjustment for multiple testing was used (familywise error rate (0.05) / the number of tests).

Comments were also analysed according to whether the comments addressed the characteristics and attributes of the cow or the video, as in Table 2.5.

Table 2.5 The comments were classified into two categories; comments on the video characteristics and the way of presentation that did not serve the ease of the evaluation and thus affected the confidence in the score, and other comments.

Video attributes/characteristics	Other comments
Cow running - not walking	Comments about lameness localisation in the limb (i.e., lame in front right leg)
Stopping many times during the clip	Udder size
Short clip duration	Limb movement
Limbs not in the frame	Arched back
Visual obstructions (i.e., rails hiding body parts)	Behaviour - exploratory, weird
Slippery surface	

After the categorisation and classification, we included the comments in the generalized linear model and the equation transformed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Where,

Y = the dependent variable – the scores

β_0 = the intercept, which is always a constant in the model

$\beta_{1,2}$ = the coefficients of each respective variable

X_1 = other comments

X_2 = comments about video characteristics

2.2.10 Role of experience in agreement

We were interested in the effect of experience on the inter-assessor agreement, so we combined the kappa values of the two on-farm assessments to obtain more observations and then visually checked their distribution with the "*descdist"* function (Delignette-Muller & Dutang, 2015) in R, and we performed a Shapiro test of normality (data samples <50). We then performed a parametric (t-test) and a non-parametric (Kruskal Wallis) test on the groups of interest (experienced vs experienced, inexperienced vs inexperienced, experienced vs inexperienced) to determine whether experience affects the agreement.

2.3 Results

2.3.1 Inter-assessor agreement

Table 2.6 lists the results of inter-assessor comparisons (kappa Fleiss) for each assessment using the AHDB-4-level scoring system, with agreement among assessors with the same experience in cow mobility assessment. Average pairwise Cohen's kappa and percentage agreements with standard deviations are presented for comparison. The best agreement was achieved between the two registered mobility scorers (RoMS) for Farm B video assessment (kappa = 0.53; per cent agreement = 65.22%), corresponding with a moderate level of agreement according to Landis & Koch (1977). The remaining kappa values ranged from 0.16 to 0.39, showing slight to a fair agreement. The highest percentage agreement (69.81%) was observed among the three veterinary students on the second visit to Farm A. We observed only slight differences in the results between Fleiss's kappa directly applied to the assessors' scores and the averaged Cohen's kappa of the pairwise comparisons, as seen in Table 2.6. Meanwhile, increased percentage values were reported after averaging scores and recalculating agreement, such as Farm A Visit1 Video, which from 2.13% agreement in the first calculation transformed to 41.9%.

Table 2.6 Inter-assessor agreement (kappa statistics and percentage agreement) for the assessments using the AHDB mobility scoring system. Comparisons with kappa Fleiss and pairwise comparisons with Cohen's kappa were made among all assessors of each assessment and assessors based on scoring experience, i.e., veterinary students or veterinarians.

Location	Assessment type	Number of assessors	Kappa	% Agreement	Average pairwise Cohen's kappa (SD)	Average % agreement (SD)
Farm A Visit-1	ON-FARM	3 (vet students)	0.29	34.69	0.29 (0.1)	47.8 (11.53)
Farm A Visit-1	ON-FARM	3 (vets)	0.19	34.69	0.22 (0.17)	50.79 (6.97)
Farm A Visit-1	ON-FARM	6 (all)	0.21	14.29	0.21 (0.12)	49.1 (6.19)
Farm A Visit-2	ON-FARM	3 (vet students)	0.24	69.81	0.24 (0.18)	47.8 (11.53)
Farm A Visit-2	ON-FARM	2 (vets)	0.39	58.49	0.39	58.49
Farm A Visit-2	ON-FARM	5 (all)	0.27	16.98	0.28 (0.10)	50.19 (6.6)
Farm A Visit-1	VIDEO	7 (all)	0.16	2.13	0.20 (0.13)	41.90 (10.1)
Farm A Visit-1	VIDEO	2 (ROMS)	0.37	55.1	0.37	55.1
Farm B	VIDEO	8 (all)	0.33	8.96	0.33 (0.1)	51.27 (8.46)
Farm B	VIDEO	2 (ROMS)	0.53	65.22	0.53	65.22

2.3.1.1 On-site assessments

The results of the pairwise comparisons for the two on-site assessments of the same cows on Farm A are presented in Table *2.7*. When the AHDB-4 level system was used for analysis, kappa results showed fair agreement after averaging the pairwise kappa indices for both visits. After converting the four-level system to binary, the kappa results improved, giving moderate agreement, and the percentage agreement increased from 50-51% to 80-85%.

Table 2.7 Average kappa values (SD), and percentage agreement of the two different scoring systems for pairwise comparisons of the two on-site assessments of the same cows at farm A.

	AHDB (4-level)		**AHDB (Binarised)**	
	Kappa (SD)	% Agree (SD)	Kappa (SD)	% Agree (SD)
Farm A on-site Visit 1	0.21 (0.12)	51.16 (6.92)	0.45 (0.21)	85.45 (5.15)
Farm A on-site Visit 2	0.33 (0.1)	50.19 (6.61)	0.5 (0.11)	80.75 (3.74)

2.3.1.2 Video assessment

Table *2.8* shows the average kappa values and percentage agreement of pairwise comparisons for both farm video assessments. The assessors were the same individuals, with the addition of one experienced evaluator. Regarding kappa statistics, the binarised-converged-AHDB system had the highest indices, followed by the binarised-AHDB system. The AHDB-4-level mobility system produced only a slight (0.17) and fair (0.33) agreement for the two farms. The percentage agreement followed the same trend as the kappa values, meaning the highest score was observed in the binarised-convergent-AHDB system, followed by the binarised -AHDB, then the convergent- AHDB and lastly, the AHDB-4-level system.

Table 2.8 Average kappa values, percentage agreement and standard deviation (SD) of the three different scoring systems for pairwise comparisons of the two video assessments.

	AHDB – 4-levels		**Binarised-AHDB**		**Convergent-AHDB**		**Binarised-Convergent-AHDB**	
	Kappa (SD)	% Agree (SD)	Kappa (SD)	% Agree (SD)	Kappa (SD)	% Agree (SD)	Kappa (SD)	% Agree (SD)
Farm A video	0.20 (0.13)	41.90 (10.1)	0.37 (0.19)	71.39 (9.23)	0.36 (0.12)	54.12 (9.74)	0.58 (0.13)	80.84 (6.79)
Farm B video	0.33 (0.1)	51.27 (8.46)	0.53 (0.12)	77.76 (7.1)	0.44 (0.09)	59.51 (7.39)	0.67 (0.15)	84.74 (7.85)

2.3.2 Intra-assessor agreement

2.3.2.1 Farm assessment

Because the same two assessors (Assessors 1 and 2) evaluated the same animals at different time points (on-site and video), we could calculate the intra-assessor agreement. The results revealed only a slight agreement for Assessor 2 (Kappa: 0.18, Agreement: 49%) and a fair agreement for Assessor 1 (kappa: 0.23, Agreement 46.9%). That means the assessors' scores differed when scoring the same animal on-farm and on video. Both intra-assessor and percentage agreement indicate more than a 50% discrepancy between the two evaluations of the same animals.

2.3.3 Relationship of confidence, video comments and video viewing times to scores

2.3.3.1 Farm A Video

Fewer viewing times were associated with higher scores for Assessors 11 and 13 (Figure 2.4). Conversely, Assessor 8 was more likely to give a high score in cases where they watched the video more times.

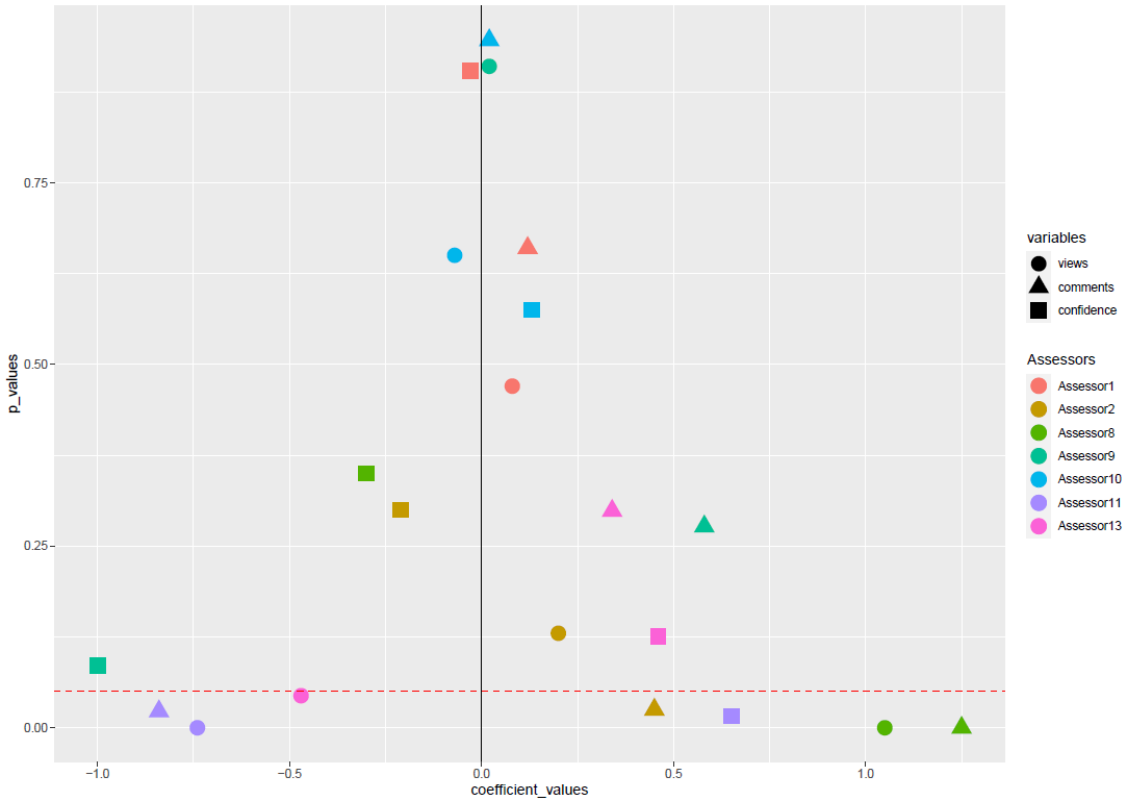


Figure 2.4 Coefficients and p-values from the generalized linear models for the associations between the individual AHDB scores and the (1) views, (2) comments, and (3) confidence for the Farm A video assessment. Colours represent the coefficients of each assessor, the shapes represent the coefficients of each category (views, comments and confidence), and the red dotted vertical line points to the p-value=0.05. Observations below the dotted vertical line are considered statistically significant.

In the two videos for which all assessors declared a high confidence level, the difference between the highest and the lowest scores given by assessors was three levels – i.e., the same cow confidently rated as 0 and 2 on the four-level scale.

After classifying the comments into two categories and using them in the generalized linear model as covariates against the scores, Assessors 2, 8, and 13 were shown to have statistically significant and positively correlated results for the comments they made about cows (Table 2.9). The comments were mainly about the localisation of the lesions and other animal characteristics (Table 2.5). Assessor 11 was the only one with a significant negative correlation between their scores and the video characteristics, meaning they were likely to comment about the video attributes and assign a lower score. Assessor 11's comments about the video were about the animal running instead of walking.

Table 2.9 GLMs coefficients and p-values of the categorised comments for Farm-A.

Assessors	Cow comments coefficient	Cow comments P-value	Video comments coefficient	Video comments P-value
R1	-0.24	0.51	0.17	0.63
R2	0.70	<0.001	-0.12	0.59
R8	1.19	<0.001	0.28	0.62
R9	-0.17	0.87	0.64	0.34
R10	0.32	0.26	-0.43	0.20
R11	-0.34	0.40	-1.47	0.01
R13	0.74	0.01	-0.52	0.18

2.3.3.2 Farm B Video

The number of video views of Assessor 11 was negatively associated with their scores, as shown in Figure 2.5. The views of all other assessors did not significantly associate with their scores.

In the video scoring of Farm B, all assessors scored 11 of the 69 cows confidently. However, in only four cows out of 11, there was full agreement on the AHDB 4-level score. The assessors agreed on the severely impaired animals – assigning a score 3. Only Assessors 2 and 13 had a significant association between their stated confidence and their assigned scores. Assessor 2 tended to be confident when assigning a low score, whereas Assessor 13 had higher confidence with higher scores.

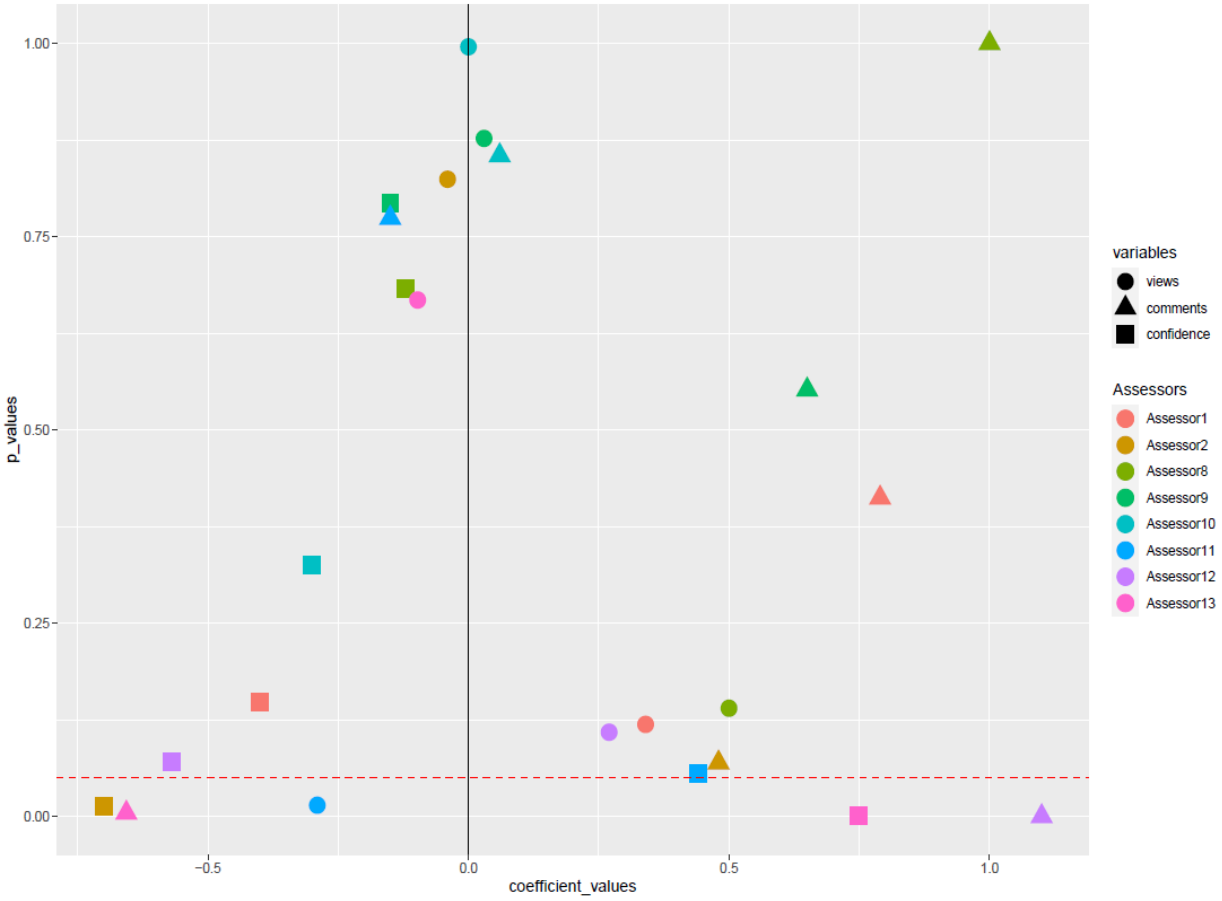


Figure 2.5 Generalized linear model coefficients and p-values from the video assessment of Farm B. The individual AHDB scores were compared to the (1) views, (2) comments, and (3) confidence of each assessor. The red dotted vertical line represents the p-value 0.05. The colours represent each assessor's coefficient point.

The comments of Assessor 13 were about the length of the video and that they would have preferred to be able to see the animals for longer.

Table 2.10 Results of the GLMs for the categorized comments and the AHDB scores for Farm-B.

Assessors	Cow comments coefficient	Cow comments P-value	Video comments coefficient	Video comments P-value
R1	-	-	0.79	0.412
R2	0.87	0.002	-0.5	0.209
R8	-	-	-	-
R9	0.65	0.55	-	-
R10	-	-	0.06	0.855
R11	-	-	0.55	0.3
R12	1.21	0.0000089	0.41	0.464
R13	-0.47	0.068	-0.69	0.009

2.3.4 Role of experience in agreement

The '*descdist*' function and the Shapiro test (p=0.69) showed that the kappa values were normally distributed (Figure 2.6). The results of both tests we performed produced p-values greater than 0.05 (average p-value of t-tests for the three groups = 0.99 and Kruskal Wallis p-value = 0.76), meaning that there were no significant differences between groups, and thus experience had no effect on the inter-assessor agreement.

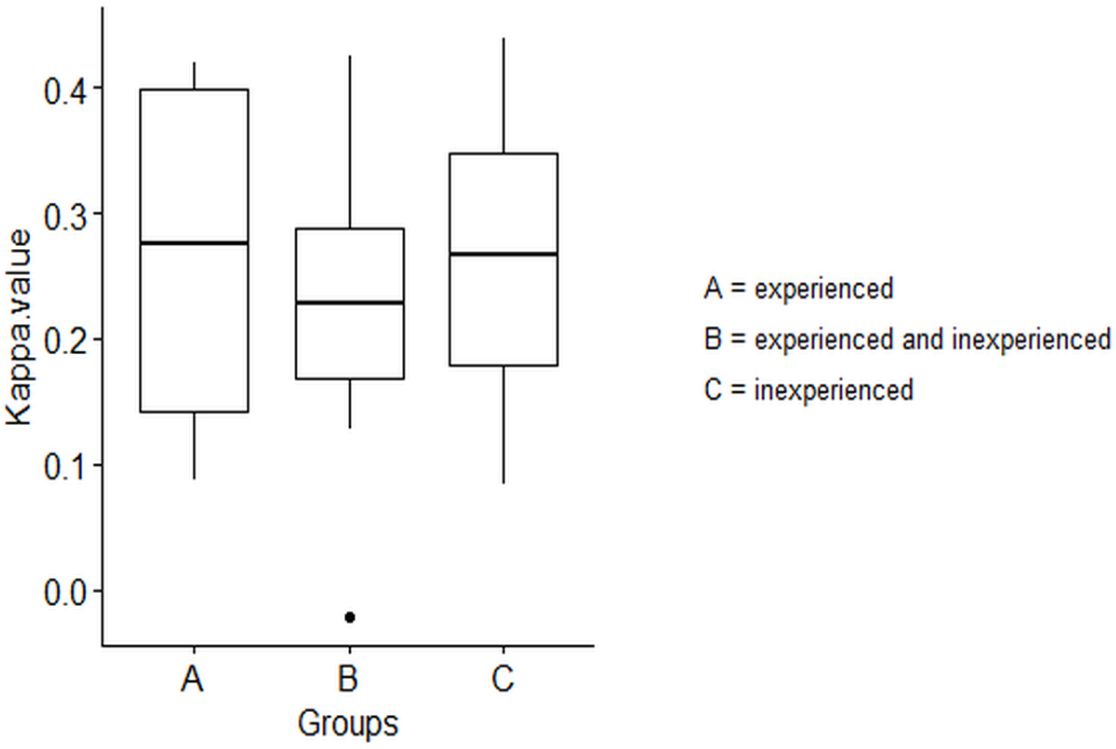


Figure 2.6 Boxplot of the kappa values of the three groups of interest divided by experience. The plot shows no statistically significant differences between the groups; thus, the experience level did not play a significant role in the inter-assessor agreement.

2.4 Discussion

This project followed our group's previous work with micro-Doppler radar, in which encouraging accuracy (>80%) was obtained using on-farm scoring by a single assessor (Busin et al., 2019; Shrestha et al., 2018). We wished to expand on those studies with longitudinal and larger cross-sectional studies. Prior to commencing these studies, we sought to confirm that the labelling was unbiased by the person undertaking the assessment and was repeatable among assessors. We hypothesised that accurate and repeatable labels would be obtained using agreed or consensus scores from multiple assessors and wished to determine the minimum number of assessors and the best way of obtaining consensus. Adaptively due to covid-19 pandemic, we planned to use video recording to allow multiple assessors to rate all the cows, with the additional benefit of minimising the logistical challenge of having multiple assessors on the farm interfering with normal cow behaviour. The initial findings of poor to fair inter- and intra-assessor agreement were unexpected and created further questions, which we tried to address in the experimental design. The main finding was variation among assessors' scores when they used the AHDB system to evaluate cow mobility on two farms by visual on-site and video assessments.

After analysing the inter-assessor agreement levels with the AHDB scorings, we observed high variation, with kappa values ranging from 0.16 - 0.53 (Table 2.6). These findings are consistent with other studies on cow mobility assessment using visual observations. For example, in Thomsen et al. (2008) study, inter-assessor agreement ranged from 0.24 to 0.68 using a 5-level visual rating system. Several other studies have presented kappa values below or near the 0.60 threshold, which indicates substantial agreement when assessing mobility and other relevant traits such as leg scores (Channon et al., 2009a; Croyle et al., 2018; Dahl-Pedersen et al., 2018; Holzhauer et al., 2005; Katzenberger et al., 2020; Schlageter-Tello et al., 2015a). However, some other studies on lameness presented pairwise kappa values that were substantial or nearly perfect (i.e., Barker et al., 2010; Garcia et al., 2015). For example, in the study of Barker et al. (2010), the kappa values ranged from 0.67 to 0.93. Still, while presenting the results of agreement using the kappa statistic in a binary system, Barker et al. 2010 did not provide details about the dichotomisation of levels or other information about the statistical analysis of the inter-assessor

agreement. In short, the agreement of the present study is at the same levels as those reported in most of the literature, showing only up to a moderate agreement between visual observation assessors.

The conversion of the four-level system to binary and convergent increased the agreement as expected. Other studies have dichotomised systems with multiple levels to estimate and improve inter-assessor agreement (Brenninkmeyer et al., 2007; March et al., 2007). The better performance when the number of classes is lower is due to the need for lower discrimination (Knierim & Winckler, 2009). The evaluator needs to decide between only two options, which, depending on the desired result, may offer advantages over a system with more levels that might have finer divisions but also have more chances of producing more score discrepancies between two or more assessors. One study (Garcia et al., 2015) showed better inter-assessor agreement at the lower (1,2) and higher (4,5) levels of a 5-level system as opposed to the middle classes (2,3 or 3,4), where differentiation may be more challenging. Another study by Schlageter-Tello et al. (2014), merging a 5-level lameness detection system into fewer levels with all possible combinations, also found difficulty among experienced assessors in identifying cows with slight variations in mobility. In the same study (Schlageter-Tello et al., 2014), the best results were obtained when the system was dichotomised. The present study also had the best agreement results in dichotomised systems, and the dichotomy precisely in the middle of the levels (0 and 1 vs 2 and 3) was logically chosen following the line of the AHDB system, suggesting additional actions for animals with scores greater than or equal to 2. The overall number of classification levels affects the kappa statistical analysis method, generally giving better results when fewer levels are used. But even lowering the levels of a system does not necessarily guarantee higher kappa values, which is why one should also state the per cent agreement of the evaluations to provide a complete picture of the assessments.

We asked the reviewers whether they were confident in the scores they gave, and the confidence level was not significantly associated with the inter-assessor agreement. Even when high confidence was claimed, the agreement was poor. Of all the cows evaluated at both visits (118 cows in total), all assessors had the same positive confidence level in their score on only 13. Even in those cases where all assessors gave their score with confidence, the scores

were the same only for score level 3, indicating that for severely lame cows, a confident agreement is more likely.

Most scores, and thus inter-assessor agreement, were not significantly or consistently affected by comments, video viewing times, or video characteristics. These results suggest that the videos and the way the cows were captured in the videos were not consistently identified as a problem by the assessors, and thus the difference in the inter-assessor agreement is probably not due to these causes. Video cow mobility assessment is commonly used for scientific research (Garcia et al., 2015; Schlageter-Tello et al., 2015a) and industry and is consistent with the examination and calibration testing methods used by RoMS. Studies have found no statistically significant differences in assessors' agreement between live/live, live/video, or video/video mobility assessments (Bernardi et al., 2009; Channon et al., 2009a). However, in our study, the two assessors' intra-assessor agreement for live/video was only fair (average kappa = 0.21 and average per cent agreement = 48%). We expected better inter-assessor agreement when using the videos for mobility assessment, as assessors had the same vantage point and could watch each video multiple times to reach a decision and even return to the video after assigning a score if they changed their minds. However, live on-farm assessment gives the advantage of being able to follow and monitor the animal potentially for longer and see it from more angles and not only the side, as in our study.

Experienced and inexperienced assessors had a similar agreement. Pre-assessment training may be important, but there appear to be conflicting results from the literature. In one study (March et al., 2007) on the effect of training on the inter-assessor agreement of lameness, even limited practical experience improved agreement and proposed intensive training procedures with animals and the presence of an experienced observer to achieve further improvement. Other studies have also concluded that training positively affects inter-assessor agreement (Gibbons et al., 2012; Vanhoudt et al., 2019). However, contrary results have also been published in medical literature, showing no difference before and after training implementation (van Tubergen et al., 2003). Other studies on the effect of training in the inter-assessor agreement of lameness scores produced ambiguous results (Engel et al., 2003; Garcia et al., 2015). For example, Engel et al. (2003) found that further

training affected the results but only sometimes positively, as some observers underestimated or overestimated some cows' conditions; at the same time, assessors improved their agreement in extreme lameness cases. Although our study did not extensively assess how training affects the agreement among and between assessors, we found that the two registered mobility scorers had the highest pairwise agreement compared to the mean, and it is a factor that deserves further investigation. Finally, it should be noted that our study only involved a limited sample size, so the results should not be generalised.

Another issue that compromises the reproducibility of the scores is the lack of exclusive and exhaustive classes within the mobility system. For example, the AHDB mobility system, at level score 3, states that a cow should be classified in this category if she "cannot keep up with the healthy herd". This phraseology, which is also a guideline for the assessment, introduces two issues; first, the evaluation is done individually for each animal, so it cannot be applied, and second, the assessment is comparative with the rest of the herd, which means that animals seen first at the time of the evaluation could be treated differently score-wise compared to animals that appeared at the end. Therefore, the system needs consistent guidelines and specifications to ensure that the mobility assessment is performed invariably and reliably, regardless of who performs it or where it is performed. When the classes or levels are unclear and vague, achieving high levels of agreement and repeatability or comparing result scores across different studies is challenging and potentially unreliable.

In addition to a system being open to subjective interpretation, assessors have been shown to have personal preferences for which traits and characteristics of the cow's mobility they choose to focus on during the evaluation (Garcia et al., 2015), which is an additional factor that directly affects levels of agreement. Other factors that act on the assessors, such as environmental stimuli, like time pressure, attentiveness, distractions during the assessment and biases, have also been proven to influence decision-making processes (Hall, 2002; Klapproth, 2008; Maule et al., 2000; Sheng et al., 2022). Generally, a human assessor can be influenced by several factors when evaluating and making decisions, which in turn shows the usefulness and necessity of a support mechanism such as an artificial intelligence system that will produce consistent results regardless of unrelated inputs.

The study's results highlight the challenge of obtaining high inter- and intra-assessor agreement and repeatability among assessors in dairy cows' mobility scoring using the AHDB system.

2.5 Conclusions

In conclusion, we achieved our initial aim to examine agreement among multiple assessors and found variation in scores when the AHDB mobility assessment system was used. We also found that training (RoMS), experience, and confidence in the scoring decision did not contribute to a better agreement. The variation and, thus, the uncertainty in the actual status of the animal makes it challenging to obtain ground truth labels for automated lameness detection using machine learning. However, transforming the scores (convergent-AHDB) and reducing the levels offer hope for better label performance. Based on these findings, it seems necessary to continue the search for high-confidence mobility labels that accurately describe the animal's state and use them as ground truth in machine learning for automating lameness detection.

Chapter 3

Improving confidence in cattle mobility scores as labels for machine learning: Analysis of video-based assessments and relationship with physical examination lesions - A longitudinal study Part 1

3.1 Introduction

This study seeks to improve animal welfare by introducing automation in modern farm production units. Various automated systems have been installed on farms over the past few years, such as robotic milking systems (M. Pastell et al., 2006) and automatic cattle feeders (Wilson et al., 2018), to improve productivity and efficiency, produce accurate information, and reduce labour costs. Although smart farming, which uses artificial intelligence systems, is a relatively new concept, it is becoming increasingly prevalent and rapidly advancing. This study aimed to find accurate labels to train machine learning and artificial intelligence to detect lameness effectively and promptly. By following the same cows over time, a longitudinal study will increase confidence in the automatic detection and classification of lameness, allowing for the detection of patterns and changes in the individual animals as well as the overall herd population, providing a comprehensive picture of mobility which can then be generalised to other herds.

Lameness is an important cause of poor welfare in dairy cows, and along with fertility and mastitis, are the main reasons for early culling (Enting et al., 1997; Leach et al., 2010c). Nevertheless, lameness has not received the same attention as the other issues, as farmers tend to underestimate the prevalence and severity of lameness incidents in their herds (Leach et al., 2010a; Šárová et al., 2011b). Foot lesions account for most lameness cases (Murray, Downham, et al., 1996), although not all foot lesions always result in mobility alterations (Manske et al., 2002a). Nevertheless, failure to prevent or detect lameness early, either in the hoof or upper leg, can cause severe problems for

animal welfare and financial loss through subtle adverse changes in animals' health. Recent research has shown that a farmer's total loss arising from a lameness incident can reach up to $533 per animal per year (Dolecheck & Bewley, 2018), with productivity (Green et al., 2002) and fertility (Garbarino et al., 2004) being impaired in severe cases without early treatment. Therefore, a robust, timely, and effective detection system should benefit both the animal and the farmer.

Several automated sensor-based systems for lameness detection, including long-distance pedometers, force platforms, and accelerometers (Byabazaire et al., 2019; Chapinal & Tucker, 2012; Rajkondawar et al., 2002; Pastell et al., 2009; Taneja et al., 2020), have been applied and tested in diverse farm environments. All are promising but also have a downside. Some disadvantages of applications that have been described are, for example, in the case of camera use, the need for relative lighting since darkness or light refraction can affect visibility and, therefore, results (Poursaberi et al., 2010). Force platforms require the animal to stand still (Mokaram Ghotoorlar et al., 2012) or to be handled to collect the necessary data, which can be a time-ineffective process in an everyday farm routine. Another disadvantage concerns the equipment (i.e., accelerometers, pedometers) that needs to be attached to the animals and replaced or increased with every increase in animal numbers, thus expanding the costs and the carbon footprint. These are a few motivations for adopting radar technology for farm applications and lameness detection. A radar such as the recently proposed micro-Doppler system (Shrestha et al., 2018) has the advantages of not requiring physical contact with the animals and not being dependent on specific lighting or environmental conditions to collect valuable data. At the same time, the system has been tested on humans and animals (cows, sheep, horses), proving that it is possible to detect and classify mobility patterns with outstanding results (overall accuracy >80%) (Busin et al., 2019; Fioranelli et al., 2019; Shrestha et al., 2017; Shrestha et al., 2018). However, like other technological systems that use supervised machine learning-based algorithms, this system's challenge is the requirement for reliable labels.

Labels are necessary for algorithms based on supervised machine learning, but they have a few potential issues. First, humans generate labels, meaning that a certain degree of bias and subjectivity are usually involved in the process.

Subjectivity and bias lead to variation in the labels generated by different assessors (see Chapter 2), impacting the performance of the machine learning algorithms (Lebovitz et al., 2021). Second, the performance of the algorithms is highly dependent on the quality and representativeness of the input training data. Models trained with labels generated from a small population without full representation usually perform poorly with new and diverse data. Third, limited datasets, with insufficient labelled data available to train machine learning algorithms, generally lead to limited performance and potential overfitting of the algorithms (Bashir et al., 2020). One example of a poor performance model is from the medical imaging field; in the study of Brinker et al. (2019), they presented a model with high accuracy in detecting melanomas trained in primarily white males, but when tested with dark skin patients (Kamulegeya et al., 2019), the diagnostic accuracy was reduced by almost half. Another example of poor labels comes from a study by Hendrycks et al. (2018) in which, in image classification algorithms, even a small amount of incorrect annotated labels used during the training decreased the accuracy and led to overfitting; thus, the model could not generalise on other data producing the desired results. All these considered, paying attention to the training data's quality and the impact that labels can have on the performance of machine learning models is a necessity if the goal is a valid, accurate and reproducible outcome.

Selecting labels as the ground truth requires consideration of their relevance, accuracy, and consistency. However, a significant challenge arises when uncertainty exists regarding the validity of a label. In such cases, identifying appropriate labels could become a complex task. Several studies (Galati et al., 2022; Liu et al., 2021; Wang et al., 2021) focus on selecting reliable ground truth labels when there is uncertainty, which is especially true in the diagnostic medical field. In medical diagnosis, experts often encounter a significant degree of uncertainty, rendering incorporating artificial intelligence (AI) tools particularly attractive as objectivity and assistance in decision-making will be introduced. The main problem is that for a model to learn to predict and classify, it needs to be trained; training is usually done using human-generated labels, which are not always reliable. However, despite the uncertainty and the lack of agreement between assessors, modern technological developments have changed the landscape for generating reliable labels. Especially in the

field of diagnostic radiology, which is pioneering in the field of AI, where ML-based tools claim to produce high-quality results and even outperform experts in some cases (Galati et al., 2022; Gulshan et al., 2016; Wang et al., 2017).

Lameness detection and classification in cattle is also a medical diagnosis task. The gold standard for detecting lameness in cattle is the individual physical examination during which the animal's limbs are lifted and carefully examined, its mobility is assessed, and its medical history is taken into account, including previous lameness incidents and hoof trimmings (Desrochers et al., 2001). The limitation to this gold standard is that the process is time-consuming and logistically impracticable when a herd consists of a few hundred animals. For this reason, evaluations have been created that are indirectly based on detecting lameness problems through the visual observation of the cows' walking patterns. There is a standard way of walking, and everything deviating from that is considered a mobility issue. However, it has been observed that there are differences in the perception of lameness characterisation among different assessors, which renders automation the next rational measure towards an objective, standardised and prompt way of detection and monitoring. In this light, we aimed to identify a reliable method to generate labels that accurately reflect animal mobility conditions and can be used to develop a supervised machine learning-based AI tool. The objectives were to combine observation and analysis systems to assess mobility, using video recordings shared with expert assessors in conjunction with clinical examination of animal hoof pathology, to strengthen confidence in labels. By designing this longitudinal study, we would like to observe changes in gait patterns over an extended period of time and gain insights from multiple sources of information into the developmental processes of lameness at the individual and herd levels.

3.2 Materials and methods

3.2.1 Data collection and handling

The study did not include any procedures regulated under the Animals (Scientific Procedures) Act 1986 but was carried out in accordance with the University of Glasgow guidelines and with local ethical approval (Ethics licence EA06 19).

This chapter is the first part of a longitudinal study in which, in addition to the data we recorded and presented here, data were recorded using a radar system which will be presented in detail and analysed in Chapter 4. This part describes the methodology employed for generating labels (cow mobility scoring), which will be utilised as labels (ground truth) in the subsequent chapter for analysing the data acquired by radar.

3.2.2 Farm visits and video recordings

We conducted nine fortnightly visits to a farm in central Scotland from April to September 2021 (Figure *3.1*). During the visits, an operator using the Kodak PlaySport (Zx5 Full HD 1080P) camera recorded all milking cows from the side, walking individually along a 7 m long by 1.5 m wide temporary railed extension to the farm's solid-walled passageway at the exit of the milking parlour. The operator had approximately a 5 m distance from the passageway, with freedom to move and track cows as they passed by.

The recorded videos were embedded in a PowerPoint file (see Chapter 2) and shared with between three to four assessors selected following the studies in Chapter 2, all of whom were experienced in scoring cow mobility, and one being certified registered mobility scorer (RoMs - https://roms.org.uk/).

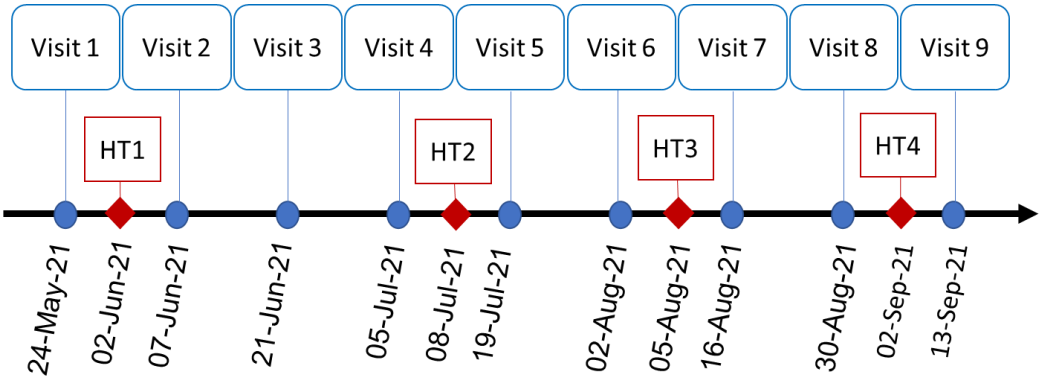


Figure 3.1 Timeline of the farm visits (video and radar data collection) and the hoof trimmings (HT).

3.2.3 Animals

The same 50 Holstein Friesian dairy cows were video recorded on each visit resulting in 393 videos being retained for scoring after eliminating unsuitable videos in which it was impossible to evaluate the cows, e.g., a cow standing still and not moving forward. In addition, ear tags (identification numbers) and lactation numbers were recorded for each animal.

Cows' lactation numbers ranged from 1 to 10, with a mean of 3 (SD=2). Nineteen cows had calved, on average, 3.4 (SD=2.5) months before the commencement of the experiment. Twelve cows calved during the study, averaging 2.2 (1) months post the commencement date. And 19 cows calved after the end of the study, within 1.9 (0.9) months after the last recording.

3.2.4 Physical examination of the hooves

The farm's trained hoof trimmer conducted 50 examinations before the study commenced. Following this, a qualified hoof trimmer (National Association of Cattle Foot Trimmer member -NACFT, Category 1) performed monthly hoof examinations (Figure *3.1*). In total, we obtained 180 instead of 200 observations over the four visits because a few animals were not available during all examinations (i.e., animals were calving).

During the examination, the hoof trimmers placed each animal in the crush and lifted each leg, cleaning the hooves and correctively trimming where needed. The intervention in the claws was minimised, trying not to affect the animal adversely. In most cases, the hoof trimmer used only a brush and a bucket of water to clean the hoof with minimal use of trimming knives. The researcher recorded lesions on the hooves, such as sole haemorrhages, hoof discolourations, cracks, separations, interdigital growths, stones, under-run soles, white line, sole ulcers (Figure 3.2 B), and digital dermatitis (Figure 3.2 A). Lesions were not scored for severity.

We then sorted the hoof examination data into two categories for statistical analysis.

- Healthy: no problems with the hooves or showed signs of only small haemorrhagic spots (superficial and up to 2mm in diameter) during the examination which could be easily removed using the knife (score 0).
- Unhealthy: one or more lesions that, due to the size or appearance, may compromise the cow’s mobility, such as white line disease, sole ulcer, digital dermatitis, stones, cracks, and under-run sole (score 1).

Figure 3.2 Two hooves classified as score 1 (unhealthy). Photo A shows digital dermatitis and photo B a sole ulcer lesion and block placement on the healthy claw.

3.2.5 Mobility Assessments

The recorded videos were shared after each visit with expert assessors: 3 assessors for visits 1-5 and 4 assessors for visits 6-9. The fourth assessor was included in the process because they were familiar with the herd and would help resolve disagreements where there were discrepancies between the three assessors, lending greater confidence to the outcome. The AHDB-Dairy 4-level mobility system was used throughout the evaluations, with scores ranging from 0 (non-lame) to 3 (severely lame). Before each evaluation, assessors were presented with calibration videos of cows in each level (AHDB 0-3). They were also asked to record a second score in cases where they were not certain of their first evaluation. After independently scoring all videos from each visit, the assessors convened to discuss each video from that visit in turn and agreed

on a consensus score for each cow, thus creating a consensus score set, as shown in Figure *3.3*.

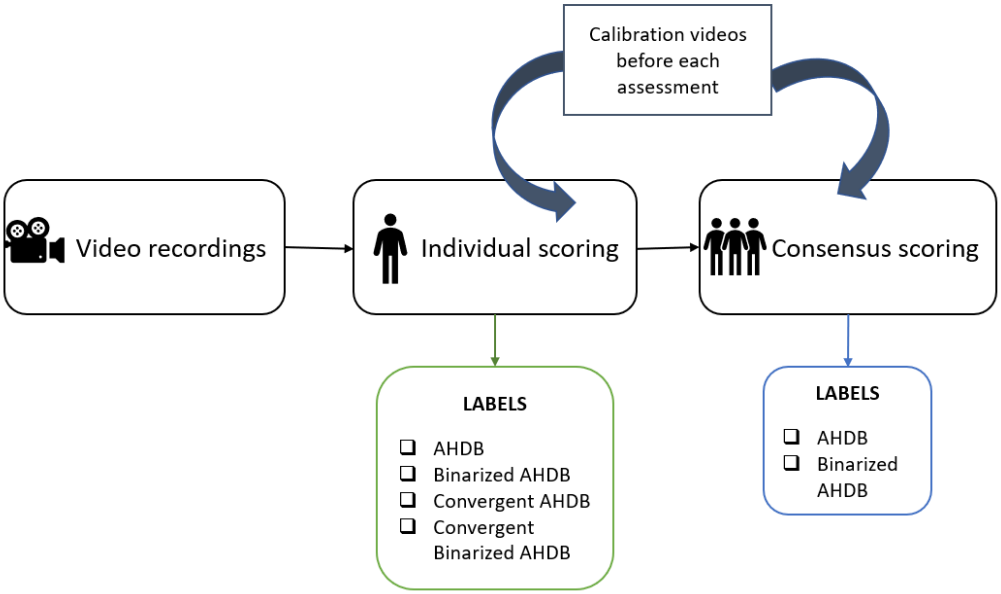


Figure 3.3 The process followed for creating labels. The cows' videos were shared with each assessor for individual evaluation, and then everyone scored the videos together, creating the listed labels.

3.2.5.1 Modified scoring systems

After collecting all the data, individual and consensus scores, we created four additional scoring sets, resulting in six systems in total, as listed and described in Table *3.1*.

Table 3.1 The table presents the six scoring sets we used for the statistical analysis and their abbreviations that will be used in the rest of the chapter. All generated systems were retrospectively created based on the AHDB mobility system, which was used for the initial evaluation of the cows through the videos.

Scoring Set	System's levels	Description	Abbreviation
Individual AHDB scores	AHDB 4-level system (0-3)	3 or 4 assessors scored all animals individually using the AHDB system	IndivAHDB
Convergent-AHDB scores	AHDB 4-level system (0-3)	When scoring independently, the assessors had the choice to assign a second score when they were not certain about their first decision. We selected the scores with the greater agreement	ConvAHDB
Consensus-AHDB scores	AHDB 4-level system (0-3)	All assessors convened in a group, and they agreed on consensus scores for each animal	ConsAHDB
Binary-AHDB individual scores	Binarised-AHDB system (0 and 1)	We merged scores 0/1 and 2/3 from the individual assessment into 0 (non-lame) and 1 (lame), respectively	IndivBin
Binary-convergent-AHDB scores	Convergent-Binarised - AHDB system (0 and 1)	We merged scores 0/1 and 2/3 from the convergent scoring set into 0 (non-lame) and 1 (lame), respectively	ConvBin
Binary-consensus-AHDB scores	Consensus-Binarised - AHDB system (0 and 1)	We merged scores 0/1 and 2/3 from the consensus scoring set into 0 (non-lame) and 1 (lame), respectively	ConsBin

3.2.6 Statistical analysis

3.2.7 Lameness prevalence

We calculated the lameness prevalence of the herd from each assessor's scores and for each visit. We then averaged the scores and estimated the average lameness prevalence. The prevalence was calculated according to the formula in Equation *3.1*

Equation 3.1 Formula of lameness prevalence calculation

$$Lameness\ prevalence\ = \frac{Number\ of\ lame\ animals\ (scores\ 2\ and\ 3)}{Total\ number\ of\ animals}$$

We also plotted each cow's consensus scores across the nine visits to make it easier to visually distinguish cows with the same continuous scores from animals whose scores changed during the study.

3.2.8 Assessors' agreement

Inter-rater agreement was determined using the statistical package "irr" (Gamer et al., 2019), with Cohen's kappa (function: "kappa2") for the pairwise comparisons and Fleiss' kappa for the comparisons among all the assessors (function: "kappam.fleiss"). The agreement was expressed as a percentage agreement for all the assessors. In all comparisons, zero tolerance was allowed, meaning that assessors must have given the same score for 100% agreement to be reached. The interpretation of the agreement was based on Landis & Koch, (1977), assuming substantial agreement when the kappa value is >0.61.

Table 3.2 Landis & Koch (1977) Kappa interpretation.

Kappa	Strength of agreement
<0	Poor
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

3.2.9 Associations of scores with hoof examinations

We used generalised linear models (GLM) in R (function "glm") to quantify associations between the consensus scores and the hoof examination outcomes. Lactation number was included in the models as a covariate because the prevalence of lameness is expected to be higher in older cows (Bran et al., 2019). The scores were the dependent variables in the model, while the hoof examination results from the closest visits before and after the mobility assessment were explanatory variables, as in the model in Equation *3.2*.

Equation 3.2 Formula of the generalised linear model with the scores and the hoof examinations.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Where:

Y = the dependent variables - scores (consensus and average individual scores of each system)
$ß_0$ = the intercept
$ß_{1,2,3}$ = the weights of $X_{1,2,3}$ variables, respectively
X_1 = Hoof trimming outcomes from the visit before the mobility assessment
X_2 = Hoof trimming outcomes from the visit after the mobility assessment
X_3 = Lactation number

For the interpretation of the GLM outcomes, we used the Bonferroni adjustment for multiple testing correction (familywise error rate (0.05) / the number of tests). We considered 25 tests, so achieving statistical significance at α = 0.05 according to the Bonferroni criterion would require a P-value < 0.05/25 = 0.002.

Hoof trimming recordings were also used as dependent variables in a Bayesian generalized mixed-effects model using the "brms" and "rstanarm" packages in R (Bürkner, 2017; Goodrich et al., 2023). A separate model was used for each of the four hoof trimmings, with each model featuring fixed effects for scores and lactation numbers, as well as random intercepts for the assessors and cows and random slope for the scores within each assessor.

3.3 Results

3.3.1 Lameness prevalence

The prevalence of lameness (scores 2 and 3) was calculated for each assessor separately and plotted with the mean, as shown in Figure *3.4*. Variation was high between assessors suggesting that depending on who evaluates the herd may yield different results. In the last farm visits, the assessors were more consistent in their evaluations of the prevalence of lameness in the herd. For example, in Farm Visit 8, the average prevalence was 0.66, which was closer to the mean value, and the standard deviation (SD) was 0.06, indicating that the scores were more tightly clustered around the average.

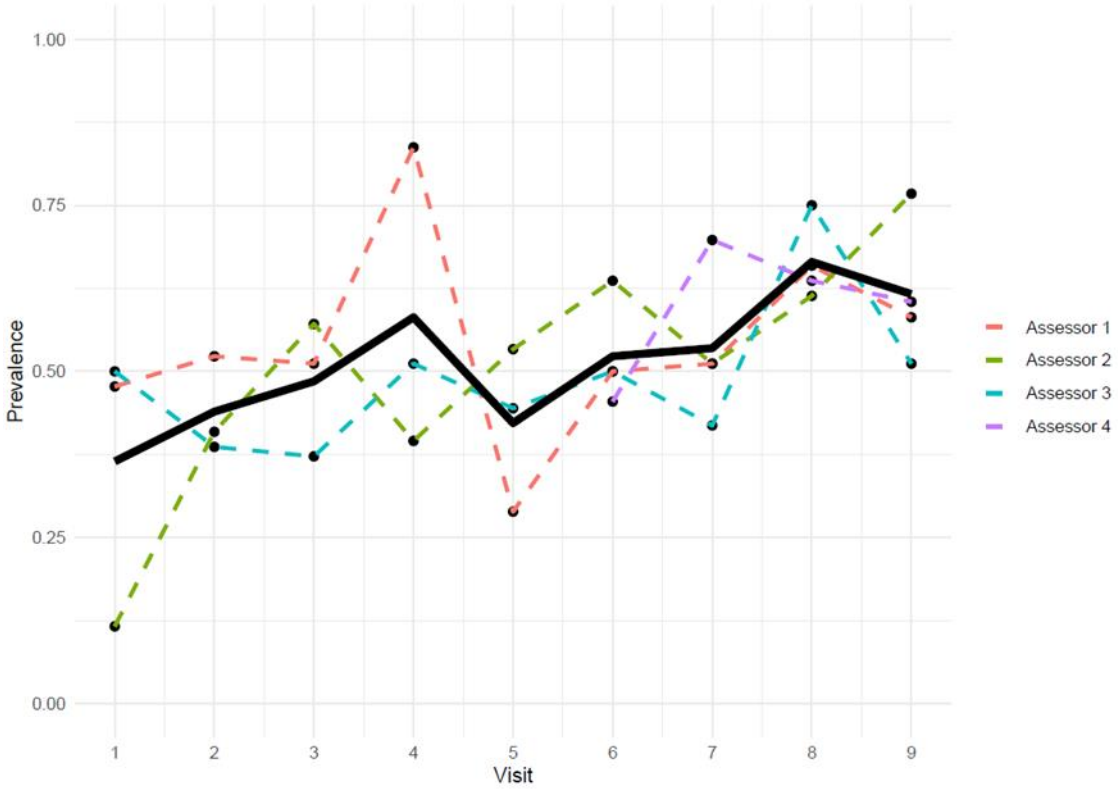


Figure 3.4 Lameness prevalence of the herd during the longitudinal study (nine scorings) for each assessor in dashed lines and the average in the black line. Rater 4 assessed visits 6 to 9, thus, the data points cover only the particular visits.

Figure *3.5* shows the agreed scores assigned by all assessors to each cow in the consensus evaluation throughout the study. In several cases, such as in cow 411, the status appeared to alternate from visit to visit between scores 1 and 2. In contrast, only a few cows (e.g., cows 12, 35) had a constant mobility state throughout.

Figure 3.5 Consensus AHDB scores for each cow during the study. Only a few cows had a constant score for all visits; some were not assessed in all visits as only milked animals were enrolled.

3.3.2 Assessors' agreement

Assessors' agreement was calculated for all systems for which we had individual ratings, namely IndivAHDB, ConvAHDB, IndivBin, and ConvBin. The results of the comparisons are shown in Table 3.3.

When the AHDB system was used for assessment, the kappa indices ranged from 0.04 to 0.4 across visits. There was an improvement in values when the convergent-AHDB 4-level system (ConvIndiv) was used, agreement ranging from 0.45 to 0.88, which according to Landis and Koch, indicates a moderate to almost perfect agreement. The results of the almost perfect agreement were produced on the seventh and eighth visits when we used three of the four assessors in the calculation. An expected increase in kappa indices was observed when the above systems were converted to binary. The highest value (0.97) occurred at the seventh visit for the ConvBin system among the three assessors. The agreement for IndivBin was better compared to IndivAHDB, but the values were much lower than the four-level ConvAHDB system.

3.3.3 Associations of scores with hoof examinations

We found weak associations between consensus scores and physical hoof assessments. Across all analysis, most variables had tendency to a positive correlation; however, most produced a high p-value ($p>0.02$ Bonferroni corrected value). Some exceptions were found between scores from visit 2 with recorded lesions in HT2 and scores from visit 9 with lesions presence in HT4 for the consensus scores (Figure 3.6 A). In the binary consensus set (Figure 3.6 B), scores from visit 6 with lesions presence in HT3 and scores from visit 9 with lesions presence in HT4 positively correlated with p-values 0.03 and 0.02, respectively. Lactation numbers did not appear to play a statistically significant role or influence the associations.

When we used the averages of assessors' scores in the generalized linear model, we found a few statistically significant (where<0.05) associations between them and some hoof examinations. As shown in figure 3.5.A there was a positive correlation between scores of Visit 5 and the second hoof examination

records (HT2 p=0.023). A positive association was also observed between Visit 6's scores with lesions from HT3 and scores from Visits 8,9 with lesions from HT4 when the convergent and binary convergent sets were used, as shown in Figures 3.5 B and D. The remaining associations with hoof pathology examinations were not statistically significant. Regarding lactation numbers and scores, the generated values did not suggest a linear association between the variables, meaning that the scores did not consistently increase or decrease according to the animals' age.

When hoof trimming outcomes (HT1, HT2, HT3, HT4) were used as dependent variables with the Bayesian models, we considered the assessor's AHDB individual scores and the cows' lactation number as fixed effects and cows and assessors as random, with the addition of a random slope for scores within each assessor. For HT1, the intercept mean was -8.2, with scores and lactation numbers mean coefficients of 0.0 and 1.8, respectively. Similarly, HT2 exhibited an intercept mean of -0.3, with assessors' scores and lactation number mean coefficients of 0.3 and 0.4. Moving on to HT3, the intercept estimate was -1.5, with mean estimates of 0.2 and 0.6 for scores and lactation numbers, respectively. In HT4, the intercept estimate was -0.1, with scores and lactation numbers mean estimates of 0.4 and 0.9. Random effects for cows and assessors indicated individual variability with varying intercepts in all models. All models also displayed excellent convergence with Markov chain Monte Carlo diagnostics.

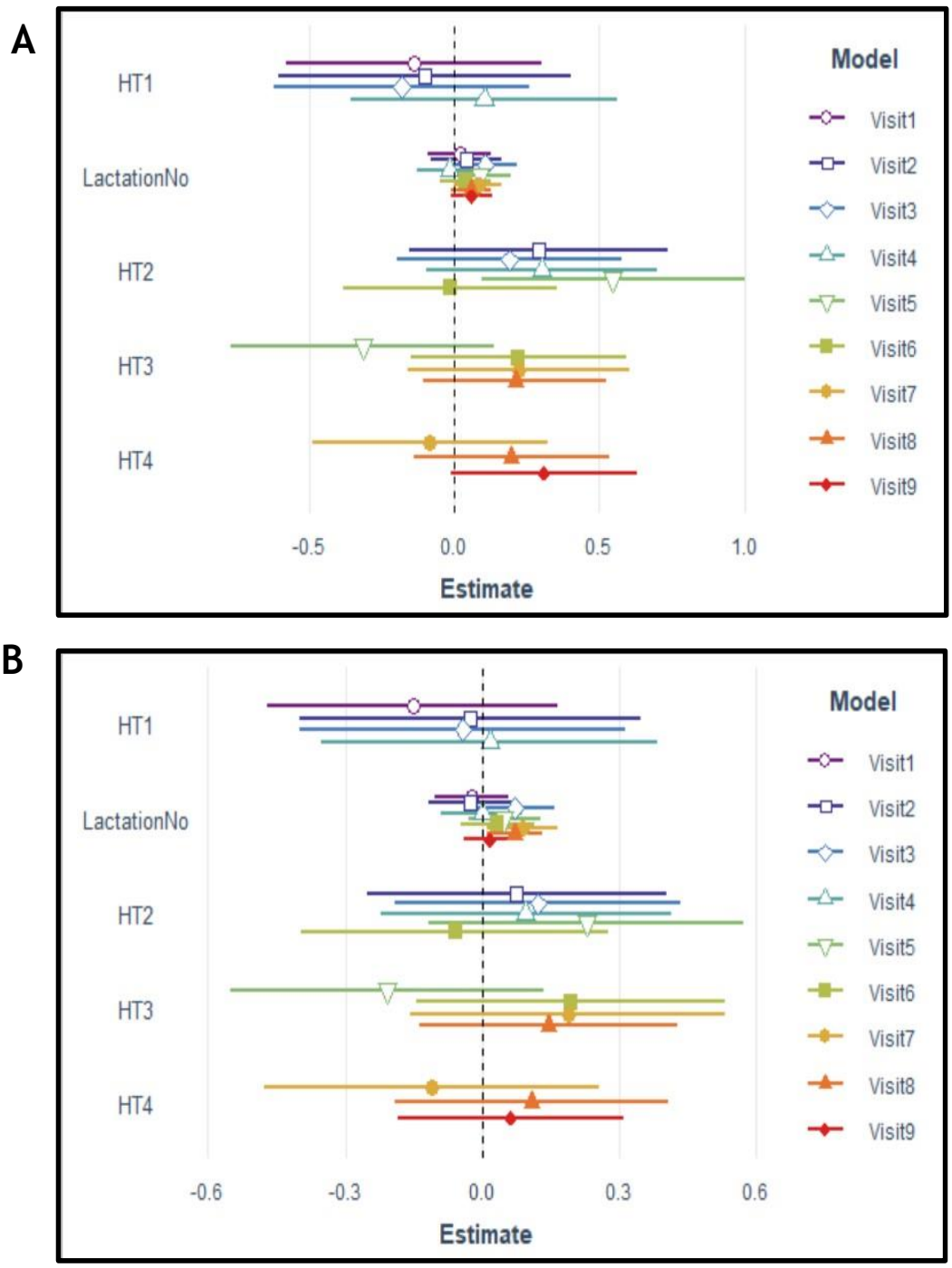


Figure 3.6 Coefficient plots of the generalised linear models for the two scoring systems against the HT before and after each visit and the cows' lactation number as per Equation 3.2. Each plot represents the different scoring systems used in the model; A: AHDB Consensus system B: Binarized AHDB Consensus system. The coloured horizontal lines represent the 0.95 confidence intervals for the coefficients.

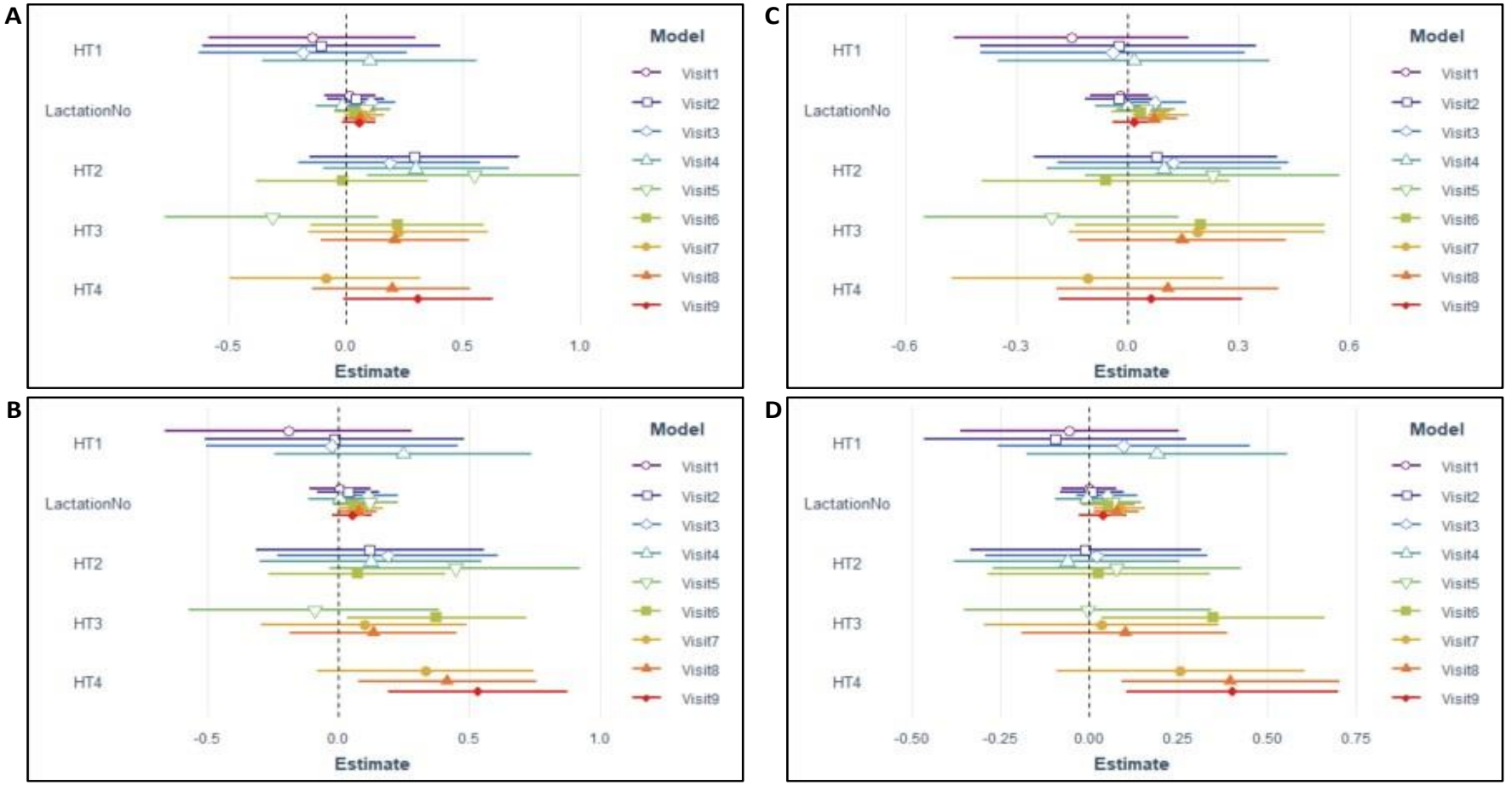


Figure 3.7 Generalized linear models for averaged scores vs Hoof Examination before and after, plus Lactation Number as per the Equation 3.2 with dependent variables A: average AHDB scores, B: average convergent-AHDB scores, C: average binarised-AHDB scores, D: average binarized-convergent-AHDB scores.

Table 3.3 Inter-rater agreement (kappa indices and percentage agreement in the parenthesis) among all assessors in each visit evaluation using four scoring systems (AHDB, Convergent, Binary AHDB, Binary Convergent).

Number of assessors	SCORING SYSTEM	VISIT1	VISIT2	VISIT3	VISIT4	VISIT5	VISIT6	VISIT7	VISIT8	VISIT9
4	IndivAHDB						0.23	0.35	0.24	0.38
							(24.4)	(37.2)	(29.5)	(41.9)
4	ConvAHDB						0.75	0.74	0.61	0.67
							(73.3)	(72.1)	(63.6)	(67.4)
4	IndivBin						0.36	0.46	0.33	0.17
							(40)	(65.1)	(45.5)	(39.5)
4	ConvBin						0.82	0.79	0.69	0.70
							(82.2)	(79.1)	(72.7)	(72.1)
3	IndivAHDB	0.10	0.21	0.30	0.04	0.27	0.26	0.40	0.17	0.26
		(18.6)	(25)	(38.1)	(18.6)	(35.6)	(42.2)	(51.2)	(36.4)	(41.9)
3	ConvAHDB	0.50	0.77	0.79	0.45	0.77	0.88	0.88	0.60	0.66
		(55.8)	(79.5)	(81)	(51.2)	(80)	(91.1)	(90.7)	(70.5)	(74.4)
3	IndivBin	0.14	0.48	0.50	0.08	0.33	0.37	0.53	0.24	0.38
		(39.5)	(61.4)	(62.8)	(32.6)	(51.1)	(53.3)	(48.8)	(50)	(55.8)
3	ConvBin	0.50	0.84	0.87	0.53	0.82	0.91	0.97	0.69	0.71
		(67.4)	(88.6)	(90.7)	(65.1)	(86.7)	(93.3)	(97.7)	(79.5)	(79.1)

3.4 Discussion

We sought to quantify the relationship between mobility evaluations on video recordings of cows with the presence of lesions in their hooves and, in doing so, to establish which method of combining multiple ratings was most closely related to pathology. The overarching aim was to improve on single-assessor judgements as labels for machine learning. We calculated the agreement among assessors using the AHDB 4-level mobility system and three modified systems (binarised-AHDB, convergent-AHDB, binarised-convergent-AHDB) derived from it, which were also used in Chapter 2. Agreement results were below moderate for the AHDB system. However, they were improved when modifications of that system were used, with the best performance occurring in the binarised-converged AHDB (ConvBin) system (kappa = 0.97, Visit 7). The same assessors gathered and scored all videos, giving a consensus score. Finally, we compared averaged scores from individual evaluations and the consensus scores with the physical examination of the cows and their lactation number. The results of the comparisons showed weak associations between scores, physical examinations, and lactation numbers.

Inter-assessors agreement varied significantly by visit and by use of different scoring systems. The AHDB 4-levels, one of the UK's most widely used assessment systems (Afonso et al., 2020), introduced variations to the assessment. The converged set we created from the assessors' second scores gave substantial results, which are convenient for agreement comparisons. However, in practice, when evaluating an animal on-farm, we need a single accurate and prompt label on which to base the next decision and action (e.g., calling the hoof trimmer, handling the lame animal, or in our case, training machine learning using the score/label as the ground truth). When we reduced the system's levels, the agreement results were promising, reaching an almost perfect agreement. We expected a significant improvement since the chances of agreement increased in a binary system compared to a four-level system. Similar results to our study on the agreement among assessors were produced in other studies when they ran an agreement analysis between assessors with reduced scoring levels (March et al., 2007; Schlageter-Tello et al., 2014). Still, while a binary system seems to deliver

desired levels of agreement, it sacrifices precision since animals that may be slightly lame are lumped into the same category as severely lame animals.

Variation in assessor scores causes variation in the estimates of herd lameness prevalence. The assessors started with differences at the first evaluation and appeared to agree more on the state of the herd towards the last scorings. This may signal a form of calibration in how they scored the animals after the first five visits, which may be the time they required to become familiar with the evaluation process. Additionally, general discussions between assessors about herd prevalence (not in individual scorings) may have reduced extensive discrepancies. In this study, the mobility status of each cow may be better indicated by calculating the average score value given by all the assessors. This is because, taking the average score of multiple assessors, all the ratings are given equal importance, which helps to avoid any potential biases (Jones 2000). However, in this study, we chose to use consensus over average as outlier scores from individual assessors could skew the results, and we considered that scores agreed by the majority would be more representative. Our objective through group scoring (consensus) was to find the best possible label for each cow, aiming to improve the accuracy of these assessments.

We expected to find significant associations between scores and physical examinations since research has shown that mobility problems are mainly caused by hoof lesions (Archer et al., 2010; Murray et al., 1996) and studies such as Brenninkmeyer et al. (2013) and Solano et al. (2015a) found correlation between lameness and hock lesion prevalence. At the same time, we expected that the likelihood of lameness is associated with increasing age (Bran et al., 2019). However, our results showed only weak and inconsistent associations between the scores, the pathologies, and the cows' ages for both consensus and averaged scores when we used the AHDB 4-level and binarised-AHDB systems. There were only a few exceptions when the convergent-AHDB and binarised-convergent-AHDB sets were used in the equations, where we found positive associations between the two scoring sets with the last two hoof examinations. When the AHDB individual scores were used as explanatory variables, they were found to be ineffective in predicting the hoof trimming outcomes. The weak association outcomes agree with the studies of Flower & Weary (2006), Logue et al. (1994), and Tadich et al. (2010b), which found that pathologies such as white line lesions and sole

haemorrhages (which were also the most prevalent pathologies found in our study) are not related to poor mobility scores. These lesions might cause discomfort to the animal, but they will only change the cow's gait if they are severe (Tadich et al., 2010b). Another explanation for the lack of consistent association between scores and hoof pathologies is that the deviation in mobility could be due to lesions not located in the hooves. An example is septic arthritis, a lesion in the cow's upper leg that can cause severe lameness (Desrochers, 2017) and, thus, locomotion alterations. However, in this study, we only examined hoof pathologies. Conformation is another reason for the absence of scores - hoof pathologies association. The deviation from the typical gait pattern could result from natural causes (i.e., born this way), and the animal could be perceived as lame without having any lesions or affecting its welfare. Finally, another justification for the poor association to be considered is that the visual mobility system may not be sensitive enough to detect subtle changes in animal movement patterns when they have a lesion at an early stage, which other researchers have suggested (Kofler et al., 2011; Manske et al., 2002a; Tadich et al., 2010b). Combined, the above reasons in addition to the limited population size could account for our study's poor association results.

The lack of consistent associations between scores and hoof pathologies did not support the original goal of increasing confidence and reducing uncertainty about the mobility status of the examined cows. Nevertheless, we were able to identify which of the proposed mobility scoring systems yielded the highest levels of agreement between multiple assessors and achieved near-perfect agreement. These longitudinal study results are satisfactory to continue research examining and improving the automatic radar lameness detection system, providing alongside hope for a way to generate valid labels with a high inter-assessor agreement in situations of uncertainty for use with supervised machine learning algorithms.

3.5 Conclusions

In the present study, we attempted to quantify the associations between hoof pathologies and scores derived from the visual assessments, concluding that there are no consistent associations. We justify the results by arguing that there was no effect of pathologies on gait patterns or that the observation system was insufficiently sensitive to capture the recorded pathologies. However, although we did not increase confidence in the mobility ratings through physical examination, we did increase confidence by obtaining sufficient score sets with substantial to perfect agreement to test with machine learning algorithms, which leads us to the following study where the use of micro-Doppler radar system will be added to data acquisition and analysis to automate lameness detection in dairy cows.

Chapter 4

Cattle mobility scorings as labels for the classification of micro-Doppler radar data using supervised machine learning - A longitudinal study Part 2

4.1 Introduction

Lameness in dairy cows is a prevalent mobility disorder manifested by deviance from the typical gait pattern, resulting in impaired welfare, reduced production, and cost of treatment (Dolecheck & Bewley, 2018; Garbarino et al., 2004; Green et al., 2014). Therefore, monitoring and rapid lameness diagnosis are essential to avoid adverse effects and maintain the animal's welfare status. However, the most common method in massive lameness screenings (herd assessment) is visual scoring systems that are labour-intensive, time-ineffective, and not highly repeatable (Schlageter-Tello et al., 2014). Automating the lameness detection process could improve animal monitoring and accurate lameness classification (Afonso et al., 2020), improving welfare and profitability.

Several automatic lameness detection methods have been proposed in recent years as a solution to the identified challenges (Byabazaire et al., 2019; Chapinal & Tucker, 2012; Taneja et al., 2020). Automating lameness detection in livestock could possibly detect lameness more accurately or in less time than humans (Kühl et al., 2020), reducing the likelihood of misdiagnosis. The need for manual monitoring would be reduced, saving time and labour costs, and animal welfare would be increased by reducing the risk of discomfort and maintaining productivity levels. Finally, the data collection and analysis would be enhanced when an automated system is involved as it has the potential of continuously monitoring, finding patterns, and assisting the farmers in improving management practices. Although a few automatic herd monitoring systems with an option for lameness detection are already available on the market, employing mainly camera systems and accelerometers attached to the animals, they have not been widely adopted

on farms yet (Van De Gucht et al., 2017). Automated systems can be expensive to purchase, install and maintain, making them potentially unaffordable for small-scale farms (Gucht et al., 2018). A few systems can be complex to operate or have performance issues, i.e., accuracy and functionality can be affected by factors such as video data quality and farm lighting conditions (Russello et al., 2022). Finally, limited farmer awareness or consideration of the negative impacts of lameness (Leach et al., 2010b; Šárová et al., 2011a) might be another potential reason for the limited adoption of automated systems.

Supervised machine learning (SML) requires each set of observations or signals from the sensors to be labelled with a ground truth state as a basis for the AI to learn and perform the classification task. In automated lameness detection based on supervised machine learning, data are manually labelled by one or multiple experts and used in the algorithm to recognise the patterns associated with lameness and make predictions on new observations. The goal is to minimise the difference between the algorithm's predictions and the input labels or ground truth. The problem with using manually labelled data is that the SML relies on the accuracy and consistency of these labels. If the labels are incorrect or inconsistent, the resulting outcomes might be biased, inaccurate or ineffective in recognising lameness patterns or generalising with new data. The variation and lack of repeatability (lack of inter- and intra-assessor agreement) involved in the visual assessment method used as ground truth and the uncertainty it creates has been noted in several studies (Afonso et al., 2020b; Engel et al., 2003; Schlageter-Tello et al., 2014; Tadich et al., 2010).

In two preliminary small-scale studies, we used micro-Doppler radar sensing for lameness detection in cows and other animals, with good specificity, sensitivity, and accuracy results (88%, 81%, >80%, respectively) (Busin et al., 2019; Shrestha et al., 2018). Micro-Doppler radar is a technology that emits electromagnetic pulses in the air and measures the pulse’s time of flight (TOF) from the radar to a target and back. When the pulses are reflected off a moving target with rotating or vibrating parts, it produces a distinct signature that can be used for detection, classification, and discrimination (Chen, 2008; Fioranelli et al., 2015). The non-contact and non-invasive nature of radar detection makes it advantageous for farm applications, as the absence of wearable devices reduces the potential for animal stress and the non-routine handling (Z. Wang et al., 2022). Another

potential benefit of the radar is that its function and the produced results do not depend on the weather or light environment; it is not affected by rain, light refraction, fog or other weather and light conditions. Together with high resolution that enables detailed analysis of complex moving patterns, micro-Doppler radar offers advantages over other motion detection methods. The overall benefits of micro-Doppler radar make it promising for a wide range of applications, including automated lameness detection in livestock (Shrestha et al., 2017), human movement analysis for medical applications (Hayashi et al., 2021; Kao et al., 2013) and detection and tracking of small autonomous aerial vehicles (Gong et al., 2022).

Previous chapters addressed the assessment and quantification of inter-assessor agreement using different scoring systems to classify cow mobility via video. A general conclusion was that binary scoring systems produced better results than the UK's most widely used four-level mobility system. The present study utilises all the scoring labels produced and described in the previous chapter to find which label-algorithm combinations provide more accurate predictions. Hence, we expected at least one system that has previously delivered acceptable levels of inter-assessor agreement (Cohen's kappa >0.6) would also deliver equally sufficient levels of accuracy when used as ground truth in machine learning classification models.

4.2 Materials and Methods

Farms, visits, animals, video recordings and labels are described in detail in Chapter 3. Micro-Doppler radar signals were collected while these animals were video recorded during these visits. This chapter describes the second part of the longitudinal study, which concerns the training and validation of machine learning from the assembled data.

4.2.1 Farm Visits and recorded animals

Nine fortnightly visits were carried out on a central Scottish farm between April and September 2021. On each visit, we recorded the micro-Doppler signature of 50 cows, the same animals that were video recorded in Chapter 3, and we obtained 393 radar signatures in total after eliminating signals with more than one cow being recorded at the same time. In cases with a recorded signal from the radar but no label from the evaluation process, the signal was not included in the analysis. This resulted in the total number of analysis samples for some evaluations being slightly less or more, as shown in Table *4.3*. The farm facilities and the stations (radar (A), video recordings (C)) set up can be seen in Figure 4.1.

4.2.2 Radar equipment and set up

In this project, we used an FMCW radar system from Ancortek operating at 5.8 GHz, with a bandwidth of 400 MHz and a pulse repetition frequency of 1 kHz. In a monostatic configuration, the radar was connected to two separate Yagi antennas, a transmitter and a receiver antenna. The transmitted power from the antenna was approximately 100 milliwatts, with an antenna gain equal to approximately 17 dBi (Decibels relative to isotropic) and a beam width of 24 degrees in azimuth and elevation. The station with all the equipment was set up on the farm with a line of sight along a passageway after leaving the milking parlour (Figure 4.1 B). The antennas were placed on tripods 40 cm apart, at 1.5 m high, with the transmitted signal directed at the animal's rear, at its torso

height. A photo of the antennae set up, the corridor, the radar angle of view and the video vantage point can be seen in Figure 4.1.

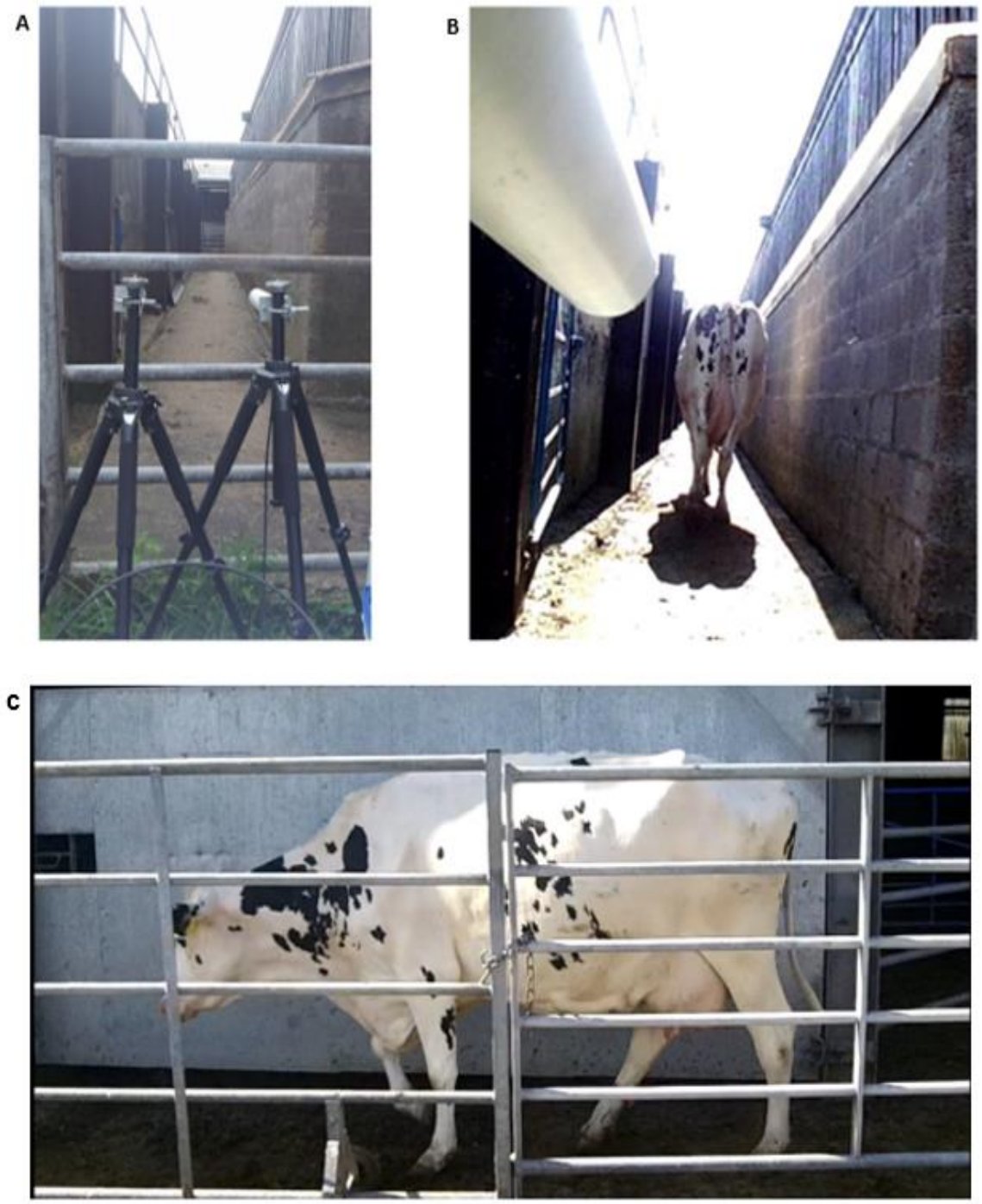


Figure 4.1 (A) Antennae set up, (B) the rear vantage point when a cow was walking in front of the antennae, and (C) the lateral vantage point of the video recordings.

4.2.3 Radar signal processing

The radar signal data, a series of numbers stored in ".dat" format, was fed into an algorithm previously developed and used in the studies by (Busin et al., 2019a) and (Shrestha et al., 2018).

The following steps were followed for the micro-Doppler signatures processing:

- First, a waveform was generated from the raw data recorded from the intermediate frequency stage of the radar containing the backscattered signals. (Figure 4.2 B)
- The data (complex in-phase and quadrature numbers - I&Q) was reshaped into a two-dimensional matrix with dimension 128×N, where 128 is the number of time samples per sweep, and N is the number of chirps in the frame. Dividing the waveform into chirps is a standard pre-processing before extracting the range information.
- A Moving Target Indicator (MTI) filter was applied to remove static clutter caused by stationary objects in the environment. The MTI was implemented using an infinite impulse response (IIR) notch filter, removing common frequencies below a set threshold of 0.0075 from one pulse repetition interval to the next.
- The range information was extracted by applying a Fast Fourier Transformation (FFT) - FMCW radar recorded signals were encoded into frequencies that are directionally proportional to the time of flight of the signal, ergo the distance of the target-generating range plots for each chirp — these plots were accumulated over time to form the range-time plot displaying the target's (cow) distance from the radar/antenna (Figure 4.2C)
- A Short Time Fourier Transformation (STFT) using a 0.2 s Hamming sliding window with a 95% overlapping factor was chosen and implemented on the acquired range-time data matrix creating a range-Doppler image which contained information about the Doppler shift of the targets over time. Next, we summed together the range information in the range-Doppler image to obtain a slice of the Doppler time representation. The resulting spectrograms (Figure 4.2 D) displayed the movement and direction of the targets, indicating whether they were moving towards or away from the

radar, allowing for feature extraction, velocity analysis and activity classification.

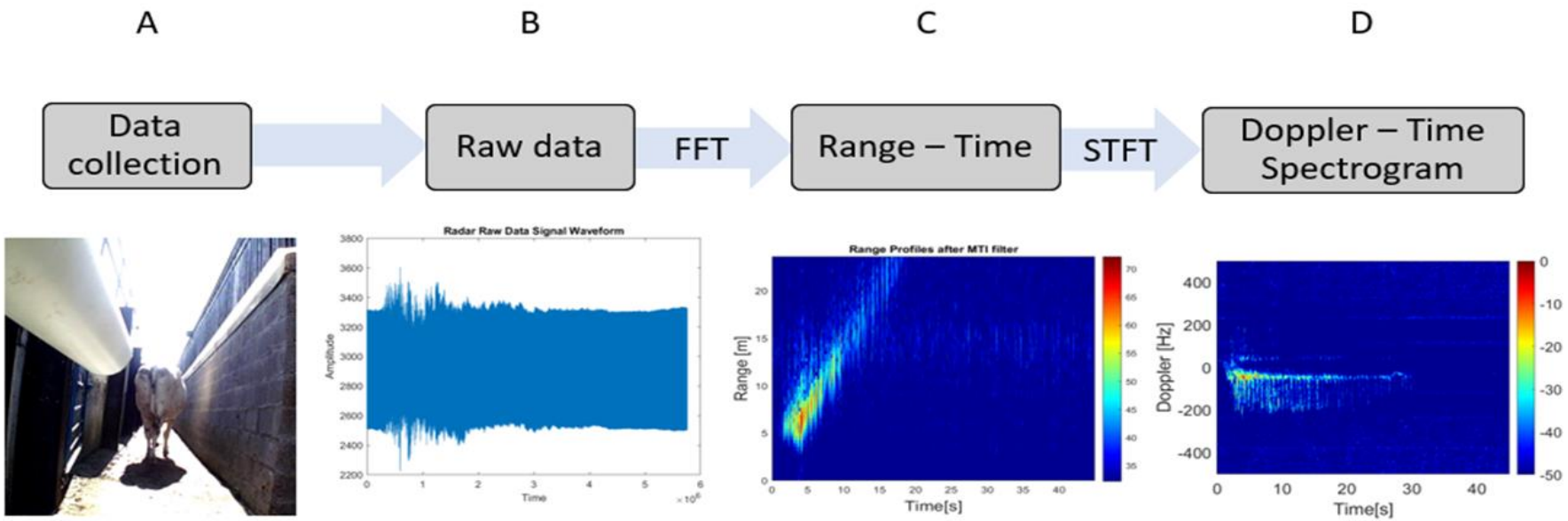


Figure 4.2 Radar signal processing chain from the moment of data recording (A) to the generation of the spectrograms (D). The raw data were visualised as a waveform (B) and filtered before performing a fast Fourier transform (FFT), generating range–time plots (C). Then, micro-Doppler - time spectrograms were generated by completing a short-time Fourier transformation on the processed data.

4.2.4 Feature extraction

For feature extraction, we divided all micro-Doppler signatures into 1.5-2 s segments from which we extracted numerical features. The numerical features represent the relevant information in the recorded signal, such as the mean and the standard deviations. We extracted twenty features from each micro-Doppler spectrogram for each cow, which we will then use for classification. More specifically, bandwidth and centroid are features based on the radar signal's frequency-domain characteristics. The singular value decomposition (SVD) belongs to the time-frequency domain features and is a mathematical technique used to decompose the spectrogram into its components, including the left (U) and right (V) eigenvectors, which are used as features for classification. The specific features were chosen because they have led to high accuracy results in similar studies using micro-Doppler radar (Busin et al., 2019a; Shrestha et al., 2017), as they can distinguish between targets and estimate their physical properties, such as the target's size, shape, and temporal behaviour.

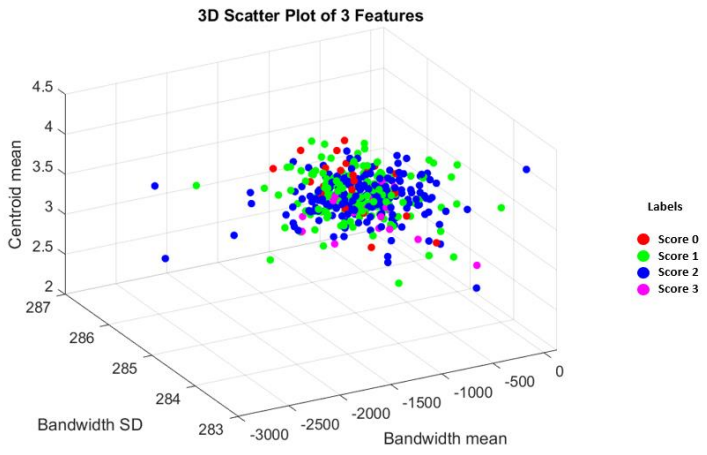


Figure 4.3 The 3D scatterplot with an example of three extracted features shows the distribution of values in space and their relationship to each other. Colours represent the algorithm's associations of the features with the 4-level Consensus scores/labels. Clutter (proximity of circles) indicates the challenge of differentiating between the features and the different level corresponding labels.

An example of a 3-Dimensional visual representation of three extracted features (example, with Bandwidth mean and SD, and Centroid mean) from our data is in Figure 4.3. By examining the data for patterns or clusters, relationships can be identified between the features. The list of all extracted features is detailed in Table 4.1

Table 4.1 Extracted numerical features from the radar recordings for each cow. We considered 20 features from each spectrogram segment, representing a statistical moment such as mean or standard deviation.

Parameters	**Extracted features**			
Centroid	mean	Standard deviation	skewness	kurtosis
Bandwidth	mean	Standard deviation	skewness	kurtosis
Spectrogram segment	mean	Standard deviation	skewness	kurtosis
First right (V) eigenvectors of the singular value decomposition (SVD) of the spectrogram segment	mean	Standard deviation	Sum of pixels for matrices V	Mean of the diagonal of the left matrix V containing eigenvectors of the spectrogram segment
First left (U) eigenvectors of the singular value decomposition (SVD) of the spectrogram segment	mean	Standard deviation	Sum of pixels for matrices U	Mean of the diagonal of the left matrix U containing eigenvectors of the spectrogram segment

4.2.5 Video recordings and labels

We used the video recordings and scores described and generated in the first part of the study (Chapter 3). The scores served as labels from the different scoring systems we used in this chapter are listed in Table 4.2.

In some instances, not all assessors scored all the cows, resulting in varying total numbers of animals assessed. For example, this happened when two animals were evaluated simultaneously in a single video clip, and one assessor scored one animal while the other scored the second. In this case, the video was excluded from the consensus scoring, but the scores provided by assessors who evaluated both animals were still considered.

Table 4.2 Labels and their descriptions used for the algorithm training. We used four different scoring systems as developed and assessed in Chapter 3.

Labels set	Levels	Labels	Description
AHDB	4	0 = good mobility 1 = imperfect mobility 2 = impaired mobility 3 = severely impaired mobility	Individual scores from 3 or 4 assessors
Binary (AHDB)	2	0 = not lame 1 = lame	Individual scores from 3 or 4 assessors merging the 4 levels of the AHDB (0,1=0 and 2,3=1)
Convergent	4	0 = good mobility 1 = imperfect mobility 2 = impaired mobility 3 = severely impaired mobility	Individual scores from 3 or 4 assessors
Binary convergent	2	0 = not lame 1 = lame	Individual scores from 3 or 4 assessors merging the 4 levels of the AHDB (0,1=0 and 2,3=1)
Consensus	4	0 = good mobility 1 = imperfect mobility 2 = impaired mobility 3 = severely impaired mobility	Agreed scores between assessors after discussion
Binary Consensus	2	0 = not lame 1 = lame	Agreed scores between assessors after discussion merging the 4 levels of the AHDB (0,1=0 and 2,3=1)

4.2.6 Test and validation data

For classification, we used 90% of the data for the training and 10% for the validation test. Using randomly chosen subsets for training and testing, we repeated the process ten times (cross-validation process), recording the average accuracy of each test.

4.2.7 Classification models

We used the Matlab classification application to train and validate the classification models. We used several supervised ML classifiers, but only support vector machine (SVM) and k-nearest neighbour (KNN) results will be presented as they cover an overall complexity (with SVM being more computationally demanding but scaling well with large datasets and KNN being more simplistic but being slower with many observations). In particular, quadratic SVM and linear KNN have produced reliable results and work well with data like those we consider in the present study.

The process followed while using the classification learning application is shown in Figure 4.4.

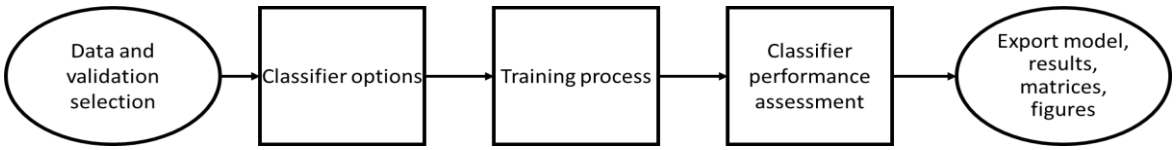


Figure 4.4 Steps in the classification learning process. First, we selected the data for testing and validation, then the classifier choices, which models we would like to use for training (i.e., KNN, SVM), observing the accuracy results, and extracting all the valuable data.

4.3 Results

4.3.1 Score distribution – labels

During the study, the lameness prevalence (scores 2 and 3) in the herd based on the different scoring systems averaged 0.5. This means that at any given time, about half of the animals in the herd were lame (score 2 or 3). Most animals had scores of either 1 or 2 in the 4-level systems, while in the 2-level systems, they were almost equally distributed (Table *4.3*).

Table 4.3 Scores distribution of each system used and for each assessor

System	Assessors	Scores				Total
		0	1	2	3	
AHDB	1	20	148	184	17	369
	2	29	165	168	15	377
	3	59	125	184	8	376
Binarised AHDB	1	176	201	-	-	377
	2	194	183	-	-	377
	3	184	192	-	-	376
Convergent AHDB	1	23	169	177	8	377
	2	29	160	177	11	377
	3	26	178	168	4	376
Binarised - Convergent AHDB	1	192	185	-	-	377
	2	190	188	-	-	378
	3	204	172	-	-	376
Consensus AHDB	all	24	144	189	18	375
Binarised Consensus AHDB	all	168	207	-	-	375

4.3.2 Performance and accuracy of the KNN and SVM models

4.3.2.1 Binary systems

The prediction accuracies from the two models, when used with the binary labels, are listed in Table 4.4. Accuracy ranged from 0.57 to 0.64, with the highest accuracy observed in the KNN model using the binary consensus scores as labels. The average sensitivity and specificity of the binary systems were 0.6.

Table 4.4 Estimations of sensitivity, specificity, and accuracy of the SVM and KNN models for each binary system label.

Assessors	SVM							KNN						
	Binary AHDB			Binary Convergent			Binary Consensus	Binary AHDB			Binary Convergent			Binary Consensus
	A1	A2	A3	A1	A2	A3	ALL	A1	A2	A3	A1	A2	A3	ALL
True positives	58.3	56.1	55.9	56.7	56.5	55.7	59.9	61.9	58.9	57.9	64.2	56.2	59.4	58.2
False positives	41.7	43.9	44.1	43.3	43.5	44.3	40.1	38.1	41.1	42.1	35.8	43.8	40.6	41.8
True negatives	57.6	61.2	67.1	60.5	61.4	64.8	66.7	55.7	62	68.4	61.1	58.1	61.3	70
False negatives	42.4	38.8	32.9	39.5	38.6	35.2	33.3	44.4	38	31.6	38.9	41.9	38.7	30
Sensitivity[1]	0.58	0.59	0.63	0.59	0.59	0.61	0.64	0.58	0.61	0.65	0.62	0.57	0.61	0.66
Specificity[2]	0.58	0.58	0.6	0.58	0.59	0.59	0.62	0.59	0.6	0.62	0.63	0.57	0.6	0.63
Accuracy[3]	0.58	0.59	0.62	0.59	0.59	0.6	0.63	0.59	0.6	0.63	0.63	0.57	0.6	0.64

Sensitivity = true positives/ (true positives + false negatives)

[2] Specificity = true negatives/ (true negatives + false positives)

[3] Accuracy = (true positives + true negatives)/count of all observations

4.3.2.2 Four-level systems

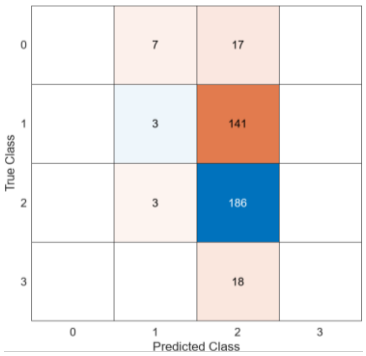
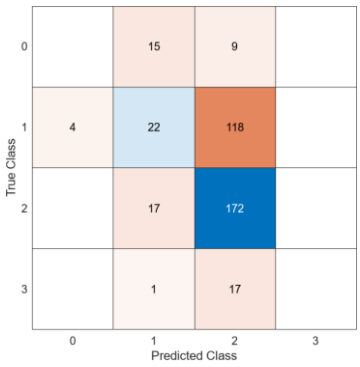
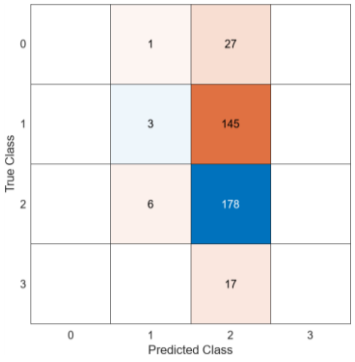
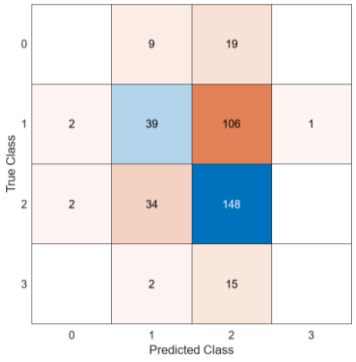
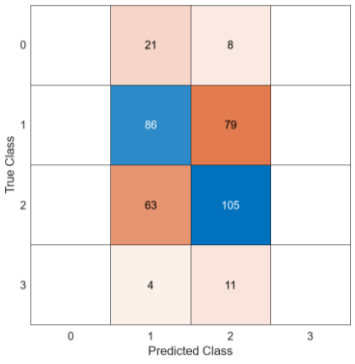
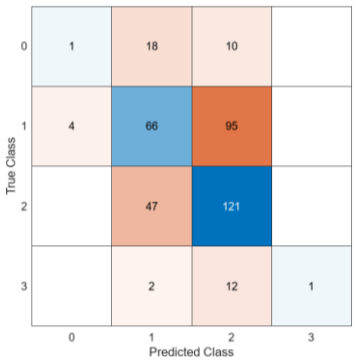
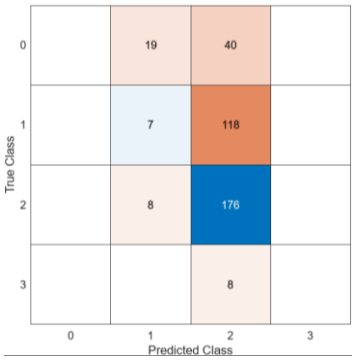
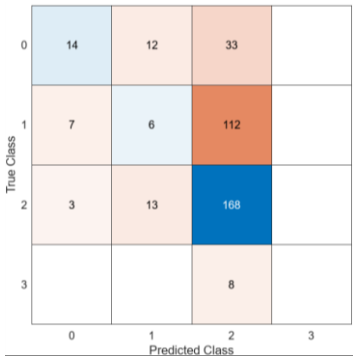
The four-level systems produced poor results. The accuracies of the systems ranged from 48% to 55.6%. The highest accuracy (55.6%) was produced by the KNN model when using the AHDB convergent system's labels of assessor 3. The accuracies for each system's levels are listed in Table 4.5

The classifications of each model can be seen in the confusion matrices in Figure 4.5. The gradient of the colours indicates the percentages of the classifications; the darker the colour, the more animals were assigned to the specific class. The desired result would be dark blue diagonal tiles, meaning that all cows would be classified in the category indicated by the labels. The bold orange squares in the figure show the misclassification.

Table 4.5 Estimation of sensitivity, specificity and accuracy for KNN and SVM models with labels from the 4-level systems. The calculations were based on the predictions of the confusion matrices in figure 4.4.

		KNN			SVM		
System	Levels	Sens[4]	Spec[5]	Acc[6]	Sens	Spec	Acc
Consensus	0	0	1	0.94	0	0.99	0.93
AHDB	1	0.02	0.78	0.1	0.15	0.86	0.59
4 levels	2	0.98	0.05	0.52	0.91	0.23	0.57
	3	0	1	0.95	0	1	0.95
AHDB	0	0	1	0.93	0	0.99	0.92
Assessor1	1	0.02	0.97	0.6	0.16	0.81	0.6
	2	0.97	0.02	0.48	0.8	0.27	0.53
	3	0	1	0.95	0	1	0.95
AHDB	0	0	1	0.92	0.03	0.99	0.92
Assessor2	1	0.52	0.58	0.56	0.4	0.68	0.56
	2	0.63	0.53	0.57	0.72	0.44	0.56
	3	0	1	0.96	0.07	1	0.96
AHDB	0	0	1	0.82	0.24	0.95	0.59
Assessor3	1	0.06	0.89	0.61	0.05	0.9	0.62
	2	0.96	0.14	0.54	0.91	0.2	0.55
	3	0	1	0.98	0	1	0.98
Convergent	0	0.09	0.98	0.93	0.04	0.99	0.93
AHDB	1	0.57	0.5	0.53	0.38	0.69	0.55
Assessor1	2	0.5	0.61	0.56	0.75	0.44	0.58
	3	0	1	0.98	0	1	0.98
Convergent	0	0	1	0.92	0	0.99	0.91
AHDB	1	0.26	0.8	0.57	0.34	0.71	0.55
Assessor2	2	0.84	0.28	0.54	0.76	0.4	0.57
	3	0	1	0.97	0	1	0.97
Convergent	0	0	0.99	0.92	0.04	0.99	0.93
AHDB	1	0.65	0.47	0.56	0.46	0.61	0.54
Assessor3	2	0.52	0.69	0.62	0.67	0.51	0.58
	3	0	1	0.99	0	1	0.99

[4] ***Sens****itivity = true positives/ (true positives + false negatives).*
[5] ***Spec****ificity = true negatives/ (true negatives + false positives).*
[6] ***Acc****uracy = (true positives + true negatives)/count of all observations.*

		KNN	SVM
Consensus			
AHDB (individual)	*R1*		
	R2		
	R3		

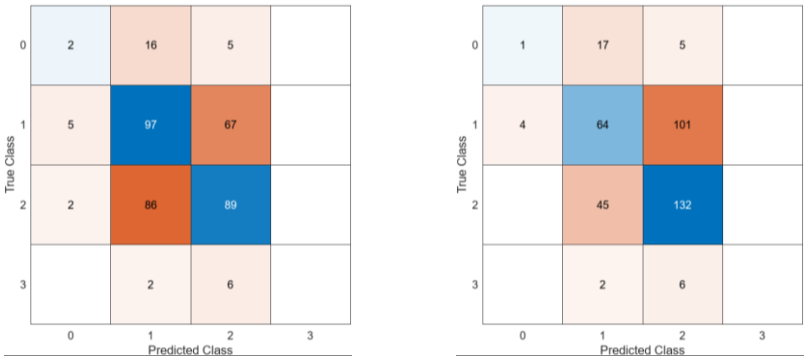
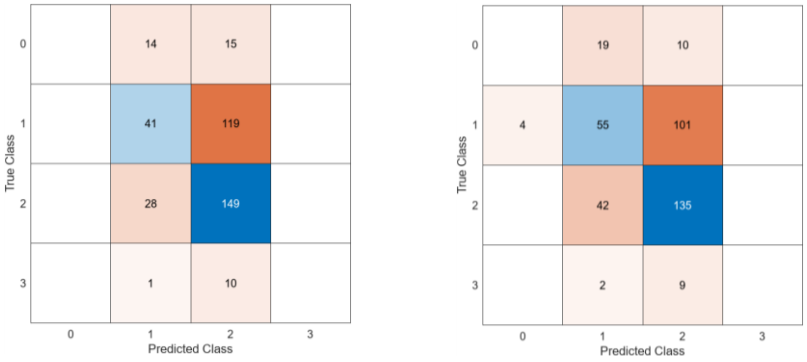
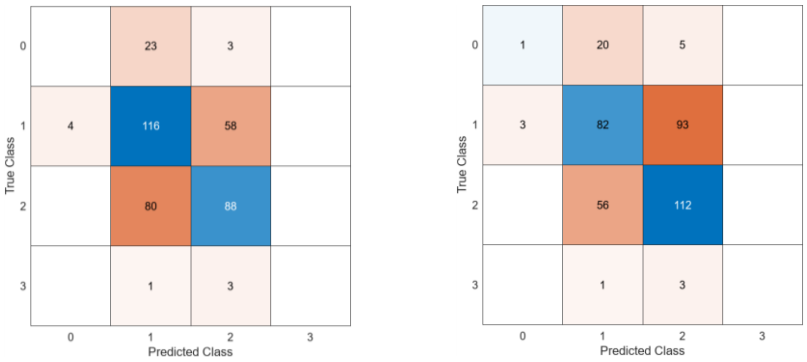
Convergent (individual)	*R1*	
	R2	
	R3	

Figure 4.5 Confusion matrices of the KNN and SVM for each label set. The blue colour tiles represent the correctly classified cases, and the orange represents the misclassification.

4.4 Discussion

We aimed to use the labels from each cow mobility evaluation system we used in the previous chapter to train an AI-based algorithm that uses radar data. In this way, we expected to find the most suitable scoring system to use as ground truth in machine learning, producing reliable mobility predictions. Unfortunately, the results were not as expected since only low accuracies were obtained for the models and labels we used and, therefore, cannot be considered yet optimal for farm use as it is.

Labels from the binary and multilevel scoring systems gave accuracy results between 40% and 64% in both KNN and SVM models. We do not consider these values good predictions since the outcome would not correspond to reality in almost half of the cases. Other studies used the same or similar radar system for data recordings, and their analysis results differed greatly from this study. For example, Busin et al. (2019) and Shrestha et al. (2018) had accuracy values of up to 88% for detecting and classifying dairy cow lameness. At the same time, studies on mobility detection in humans using micro-Doppler radar also had high classification accuracy rates (Li et al., 2020, 2021). We recognise that our study's main difference lies in how the animals were assessed (label generation). In the two studies (Busin et al., 2019a; Shrestha et al., 2017), the cows were scored live and from the rear of the animals by a single assessor, whereas multiple assessors in our study scored the cows from the side using video recordings of the cows. The second difference we identified was the uncertainty of the labels for use as ground truth, which is the major and possibly the main challenge in having accurate and reproducible machine learning results and process automation. In human mobility classification (Li et al., 2020; Li et al., 2023), the researchers had high confidence in the labels they used for ML training as they dictated to the participants the performed actions (i.e., gait and motion patterns). According to our previous studies, there is considerable variation among assessors regarding cow mobility scoring, hence uncertainty regarding the true label of a cow. Only a few animals maintained the same mobility score throughout the longitudinal study, for which we have confidence and may be promising markers for training a model. Therefore, a subsequent chapter will focus on only these animals for the model’s training.

Machine learning studies have often reported challenges in creating labels (Cruciani et al., 2018; Schröder et al., 2016). It is challenging to have a generalisable label, especially when there is no single objective source of truth. Examples of classification challenges have been presented in Chapters 2 and 3, where assessors perceived and interpreted the given data (videos of cows for scoring) differently and subjectively, making it impractical to create a ground truth. This difficulty also becomes apparent in the present study. For example, when we plotted three extracted features and used colours corresponding to the consensus labels (Figure 4.3), they appeared all clustered without clear class separations, making the lack of exclusiveness (categories within classification do not overlap) and exhaustiveness (all cases are covered or accounted for in the classification system) of the levels obvious. Animals also introduce difficulty in creating a generalisable label due to their unique gait and body structure that might deviate from a typical pattern but not adversely affect their well-being. Antithetically, because cows have been characterised as stoic animals (Weary et al., 2009), they may have a mobility problem and not exhibit any obvious sign of discomfort until severe. However, recorded hoof lesions were also used as labels with the ML algorithms, and the prediction characteristics (average accuracy 58.3%) did not exceed the accuracy achieved by using assessors' scores as labels (this analysis is not presented here). These highlight the difficulty in the label creation and machine learning part of an animal attributes' classification, which applies to most systems that use supervised machine learning, not just the micro-Doppler radar system. In all cases, supervised machine learning requires well-defined and precise labels to produce results with high accuracy.

Accuracy, however, is not always the most appropriate way to judge a classification model algorithm (Harrell, 2001). Accuracy, in this case, is the percentage of correct classifications among all classifications, and it is a straightforward and intuitive measure to assess each model's performance. However, suppose we have probabilities (0.2, 0.8) for a prediction outcome; then accuracy will be 80% if we classify everything in the second category and completely ignore the 20% chance that any result could be in the first category. On account of this, except for the probabilistic output, one needs to consider some more aspects, such as the consequences of the predictions (what happens if a cow is classified as healthy but is not?) and whether the classes are exhaustive

and exclusive enough for each level. When the thresholds of a system's levels are unclear, a cow may be classified in more than one category, and the classification ends up depending on characteristics other than the animal, such as the assessor's experience. Finally, cows cannot always be classified according to defined visual observation thresholds in mobility characterisation, which is basically what mobility assessment systems try to achieve. A cow can range from perfectly healthy to unable to take a step forward or anything in between. There is a continuum in movement and lameness characterisation, and the set thresholds are essentially cognitive shortcuts which must be passed to machine learning to produce valid results. The overall performance of machine learning tools needs to be considered, and a way to pass complex decision-making tasks related to mobility characterisation to a machine must be found to have a reliable accuracy outcome.

This study was an attempt to examine the performances of labels and algorithms. Unfortunately, the predictions from the supervised machine learning classification models were not significantly reliable with the labels we used. In a future study, it would be helpful to make some changes either to the way the labels are collected or to the machine learning analysis so that we can observe different outcomes, for example, considering different features and classifiers.

4.5 Conclusion

This study addresses the challenge of creating cows' mobility classification labels for use as ground truth in machine learning. Labels from the binary systems performed better with machine learning models than the 4-level systems. Combinations of machine learning models and binary systems had no consistent difference in performance. There were many misclassifications, but it was not unexpected since the agreement among human assessors was low in most sets. There is a need for studies investigating early lameness cases and a way to have confidence in the assigned scores and then pass accurate labels to machine learning algorithms.

Chapter 5

Validation and Enhancement of an AI Tool for Automated Lameness Detection: A Cross-Sectional Study with Alternative Labels and Pre-Processing Techniques

5.1 Introduction

This chapter aimed to improve an AI tool for automated lameness detection developed in Chapter 4. To follow up on the longitudinal study, a cross-sectional study was designed to collect data from a different cow herd and analyse it using the same method. The objective was to obtain additional evidence to test the system on a large number of animals, thereby testing its generalisability and acquiring data to improve its performance. Specifically, we investigated if the discouraging results of our previous studies (Chapters 3 and 4) were replicated in other farms with a larger number of animals. The chapter details our attempts to enhance the machine learning prediction outcomes through an alternative method of labelling and data pre-processing before loading them into the supervised machine learning classification model.

Micro-Doppler radar has shown encouraging results for mobility classification in cattle (Busin et al., 2019; Shrestha et al., 2018) and humans (Fioranelli et al., 2019), presenting advantages over other systems. Micro-Doppler systems offer benefits including high-resolution measures in range and velocity, real-time monitoring, low power consumption for classification of moving targets while also providing distance monitoring without interfering with targets' operations, immunity to environmental conditions, suitability for challenging conditions such as farms, and relative affordability as they do not need to be replaced or increased numbers over time, making them a valuable tool for monitoring mobility classification in various applications such as cattle lameness detection.

A micro-Doppler radar acquires data on the motion and vibration of objects in the field of view of an antenna (Chen et al., 2014). The obtained data can then be

analysed using time-frequency analysis and machine learning (Chen, 2008). In machine learning and pattern recognition, the raw data are transformed into a set of features so the algorithms can use the data information. The feature extraction process is necessary to reduce the dimensionality of the data and any possible noise or irrelevant information that can negatively impact the algorithm's performance. Feature extraction methods include numerical, statistical, and dimensionality reduction techniques such as principal component- or linear discriminant analysis (Nisbet et al., 2018; Salau & Jain, 2019). The process depends on the problem and the requirements of the ML algorithm. In our study, using feature extraction and statistical learning instead of deep learning, where the algorithm automatically extracts features and classifies them in one step, ensures that an embedded platform can run both the signal processing and machine learning algorithms on a resource-limited platform. After extracting features from the raw data, the numerical information is used along with a set of labels in algorithmic classification models. The model outcomes, i.e., the predictions, depend highly on the link between the labels and the extracted features (Parsons, 2010). Labels, essentially, provide the ground truth that the algorithm is trying to predict and define the classes into which the data is to be divided. A problem is introduced in cases where the ground truth is uncertain, as in some of our previous studies. Similar limitations have been demonstrated in the literature, particularly in the broad medical field, where there is a high degree of uncertainty in label assignment, and artificial intelligence tools to aid decision-making are becoming indispensable. A typical example comes from radiology and computer-aided detection (CAD). Until a few years ago, its usefulness was questioned due to poor results produced out of training with uncertain labels. A high number of apparently false positive results from CAD resulted in increased workload and associated costs (Oakden-Rayner, 2019). However, technology has advanced, and the interest in automation and computer-aided diagnosis and detection has increased again, increasing reliability in labels and producing encouraging results like those in the study by Rajpurkar et al. (2018).

Since technology and automation is prevalent, and popularity is constantly increasing, significant initiatives have been made to improve and tackle limitations with labeling issues. Some examples of recent studies (Khetan et al., 2017; Reamaroon et al., 2019; Shi & Wu, 2021; Zheng et al., 2021) suggest

different methods to deal with noisy or uncertain labels and describe techniques such as active learning, Bayesian model calibration and ensemble methods to adjust the predicted probabilities of a machine learning model so that they accurately reflect the labels' uncertainty. In our study, we used a pre-processing calibration technique described by (Li et al., 2023) in their closely related research which used the same micro-Doppler radar system for data acquisition as ours. Their results were optimistic, reaching up to 93.1% classification model accuracy using an adaptive thresholding method with a support vector machine (SVM) classifier for human activity recognition. Motivated by their encouraging results (prediction accuracy of more than 90%), we thought masking would be an appropriate and straightforward approach to improve our classification model's accuracy.

Therefore, the present study aimed to investigate the effectiveness of different data labelling systems, some of which have been seen in our previous studies and some that we have not used before, as well as a data pre-processing technique to obtain reliable data with high accuracy from a supervised machine learning algorithmic model. A new labelling system developed by Lorenzini et al. (2017) was selected, as it has a few exclusive levels, presented promising results regarding the inter-assessor agreement, and was developed to assess animals that do not show lameness traits used in other mobility systems such as in Sprecher or AHDB. The study will present comparative results before and after applying the data pre-processing approach and comparisons of accuracy results using the different labelling systems. In this way, we hope for accurate and valid predictions that will benefit mobility classification automation and possibly establish a reference point for future studies in ML labelling and processing.

5.2 Materials and methods

5.2.1 Farms and animals

For this study, we dealt with two central Scottish dairy farms; from now on, they will be referred to as farms A and B. The visits to the two farms took place in April and July 2022, where we recorded 369 and 59 cows, respectively. During the afternoon milking, we obtained radar measurements from the cows' rear as they returned to their cubicles. Following this, we captured videos of each cow from a lateral distance of approximately 5 meters as they walked a passageway (approximately 7 meters long and 1.65 meters in average width) enclosed by steel rails. This enabled us to gather data for labelling purposes, similar to previous studies, and to acquire each animal's micro-Doppler signatures using radar. Because of a failure in video recording on Farm A, a second visit was conducted 5 days later to collect videos of all cows from the same vantage point as the first visit.

5.2.1.1 Labels – scoring

The videos were recorded with the Kodak PlaySport Zx5 Full HD 1080P camera (same device as in previous studies). The videos were transposed into an Access Database file and were shared with three assessors used in previous chapters for evaluation. This study used more than one system to evaluate cows' mobility. First, assessors were asked to individually score the animals using the AHDB 4-level mobility system; then, they met twice to produce the consensus scoring sets. Once, they evaluated based on the AHDB four-level system, and the second time, they used the three-level system proposed by Grimm and Lorenzini (Lorenzini et al., 2017) with scores as described in Figure *5.1*.

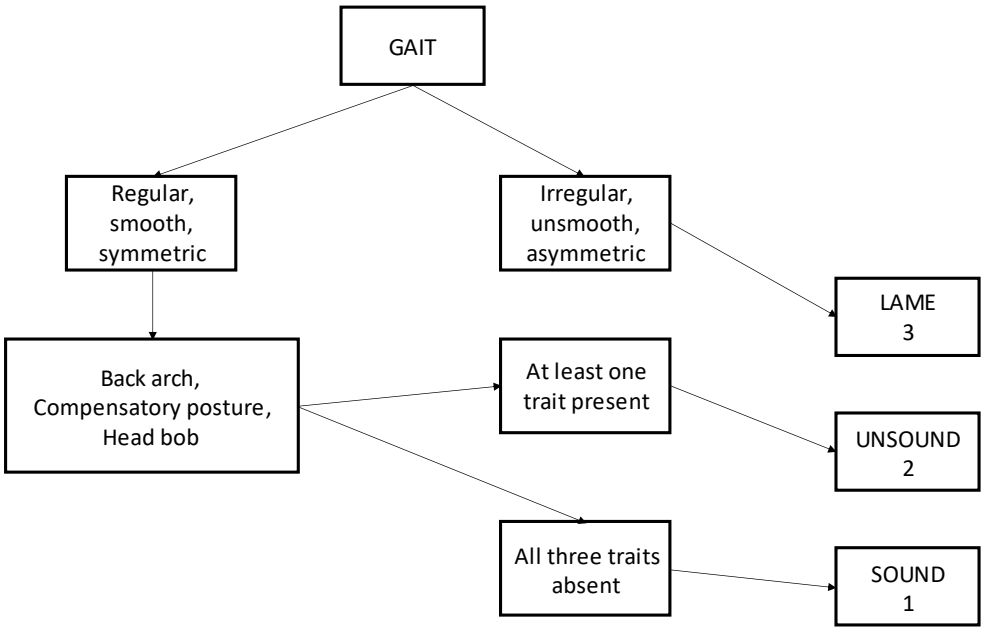


Figure 5.1 Diagram of the 3-point locomotion score by Grimm and Lorenzini. Adapted from "Using a three-point lameness scoring system combined with a clinical examination to increase the reliability of locomotion scoring" by Lorenzini et al., (2017).

As in previous chapters, the scores of the four-level systems were also converted to a binary system by combining the scores of 0,1 and 2,3.
With the above systems, we had fourteen sets of labels to use in the classification algorithm for Farm A and five sets of labels for Farm B. More details follow in Table **5.1**.

There is a difference between the farms and the labelling systems used in each. Due to a significant difference in the number of videos and time constraints, we could not assess both farms using all the systems.

Table 5.1 After scoring and modifying the systems, we obtained scores from 14 systems to be used as labels with the machine learning classifier. Not all systems were used for both farms due to limited evaluation time during the experiment.

System	Description	Labels	Farm
AHDB (individual scoring)	3 assessors – 4 levels scores	0, 1, 2, 3	A
AHDB average	The rounded average of the individual scoring	0, 1, 2, 3	A
Binary AHDB (individual scoring)	3 assessors – 2 levels scores	0, 1	A
Binary AHDB average	The rounded average of the individual scoring	0, 1	A
Convergent (Individual)	3 assessors – 4 levels	0, 1, 2, 3	A
Convergent Binary (individual)	3 assessors – 2 levels	0, 1	A
Convergent (average)	The rounded average of the individual scoring	0, 1, 2, 3	A
Convergent Binary (average)	The rounded average of the individual scoring	0, 1	A
Grimm & Lorenzini (individual)	3 assessors – 3 levels	1, 2, 3	A and B
Grimm & Lorenzini Binary	3 assessors – 2 levels (1,2 vs 3)	0, 1	A and B
Grimm & Lorenzini average	The rounded average of the individual scoring	1, 2, 3	A and B
Grimm & Lorenzini Binary average	The rounded average of the individual scoring – (1,2 vs 3)	1, 2, 3	A and B
Grimm & Lorenzini Consensus	One score from 3 assessors	1, 2, 3	B
Grimm & Lorenzini Consensus Binary	One score from 3 assessors (1,2 vs 3)	0, 1	B

5.2.1.2 Statistical analysis – agreement (different ways)

For the statistical analysis, we used further formulas than in previous studies to compare our results with other research. All analyses were performed in R (version 3.212).

We first calculated pairwise inter-assessor agreement for individual assessors. We used:

- Cohen's kappa for pairwise and Fleiss's kappa for multiple assessors (kappa2 and kappam.fleiss – (Gamer et al., 2019))

Equation 5.1 Cohen's kappa formula

$$\text{Cohen's kappa} = \frac{P_0 - P_e}{1 - P_e}$$

Equation 5.2 Fleiss's kappa formula

$$Fleiss's\ kappa = \frac{(N * (P_0 - P_e))}{N - 1}$$

Where P_0 is the relative observed agreement among assessors, P_e is the probability of chance agreement, and N is the number of assessors.

- First-order agreement coefficient (AC1) by Gwet (2008)

Equation 5.3 AC1 formula

$$\mathrm{AC1} = \frac{P_a - P_{e\gamma}}{1 - P_{e\gamma}}$$

Where P_α is the overall agreement probability including by chance/not by chance, and $P_{e\gamma}$ is the chance-agreement probability.

- Kendall's coefficient of concordance (Kendall, 1938)

Equation 5.4 Kendall's tau formula

$$Kendall's\ Tau\ =\ (C\ -\ D\ /\ C\ +\ D)$$

Where C is the number of concordant pairs and D is the number of discordant pairs.

Cohen's kappa is one of the most popular statistics for inter-assessor agreement, but it is also considered too strict regarding the chance agreement (Bexkens et al., 2018; Feinstein & Cicchetti, 1990). The main differences between Cohen's kappa and Gwet are that Cohen's kappa is directly affected by prevalence and marginal probability, while Gwet remains less affected and relatively constant (Wongpakaran et al., 2013). The difference between kappa and Gwet statistics compared with Kendall's coefficient is that the first two calculate the absolute agreement between ratings while Kendall's measures the correlation between ratings. We used these different methods because there is no single standard for analysing inter-assessor agreement in the literature, and we wanted to be able to compare our values with those of other studies. Finally, an analysis of variance (ANOVA) among the three statistical methods was performed to examine if differences in observations were statistically significant.

5.2.2 Radar

5.2.2.1 Radar setup

The radar equipment is the same as described previously in Chapter 4: an FMCW radar system from Ancortek operating at 5.8 GHz, with a bandwidth of 400 MHz and a pulse repetition frequency of 1 kHz. We used the same two Yagi antennae, a transmitter, and a receiver, with approximately 100mW of transmitted power. The height of the antennas and the distance between them were the same as in every other experimental design, 1.5 m high and 40 cm distance between them, so the radar setup did not affect the results. The recording time for each cow was 12 s (reduced compared to the longitudinal study of Chapters 3 and 4, which was 45 s), trying to imitate the everyday time routines of cows exiting the milking parlour. We set up the equipment in critical locations within the premises of each farm, where we would be allowed to record the animals without disrupting daily routines (Figure 5.2).

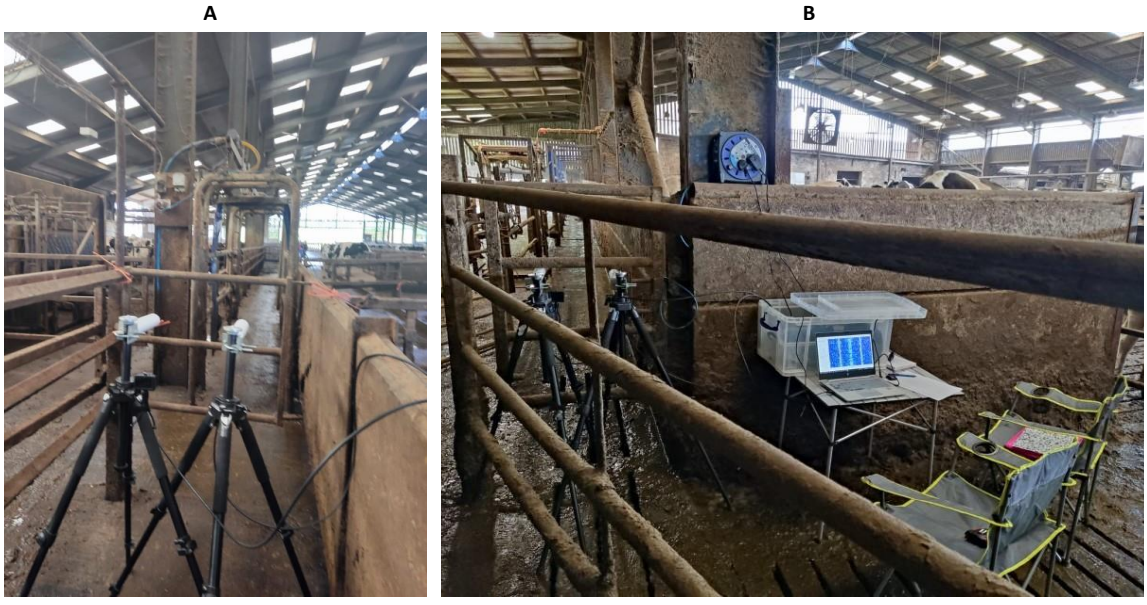


Figure 5.2 Micro-Doppler radar system set-up in farms A and B. The equipment (radar and antennas) was set up at critical points so as not to interfere with the daily routine of the farm and the animals. The antennas were pointing at the rear of the cows exiting the milking parlour.

5.2.2.2 Radar data processing

We followed the same method as in chapter 4 (Figure *5.3*) to process the radar data with MATLAB (*MATLAB R2022b*, 2022). We first applied a Hamming windowed fast Fourier transformation to each chirp creating a range-time map. Then, because the moving target and the surrounding objects reflect the transmitted radar wave, we applied a moving target indicator (MTI) filter to remove noise and static clutter. Then a short-time fast Fourier transformation followed by a 0.2 s Hamming window with a 95% overlapping factor, creating the micro-Doppler signature (Spectrogram).

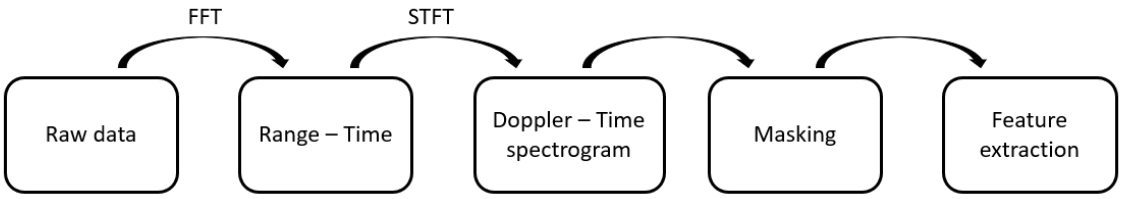


Figure 5.3 The steps followed in extracting numerical features for classification from the raw radar data with the masking application. First, a Fast Fourier Transformation followed by a Short Time Fourier Transformation were applied to the data to create range-time and then Doppler-time data. Then, the masking pre-processing operation and the feature extraction were performed.

Because the spectrograms have a significant area of non-useful information (speckled blue colour - as shown in Figure 5.4 A), we wanted to exploit only relative information. Thus, we pre-processed the data by first converting the spectrograms to grayscale images Figure 5.4 B. Then, we applied a mask to the grey scale spectrograms to further filter the signal and focus on the region of interest. For the masking process, we compared the grey-scale micro-Doppler spectrograms to a set threshold created with an adaptive method for each spectrogram (further details on the threshold generation process can be found in Li et al. 2021 work). Then the spectrogram values were compared to the

threshold, and 0s and 1s replaced the signal values, so the final image was binarised - black and white (Figure 5.4 C). The development of the generated spectrograms following these steps can be seen in Figure 5.4.

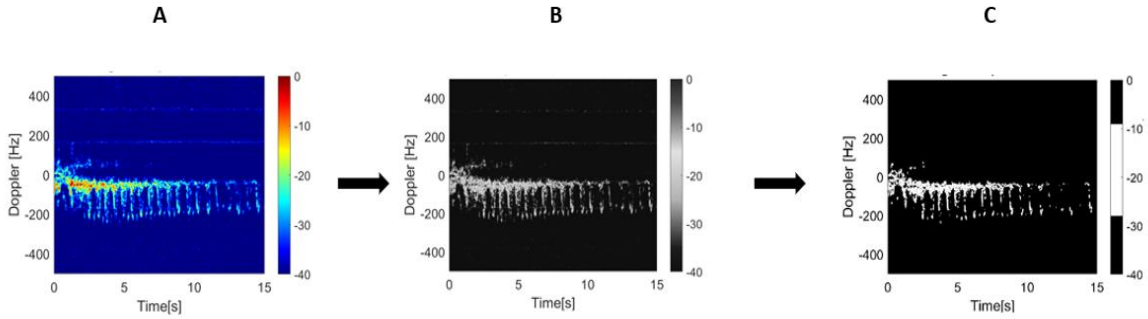


Figure 5.4 Mask pre-processing technique. The spectrogram (A) was first converted to a grey scale (B) and then was binarised, creating only black and white (0 and 1) figures (C), thus retaining only the spectrogram's area of interest with the most helpful information.

Finally, we extracted the features in Table *5.2* from the masked signals to use in machine learning classification. The selection of numerical features we chose to extract has been proposed and used in other studies (Li et al., 2021; Sharma et al., 2018), and we thought they would convey helpful information that could enhance the classification task.

Table 5.2 Extracted features from the masked spectrograms.

Category	Features	Brief description
Radar spectrogram features	Entropy, skewness, centroid (mean and SD), bandwidth (mean & SD), energy curve (mean, SD and trapezoidal numerical integration), singular vector decomposition (mean and SD of the first three vectors of components)	These features are used to analyse radar data by extracting characteristics such as the randomness, asymmetry, centre of mass, bandwidth, and energy distribution of the signal, as well as the first three vectors of the Singular Vector Decomposition. [1]
Region of interest (ROI) features	Perimeter area, centroid, eccentricity, orientation, major and minor axis length	These features are used to describe specific regions of an image by measuring the length of the boundary, the total area, the centre of mass, the elongation, the orientation, and the length of the major and minor axis of an ellipse that fits the region. [2]
Textural features	Local binary pattern (LBP) of image, moment of image	These features are used to describe an image's texture by analysing the local texture patterns using Local Binary Patterns and measuring the statistical descriptors of the spatial distribution of the intensity values using Moment. [3]

[1] (Mahafza, 2016)
[2] (Gonzalez & Woods, 2006)
[3] (Sonka et al., 2008)

5.2.3 Classification

We used a quadratic support vector machine (SVM) with a 10-fold cross-validation method for classification. The selection of the quadratic polynomial kernel (SVM) was based on its consistently high accuracy demonstrated in various radar data classification studies using similar features. SVMs are effective in high-dimensional spaces with many features, as they can find linear and non-linear boundaries to separate classes (Hsu et al., 2008). SVM classifiers are also robust to outliers and are not biased towards the majority values, making the model accurately predict less common values in the classification (Cristianini & Shawe-Taylor, 2000). All computations were performed in the Matlab R2022b classification application.

The results section will present the false positives and false negatives, positive and negative predictive values, and the classification models' specificity, sensitivity, and accuracy outcomes. Accuracy, specificity, and sensitivity were calculated according to Equation 5.5, 5.6, and 5.7, respectively. Receiver operating characteristic curve plots are included in Appendix A.

Equation 5.5 Accuracy formula

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Equation 5.6 Sensitivity formula

$$Sensitivity = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

Equation 5.7 Specificity formula

$$Specificity = \frac{True\ Negatives}{(True\ Negatives + False\ Positives)}$$

To assess if the masking pre-processing technique affected the classification accuracy of the ML predictions, a paired t-test was used, as we had measures of the same data before and after applying the masking. Since the sample size was not large enough (less than 20), we checked whether the differences between the pairs followed a normal distribution by performing a Shapiro-Wilk normality test.

5.3 Results

5.3.1 Labels – scoring agreement

After filtering out videos that were difficult to assess due to cows that were not in the frame long enough to be scored or two cows in the video simultaneously, 325 videos from Farm A and 55 videos from Farm B were left for scoring.

The results of the agreement levels among assessors are presented in *Table 5.3* and Table *5.4*. Analysis with Cohen's kappa punishes the discrepancy between evaluators over the other two systems to a greater extent. Gwet (AC1) analysis produced higher values when used to calculate agreement with binary systems, while Kendall had higher correlation values in multilevel system analysis. The last two systems produced higher values for inter-assessor agreement than Cohen's kappa analysis. The best agreement for farm-A (Table 5.3) was achieved in the Binary Convergent system (0.79) and for farm-B (Table *5.4*) in the Binary Constructed system from Grimm & Lorenzini (0.60) with the Gwet (AC1) statistical way of analysis. However, according to the ANOVA test, the differences among the different statistical analysis systems were not statistically significant (Figure 5.5).

Table 5.3 Mean (and SD) of the pairwise comparisons between the three assessors for Farm A for each statistic.

	Cohen's Kappa	AC1 coefficient	Kendall tau	% Agreement
AHDB	0.19 (0.03)	0.29 (0.04)	0.44 (0.13)	44.37 (2.75)
Binary Constructed AHDB	0.27 (0.07)	0.52 (0.16)	0.34 (0.02)	70.47 (7.12)
Convergent	0.58 (0.12)	0.66 (0.11)	0.74 (0.05)	73.4 (8.43)
Binary Convergent	0.61 (0.09)	0.79 (0.07)	0.65 (0.07)	86.17 (3.95)
Grimm & Lorenzini	0.27 (0.1)	0.37 (0.13)	0.4 (0.14)	55.37 (7.7)
Binary Constructed Grimm & Lorenzini	0.24 (0.04)	0.72 (0.13)	0.31 (0.04)	79.4 (7.67)

Table 5.4 Average values and standard deviations of pairwise comparisons for Farm-B using three statistical analysis systems, providing a summary for comparing the performance of different scoring systems.

	Cohen's kappa	AC1 coefficient	Kendall's tau	% Agreement
Grimm & Lorenzini	0.38 (0.12)	0.42 (0.13)	0.48 (0.17)	58.83 (7.82)
Binary Constructed Grimm & Lorenzini	0.45 (0.20)	0.60 (0.10)	0.48 (0.18)	77.17 (7.83)

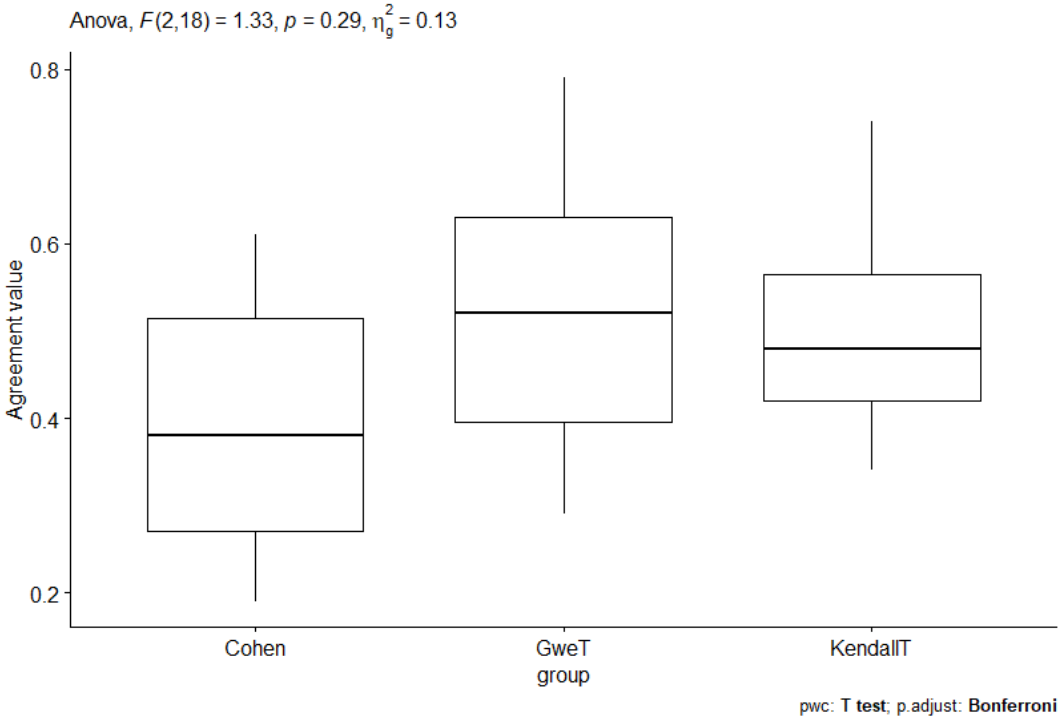


Figure 5.5 Analysis of variance (ANOVA) results to determine differences among the three statistical analysis groups used in farms A and B. ("pwc" stands for pairwise comparisons).

5.3.2 Classification with and without data pre-processing

The data pre-processing technique used, i.e., spectrogram masking, did not substantially affect the accuracy results of the classification models (Table *5.5*). However, in some instances, accuracy was reduced after masking (i.e., Table *5.5* Binarised convergent averaged Farm A - from 75.8 reduced to 68.9). Best accuracy was achieved with Binarised Averaged Grimm and Lorenzini for farm A with no masking and Binarised Consensus Grimm and Lorenzini for farm B after masking (Table *5.5*). In individual scoring, Assessor 1 labels had some of the highest accuracies with and without masking. However, in the example of the confusion matrices for unmasked binarised Grimm and Lorenzini (*Table 5.6*), the accuracy for Assessor 1 is seemingly good (93.5), but the model classified almost all animals in only one of the classes (Figure 5.7). Conversely, the accuracy value of Assessor 2 was lower (69.9) at the confusion matrix (Figure 5.7) but with a slightly better prediction distribution as a few data were classified in the second class. However, in this case, also, 97 animals were misclassified. The individual ROC curves corresponding to Tables 5.5 and 5.6 can be found in Appendix A.

Table 5.5 Comparison of accuracy results of models using different annotation systems before and after the data pre-processing (masking) for the two farms.

	Unmasked Spectrogram		**Masked Spectrogram**	
	Farm A	Farm B	Farm A	Farm B
AHDB average	50.3	-	48.1	-
Binary AHDB average	73.9	-	76.1	-
Convergent Average	48.1	-	47.2	-
Binary Convergent Average	75.8	-	68.9	-
Grimm & Lorenzini Average	49.7	53.7	51.6	40.7
Binary Grimm & Lorenzini Average	87.3	51.9	85.7	51.9
Grimm & Lorenzini Consensus	-	37	-	31.5
Binary Grimm & Lorenzini Consensus	-	48.1	-	59.3

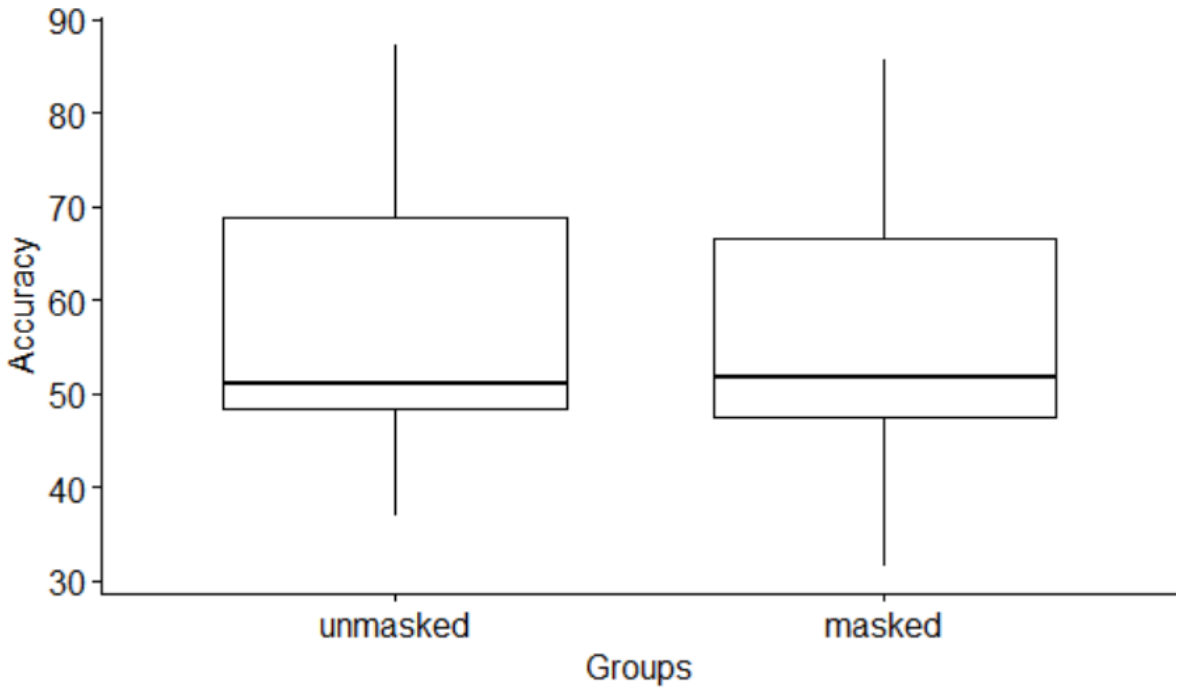


Figure 5.6 No statistical differences were found in the accuracy t-test results for the two groups - before and after applying the masking pre-processing technique. (t = 0.73285, df = 9, p-value = 0.4823)

Table 5.6 Comparison of models' accuracy results using each assessor's scorings for each system before and after data pre-processing (masking).

		Unmasked Spectrogram			**Masked Spectrogram**		
		Assess1	Assess2	Assess3	Assess1	Assess2	Assess3
Farm A	AHDB	57.5	36.3	36.3	49.7	34.5	33.5
	Binary AHDB	89.8	57.8	66.8	88.5	53.1	66.5
	Convergent	47.8	45.3	48.8	43.5	39.1	49.4
	Binary Convergent	82.3	62.7	76.4	82.3	61.8	76.4
	Grimm & Lorenzini	61.5	43.5	42.8	63.7	39.4	43.1
	Binary Grimm & Lorenzini	93.5	69.9	81.6	94.1	67.4	83.8
Farm B	Grimm & Lorenzini	48.1	51.9	35.2	44.7	42.6	40.7
	Binary Grimm & Lorenzini	63	55.6	46.3	68.5	55.6	51.9

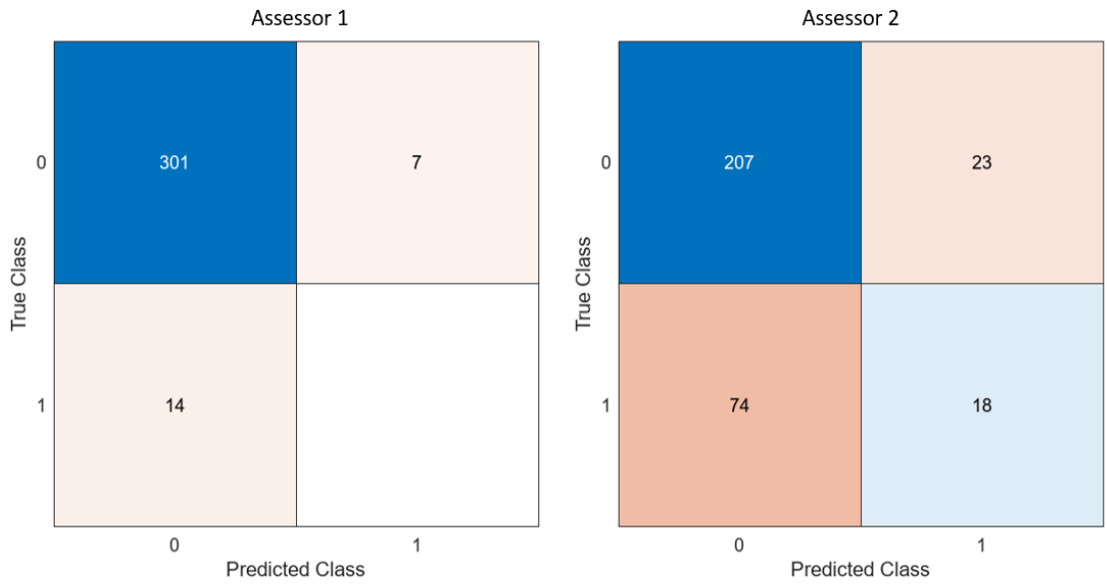


Figure 5.7 An example: Confusion matrices of models' accuracy when using the Binarised Grimm and Lorenzini labels of Assessor 1 (left) and Assessor 2 (right). Model accuracy when using Assessor 1 labels was 93.5, and accuracy with Assessor 2 labels was 69.9. Despite the greater accuracy of the one model, the model would always classify the data into only one category.

5.4 Discussion

The study aimed to follow on previous research and use the developed supervised machine learning algorithm to test the system on a larger number of animals, thereby testing its generalisability and acquiring data to improve its performance. At the same time, we introduced an additional suggested annotation system for label generation and a pre-processing technique to improve ML classification. Unfortunately, the results we obtained were not as expected, with mainly low-accuracy predictions from the algorithms, regardless of the application of the pre-processing masking method. Inter-assessor agreement varied among the labelling systems and statistical computation methods but was consistently below any reasonably acceptable level (i.e., ideally, a false negative rate of less than 5% and a false positive rate of less than 10%) for an effective automated system. The binarised convergent-AHDB was the system that produced better overall results for the inter-assessor agreement, but the performance was not transferred in the ML predictions.

We found a lack of consistency among the scores provided by the different assessors when we used Cohen's kappa statistics to analyse the data. The highest strength of agreement (Cohen's kappa 0.61) was produced when we compared scores from the binarised convergent-AHDB system in farm A. According to Landis & Koch's (1977) interpretation, this is substantial agreement as it exceeds the value of 0.6. The results of the compared analysis are consistent with the results of other studies, such as Rutherford et al. (2009), which also had kappa values between 0.42-0.73 when using a binary system to assess cattle mobility. When we used the rest of the mobility assessment systems, the results were poor, in agreement with our studies described in previous chapters and with other research that analysis of agreement with the kappa statistic yielded low results (kappa among 3 assessors = 0.42 - Lorenzini et al., 2018). However, even if not statistically significant, when Gwet statistical analysis was used, there was an expected overall improvement in inter-assessor agreement values, as it is more stable than the kappa coefficient and "paradox-resistant", as Wongpakaran et al. (2013) have put it in their study. When the Kendall analysis was used, it consistently had higher values than the kappa statistics results. Still, it did not always have higher values than those produced by Gwet. Agreement with the

Kendall method ranged from 0.31 when using the scores from the binarised Grimm and Lorenzini system on farm A to 0.74 with the convergent-AHDB scoring system. Our results using the Kendall analysis with the Grimm and Lorenzini evaluation system were 0.4 for both farms. These values show that assessors' scores are positively but not highly correlated since tau < 0.5. Considering the percentage agreement using the Grimm and Lorenzini system, one can see that the assessors assigned the same score to only about half of the assessed animals. The highest percentage agreement (86.17) was produced with the binarised convergent-AHDB system, while the other systems had 44.7 to 79.4% agreement. These results, especially after the analysis with Kendall's tau, surprised us as they disagreed with the results of Lorenzini et al. (2017), where tau=0.7. While we followed the same methods to produce the results (i.e., video cow evaluation and their proposed three-level system), the inter-assessor agreement in our study was considerably lower. While Lorenzini et al. (2019) suggested that breed may affect the expression of lameness traits in cows, our study did not directly investigate this factor. Therefore, it remains unclear whether differences in breed composition and in assessors may have contributed to the discrepancies observed in our results compared to those in their study. The implications of these findings suggest that the choice of statistical analysis method does not play a critical role if the assessment system is not consistent and reliable. The study's results highlight the limitations of using specific evaluation systems in producing labels for machine learning and the need for further investigation into the factors that may affect the accuracy and reliability of cattle mobility assessments.

When we used the individual scores of the three raters as labels in the ML model, we found differences between assessors and scoring systems. These results reproduce the same outcomes as in previous chapters, where there were variations and uncertainty for the most accurate assessment. The labels given by Assessor 1 yielded better accuracy when used as ground truth in the model than the other assessors' scores. However, although accuracy is a commonly used metric to evaluate machine learning models, it is not always the best since it does not always reflect the model's performance. In our analysis, the models perform well in one class but poorly in another class, and the overall accuracy may be high, but the performance of the model in the poorly performing class is not acceptable. There is not a single acceptable performance threshold; the acceptable levels of

false negatives and false positives for lameness detection may vary depending on several factors, such as the severity of the condition and the purpose of the detection. In general, it is desirable to have a high accuracy rate for detection, with around 5% false negatives and 10% false positives. These performance thresholds are based on the specific context and purpose of the lameness detection in our study, which would not cause unreliability or significant damage to the system user with the implications of misleading outcomes. However, in practice, we could accept a slight compromise between sensitivity (low false negatives) and specificity (low false positives) depending on the circumstances. This means that sometimes it may be acceptable to have a slightly higher rate of false positives or negatives depending on the situation. Classification of imbalanced data is the case with binary classification labels in our study, where the majority is gathered in only one data set class. Then the model would achieve relatively high accuracy by predicting the majority class each time. This happened in the cases where we used the Assessor's 1 labels - seemingly high accuracy but no or only minimal correct predictions for some of the classes. Studies have identified the problem of imbalanced data allocation concerning accuracy, which can result in a biased performance of models since classifiers prioritize error rates over data distribution (Patel et al., 2020; Tanha et al., 2020; L. Wang et al., 2021). Patel et al., (2020) have recommended some strategic solutions, including pre-processing techniques, algorithm modifications, cost-sensitive and ensemble approaches, and feature selection. We implemented two of these methods (pre-processing and feature selection) in our study, but the outcomes indicate that additional investigation is necessary.

Systems with good inter-assessor agreement resulted in higher accuracy when used in the machine learning models with only a few exceptions. For example, the convergent- and binarised convergent-AHDB systems had some of the highest agreement values. However, the SVM machine learning models did not have a statistically significant difference with different labels. When an average or consensus score was used as labels in the model, accuracy results ranged from 31.5 to 87.3, with only results from three models exceeding 73%. These values signal a poor connection between the labels we provided, the analysed data, and the patterns detected by the model. The poor model performance could mean that the labels are not representative of the data (i.e., errors or ambiguities in

the labels) or the data are noisy. There is much research on labelling misclassification that results in invalid predictions. A few examples from the medical literature (Brenner et al., 2016; Hubbard et al., 2017) demonstrated how diagnostic labelling misclassification could impact the accuracy of predictive models, leading to incorrect predictions and suboptimal treatment decisions. Data misinterpretation by the model because of the prevalence of only specific groups (i.e., cows with score-1 as a label in a binary system) could also be the case in our study. In most binary classifications, the model correctly classified cows in the majority class but misclassified most animals belonging to the other level. Similar issues have been reported in another recent study (Shahinfar et al., 2021) using machine learning to predict cow lameness incidents. While some of the models used had high accuracy in that study, some classes' binary classification predictions were poor. Their justification for their suboptimal results was due to the unbalanced training dataset or the general complexity of lameness, which are some of the common challenges we also deal with in our study.

Addressing the poor association between labels and data is essential, as it leads to incorrect prediction results that cannot be generalised. The two main approaches to improved predictions are improving the labels' quality and applying pre-processing techniques to the data. These two ways we followed in the present study: a labelling system that has shown promising inter-assessor agreement results and a masking procedure with radar data with high model accuracy outcomes. Despite using new labels and the data pre-processing technique, the results did not significantly improve compared to previous attempts. The new labels did not perform better than labels from other systems, and the masking approach did not improve the prediction accuracy of the models. Our hypothesis to explain the observed lack of performance in our study is that it may be attributed to the assessment process employed for label creation. Regarding the study by Li et al. (2023), where the masking pre-processing technique was first used, the main difference, apart from the study referring to human motion classification, is again in the labels. The data they worked with was a sample of people asked to do specific actions that the radar would record (Li et al., 2023). This means that the researchers had a solid reference and, therefore, ground truth to provide to the classification model. In our study, this was not the case, which

is quite common when dealing with a diagnosis with high uncertainty and even more when dealing with animals with no verbal and direct communication.

While our video recordings of animals in farm B were conducted 5 days after the radar recordings, a comparative analysis of these videos with some from the initial visit suggested consistent mobility statuses among cows. However, we acknowledge the potential for altered mobility in the interim 5 days for certain cows, introducing a temporal limitation to the study's observations.

Considering all our approaches and efforts to improve the machine learning predictions, we wish to reflect on the elements that might have affected the process and the results—from the beginning, i.e., recordings of videos and radar signals, till the final model's outcome. The analysis of this process will be presented in detail in the following chapter, emphasising the evaluation process of the mobility status of the animals resulting in labels that do not match the patterns detected by the radar system. In this chapter, we tried to process all the parameters we could from the data we already had - that is, new labels for the already obtained videos and data pre-processing for improving the link between labels and data. The next step will include a different way of acquiring data for analysis, with the expectation of different results.

5.5 Conclusions

Automating lameness detection using standard visual mobility assessment systems as labels in machine learning did not yield satisfactory results in our study. Regardless of the analysis method and the system used, there was always a degree of variation between the assessors' decisions, leading to uncertainty about the animal's true state. Although the models produced an acceptable accuracy in some cases, a class imbalance in most predictions means poor performance. The techniques we used, namely different labels and data pre-processing, did not help to improve the results. Further studies are recommended, and we plan to address the issue in the following chapter.

Chapter 6

Improving Automated Lameness Detection in Cattle using a rear assessing vantage point

6.1 Introduction

The previous chapters have highlighted the challenges associated with training machine learning models to accurately classify cow mobility. Although previous studies have reported satisfactory accuracy in predictions (Busin et al., 2019; Shrestha et al., 2018), our attempts to replicate and refine these methods did not perform well. Our objective was to enhance the automated detection of lameness in cattle by implementing pre-processing techniques and devising scoring systems to generate labels. Unfortunately, our prior efforts proved to be ineffective. This chapter aims to retrospectively evaluate our previous methodology and investigate potential causes of the unsatisfactory outcomes. We propose a novel approach for label generation and elaborate on the data acquisition process from the farm. Furthermore, we present three studies that provide a distinct perspective on the data and deliberate on potential directions for future research.

Various factors (Figure 6.2) could affect the assessor’s classification and label assignment process and can be broadly categorised into four main areas:

1- the environment in which the assessment takes place that can impact the assessor's decision-making,
2- the nature and characteristics of the animal which can affect the severity and presentation of lameness,
3- the assessment system used to classify lameness
4- the assessors themselves, in terms of experience, training and mood.

Understanding how these factors affect the classification process and decision-making is essential for reflecting on previous studies' outcomes and developing

accurate and effective strategies for generating useful labels for automated machine learning classification.

The environment and conditions prevailing on a farm can affect the outcome of the lameness classification. Firstly, weather conditions such as rain can affect the cow's gait; for example, a wet or muddy surface may make it more difficult for an animal to move in a typical way leading to a possible inaccurate assessment. Similarly, poor lighting conditions can make it difficult for the assessor to see the animal's movements clearly and contribute to misclassification. The assessor's vantage point can also affect lameness detection, with viewing distance and angle potentially exacerbating or minimising the appearance of gait abnormalities. Then, distractions, such as noise from machinery or other animals, and obstacles, such as herd mates or an uneven floor, can alter cows' behaviours and complicate the assessment. Finally, farms have demanding procedures and schedules to follow, and there is usually limited time for the mobility assessment process. A limited mobility assessment duration can impact the thoroughness of an assessment, potentially leading to rushed decisions on the scores. When individuals are required to decide within a limited timeframe, they tend to prioritize efficiency or opt for what appears to be the optimal choice, resulting in reliance on heuristics or cognitive shortcuts (Finucane et al., 2000). While utilizing heuristic approaches does not automatically entail incorrect decisions, it can potentially result in severe and systematic errors (Tversky & Kahneman, 1974)

Various animal-specific factors can also influence the classification of lameness in cattle, such as the animal's age, breed, and conformation. For example, certain types of cattle breeds like Holstein Friesian (Chawala et al., 2013) have been shown to be predisposed to clinical lameness incidents more often than other breeds, which may need to be considered during assessment. Additionally, older animals may have joint or other health issues that could impact their gait, leading to potential misclassification. The animal's physical structure or conformation is also an element to consider during the assessment, as it can be misleading. Finally, the level of familiarity the animal exhibits towards the observer could affect the cues the assessor receives during the evaluation. The literature provides examples of the effect of human presence on the animal behaviour and production (Hemsworth et al., 2000; Lange et al., 2020; Titterington et al., 2022), which can lead to fear and stress. Thus, animals evaluated either in vivo or through video

recording might yield different results, with the behaviour in the presence of assessors varying and exhibiting stoic behaviours and masking signs of lameness (Hudson et al., 2008). Therefore, it is essential to consider the potential impact of these animal-specific factors and take measures to minimise them, such as reducing the stress on the animals during the assessment.

The mobility system used for lameness classification typically involves a set of criteria used to evaluate the cow's gait and determine the presence and severity of lameness, which is another factor that influences the assessment outcomes. An exhaustive, exclusive, and coherent system that includes all relevant aspects such as gait, weight-bearing, swelling, and cow's other issues (i.e., metabolic disorders), and provides a cohesive picture of the cow's lameness which can all lead to a more accurate, precise, and unified understanding of the cow's condition and thus classification. An ideal system would be simple and efficient, easily adopted, consistent across different evaluators, not highly complex, and practical in a real-world setting. The accuracy and efficiency of a system can save time and resources by quickly identifying the problem, leading to a precise treatment and management of the lame animals.

The human factor is the next and perhaps one of the most critical factors affecting lameness classification. The process is subject to various human characteristics, including bias, experience, training, attentiveness, and mood. For example, Garcia et al., (2015) reported that assessors preferred to use different cow characteristics during mobility evaluation, and these preferences were associated with the agreement probabilities. We also identified that assessors had preferable attributes, such as speed of the cow and ability to localise the lame limb, which carried more weight during the consensus scorings in our studies. Then, attentiveness is necessary to identify subtle signs of lameness, and human mood can affect this attentiveness. Mood and emotions can influence the way individuals interpret and respond to information (Kahneman et al., 2021; Lerner et al., 2015). For example, if the assessors experience stress or excitement, they may be more likely to make impulsive or irrational decisions about the lameness scoring, which they would not make under other circumstances. Finally, assessors with different experience levels might yield different classification outcomes (Kristensen et al., 2006), whereas inadequate training might lead to inconsistencies and inaccuracies (Polderman et al., 2001).

Reflecting on the described factors (Figure *6.1*), the processes followed (Figure *6.2*), and reviewing the differences between our current experiments and previously published studies (Busin et al., 2019; Shrestha et al., 2018), we noted that the main difference was the use in the current studies of videos or live observation from a lateral vantagepoint, whereas in previous studies, a rear vantagepoint was used. To address this, we first chose to ensure the algorithms work, focusing on clear and confident examples (scores 0 and 3 from a 4-level mobility system) and to evaluate cows live and via video from a rear vantage point - the same perspective the micro-Doppler radar collects data. The studies presented in this chapter aimed to determine if changing the assessment vantage point would lead to improved results in automating lameness detection using micro-Doppler radar.

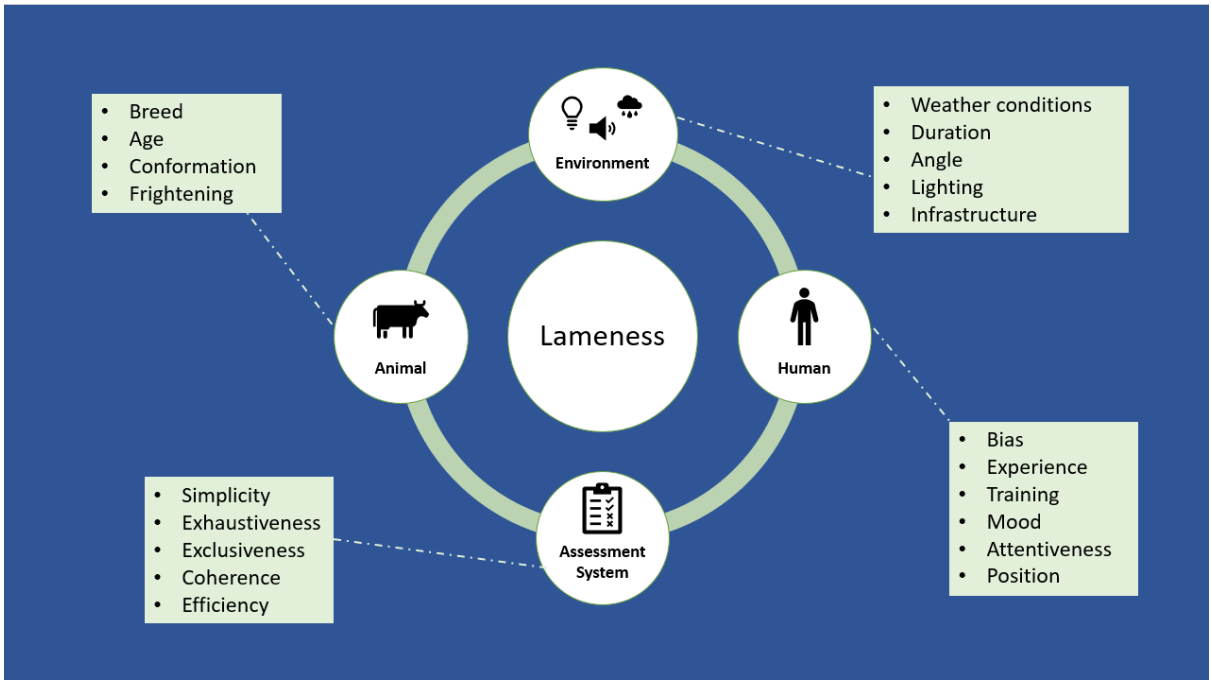


Figure 6.1 Factors affecting lameness classification related to the generation of labels for an ML system from human visual observations, categorised into four main areas; environmental factors, human factors, assessment systems, and animal factors.

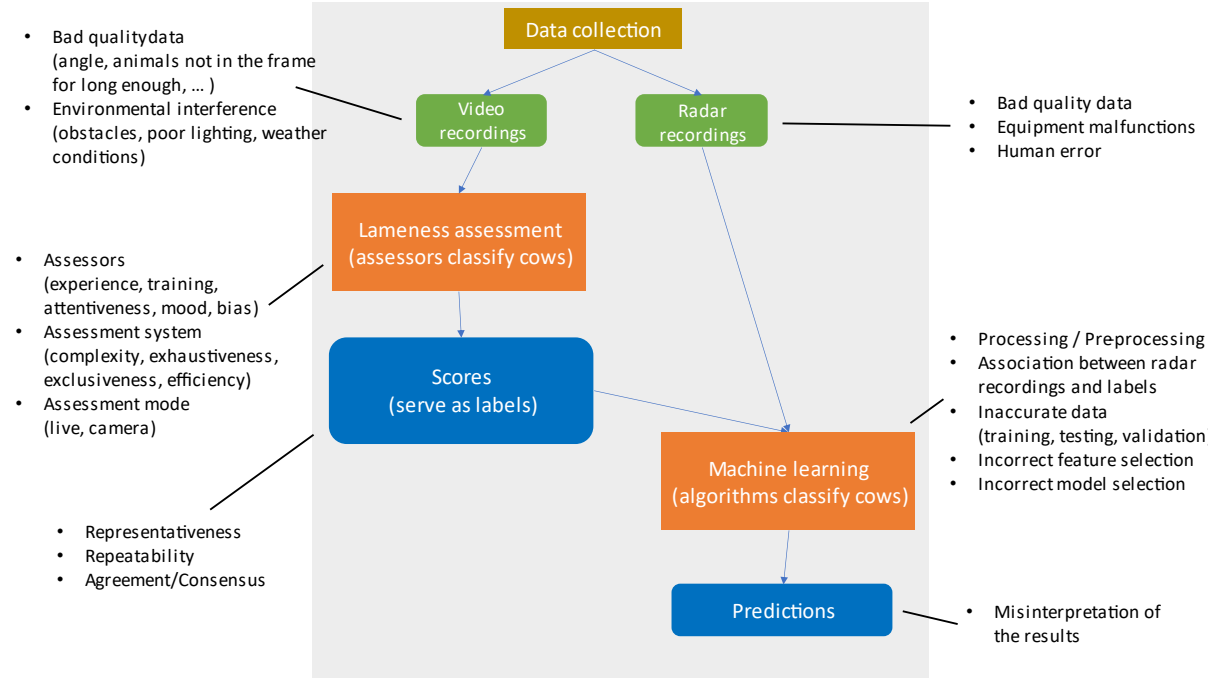


Figure 6.2 Points where errors may be introduced in the automation of the lameness classification process.

6.2 Materials and Methods

The chapter describes 3 studies (Figure 6.3). Studies A and B used a subset of the data collected during the longitudinal study described in chapters 3 and 4. For study C, new data were collected on the same farm using the same animals as in studies A and B, more than 6 months later.

6.2.1 Study A

For study A, the data from the cows with the extreme 3-assessor-consensus scores from the 4-level mobility scoring system (scores of 0 or 3) at any of the nine visits were used to examine the algorithms' performance. Thirty-nine cows were included: 18 with score 3 (severely lame) and 21 with score 0 (healthy).

6.2.2 Study B

For study B, videos collected from an action camera mounted on the tripod of one of the radar antennas during the longitudinal study were used. The camera's vantage point was to the rear of the cows. The data collected by this camera had not been intended for analysis in the previous experimental procedures but was for backup to reference in case there was a problem with the radar signal such as unexpected reflections. Only videos from Visit 5 were used, with 45 evaluated animals. Visit 5 was chosen because it was in the middle of the nine assessments, were clear. The videos were evaluated by three assessors using the 4-level AHDB mobility scoring system.

6.2.3 Study C

For study C, we returned to the same farm in central Scotland in December during an afternoon milking and collected radar data and mobility scores of 45 cows from a live assessment by a single assessor (the same assessor as provided the scores

from the previously published studies – (Busin et al., 2019; Shrestha et al., 2018) from a rear vantage point, using the AHDB 4-level mobility system. The micro-Doppler radar data signatures of each animal were collected in the same way and with the same tools as in chapters 4 and 5 (Frequency Modulated Continuous Wave radar operating at 5.8 GHz - bandwidth of 400 MHz and 1 kHz pulse repetition, 2 Yagi antennas (17 dBi gain) and a transmitted power of 100 mW.

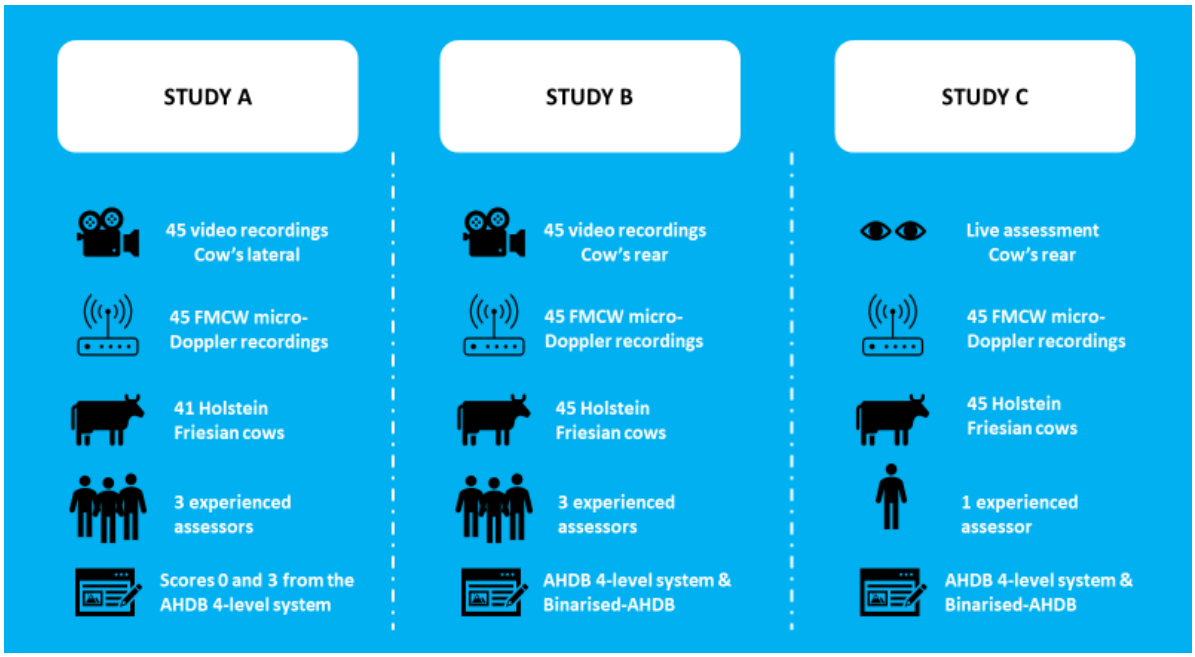


Figure 6.3 Brief visual description of the mean, numbers of animals & assessors, and scoring systems used in the three studies.

6.2.4 Statistical analysis

Statistical analysis for agreement quantification was performed only for study B in this chapter. We compared the scores given by the three assessors using Fleiss's kappa and percentage agreement on 4-level scores and 2-level scores derived from merging scores 0,1 and 2,3 as in previous chapters. In the previous chapter, it was found that Kappa statistics had no significant difference compared to other statistical analysis methods, such as Gwet. The strictness of Kappa statistics ensures that a high level of agreement implies high confidence; hence, it was chosen for the current analysis.

In all studies, accuracy, specificity and sensitivity for the algorithmic models' performance were calculated using the following formulae.

Equation 6.1 Sensitivity formula

$$Sensitivity = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

Equation 6.2 Specificity formula

$$Specificity = \frac{True\ Negatives}{(True\ Negatives + False\ Positives)}$$

Equation 6.3 Accuracy formula

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{count\ of\ all\ observations}$$

6.2.5 Machine learning

First, for study C, the same procedure as in chapter 4 preceded the classification to extract numerical features from the radar signals using the previously developed algorithm. Initially, a waveform was produced by processing the raw data. This was followed by the application of a pre-processing technique called Fast Fourier Transformation, which was utilized to extract range information, and then a Moving Target Indicator (MTI) was applied to remove static clutter by objects in the environment. Subsequently, a Short Time Fourier Transformation was employed, utilizing a 0.2 s Hamming sliding window with a 95% overlapping factor to extract Doppler-time signatures. Finally, numerical features were extracted from the resulting spectrograms for classification.

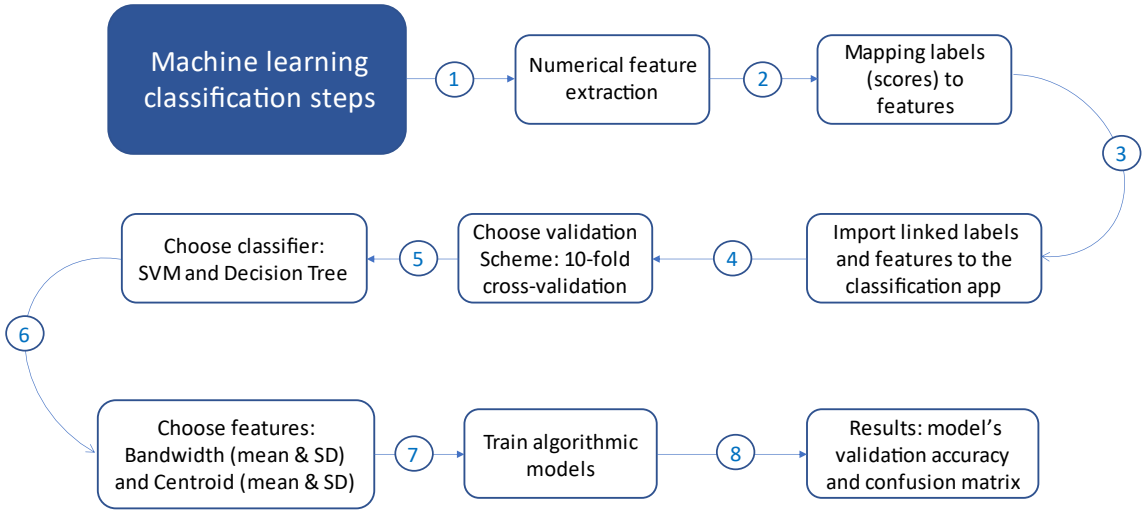


Figure 6.4 Machine learning classification process described in 8 steps. First, features are extracted from the radar signal, combined with the labels, and then loaded into the classification application, then, the desired analysis features are selected, and the prediction accuracy results are generated.

Then, the Matlab Classification Learner Application (*MATLAB R2022b*, 2022) was used to train (test and validate) the support vector machine (SVM) or decision tree classifier models for the three studies. A 10-fold cross-validation technique was used in all the studies' analyses for evaluating the models' performance as it provides a robust estimate and a good balance between accuracy and computational efficiency. The choice of classifiers and extracted features used in the signal analysis was based on results of previous chapters and published studies that used the same or similar experimental procedure with good results (e.g., Busin et al., 2019; Shrestha et al., 2018). The steps followed in the machine learning analysis are described in Figure *6.4*.

6.3 Results

6.3.1 Study A

The SVM model's accuracy when we used only the extreme values of the 4-level AHDB system was 0.82, with a specificity of 0.79 and a sensitivity of 0.85. More details are shown in the confusion matrix in Figure 6.5 A.

A scatter plot in the same figure (Figure 6.5 B) with two extracted numerical features as an example, presenting the model's correct/incorrect predictions for each level for the specific features.

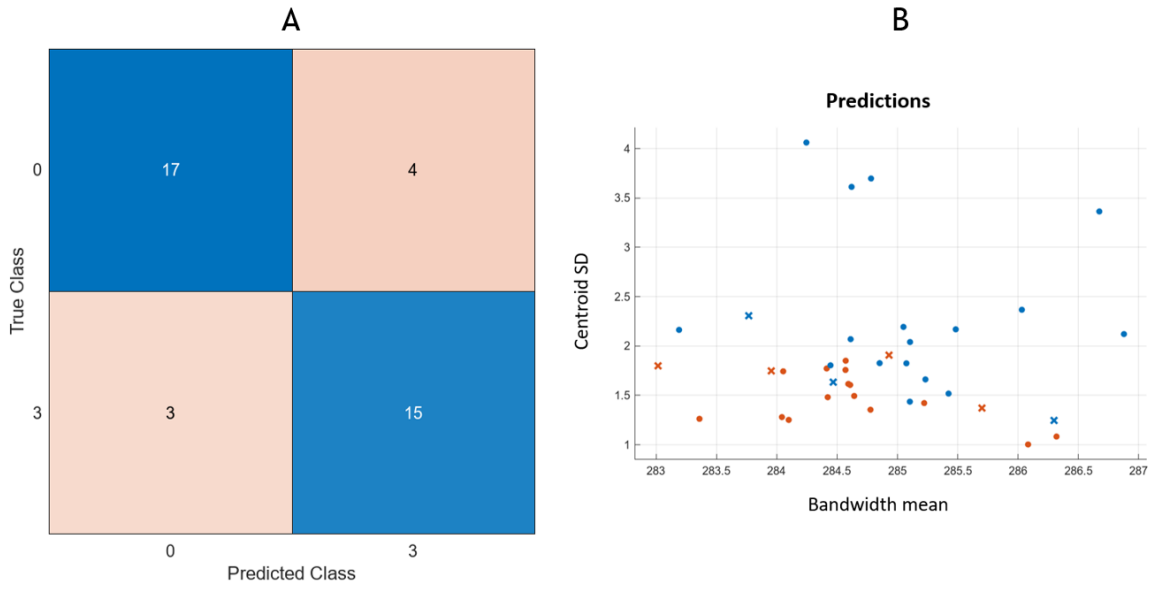


Figure 6.5 (A) confusion matrix of the SVM model when 4 features were used for classification. The used labels were scores 0 and scores 3 from the AHDB 4-level mobility system. The model's accuracy was 82.1%. On the right (B) is the scatterplot of the model's predictions with two selected features (Centroid SD and Bandwidth mean) as examples. The data point colours represent the two classes (red for score 3 and blue for score 0), and the 'x' marks represent the misclassifications.

6.3.2 Study B

Table 6.1 and Table 6.2 show the distributions of the three assessors' AHDB and binarised AHDB scores after scoring the animals from the rear vantage point. In the binary system, animals were more evenly divided into the two levels.

Table 6.1 Score distribution of AHDB scores of the three assessors from the rear vantage point evaluation.

AHDB - Rear	Score 0	Score 1	Score 2	Score 3
Assessor 1	18	8	15	3
Assessor 2	14	14	15	1
Assessor 3	4	21	18	1

Table 6.2 Score distribution of binarised AHDB scores of the three assessors from the rear vantage point evaluation.

Binarised AHDB - Rear	Score 0	Score 1
Assessor 1	26	18
Assessor 2	28	16
Assessor 3	25	19

Pairwise inter-assessor agreement when scoring animal mobility from the rear ranged from 0.21 to 0.43 for the AHDB 4-level system and from 0.33 to 0.57 for the binarised AHDB system. According to (Landis & Koch, 1977), these kappa scores translate as fair to moderate strength of agreement. Both kappa and percentage values are listed in Table *6.3*.

Table *6.4* shows the results of estimating the models' accuracy, sensitivity, and specificity using each assessor's binary-transformed-AHDB scores as labels from the evaluation of the rear vantage point.

Table 6.3 Agreement results of the pairwise comparisons with kappa statistics and percentage agreement.

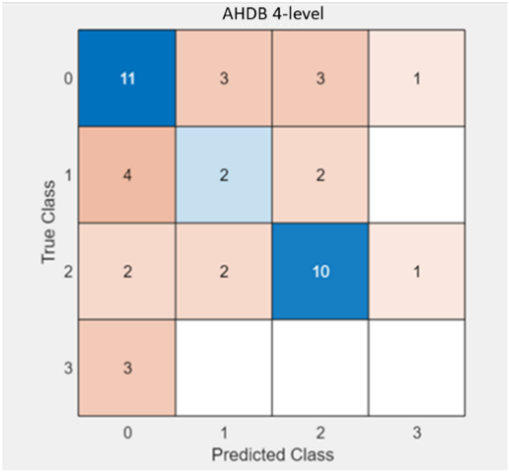
	AHDB		Binarised AHDB	
	Cohen's kappa	% Agreement	Cohen's kappa	% Agreement
Assessors 1-2	0.214	45.5	0.33	68.2
Assessors 1-3	0.258	45.5	0.487	75
Assessors 2-3	0.431	61.4	0.575	79.5
Average	0.30	50.8	0.46	74.23

Table 6.4 Estimation of Decision Tree models' specificity, sensitivity, and accuracy with twenty extracted numerical features from Visit 5 of the longitudinal study and each assessor's labels from the Binarised AHDB rear assessment.

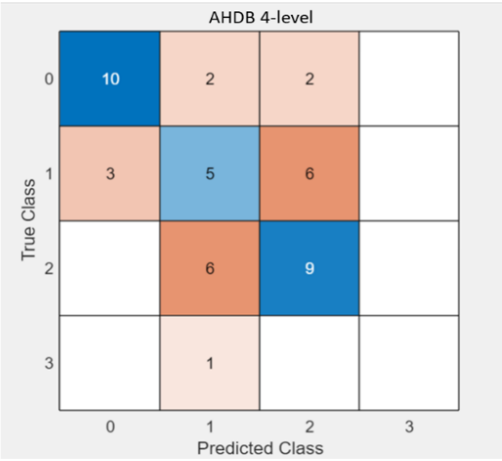
Binarised AHDB - REAR	Assessor 1	Assessor 2	Assessor 3
True positives	25	18	18
True negatives	14	7	11
False positives	1	10	7
False negatives	4	9	8
Specificity	0.93	0.7	0.61
Sensitivity	0.96	0.41	0.69
Accuracy	0.89	0.57	0.66

The produced outcomes of the 4-level scores from the individual assessments used as labels are presented in Figure 6.6. All three assessors scored only a small number of cows with a score of 3, and none of the severely lame cows was correctly predicted by the machine learning model. Assessor 3 appeared to achieve higher accuracy for each class than the other assessors.

Assessor 1	Class 0	Class 1	Class 2	Class 3
True positives	11	2	10	0
True negatives	17	31	24	39
False positives	9	5	5	2
False negatives	7	6	5	3
Sensitivity	0.61	0.25	0.67	0
Specificity	0.65	0.86	0.83	0.95
Accuracy	0.64	0.75	0.77	0.89



Assessor 2	Class 0	Class 1	Class 2	Class 3
True positives	10	5	9	0
True negatives	27	21	21	43
False positives	3	9	8	0
False negatives	4	9	6	1
Sensitivity	0.72	0.36	0.6	0
Specificity	0.9	0.7	0.72	1
Accuracy	0.84	0.59	0.68	0.98



Assessor 3	Class 0	Class 1	Class 2	Class 3
True positives	1	14	12	0
True negatives	37	17	19	42
False positives	3	6	7	1
False negatives	3	7	6	1
Sensitivity	0.25	0.7	0.67	0
Specificity	0.93	0.74	0.73	0.98
Accuracy	0.86	0.7	0.70	0.95

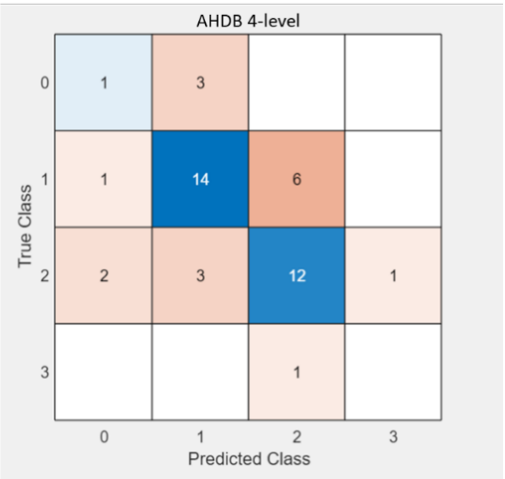


Figure 6.6 On the left side of the page are presented the tables with the estimations of accuracy, sensitivity and accuracy of the models' classes when the AHDB 4-level score of each assessor was used. On the right of the page, the confusion matrices of the models for each assessor are shown respectively. The scores/labels were derived from the rear-side evaluation of cows, and 20 numerical extracted features were used for the classification.

6.3.3 Study C

When we used as labels the scores from the 4-level system of a single assessor who evaluated the cows from the rear vantage point, the accuracy of the SVM model was 52.3%. After the binary conversion of the scores, the accuracy results were 86.4% using a decision tree classifier and four numerical features (Bandwidth mean & SD, Centroid mean & SD) for the predictions. Machine learning correctly classified most animals according to the labels (Figure *6.7*).

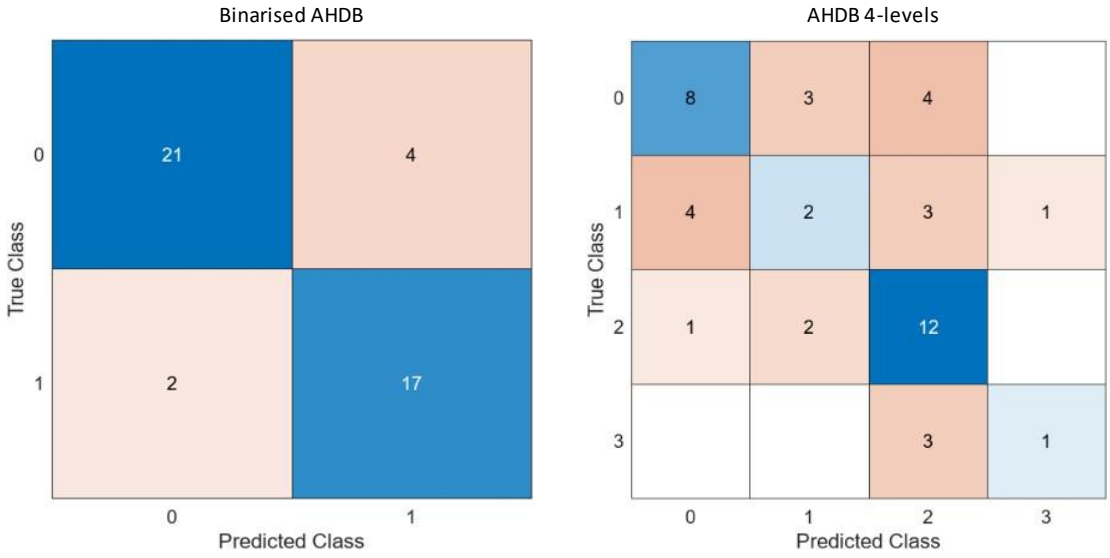


Figure 6.7 Confusion matrices from the models' predictions with the 2- and 4-level systems used as labels. On the left, a Tree model produced 86.4% accuracy with the binary labels; on the right, an SVM model had 52.3% accuracy with the 4-level labels.

6.4 Discussion

The aim of this chapter was to optimise the parameters of the lameness detection automation process and enhance the accuracy of machine learning predictions. We initially examined the algorithms and extracted features using highly reliable labels to identify whether the problem was attributed to the machine learning analysis or the ground truth. Based on the results of study A, we concluded that the algorithmic model and features functioned satisfactorily and were consistent with the findings of other studies (Busin et al., 2019; Shrestha et al., 2018). Hence, we hypothesised that the inadequate performance in previous chapters was due to the labels used as ground truth. Despite attempts to improve the results using a 4-level and 3-level mobility assessment system and their modifications (binarised systems) in previous chapters, we could not achieve the desired outcomes. Therefore, we chose to alter the vantage point from which the evaluation was conducted, sifting from assessing the cows from their side to the cow's rear. Both studies, B and C, followed the same procedure, evaluating cows' mobility from their rear, utilising either video or live evaluations on the farm, and the results were satisfactory.

It is not uncommon for machine learning models to exhibit differential performance across label categories. A study by Lam et al. (2018) is an example from the medical literature, where image classification algorithms demonstrated high accuracy in discriminating between extreme categories of diabetic retinopathy diagnosis but struggled with intermediate categories. However, the difficulty of distinguishing middle levels is not unique to machine learning. The difficulty has first been observed in assessors who seem to have trouble differentiating intermediate levels (Schlageter-Tello et al., 2014). Since ground truth is based on labels produced this way, the same problem applies to artificial intelligence. In Study A, our classification algorithm performed well when restricted to only two extreme label categories but poorly when all four categories or a binary transformation of the 4-level system was employed. These outcomes could potentially be explained by factors, such as insufficient data and label ambiguity (Domingos, 2012). However, our longitudinal and cross-sectional studies, which involved several hundred samples with the algorithms, failed to produce desired results. While the sample size may seem substantial, it is crucial

to acknowledge the potential concern regarding data sufficiency. We recognise that in the realm of data analysis, larger sample sizes are often desirable for robust conclusions. In our case, the unexpected outcomes prompted us to reevaluate the data collection process and labelling methods. Therefore, our realisation that the crux of the issue lies in the labelling process and not necessarily in the volume of data collected. Our labelling methodology involves a visual observation-based assessment that is susceptible to uncertainty, making it challenging for both the assessors and the ML model to distinguish between different label categories. This is due to factors such as the overlap between animal characteristics across different categories (the visual assessment system falls short in providing exhaustive and exclusive label categories) and variability in how human experts classify animals.

There are ways to improve the performance of a supervised machine learning-based classification algorithm. The previously followed methods were the features-model selection and pre-processing techniques, but we did not receive any improved prediction outcome. Another way of performance enhancement is label refinement as some studies (Jacquin et al., 2019; Lallich et al., 2002) have suggested. In our previous chapters, we explored ways to refine the labelling process, such as using different scoring/labelling systems and modifications of them, gathering experts' opinions, and using other criteria for assigning animals to different levels, but no significant improvements were observed. This led us to studies B and C, where we chose to go back to the way of label acquisition and change the vantage point, we observed the animals. Initially, with study B, we noticed a positive change and improvement in results in terms of classification and accuracy. A good classification and separation were observed between the intermediate classes (labels from scores 1 and 2), but a lack of animals in the extreme classes (scores 0 and 3) implied an unbalanced class distribution. After level reduction and conversion to a binary format, overall improvement was observed in classification and prediction accuracy. Particularly, when the labels provided by Assessor 1 were utilised in our study, the machine learning algorithms displayed a high degree of accuracy, specificity, and sensitivity, with values above 85%. This indicates a strong correlation between the ground truth and the classification patterns generated by the machine learning system. However, when the labels from the other two assessors were employed, the predictions were

superior to other times but not as good as those obtained with assessor 1's labels. These findings align with the inter-assessor agreement evaluation, where it was observed that assessor 1's ratings differed from those of assessors 2 and 3, while the latter two displayed a higher level of agreement in their scores. Further studies are required to evaluate more animals and achieve a balanced distribution of scores at all levels to confirm these results.

In study C, we achieved improved results compared to previous studies by scoring animals from their rear side. Although the 4-level system did not display high accuracy, its performance was better since some animals were correctly classified at all levels. Previously, in Chapter 4, accuracy ranged from 0.58 to 0.6 using the AHDB binarised system from a lateral vantagepoint of scoring. Here in study C, after converting the system into a binary format, the two-level labels generated sufficiently good results with a classification accuracy of 86.4% and a well-distributed set of predictions. The results can be directly compared to other published studies with a similar accuracy value range (Busin et al., 2019; Shrestha et al., 2018). Although there is the limitation of only one assessor evaluating the animals, which was the concern we had at the beginning of the project, the improved results emphasise that the low-performance issue of the system primarily stems from the label acquisition method. Further studies are necessary to reach a definitive conclusion. Nevertheless, both studies, B and C, show a better connection between mobility patterns detected by machine learning with labels generated from a rear evaluation of cows.

While using a classification system based on visual observations as a gold standard is a widely accepted practice, it is essential to continuously evaluate and critique its performance to ensure its validity and reliability. In the case of the AHDB 4-level mobility classification system, several issues have been identified that call into question its effectiveness as a ground truth for machine learning. One major issue is the ambiguity of the classification system, which leads to unclear and inconsistent results. For example, the levels are not well-defined, and overlaps between classes make it difficult for assessors to assign animals accurately to a particular level. Furthermore, there is a lack of clear guidance on performing the mobility assessment, contributing to further confusion and inconsistency among assessors. This ambiguity is problematic because it undermines the accuracy and

reliability of the system, which in turn affects the effectiveness of any machine learning algorithm trained on these labels.

Another problem is the low inter- and intra-rater agreement among assessors using the system. This means that different assessors may assign different levels to the same animal, leading to inconsistent and unreliable data. While some degree of variability in assessments is expected, the level of disagreement among assessors using this system is a cause for concern. This highlights the need for a more standardised approach that can produce consistent and reliable results. The shortcomings of the 4-level mobility classification system have significant implications for animal welfare. Lameness in cattle is a severe concern, as it can lead to pain, discomfort, and decreased mobility, which in turn can affect an animal's overall well-being. Using an unreliable and inconsistent gold standard classification system not only undermines the accuracy of research studies but also hinders efforts to address and alleviate the problem of lameness in livestock. Given these concerns, it is essential to review the gold standard and consider alternative approaches that are more effective in detecting and assessing animal lameness. This could involve a more rigorous and standardised approach to mobility assessment, including more objective measures such as statistical analysis from the evolution of the gait parameters and the use of semi-supervised, self-supervised or unsupervised learning methods.

Considering the limitations mentioned, adopting an alternative method employing a different vantage point for evaluating lameness appeared advantageous. As such, in this chapter, cows were assessed from their rear rather than their lateral, and these evaluations were used as inputs for machine learning training. This alternative approach exhibited promise and appeared to overcome some of the prior limitations by presenting a different perspective that was better aligned with the attempt for automated detection of lameness using micro-Doppler radar technology.

6.5 Conclusions

In conclusion, the hypothesis was tested that evaluating cows in a short time and from the recommended evaluation lateral vantage point may not be sufficient for

accurate predictions from a supervised machine learning model. The study results suggest that the algorithms worked well with the extreme – clearly defined levels, indicating that the problem was in the labels. The studies also revealed that changing the assessment vantage point from lateral to rear improved classification results, indicating that ML algorithms were better linked with the data. Overall, the findings of this study highlight the importance of carefully considering the quality of the labels used in machine learning. Further research can build on these findings and explore the potential of using alternative ways of assessments to enhance automated lameness detection accuracy in cows.

Chapter 7 General Discussion

7.1 Aims and objectives of the thesis

Lameness is a significant issue that affects the welfare and productivity of cattle and can cause substantial economic losses for stakeholders. A micro-Doppler radar sensing system has been previously explored for automatically identifying lameness in cattle (Busin et al., 2019; Shrestha et al., 2018). The primary objective of this project was to develop, refine, and optimise the proposed system incorporating supervised machine learning techniques to analyse and interpret micro-Doppler signals collected from the animals, enabling it to identify subtle changes in gait and posture associated with lameness. Ultimately, this project aimed to provide a reliable, cost-effective, and non-invasive tool for automating lameness detection.

7.2 Summary of results

In Chapter 2, we aimed to investigate the agreement among assessors in evaluating lameness in dairy cattle to ensure that the algorithm used in previous studies, which relied on a single assessor, was objective. The findings revealed a significant variation in the scores of different assessors, including experienced and trained mobility scorers in the AHDB mobility assessment system. As a countermeasure, various modifications were made to the assessment system to increase the agreement levels, such as merging and binarising levels and using second-given scores. Despite these modifications, high agreement was not achieved. Several other studies have examined the evaluation of mobility systems and the agreement among assessors with varying levels of experience and training (Channon et al., 2009a; Croyle et al., 2018; Dahl-Pedersen et al., 2018; Holzhauer et al., 2005; Katzenberger et al., 2020; Schlageter-Tello et al., 2015a). Most of these studies reported high levels of agreement (kappa values > 0.6), and only Thomsen & Baadsgaard (2006) have found lower levels (kappa values ranged from 0.4 to 0.88), similar to the present study's results. Differences in the study

designs, such as assessors talking to each other about the scores during the evaluation or extensive training of all assessors at the same time, might have played a role in the different outcomes. This chapter highlights the difficulty of obtaining objective scores for lameness evaluation and identifies a problem with the current assessment system, as even experienced assessors had low agreement levels.

Chapters 3 and 4 described a longitudinal study that aimed to improve the inter-assessor agreement levels in evaluating lameness in dairy cattle and use the scores as labels in machine learning training. The study used the same assessment systems as in the previous chapter (AHDB and modifications), adding a hoof physical examination to increase confidence in the decisions. The inter-assessor agreement was poor with the original 4-level system (average kappa for the AHDB system = 0.22, SD = 0.11) but improved (kappa for the convergent binarised AHDB ranged from 0.5 to 0.97) when the modified systems were used. Despite this improvement, applying the high-agreement scores as input labels for the supervised machine-learning classification produced unsatisfactory results (accuracy ranged from 0.57 to 0.63). The performance of the 4-level system was poor, and the reduced-level (binary) systems gave the impression of better results due to the classifier's tendency to classify most animals at the majority level. This means that the majority of the scores matched one of the two levels, resulting in higher estimated accuracy since the classifier accurately predicted most animals within that category. However, this led to a high frequency of false negatives or false positives. This situation resembles the problem of uncertain labels in medical diagnosis studies (Dimitrovski et al., 2015; Hao et al., 2020) and other fields (Bouveyron & Girard, 2009; He et al., 2011; Shin et al., 2018), where a classifier needs to be trained, but the labels include levels of uncertainty or inconsistency rendering unreliable predictions. Finally, there was no association between the scores and evidence of hoof pathology or attributes of the assessors and their agreement levels, results consistent with other studies (Flower & Weary, 2006; Logue et al., 1994; Tadich et al., 2010b).

In Chapter 5, we attempted to improve the classification predictions of machine learning by employing a different labelling system and a pre-processing technique. Although these approaches had shown promising results in other studies (Li et al., 2023; Lorenzini et al., 2017), our results were unsatisfactory. The scoring system

did not yield high agreement among assessors (average kappa = 0.37, SD = 0.16) when statistically analysed, and the pre-processing technique did not produce statistically significant results compared to the non-pre-processed data. While our study's machine learning overall accuracy was high in some cases (i.e., accuracy = 93.5 binarised Grimm & Lorenzini scores by Assessor 1), we encountered the same issue as the previous chapter: correctly classifying cows at the majority level but poor performance at the other levels. We considered that the method of obtaining labels, through the way of assessment instead of the scoring system (i.e., assessment via short video clips from the animals' side instead of the AHDB/Grim & Lorenzini scoring system), was the root cause of this issue and we continued the investigation through three studies described in Chapter 6.

In Chapter 6, we initially verified that the algorithm performed well (accuracy = 0.82, specificity = 0.79, sensitivity = 0.85) with high certainty labels, i.e., scores 0 and scores 3 derived from the 4-level AHDB system, using data collected previously in Chapter 3. Then, we conducted two studies to evaluate the mobility of animals using a different vantage point by assessing cows from their rear, which corresponded with the vantage point for radar signal collection. With this change, we observed improved classification results (average accuracy = 0.7 & SD = 0.17, average sensitivity = 0.68 & SD = 0.27, average specificity = 0.75 & SD 0.17) and a better link of the labels with the data. The results were promising, suggesting that assessing animals from their rear can improve the classification predictions of machine learning algorithms with micro-Doppler radar data.

7.3 Visual Assessment and Micro-Doppler Radar for Lameness Detection

Lameness detection is a crucial aspect of animal welfare, and visual assessment has been the most common method used by farmers and veterinarians. During a visual examination, observers typically evaluate the animals' movements while walking, paying close attention to limb placement, stride length, arching of the back, and gait symmetry, depending on the scoring system. Mobility scoring systems, such as the AHDB dairy mobility and Grimm and Lorenzini systems we used in this project, are often utilised to provide a standardised approach to

lameness detection in herds. However, the requirement for a human observer to decide on the state of the animal by observation makes these systems rather subjective, potentially influenced by the observer's biases, mood, skills and experience. Additionally, visual assessments require considerable time and resources, particularly in large herds, which can limit their practicality and can be logistically difficult. In recent years, there has been increasing interest in using objective measures, such as accelerometers, pressure plates, and computer vision systems, to supplement or replace subjective visual assessments. These objective measures offer the potential for more consistent and reliable evaluations of lameness, but their widespread adoption in practical settings remains limited due to various challenges, including cost and practicality.

Our suggestion, Micro-Doppler radar, which measures the Doppler shift in reflected signals, has the potential to provide objective and automated lameness detection and be cost-effective and environmentally sustainable. The system requires only one-time installation costs and the standard ongoing expenses (i.e., electricity, server storage), usually covered by the manufacturer subscriptions, making the radar system a potentially financially feasible solution for long-term lameness detection. Using the non-invasive radar system could also help reduce the farm's environmental impact by avoiding the need for additional resources and materials for wearable devices. The system has the potential to monitor cows daily without disrupting the regular farm routine. And finally, unlike other methods, such as systems that use cameras, the micro-Doppler radar is not affected by weather conditions or other factors that may impact visibility, thus providing consistent and reliable animal health monitoring. If trained well, machine learning algorithms can analyse radar-generated data and identify abnormal gait patterns with high accuracy, providing quick, consistent, and reliable assessments of lameness, which can help improve animal welfare and reduce production losses. However, this technology is still developing, as in this project, and there are still challenges to overcome, including the need for accurate labels for training and validation.

7.4 Labels and machine-learning classification

Using machine learning algorithms for mobility classification in dairy cattle has previously shown promising results (Busin et al., 2019; Shrestha et al., 2018). This project employed supervised machine learning approaches, utilising feature extraction and pre-processing techniques to train and classify mobility patterns. Models such as support vector machines (SVM), K-nearest neighbours (KNN), and decision trees were used for classification throughout the project. These models were trained and validated on data obtained from micro-Doppler radar systems, which were used to detect and track cattle movement patterns from their radar signatures. However, initially, incorporating veterinary knowledge and state-of-the-art lameness detection techniques as labels in the machine learning model did not yield promising results. In fact, these inputs resulted in poor prediction and classification accuracy. Similar problems concerning the use of labels with uncertainty and machine learning are described in the studies conducted by Algan & Ulusoy (2019) and Shi & Wu (2021). In their systematic review of methodologies to handle label uncertainty in machine learning (Algan & Ulusoy, 2019), the authors examined various approaches such as active learning, crowd labelling, Bayesian methods, and self-supervised learning. The first three methods involve acquiring labelled data samples through either algorithmic selection or crowdsourcing-based annotation and modelling probability distributions to estimate class probabilities. The self-supervised learning approach enables learning from unlabelled data by designing tasks that require the model to understand the input data and then perform classification tasks. The second paper (Shi & Wu, 2021) proposes a method for training a medical image segmentation model using noisy labels. The authors propose a two-stage approach where a model is trained on clean data and used to transfer the knowledge to a second model trained only on noisy labels. The authors demonstrate that their method outperforms several state-of-the-art methods for training segmentation models with noisy labels on various medical image datasets. Although both studies are about image classification, the task the algorithms are asked to perform is similar to the classification of radar data. Both studies highlight the importance of addressing label uncertainty in machine learning, and the proposed methods could

potentially work well with our data and will be considered for the future continuation of the current project.

In supervised machine learning, accurate and representative labels are crucial to ensure the model's predictions are reliable. To achieve this, labels must accurately reflect the class of the data, capturing essential features while minimizing individual variations and noise. Machine learning algorithms may fail to produce accurate classification results without high-quality ground truthing, leading to potentially biased or unreliable outcomes. Two familiar sources of inaccuracy in ground truth data are lack of standardization and uniformity in label definition and a lack of exclusive and exhaustive classes (Chawla, 2010; Vuttipittayamongkol et al., 2021).

One example highlighting these issues pertains to the scoring criteria for lameness severity in the widely used AHDB dairy mobility system, which was created to address concerns about the welfare of dairy cows and the economic impact of lameness (AHDB 2015, https://ahdb.org.uk/about-ahdb, accessed in March 2023). Specifically, there is ambiguity in the interpretation of score 3, which involves evaluating an individual cow's mobility relative to the rest of the herd. This criterion may be challenging to apply consistently and objectively, as the assessment is made individually for each cow. In this project, we encountered and identified limitations in the subjectivity derived from the visual assessment, leading to discrepancies between and within assessors bringing to attention the lack of standardization and uniformity. Additionally, the system may not be sensitive enough to detect early stages of lameness or subtle changes in gait indicating pain or discomfort in cows. Although the AHDB dairy mobility scoring system offers potential for assessing cow mobility, it has limitations, including discrepancies in labels generated by different assessors. This inconsistency may compromise its reliability as a ground truth for machine learning. However, implementing the system from a rear vantage point could help address these issues and improve its usefulness.

The choice of the extracted features and the validation techniques used can also impact the classification model’s performance, as some studies (Karabulut et al., 2012; Szeghalmy & Fazekas, 2023; Tougui et al., 2021) have demonstrated. These three studies investigated and compared validation and feature selection methods

and suggested that careful consideration of the approaches is essential for ensuring reliable and accurate diagnostic applications in the context of machine learning. The features can be selected manually or automatically, and their quality and relevance to the classification task are critical determinants of the model's accuracy. If the extracted features are determined not to be relevant to the classification task, the model will not be able to differentiate between classes, resulting in poor classification accuracy. If the extracted features are too complex or too many, the model may overfit the training data and perform poorly on new, unseen data.

In addition to selecting the appropriate features, the choice of validation technique can also affect classification accuracy. For example, suppose we use a simple hold-out validation technique, splitting the data into training and testing sets. In that case, the model's performance may depend on the specific samples in the training and testing sets. If the algorithm is sensitive to the particular training samples, the model may overfit or underfit the training data, resulting in poor performance on the testing set. If a more advanced validation technique is used, such as the 10-fold cross-validation we used throughout the project, we can obtain a more reliable estimate of the model's performance, which is less sensitive to the specific samples in the training and testing sets. This can help us select the appropriate algorithm and its implementation, as we can compare the performance of different algorithms across multiple folds of the data. Having a larger dataset can partially mitigate the problems of overfitting or underfitting, as it provides more diverse examples for the model to learn from, reducing its sensitivity to specific training samples. However, it's important to note that increasing the dataset size does not solve the issue of incorrect or inaccurate labels. If the training data contains labels not corresponding to the truth, adding more data will not fix this problem. In such cases, it is necessary to address the quality and accuracy of the labels.

The subjective nature of the visual scoring and the importance of objective labels for machine learning training have practical implications for the livestock farming industry. The current project highlights the need for alternative systems or a re-assessment of the current evaluation criteria. Fortunately, we also found an improvement when changing the vantage point of assessment, suggesting that scorings conducted from a rear vantage point can provide practical advantages,

link better with the micro-Doppler data, and improve the accuracy of lameness detection.

7.5 Impact

One of the impacts of this research is the generation of an extensive data set of cattle mobility radar data from different farm facilities. The data collected from the longitudinal study described in Chapters 3 and 4 concern data of the same animals we monitored for six months. We plan to make these datasets or part of them publicly available, allowing other researchers to access and facilitate new research questions, enable meta-analyses, and possibly provide a long-term impact in the lameness field.

This project has also allowed an interdisciplinary collaboration between University of Glasgow departments: the School of Biodiversity, One Health and Comparative Medicine and the James Watt School of Engineering. Combining these disciplines led to the developing of new and innovative ideas about automating lameness detection and supported addressing complex and multidimensional problems such as transferring human knowledge and the decision-making process of diagnosis to machine learning.

Finally, the research discussed in this thesis has helped to increase our understanding of how radar can recognise animal activity, particularly in identifying signs of lameness. This has involved addressing challenges discussed in previous chapters and developing new methods to overcome them. Overall, this project has expanded our knowledge of the potential uses of radar technology for animal monitoring and introduced the idea of automated lameness detection using micro-Doppler radar.

7.6 Challenges faced during the project

One of the main challenges was the unexpected variability among the assessors' scores, which resulted in the need to focus on creating accurate labels. This highlights the importance of developing clear and consistent labelling protocols to

ensure the accuracy and reliability of data used for machine learning. While this was an important step to make the system more rigorous and reliable, it also shifted the focus away from other aspects of the study, potentially limiting the scope of the results.

Another limitation was the Covid pandemic, which appeared a few months after the project started and lasted most of the duration. Although difficult to quantify, the impact on mental health resulting from the pandemic has undoubtedly been a significant challenge. While the impact on the practical aspect of the project was indirect (i.e., video scoring instead of live scoring), adapting to new circumstances may have introduced additional variability or limitations in the data. This highlights the importance of adapting to unexpected circumstances and implementing appropriate measures to maintain the integrity of the data.

7.7 Lessons Learned: Reflections on Opportunities for Improvement

Upon reflection on the choices and results obtained in the project, there were identified areas for improvement if the study were to be repeated. Firstly, a multi-angle approach for capturing videos of the animals would be implemented to provide a more comprehensive perspective of the cows' gaits and movements. It could help identify asymmetries that may not be apparent from a single angle, and it would help reduce environmental factors' impact on gait assessment, such as uneven ground or lighting conditions, which may affect the visibility of lameness from a single angle. This would potentially improve the accuracy and reliability of visual lameness scoring by minimising observer bias, as different angles may highlight particular features or characteristics of the animal's gait.

Including a more extensive sample of undoubtedly healthy and severely lame animals would also be implemented. This choice would enable the ML model to be sufficiently trained with the necessary data and subsequently tested with cases of mild to moderate lameness (i.e., AHDB scores between 1 and 2 on the 4-level system). Training the model on a broader range of data, possibly including lesion

data, could improve the generalisability of the model to new lameness cases and enhance its usefulness and practical application in diverse farm environments.

Finally, alternative machine learning methods, such as unsupervised or semi-supervised learning, would be investigated as part of potential improvements to the study. The main advantage of the unsupervised ML approach is that it does not require labelled data for the training and can be used to identify patterns or structures that might not be easily identified through manual feature extraction and visual assessment methods.

7.8 Future Directions

The research project has received funding and a grant, allowing the investigator and the supervisory team to continue the research. In addition, a collaboration with a European university specialising in artificial intelligence and biomedical applications for human and veterinary medicine has been established. This collaboration provides access to well-equipped laboratories for experimental and computational work using radar sensing systems, which will be used in the ongoing research project. The two main objectives of the future plan are to improve the accuracy and reliability of the lameness detection system. First, applying an unsupervised machine learning technique to compare the outcomes with this thesis's findings, aiming to determine which labels (assessors or hoof examinations) are more reliable. This would help reduce the variability in the labelling process and increase the system's accuracy. Secondly, we will attempt to define a gait signature for each cow, and the gait parameters will be extracted to monitor changes over time. The researchers will explore how these variations are associated with abnormal mobility, which will improve the understanding of the early stages of lameness in cows. By developing a gait signature for each cow, the researchers can track changes in their mobility over time, allowing for early detection of lameness before it becomes visually apparent. This will enable prompt intervention and treatment, improving animal welfare and reducing the financial costs for farmers.

7.9 Conclusions

Lameness assessment is a crucial task in managing the welfare and productivity of dairy cows, but the current visual assessment methods, which typically involve scoring the animals from the side using a multi-level system, have been shown to lack consistency among and within assessors. Our study found that this inconsistency is extended in machine learning predictions when the scores are used in training, despite attempts to improve agreement through binarisation, transformations of the data (convergent scores), and pre-processing techniques. However, changing the vantage point from the side to the rear of the cow for assessment showed promising results regarding machine learning training and classification. Moreover, we observed that using labels from animals with indisputable lameness status, such as healthy or severely lame, improved machine learning classification. Nevertheless, our results suggest that further research is needed to explore the effectiveness of the suggested evaluation vantage point and to involve a larger number of healthy and severely lame animals. Future research should explore the suggested approaches further.

Appendix A

Receiver operating characteristic curves (ROC) from the SVM models validation corresponding to Table 5.6. Each figure includes three plots (A1, A2, A3) representing each assessor's (Assessor 1, 2, 3) scores. The different mobility scoring systems used as labels are shown on each page. A perfect ROC curve would be one that reaches the top-left corner of the plot, indicating a sensitivity of 1 (no false negatives) and a specificity of 1 (no false positives). The Area Under the Curve (AUC) is a quantitative measure of the overall performance of a classification model. A perfect model would have an AUC of 1.0. The closer the AUC is to 1.0, the better the model's ability to discriminate between positive and negative instances.

Figure A.17 to Figure A.25 correspond to the results presented in Table *5.5*.

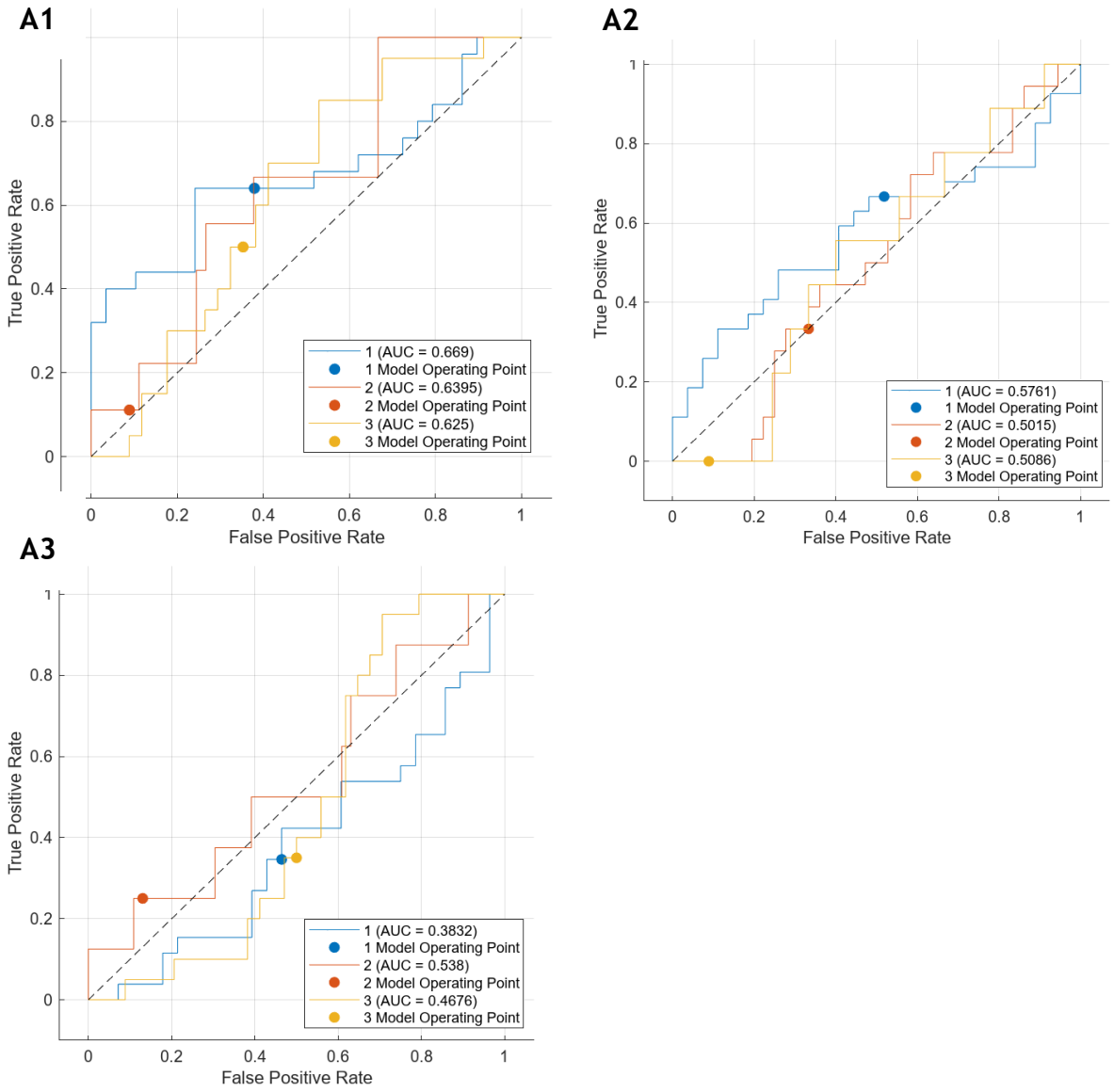


Figure A.1 Farm B – Grimm & Lorenzini scores- extracted features without the pre-processing technique (unmasked).

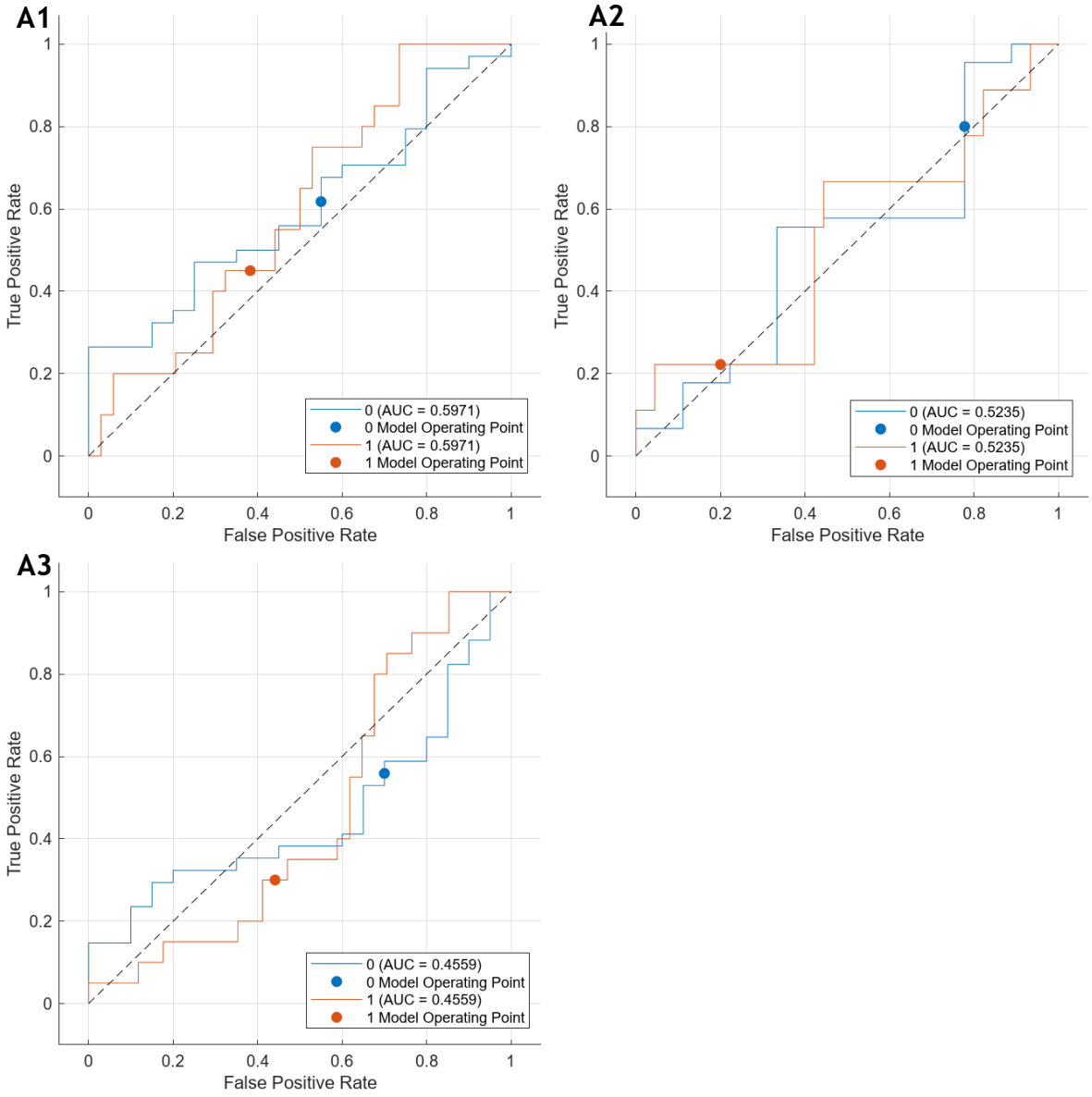


Figure A.2 Farm B - Binarised Grimm & Lorenzini scores (scores 0 & 1 / scores 3) - extracted features without the pre-processing technique (unmasked).

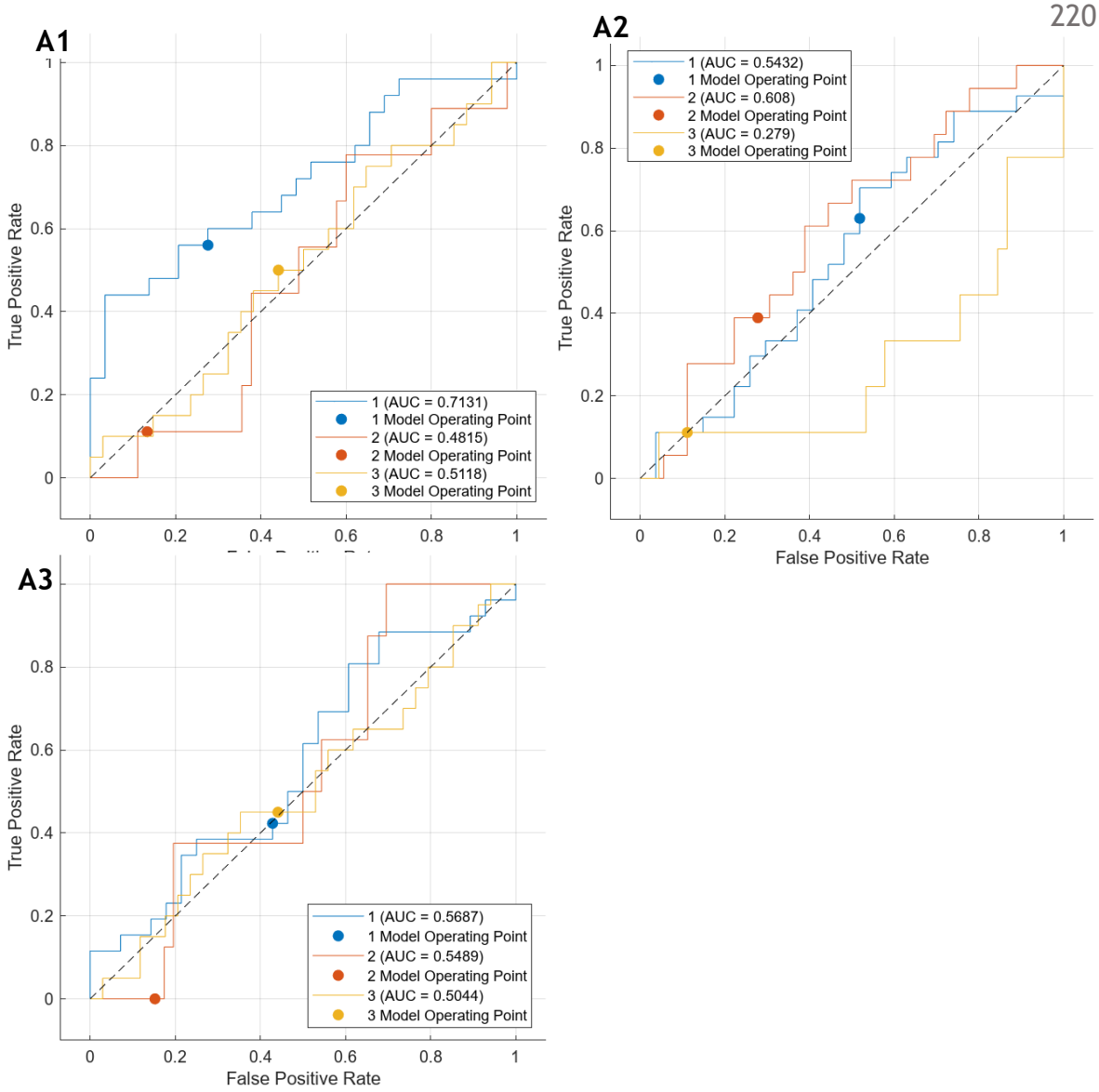


Figure A.3 Farm B - Grimm & Lorenzini scores- extracted features after the pre-processing technique (masked).

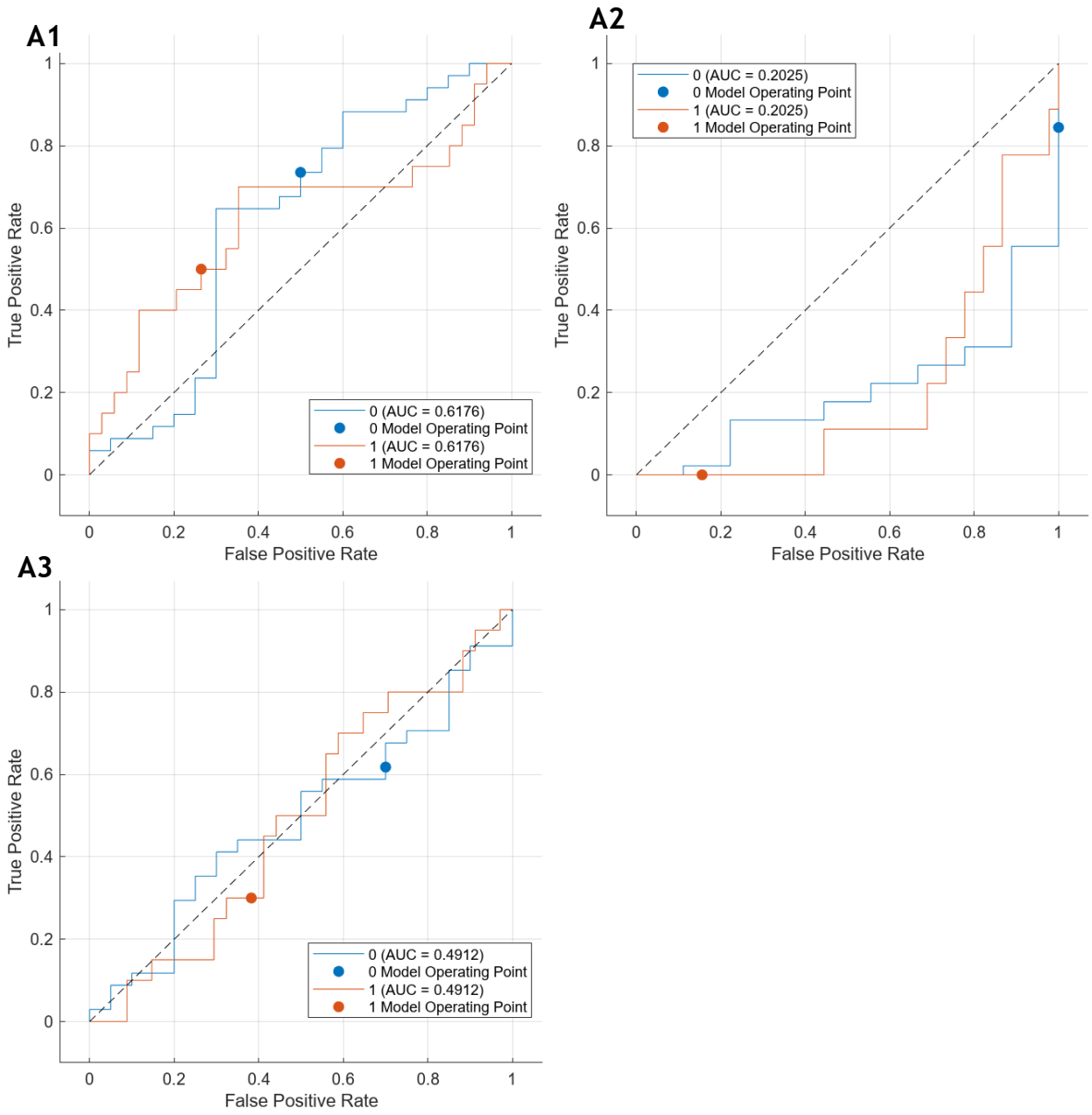


Figure A.4 Farm B – Binarised Grimm & Lorenzini scores- extracted features after the pre-processing technique (masked).

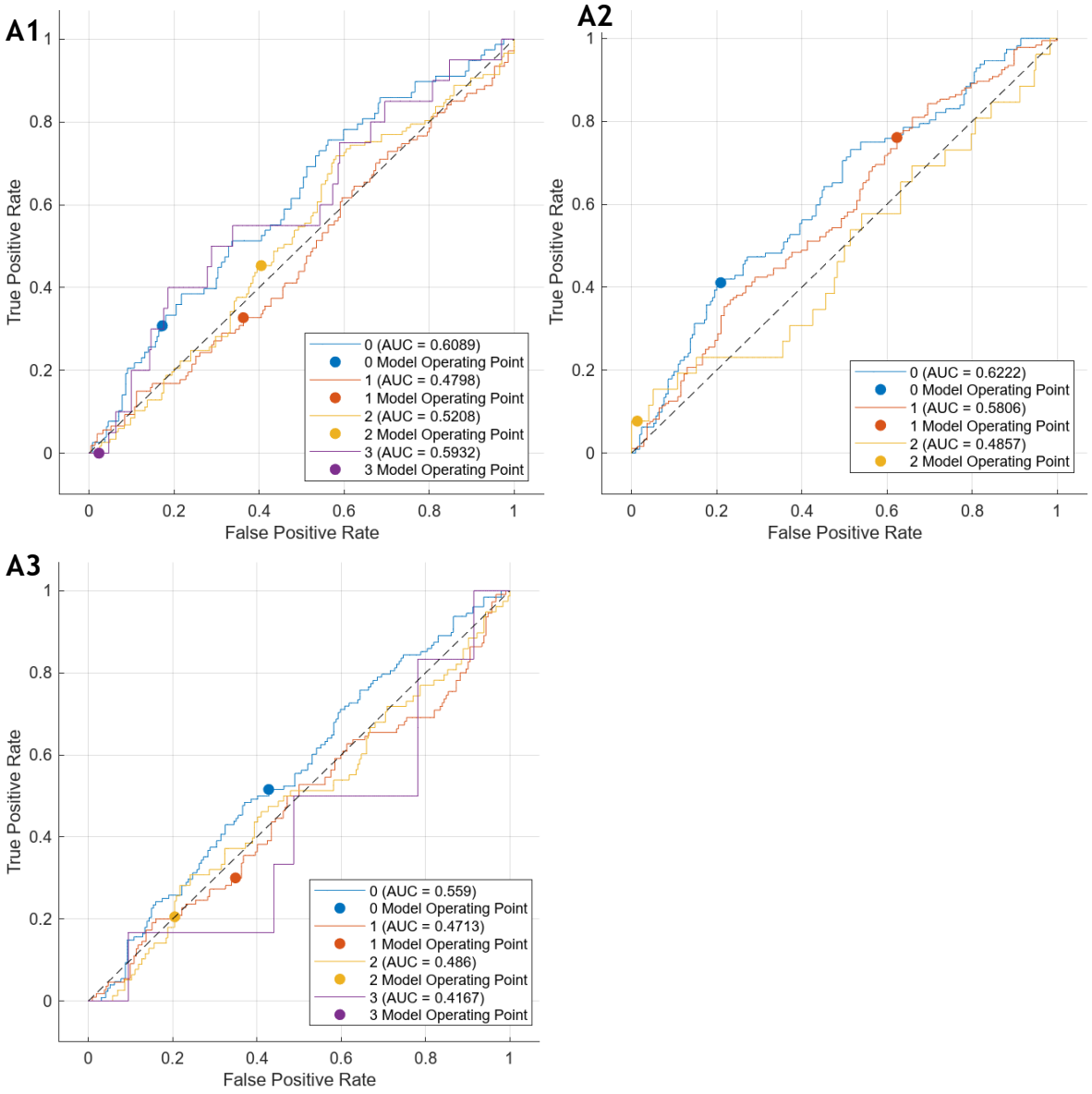


Figure A.5 Farm A - AHDB (4-levels) scores- extracted features without the pre-processing technique (unmasked).

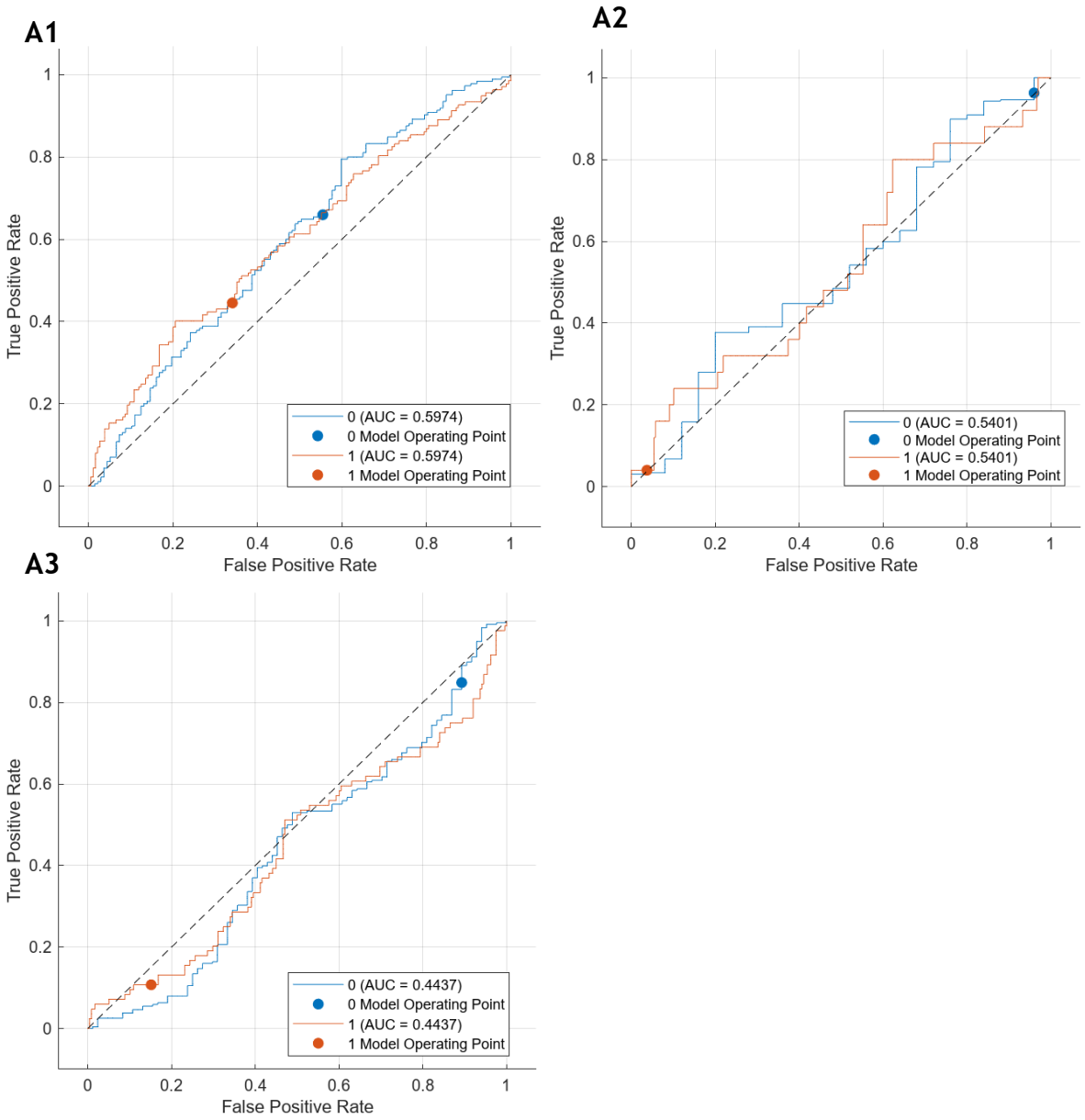


Figure A.6 Farm A - Binarised AHDB (2-levels) scores- extracted features without the pre-processing technique (unmasked).

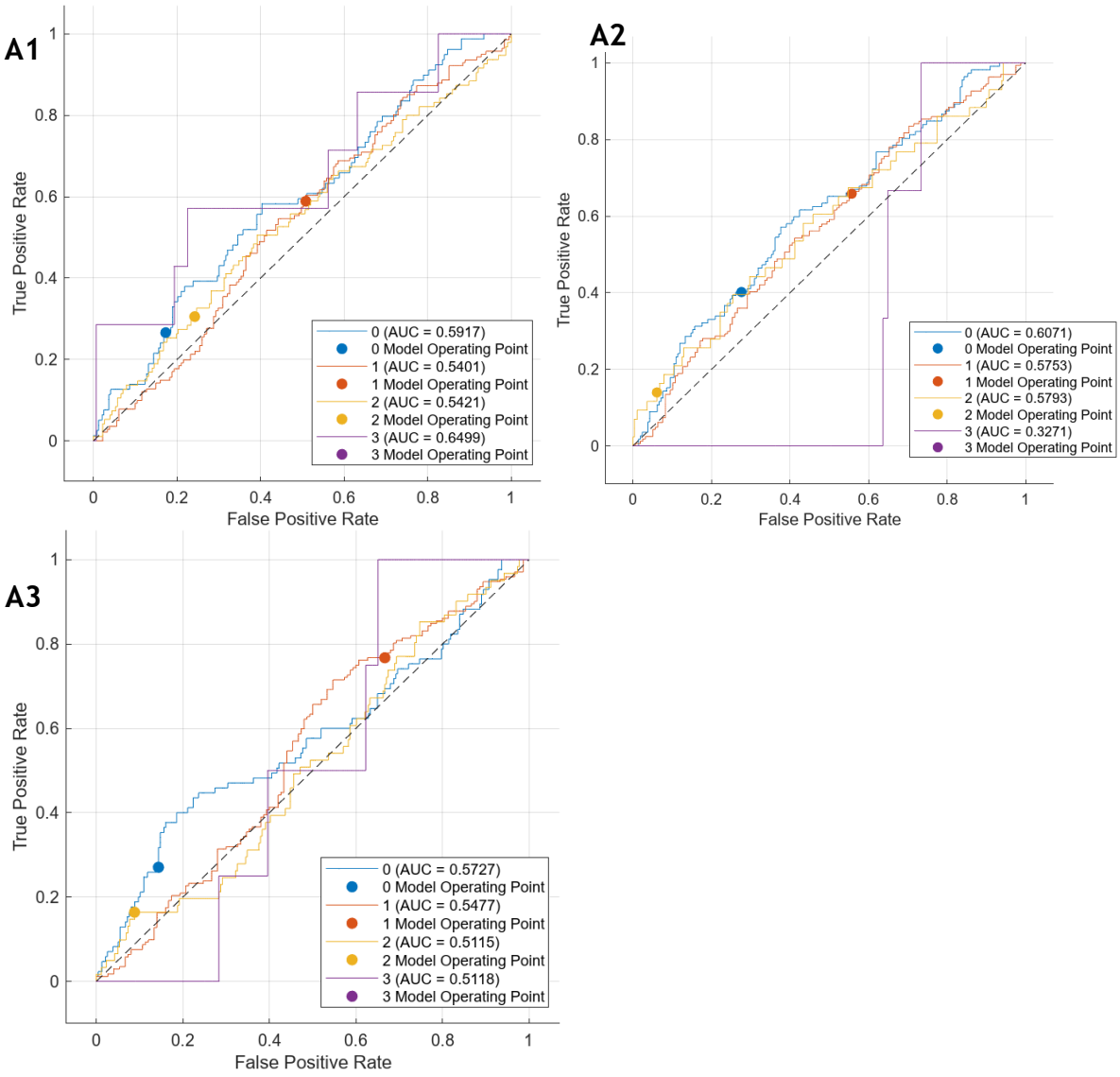


Figure A.7 Farm A – Convergent AHDB (4-levels) scores- extracted features without the pre-processing technique (unmasked).

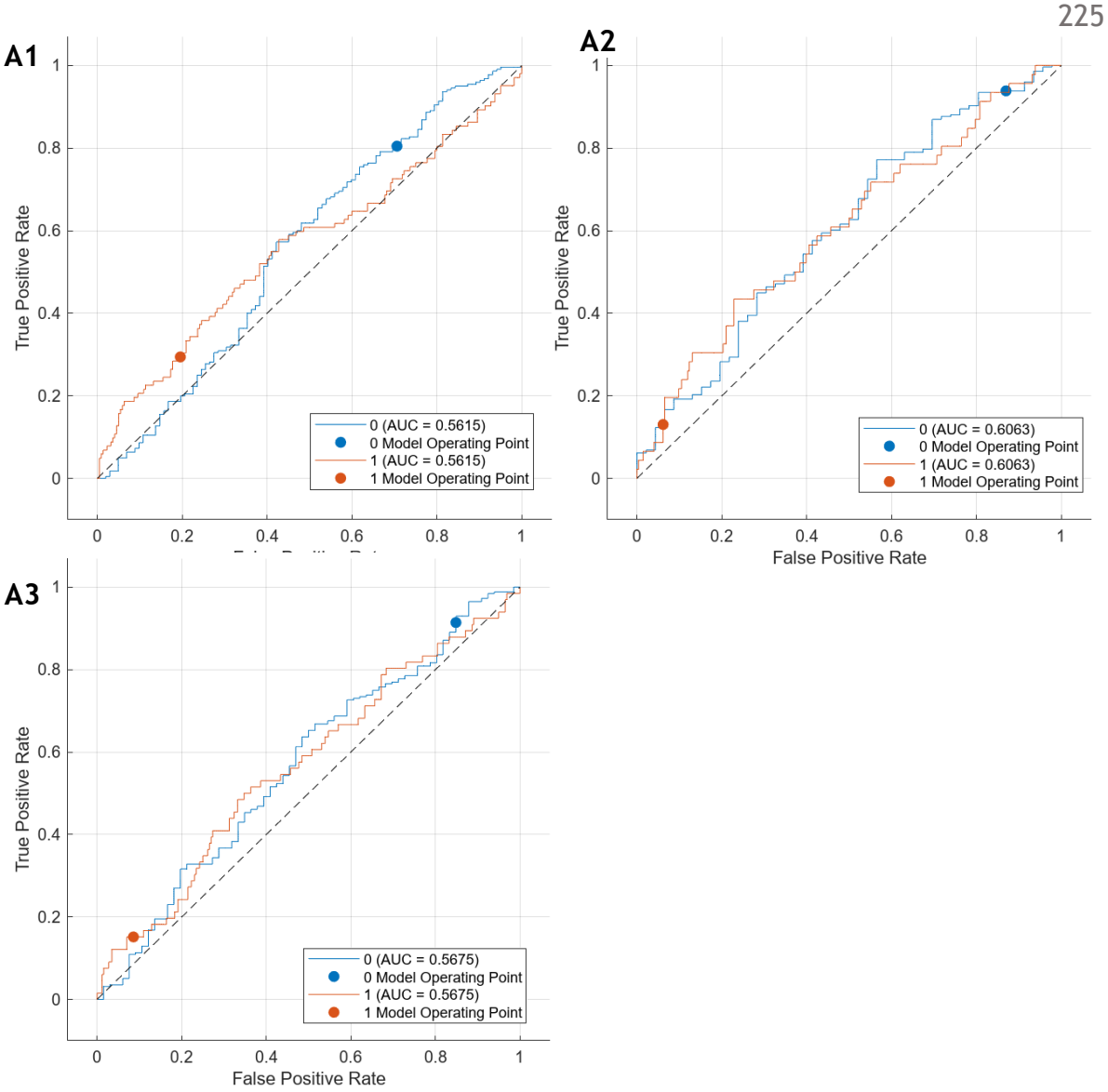


Figure A.8 Farm A – Binarised convergent AHDB (2-levels) scores- extracted features without the pre-processing technique (unmasked).

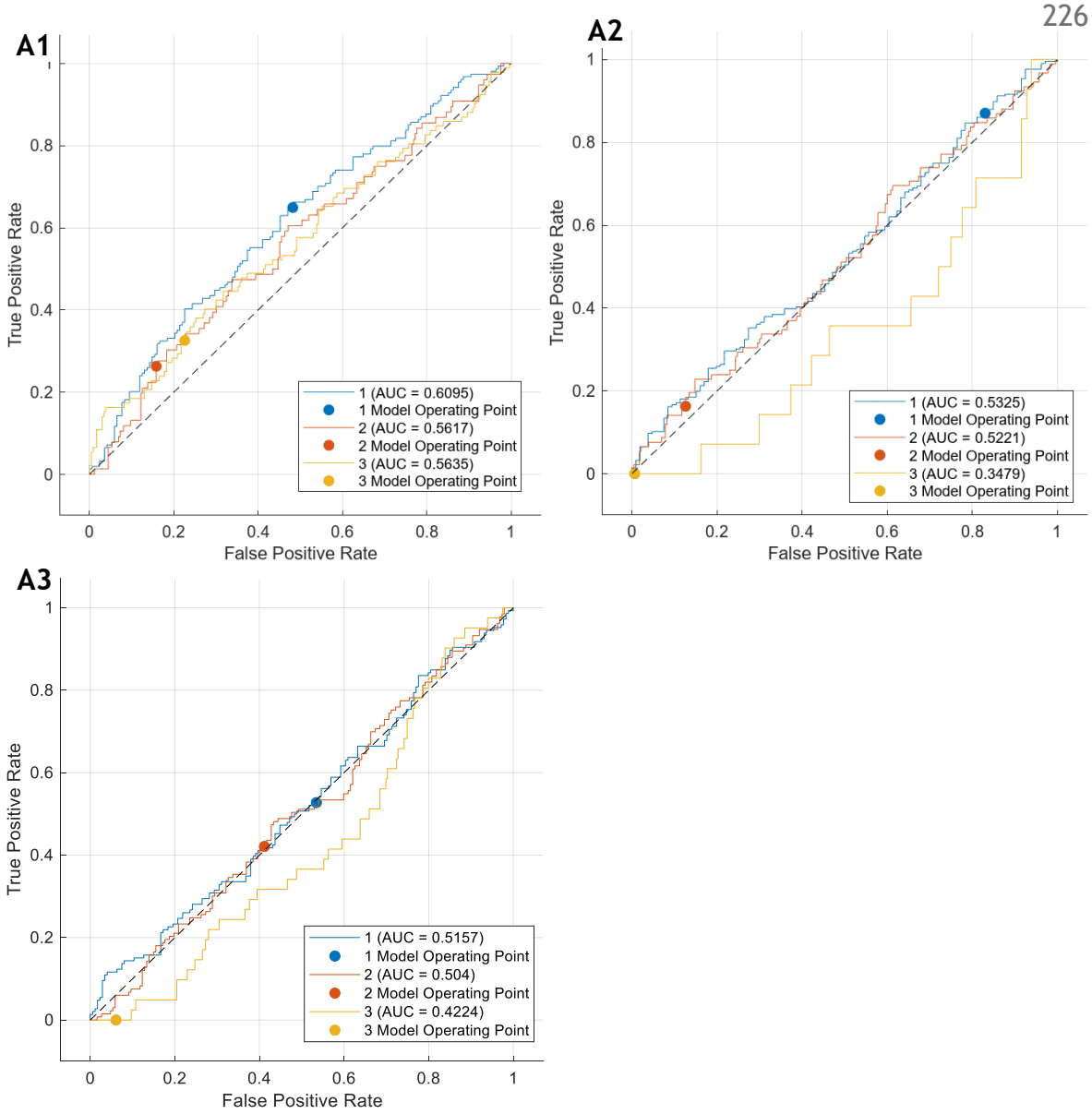


Figure A 9 Farm A – Grimm & Lorenzini (3-levels) scores- extracted features without the pre-processing technique (unmasked).

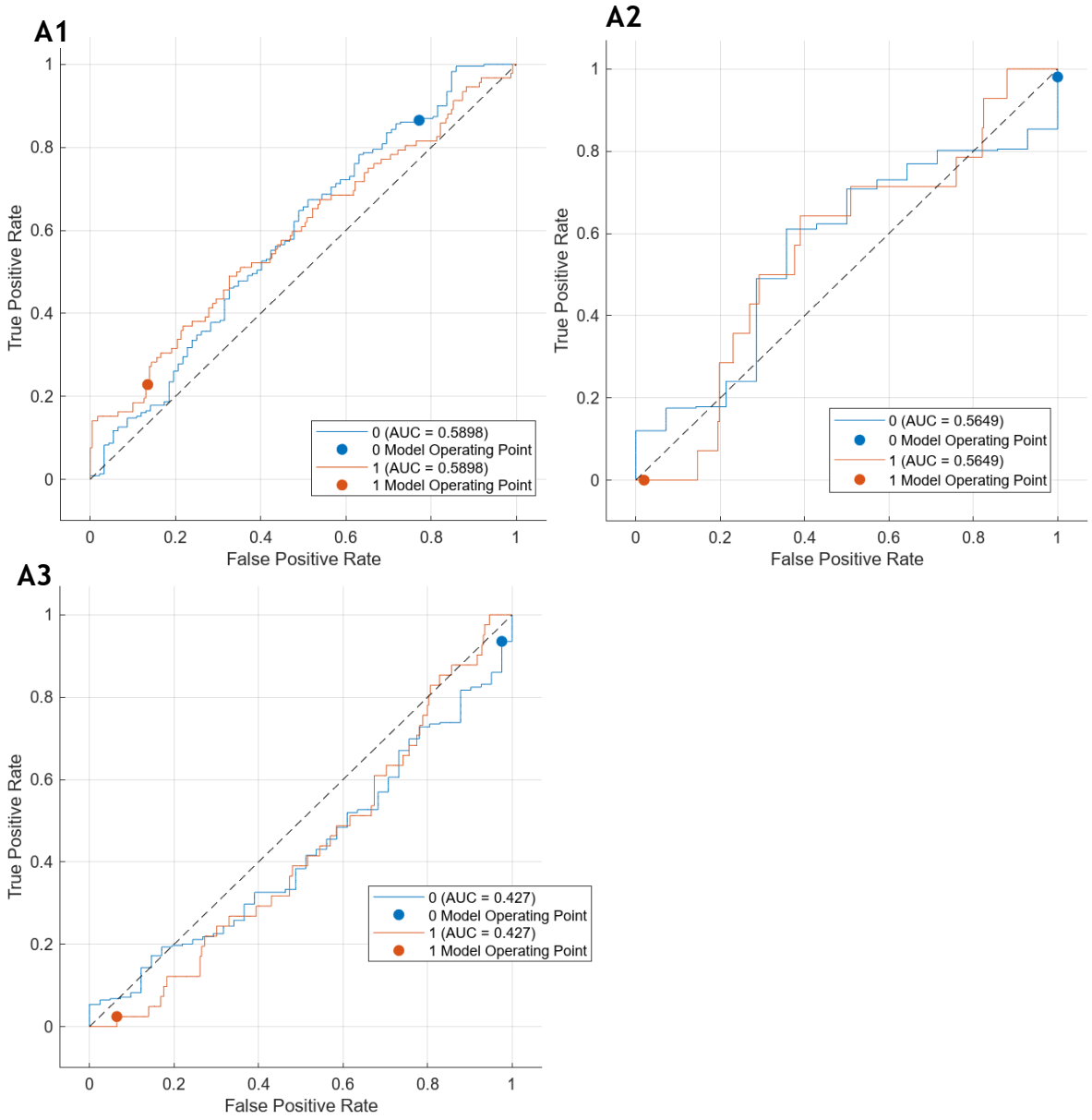


Figure A.10 Farm A – Binarised Grimm & Lorenzini (2-levels) scores- extracted features without the pre-processing technique (unmasked).

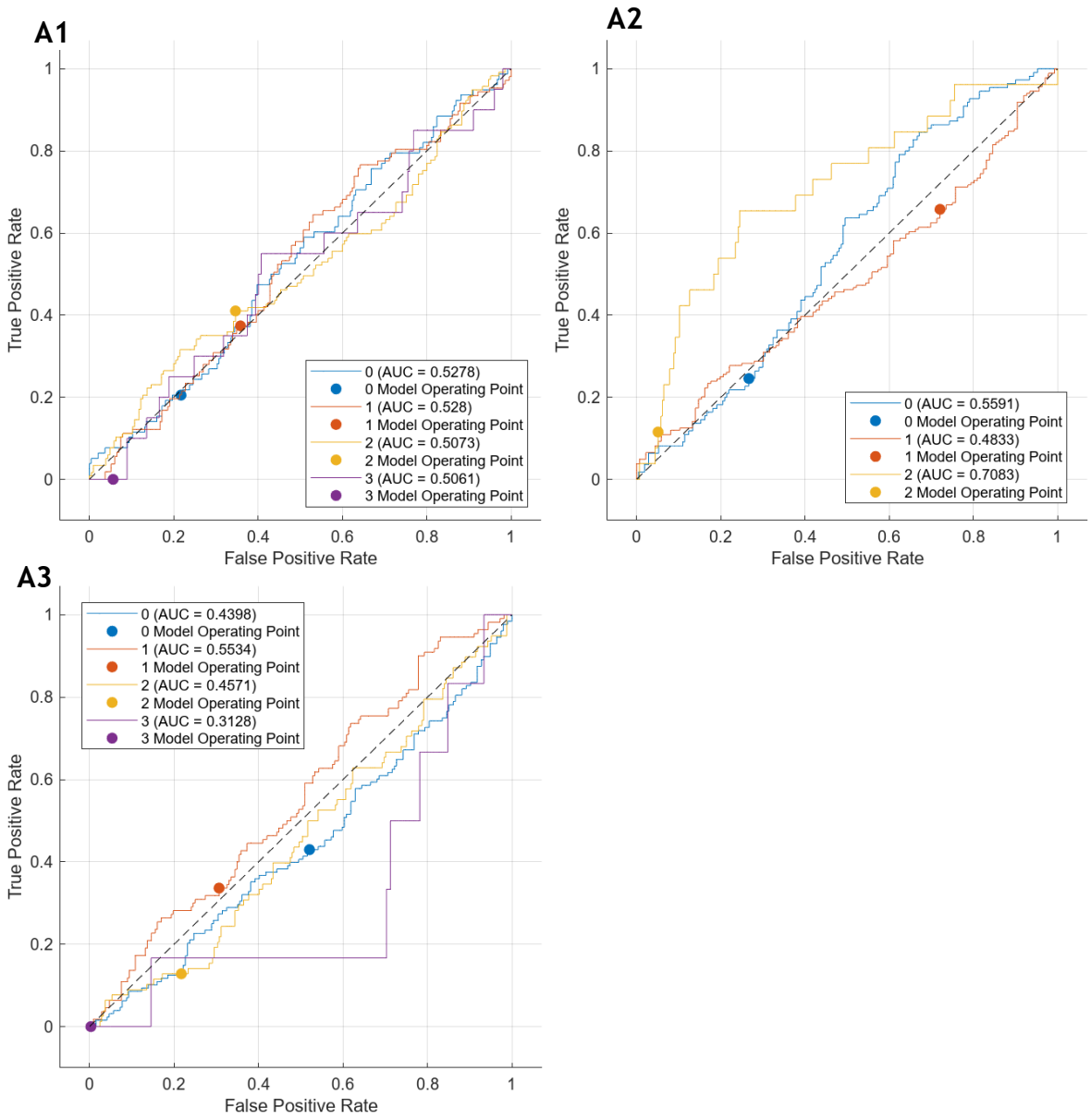


Figure A.11 Farm A – AHDB (4-levels) scores- extracted features after the pre-processing technique (masked).

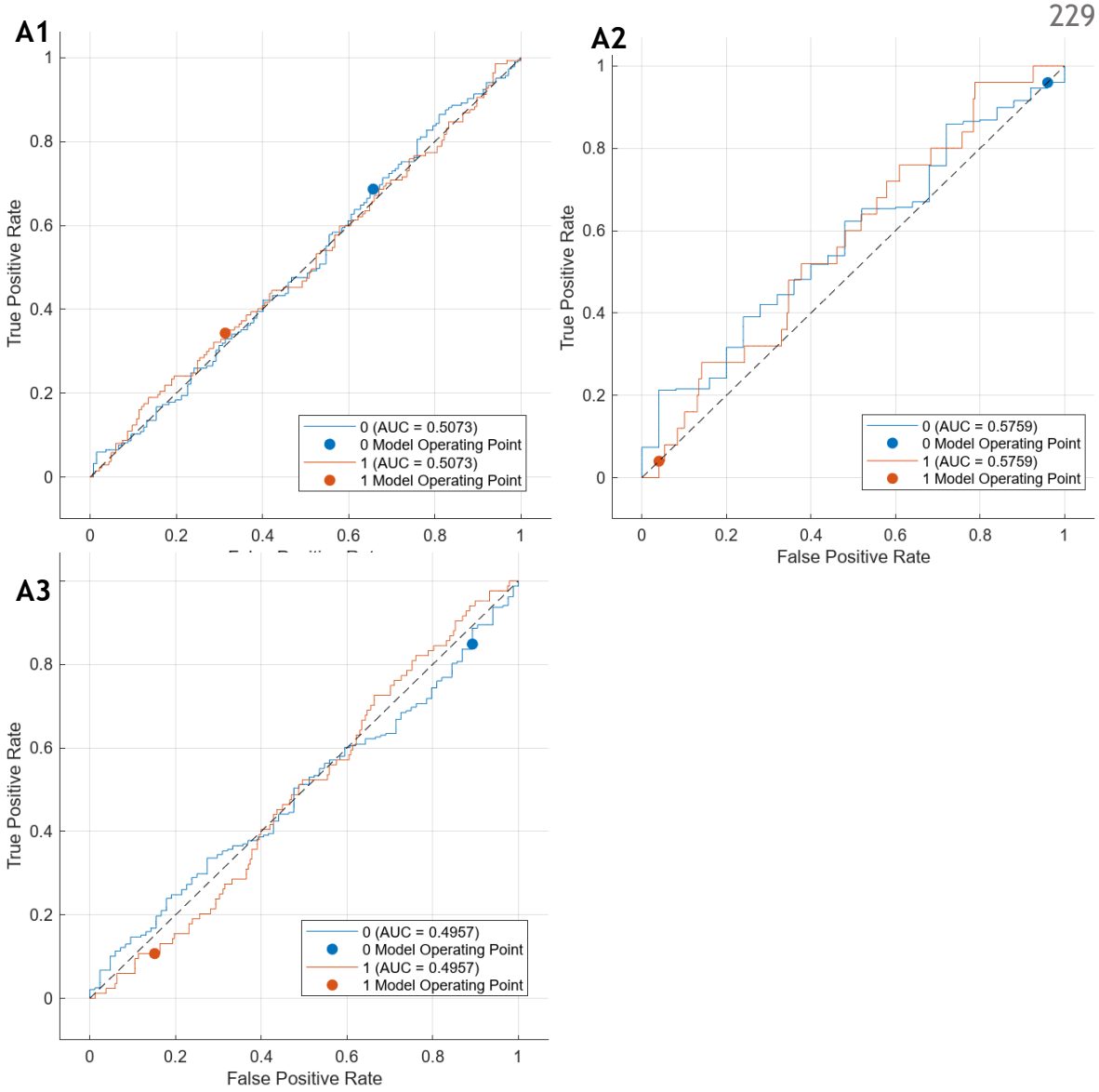


Figure A.12 Farm A - Binarised AHDB (2-levels) scores- extracted features after the pre-processing technique (masked).

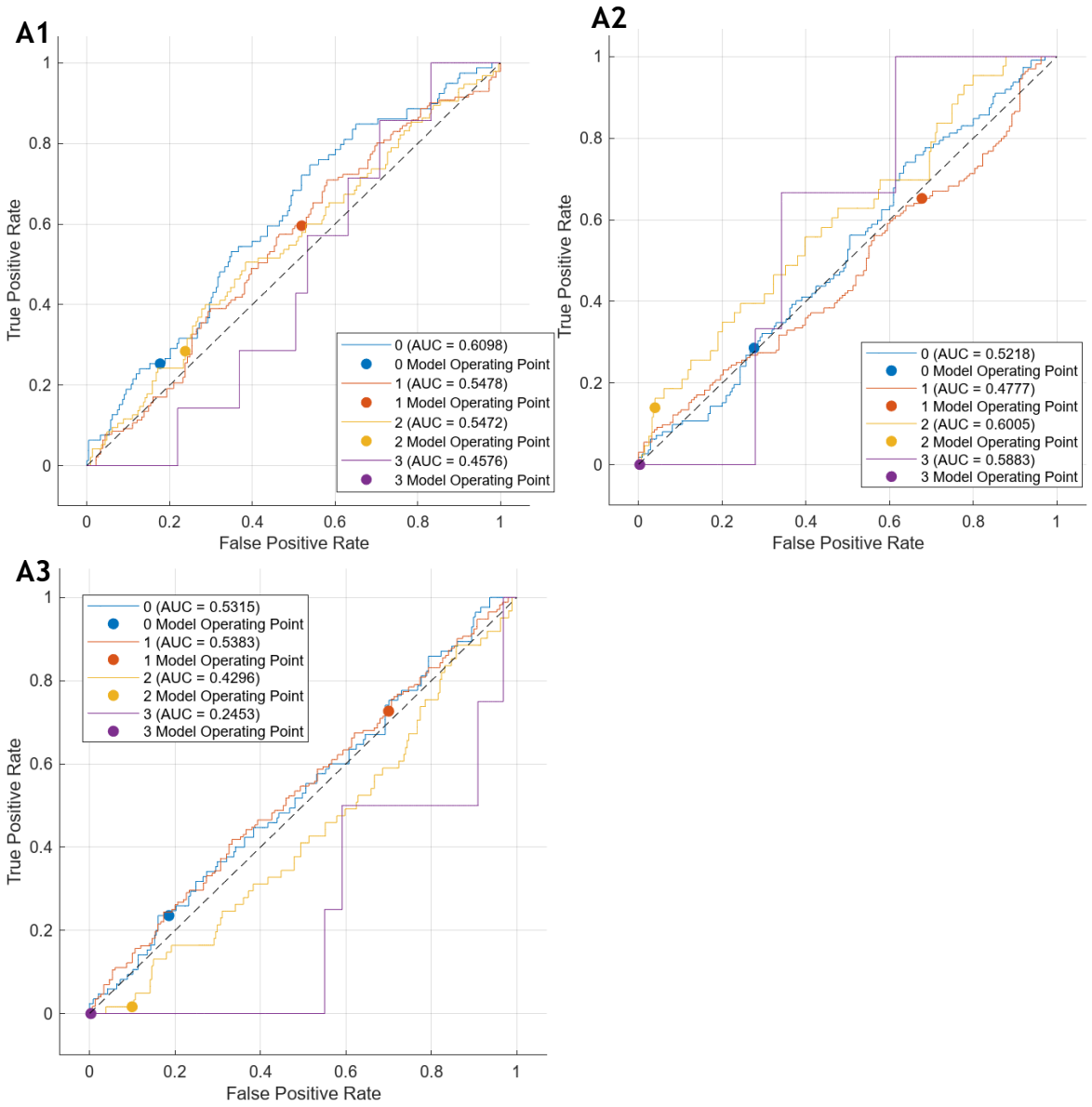


Figure A.13 Farm A – Convergent AHDB (4-levels) scores- extracted features after the pre-processing technique (masked).

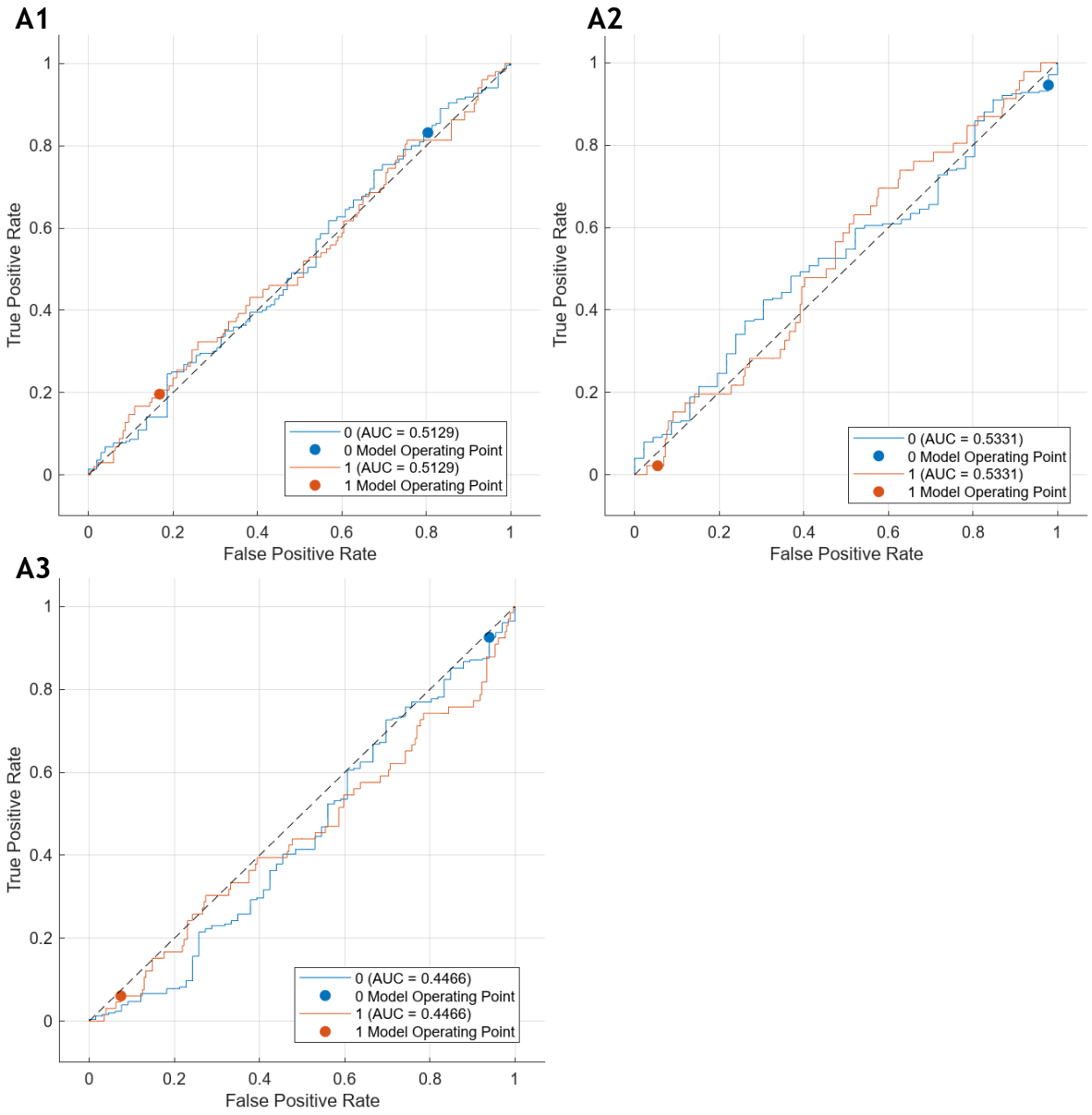


Figure A.14 Farm A – Binarised convergent AHDB (4-levels) scores- extracted features after the pre-processing technique (masked).

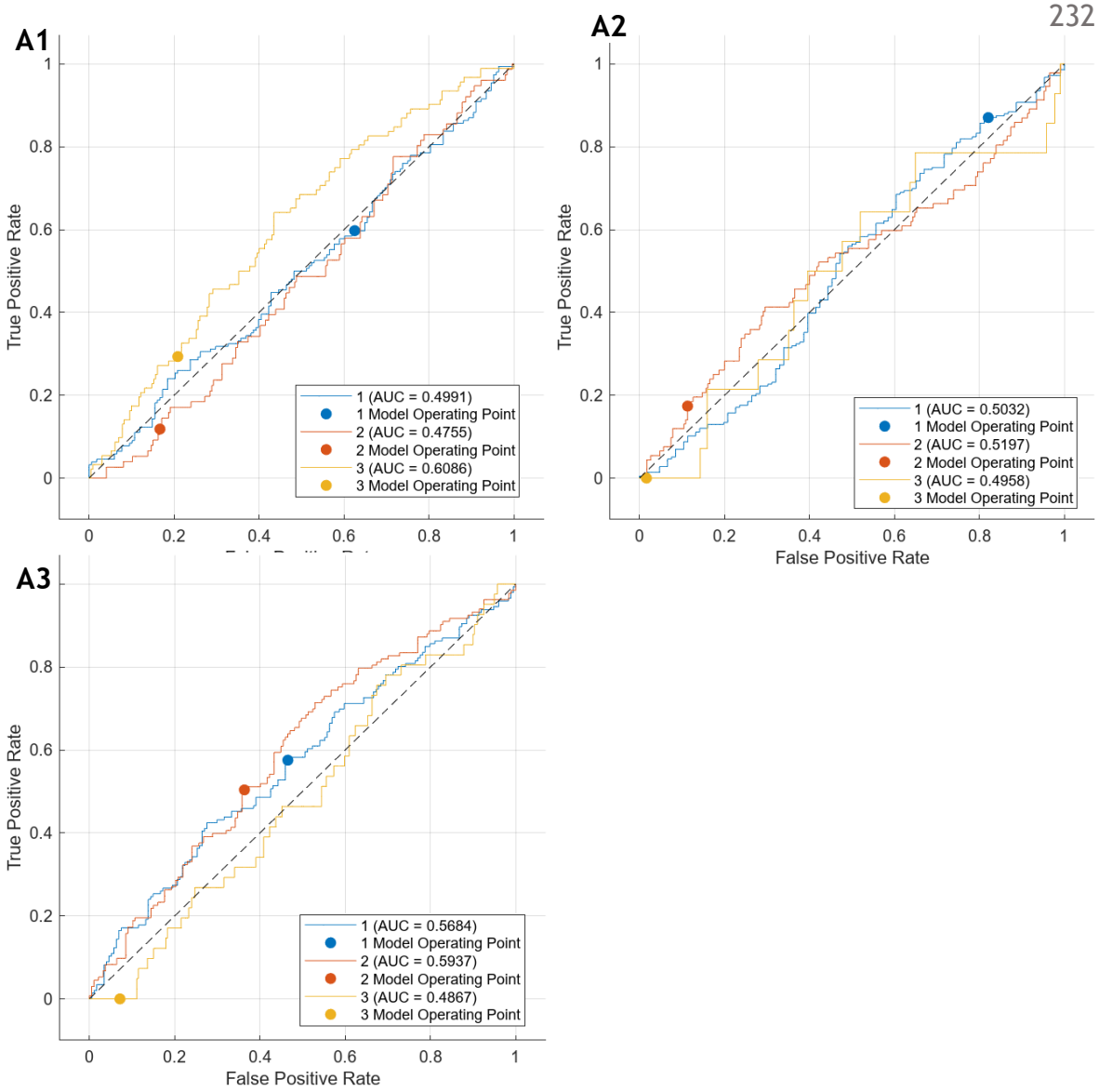


Figure A.15 Farm A – Grimm & Lorenzini (3-levels) scores- extracted features after the pre-processing technique (masked).

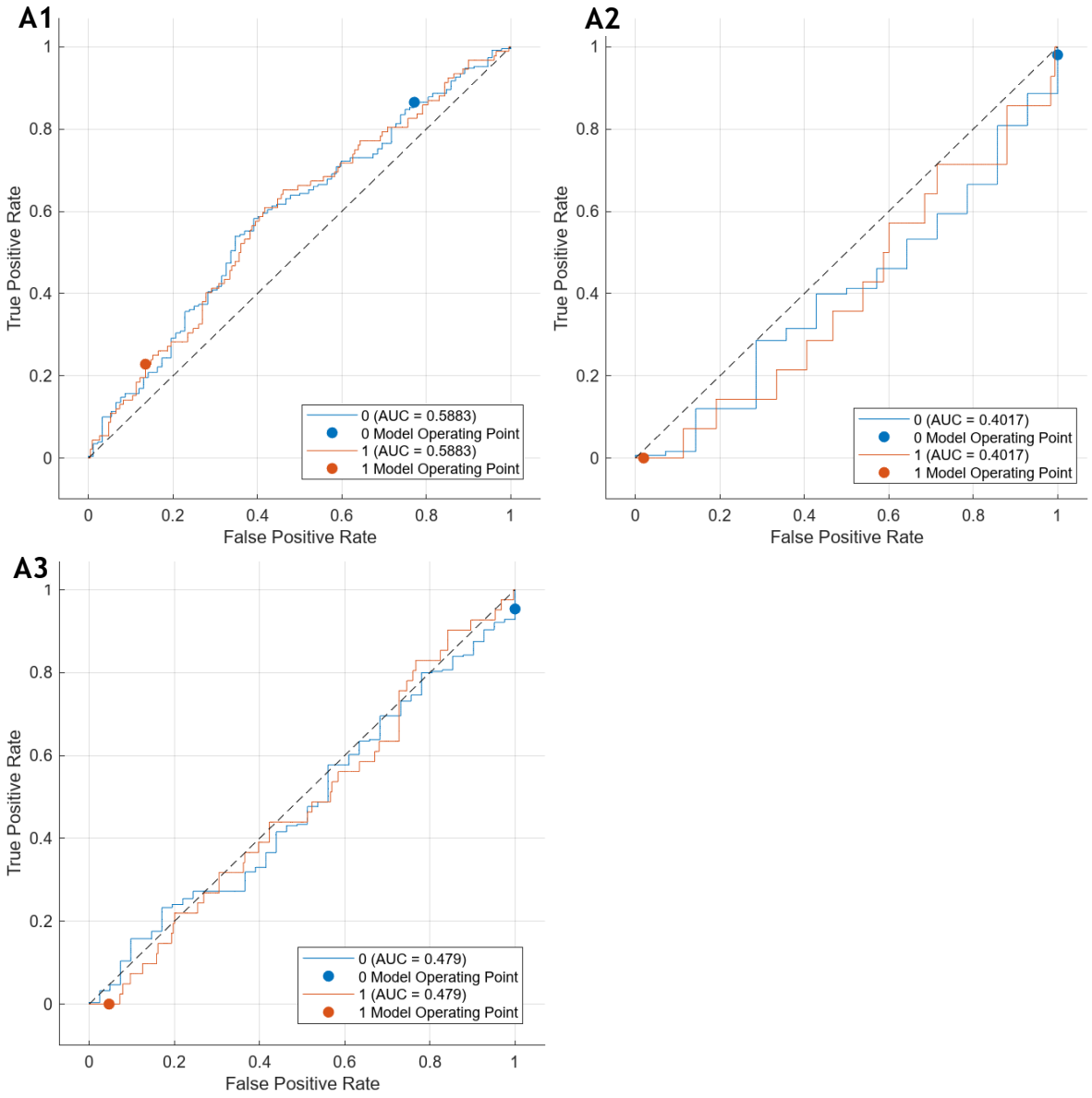


Figure A. 16 Farm A – Binarised Grimm & Lorenzini (2-levels) scores- extracted features after the pre-processing technique (masked).

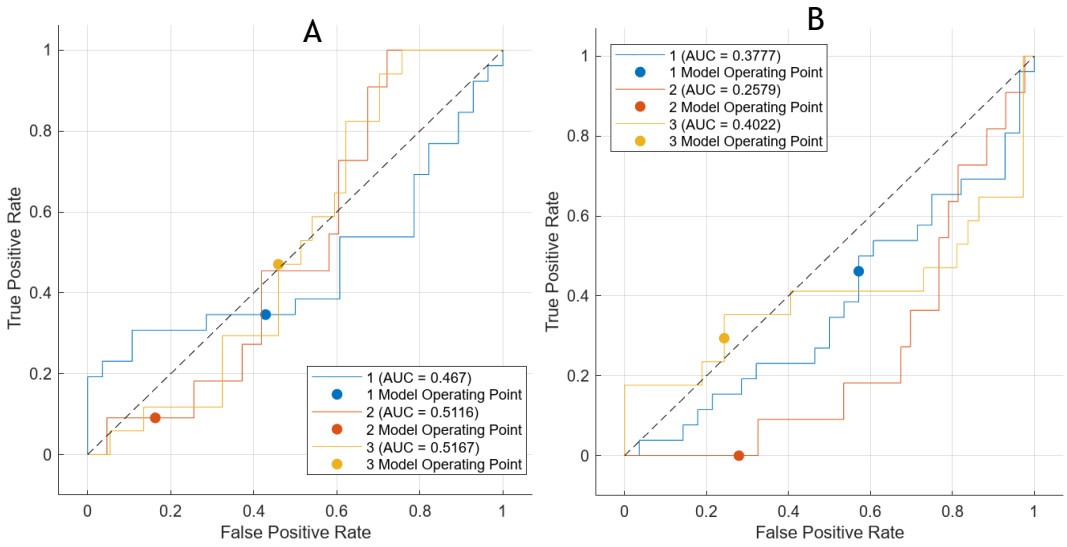


Figure A.17 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the Grimm & Lorenzini 3-levels consensus scoring for farm B assessment.

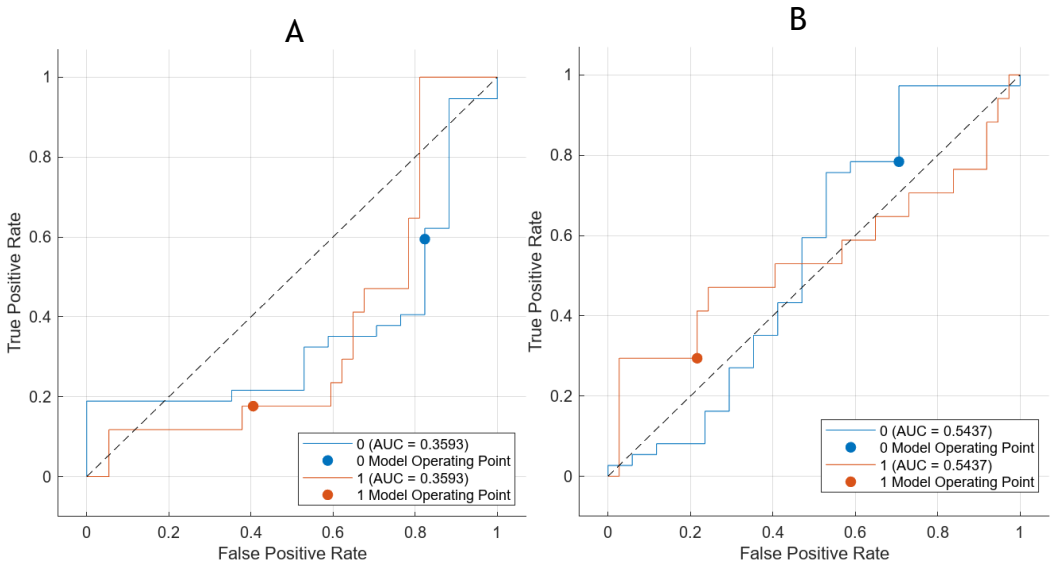


Figure A.18 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the binarised Grimm & Lorenzini (scores 1,2 vs 3) consensus scoring for farm B assessment.

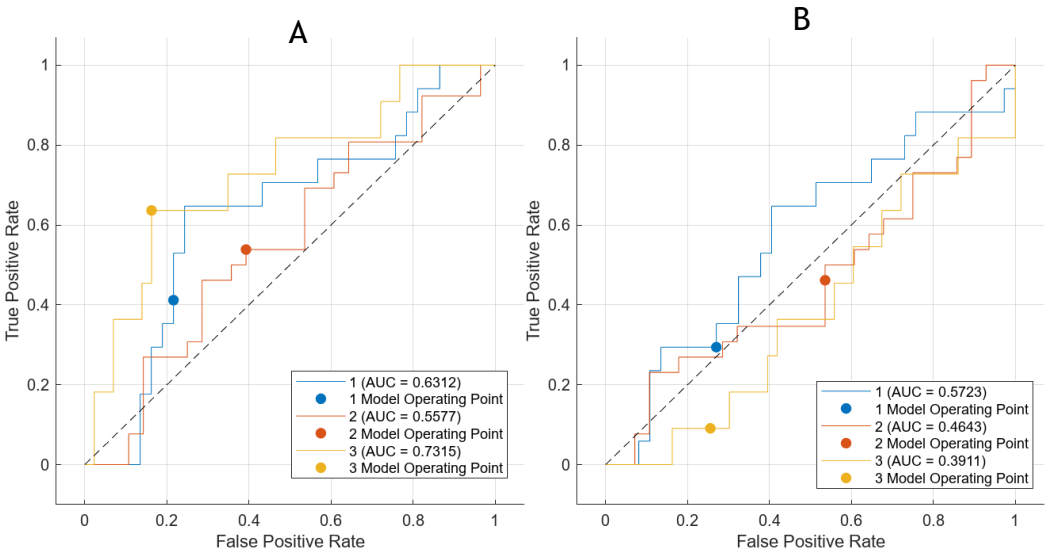


Figure A.19 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the averaged Grimm & Lorenzini individual scoring for farm B assessment.

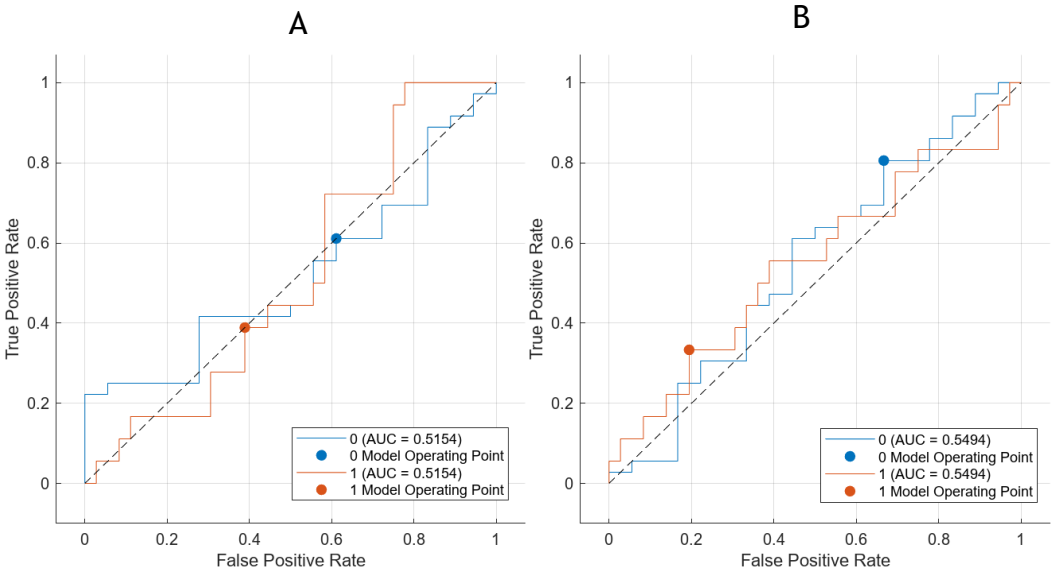


Figure A.20 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the averaged binarised Grimm & Lorenzini individual scoring for farm B assessment.

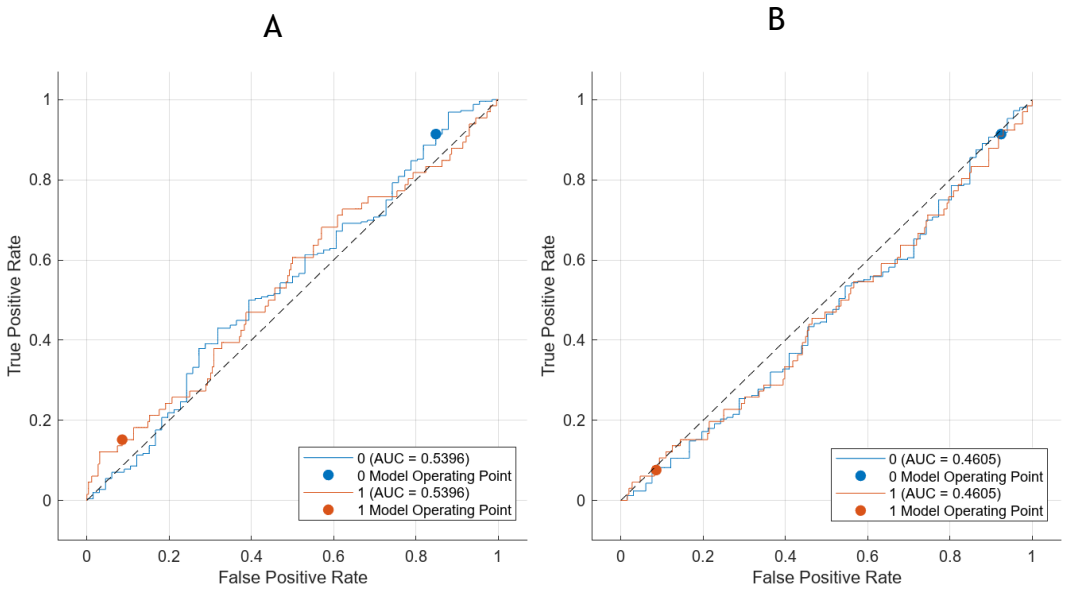


Figure A.21 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the binarised AHDB consensus scoring for farm A assessment.

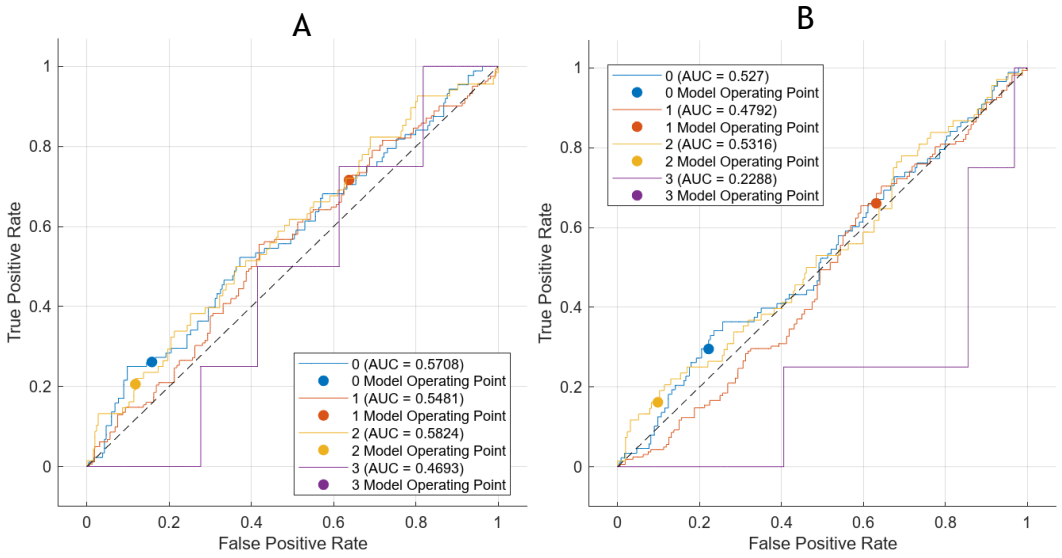


Figure A.22 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the averaged convergent AHDB scoring for farm A assessment.

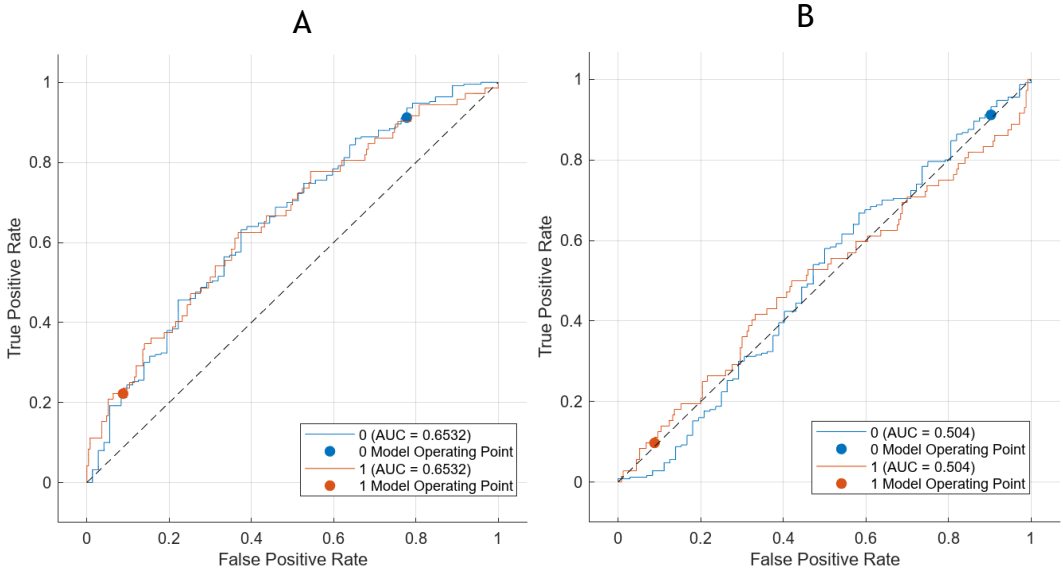


Figure A.23 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the averaged binarised convergent AHDB scoring for farm A assessment.

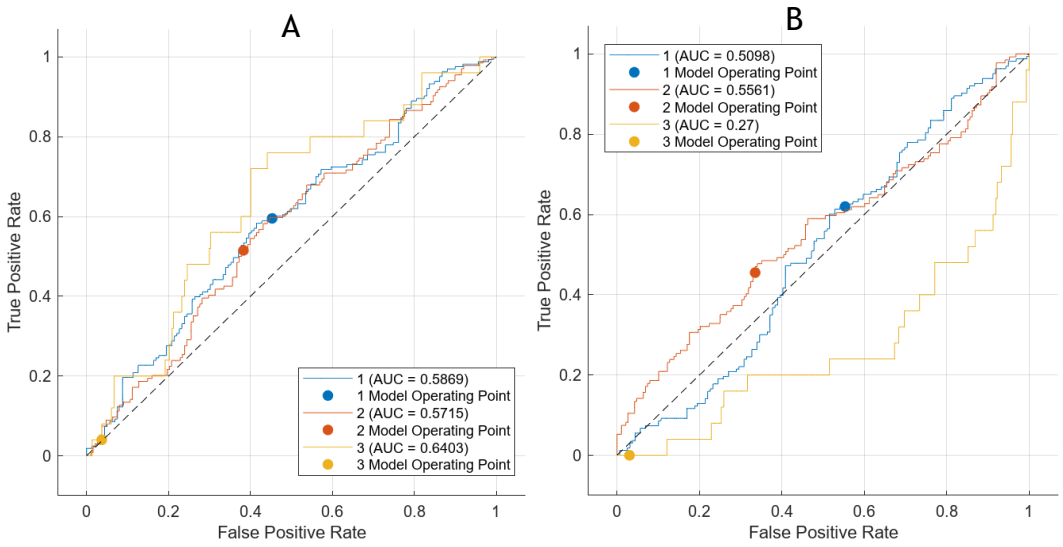


Figure A.24 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the averaged Grimm & Lorenzini scoring for farm A assessment.

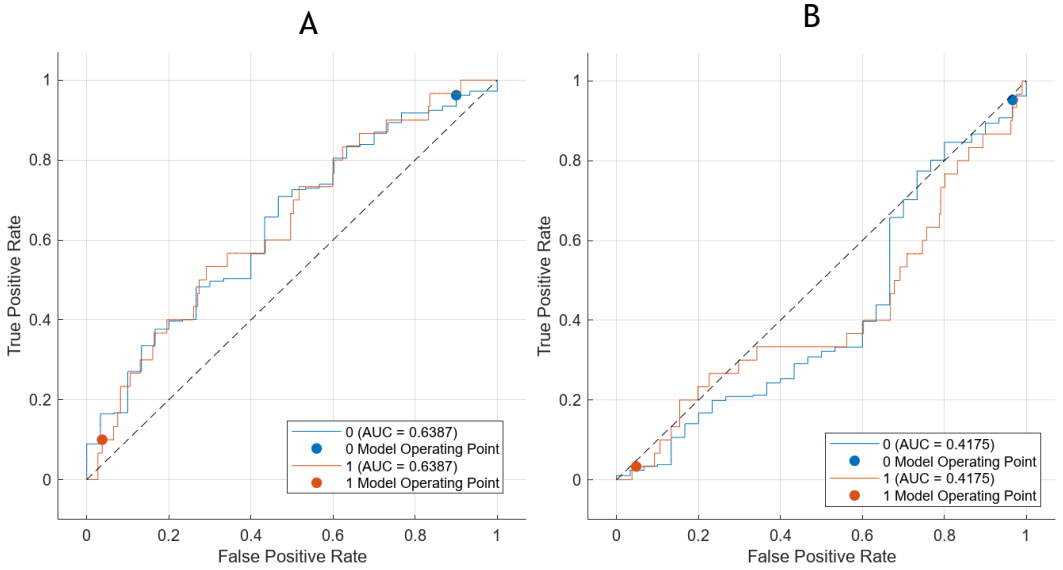


Figure A.25 ROC curves from the validation SVM models and the unmasked (A) - masked (B) data. Labels retrieved from the averaged binarised Grimm & Lorenzini scoring for farm A assessment.

References

123apps LLC. (n.d.). *online-video-cutter*. https://online-video-cutter.com/

Adriaense, J. E. C., Koski, S. E., Huber, L., & Lamm, C. (2020). Challenges in the comparative study of empathy and related phenomena in animals. In *Neuroscience and Biobehavioral Reviews* (Vol. 112, pp. 62–82). Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2020.01.021

Afonso, J. S., Bruce, M., Keating, P., Raboisson, D., Clough, H., Oikonomou, G., & Rushton, J. (2020). Profiling Detection and Classification of Lameness Methods in British Dairy Cattle Research: A Systematic Review and Meta-Analysis. In *Frontiers in Veterinary Science* (Vol. 7). Frontiers Media S.A. https://doi.org/10.3389/fvets.2020.00542

Afonso, J. S., Oikonomou, G., Carter, S., Clough, H. E., Griffiths, B. E., & Rushton, J. (2021). Diagnosis of Bovine Digital Dermatitis: Exploring the Usefulness of Indirect ELISA. *Frontiers in Veterinary Science*, *8*. https://doi.org/10.3389/fvets.2021.728691

AHDB. (2015). *How to score your herd Key benefits of scoring*. 0–1.

Ajdani, M., & Ghaffary, H. (2021). Design network intrusion detection system using support vector machine. *International Journal of Communication Systems*, *34*(3). https://doi.org/10.1002/dac.4689

Alawneh, J. I., Laven, R. A., & Stevenson, M. A. (2012). Interval between detection of lameness by locomotion scoring and treatment for lameness: A survival analysis. *Veterinary Journal*, *193*(3), 622–625. https://doi.org/10.1016/j.tvjl.2012.06.042

Albalate, A., & Minker, W. (2013). *Semi-Supervised and Unsupervised Machine Learning*. Wiley. https://doi.org/10.1002/9781118557693

Albright, K. (2004). Environmental scanning: radar for success. *Information Management*, *38*(3), 38.

Alejo, R., Valdovinos, R. M., García, V., & Pacheco-Sanchez, J. H. (2013). A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognition Letters*, *34*(4), 380–388. https://doi.org/10.1016/j.patrec.2012.09.003

Algan, G., & Ulusoy, I. (2019). *Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey*. https://doi.org/10.1016/j.knosys.2021.106771

Alhalabi, R. A., & Rebeiz, G. M. (2009). High-gain yagi-uda antennas for millimeter-wave switched-beam systems. *IEEE Transactions on Antennas and Propagation*, *57*(11), 3672–3676. https://doi.org/10.1109/TAP.2009.2026666

Alpaydin, E. (2014a). *Introduction to machine learning* (Third). The MIT Press. https://go.exlibris.link/Qk2sWyLr

Alpaydin, E. (2014b). Supervised Learning. In *Introduction to Machine Learning* (3rd ed., pp. 21–47). MIT Press. http://link.springer.com/10.1007/978-3-642-11210-2_2

Alsaaod, M., & Büscher, W. (2012). Detection of hoof lesions using digital infrared thermography in dairy cows. *Journal of Dairy Science*, *95*(2), 735–742. https://doi.org/10.3168/jds.2011-4762

Alsaaod, M., Luternauer, M., Hausegger, T., Kredel, R., & Steiner, A. (2017). The cow pedogram — Analysis of gait cycle variables allows the detection of lameness and foot pathologies. *Journal of Dairy Science*, *100*(2), 1417–1426. https://doi.org/10.3168/jds.2016-11678

Alsaaod, M., Schaefer, A. L., Büscher, W., & Steiner, A. (2015). The role of infrared thermography as a non-invasive tool for the detection of lameness in cattle. In *Sensors (Switzerland)* (Vol. 15, Issue 6, pp. 14513–14525). MDPI AG. https://doi.org/10.3390/s150614513

Alsaaod, M., Schmid, R. M., Zwahlen, N., Soto, S., Wildi, N., Seuberlich, T., & Steiner, A. (2023). First description of interdigital hyperplasia associated with contagious ovine digital dermatitis in two sheep. *Frontiers in Veterinary Science*, *9*. https://doi.org/10.3389/fvets.2022.1028880

Amstel, S. van, & Shearer, J. (2008). *Manual for Treatment and Control of Lameness in Cattle*. John Wiley & Sons.

Anagnostopoulos, A., Griffiths, B. E., Siachos, N., Neary, J., Smith, R. F., & Oikonomou, G. (2023). Initial validation of an intelligent video surveillance system for automatic detection of dairy cattle lameness. Frontiers in Veterinary Science, 10. https://doi.org/10.3389/fvets.2023.1111057

Andersson, L., & Lundström, K. (1981). The Influence of Breed, Age, Body Weight and Season on Digital Diseases and Hoof Size in Dairy Cows3. *Zentralblatt Für Veterinärmedizin Reihe A*, *28*(2), 141–151. https://doi.org/10.1111/j.1439-0442.1981.tb01174.x

Anees, R., Dinesh, P. T., Nithin, C. J., Sooryadas, S., Chandy, G., & David, P. V. (2022). Radiographic evaluation of hoof affections in dairy cattle. *Journal of Veterinary and Animal Sciences*, *53*(2). https://doi.org/10.51966/jvas.2022.53.2.158-162

Animals (Scientific Procedures) Act 1986. (1986).

Archer, S., Bell, N., & Huxley, J. (2010). Lameness in UK dairy cows: A review of the current status. *In Practice*, *32*(10), 492–504. https://doi.org/10.1136/inp.c6672

Archer, S., Bell, N., Huxley, J., & Archer, S. (2010). *Lameness in UK dairy cows : a review of the current status*. *32*(December), 492–504. https://doi.org/10.1136/inp.c6672

Atayants, B. A., Davydochkin, V. M., Ezerskiy, V. V, Atayants, B. A., & Eserskiy, V. V. (2014). *Precision FMCW Short-Range Radar for Industrial Applications*. Artech House. http://ebookcentral.proquest.com/lib/gla/detail.action?docID=1675145

Baggott, D. G. ;, Bunch, J. ;, & Gill, K. R. (1988). VARIATIONS IN SOME INORGANIC COMPONENTS AND PHYSICAL PROPERTIES OF CLAW KERATIN ASSOCIATED WITH CLAW DISEASE IN THE BRITISH FRIESIAN COW. *BRITISH VETERINARY JOURNAL*, *144*, *6*.

Bahr, C., Koppenol, A., Leroy, T., Song, X., Vranken, E., Maertens, W., Vangeyte, J., Nuffel, A. Van, Sonck, B., & Berckmans, D. (2008). *Early lameness detection in dairy cattle – Hoof movement analysis by using small camera distance*.

Balanis, C. A. (2016). *Antenna Theory : Analysis and Design*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/gla/detail.action?docID=4205879

Barker, Z. E., Leach, K. A., Whay, H. R., Bell, N. J., & Main, D. C. J. (2010). Assessment of lameness prevalence and associated risk factors in dairy herds in England and Wales. *Journal of Dairy Science*, *93*(3), 932–941. https://doi.org/10.3168/jds.2009-2309

Bashir, D., Montañez, G. D., Sehra, S., Segura, P. S., & Lauw, J. (2020). *An Information-Theoretic Perspective on Overfitting and Underfitting* (pp. 347–358). https://doi.org/10.1007/978-3-030-64984-5_27

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Beaudeau, F., Frankena, K., Fourichon, C., Seegers, H., Faye, B., & Noordhuizen, J. P. T. M. (1994). *Associations between health disorders of French dairy cows and early and late culling within the lactation level of milk production and occurrence of both reproductive disorders and poor repro-ductive performance were risk factors for late culling* (Issue 93).

Beer, G., Alsaaod, M., Starke, A., Schuepbach-Regula, G., Müller, H., Kohler, P., & Steiner, A. (2016). Use of extended characteristics of locomotion and feeding behavior for automated identification of lame dairy cows. *PLoS ONE*, *11*(5). https://doi.org/10.1371/journal.pone.0155796

Bell, N., & Huxley, J. (2009). Letters: Locomotion, lameness and mobility in dairy cows. In *Veterinary Record* (Vol. 164, Issue 23, p. 726). British Veterinary Association. https://doi.org/10.1136/vr.164.23.726

Bergsten, C., Hultgren, J., & Manske, T. (1998). Claw traits and foot lesions in Swedish dairy cows in relation to trimming interval and housing system. A preliminary report. In *Proc. 10th International symposium on lameness in ruminants* (pp. 46–46).

Bergsten, C., & Telezhenko, E. (2005). Walking comfort of dairy cows in different flooring systems expressed by foot prints and preference. *Cattle Practice*, *13*, 121–126.

Bernardi, F., Fregonesi, J., Winckler, C., Veira, D. M., von Keyserlingk, M. A. G., & Weary, D. M. (2009). The stall-design paradox: Neck rails increase lameness but improve udder and stall hygiene. *Journal of Dairy Science*, *92*(7), 3074–3080. https://doi.org/10.3168/jds.2008-1166

Berry, S. L. (2001). *DISEASES OF THE DIGITAL SOFT TISSUES DIGITAL DERMATITIS (PAPILLOMATOUS DIGITAL DERMATITIS, HAIRY HEEL WARTS, FOOT WARTS)*.

Bexkens, R., Claessen, F. M. A. P., Kodde, I. F., Oh, L. S., Eygendaal, D., & den Bekerom, M. P. van. (2018). The kappa paradox. In *Shoulder and Elbow* (Vol. 10, Issue 4, p. 308). SAGE Publications Inc. https://doi.org/10.1177/1758573218791813

Beyi, A. F., Hassall, A., Phillips, G. J., & Plummer, P. J. (2021). Tracking reservoirs of antimicrobial resistance genes in a complex microbial community using metagenomic hi-c: The case of bovine digital dermatitis. *Antibiotics*, *10*(2), 1–15. https://doi.org/10.3390/antibiotics10020221

Bicalho, R. C., Machado, V. S., & Caixeta, L. S. (2009). Lameness in dairy cattle : A debilitating disease or a disease of debilitated cattle ? A cross-sectional study of lameness prevalence and thickness of the digital cushion. *Journal of Dairy Science*, *92*(7), 3175–3184. https://doi.org/10.3168/jds.2008-1827

Bicalho, R. C., Vokey, F., Erb, H. N., & Guard, C. L. (2007). Visual locomotion scoring in the first seventy days in milk: Impact on pregnancy and survival. *Journal of Dairy Science*, *90*(10), 4586–4591. https://doi.org/10.3168/jds.2007-0297

Bicalho, R. C., Warnick, L. D., & Guard, C. L. (2008). Strategies to analyze milk losses caused by diseases with potential incidence throughout the lactation: A lameness example. *Journal of Dairy Science*, *91*(7), 2653–2661. https://doi.org/10.3168/jds.2007-0744

Blowey, R. (2005). Factors associated with lameness in dairy cattle. *In Practice*, *27*(3), 154–162. https://doi.org/10.1136/inpract.27.3.154

Blowey, R. W. (2015). *Cattle Lameness and Hoofcare : An Illustrated Guide (3rd Edition)*. 5m Publishing. http://ebookcentral.proquest.com/lib/gla/detail.action?docID=5389604

Boelling, D., & Pollott, G. E. (1998). Locomotion, lameness, hoof and leg traits in cattle II. Genetic relationships and breeding values. *Livestock Production Science*, *54*(3), 205–215. https://doi.org/10.1016/S0301-6226(97)00173-5

Böer-Auer, A., Kittler, H., & Tschandl, P. (2022). Bias, Noise, and Error. In *Pattern Analysis for Histopathologic Diagnosis of Melanocytic Lesions* (pp. 257–259). Springer International Publishing. https://doi.org/10.1007/978-3-031-07666-4_8

Boettcher, P. J., Dekkers, J. C. M., Warnick, L. D., & Wells, S. J. (1998a). Genetic Analysis of Clinical Lameness in Dairy Cattle. *Journal of Dairy Science*, *81*(4), 1148–1156. https://doi.org/10.3168/jds.S0022-0302(98)75677-2

Boettcher, P. J., Dekkers, J. C. M., Warnick, L. D., & Wells, S. J. (1998b). Genetic Analysis of Clinical Lameness in Dairy Cattle. *Journal of Dairy Science*, *81*(4), 1148–1156. https://doi.org/10.3168/jds.S0022-0302(98)75677-2

Bonser, R. H. C., Farrent, J. W., & Taylor, A. M. (2003). Assessing the Frictional and Abrasion-resisting Properties of Hooves and Claws. *Biosystems Engineering*, *86*(2), 253–256. https://doi.org/10.1016/S1537-5110(03)00136-3

Booth, C. J., Warnick, L. D., Gröhn, Y. T., Maizon, D. O., Guard, C. L., & Janssen, D. (2004). Effect of lameness on culling in dairy cows. *Journal of Dairy Science*, *87*(12), 4115–4122. https://doi.org/10.3168/jds.S0022-0302(04)73554-7

Bouveyron, C., & Girard, S. (2009). Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, *42*(11), 2649–2658. https://doi.org/10.1016/j.patcog.2009.03.027

Boyle, L. A., Mee, J. F., & Kiernan, P. J. (2007). The effect of rubber versus concrete passageways in cubicle housing on claw health and reproduction of pluriparous dairy cows. *Applied Animal Behaviour Science*, *106*, 1–12. https://doi.org/10.1016/j.applanim.2006.07.011

Bran, J. A., Costa, J. H. C., von Keyserlingk, M. A. G., & Hötzel, M. J. (2019). Factors associated with lameness prevalence in lactating cows housed in freestall and compost-bedded pack dairy farms in southern Brazil. *Preventive Veterinary Medicine*, *172*(April), 104773. https://doi.org/10.1016/j.prevetmed.2019.104773

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Brenner, S. K., Kaushal, R., Grinspan, Z., Joyce, C., Kim, I., Allard, R. J., Delgado, D., & Abramson, E. L. (2016). Effects of health information technology on patient outcomes: A systematic review. *Journal of the American Medical Informatics Association*, *23*(5), 1016–1036. https://doi.org/10.1093/jamia/ocv138

Brenninkmeyer, C., Dippel, S., Brinkmann, J., March, S., Winckler, C., & Knierim, U. (2013). Hock lesion epidemiology in cubicle housed dairy cows across two breeds, farming systems and countries. *Preventive Veterinary Medicine*, *109*(3–4), 236–245. https://doi.org/10.1016/j.prevetmed.2012.10.014

Brenninkmeyer, C., Dippel, S., March, S., Brinkmann, J., Winckler, C., & Knierim, U. (2007). Reliability of a subjective lamess scoring system for dairy cows. *Animal Welfare*, *16*(2), 127–129.

Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., von Kalle, C., Fröhling, S., Schilling, B., & Utikal, J. S. (2019). Deep neural networks are

superior to dermatologists in melanoma image classification. *European Journal of Cancer*, *119*, 11–17. https://doi.org/10.1016/j.ejca.2019.05.023

Buch, L. H., Sørensen, A. C., Lassen, J., Berg, P., Eriksson, J.-Å., Jakobsen, J. H., & Sørensen, M. K. (2011). Hygiene-related and feed-related hoof diseases show different patterns of genetic correlations to clinical mastitis and female fertility. *Journal of Dairy Science*, *94*(3), 1540–1551. https://doi.org/10.3168/jds.2010-3137

Buler, J. J., & Diehl, R. H. (2009). Quantifying bird density during migratory stopover using weather surveillance radar. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(8), 2741–2751. https://doi.org/10.1109/TGRS.2009.2014463

Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov. https://glasgow.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwfV1NTwIxEJ0gxMQTChrxg_TkbUm33ba7R4K7MZw5GC-ktF24iCf_vzO1S4BEj5NJ2vRr-trMvAcgxYxnZzHByiC8460qvJMbKbyqrCVhJL7hZYjcz82rrhd6-aHqVLtOpTEYeu3-K1InzhBOnpHoSymN0uYCBgKBPjHnr-Y5VW8h6lGmUvz9yMjLxPDUOfmpTZJllwRZtx05ZLximmvoU9nBDfTCfgTDTmyBpbM3hhdcULb7pvxDn1EUYJ8xDzKwJPywZQSYb4E19WrxlqU-1ul_Zt0Nw4g76OObP9wD0xoDmLK2yJ0onHGlsHgSvW-9E8a7YgKTP5t5-Mf3CFd441e_fwhPMGhxj4fnw7incQZ_AJMudpU

Busin, V., Viora, L., King, G., Tomlinson, M., Lekernec, J., Jonsson, N., & Fioranelli, F. (2019a). Evaluation of lameness detection using radar sensing in ruminants. *Veterinary Record*, *185*(18), 572. https://doi.org/10.1136/vr.105407

Busin, V., Viora, L., King, G., Tomlinson, M., Lekernec, J., Jonsson, N., & Fioranelli, F. (2019b). Evaluation of lameness detection using radar sensing in ruminants. *Veterinary Record*, *185*(18), 572. https://doi.org/10.1136/vr.105407

Byabazaire, J., Olariu, C., Taneja, M., & Davy, A. (2019). Lameness Detection as a Service: Application of Machine Learning to an Internet of Cattle. *2019 16th*

IEEE Annual Consumer Communications and Networking Conference, CCNC 2019. https://doi.org/10.1109/CCNC.2019.8651681

Calderón-Amor, J., Hernández-Gotelli, C., Strappini, A., Wittwer, F., & Sepúlveda-Varas, P. (2021). Prepartum factors associated with postpartum diseases in pasture-based dairy cows. *Preventive Veterinary Medicine*, *196*(February). https://doi.org/10.1016/j.prevetmed.2021.105475

Callan, R., & Garry, F. B. (2001). *Examination of the Musculoskeletal System in Recumbent Cattle Infectious Diseases of Ruminants View project*. https://www.researchgate.net/publication/266330726

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311. https://doi.org/10.1109/LSP.2006.870086

Carvalho, V. R. C., Bucklin, R. A., Shearer, J. K., & Shearer, L. (2005). Effects of trimming on dairy cattle hoof weight bearing and pressure distributions during the stance phase. *Transactions of the American Society of Agricultural Engineers*, *48*(4), 1653–1659. https://doi.org/10.13031/2013.19166

Ceballos, A., Sanderson, D., Rushen, J., & Weary, D. M. (2004). Improving stall design: Use of 3-d kinematics to measure space use by dairy cows when lying down. *Journal of Dairy Science*. https://doi.org/10.3168/jds.S0022-0302(04)70022-3

Cha, E., Hertl, J. A., Bar, D., & Gröhn, Y. T. (2010a). The cost of different types of lameness in dairy cows calculated by dynamic programming. *Preventive Veterinary Medicine*, *97*(1), 1–8. https://doi.org/10.1016/j.prevetmed.2010.07.011

Cha, E., Hertl, J. A., Bar, D., & Gröhn, Y. T. (2010b). The cost of different types of lameness in dairy cows calculated by dynamic programming. *Preventive Veterinary Medicine*, *97*(1), 1–8. https://doi.org/10.1016/j.prevetmed.2010.07.011

Channon, A. J., Walker, A. M., Pfau, T., Sheldon, I. M., & Wilson, A. M. (2009a). Variability of manson and leaver locomotion scores assigned to dairy cows by

different observers. *Veterinary Record*, *164*(13), 388–392. https://doi.org/10.1136/vr.164.13.388

Channon, A. J., Walker, A. M., Pfau, T., Sheldon, I. M., & Wilson, A. M. (2009b). Variability of manson and leaver locomotion scores assigned to dairy cows by different observers. *Veterinary Record*, *164*(13), 388–392. https://doi.org/10.1136/vr.164.13.388

Chapinal, N., Liang, Y., Weary, D. M., Wang, Y., & Von Keyserlingk, M. A. G. (2014). Risk factors for lameness and hock injuries in Holstein herds in China. *Journal of Dairy Science*, *97*(7), 4309–4316. https://doi.org/10.3168/jds.2014-8089

Chapinal, N., & Tucker, C. B. (2012). Validation of an automated method to count steps while cows stand on a weighing platform and its application as a measure to detect lameness. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2012-5742

Chawala, A. R., Lopez-Villalobos, N., Margerison, J. K., & Spelman, R. J. (2013). Genetic and crossbreeding parameters for incidence of recorded clinical lameness in New Zealand dairy cattle. *New Zealand Veterinary Journal*, *61*(5), 281–285. https://doi.org/10.1080/00480169.2013.763751

Chawla, N. V. (2010). Data Mining for Imbalanced Datasets: An Overview. In L. Maimon Oded and Rokach (Ed.), *Data Mining and Knowledge Discovery Handbook* (pp. 875–886). Springer US. https://doi.org/10.1007/978-0-387-09823-4_45

Chen, M., Mozaffari, M., Saad, W., Yin, C., Debbah, M., & Hong, C. S. (2017). Caching in the Sky: Proactive Deployment of Cache-Enabled Unmanned Aerial Vehicles for Optimized Quality-of-Experience. *IEEE Journal on Selected Areas in Communications*, *35*(5), 1046–1061. https://doi.org/10.1109/JSAC.2017.2680898

Chen, V. C. (2003). Micro-Doppler effect of micromotion dynamics: a review. *Independent Component Analyses, Wavelets, and Neural Networks*, *5102*(April 2003), 240. https://doi.org/10.1117/12.488855

Chen, V. C. (2008). Doppler signatures of radar backscattering from objects with micro-motions. *IET Signal Processing*, *2*(3), 291. https://doi.org/10.1049/iet-spr:20070137

Chen, V. C., Tahmoush, D., & Miceli, W. J. (2014a). *Radar Micro-Doppler Signatures Processing and Applications*.

Chen, V. C., Tahmoush, D., & Miceli, W. J. (Eds.). (2014b). *Radar Micro-Doppler Signatures: Processing and Applications*. Institution of Engineering and Technology. https://doi.org/10.1049/PBRA034E

Chesterton, R. N. ;, Morris, R. S., & Pfeiffer, D. U. (1989). Environmental and behavioural factors affecting the prevalence of foot lameness in new zealand dairy herds — a case-control study. *New Zealand Veterinary Journal*, *37*(4), 135–142. https://doi.org/10.1080/00480169.1989.35587

Chhatpar, K., Jora, G., & Chudasama, P. (2012). Carpal hygroma and its surgical excision in a cow. *Intas Polivet*, *13*(II), 279–280.

Clark, C. R., Petrie, L., Waldner, C., & Wendell, A. (2004). Characteristics of the bovine claw associated with the presence of vertical fissures (sandcracks). *The Canadian Veterinary Journal = La Revue Veterinaire Canadienne*, *45*(7), 585–593. http://www.ncbi.nlm.nih.gov/pubmed/15317390

Clarkson, M. J., Downham, D. Y., Faull, W. B., Hughes, J. W., Manson, F. J., Merritt, J. B., Murray, R. D., Russell, W. B., Sutherst, J. E., & Ward, W. R. (1996a). Incidence and prevalence of lameness in dairy cattle. *Veterinary Record*, *138*(23), 563–567. https://doi.org/10.1136/vr.138.23.563

Clarkson, M. J., Downham, D. Y., Faull, W. B., Hughes, J. W., Manson, F. J., Merritt, J. B., Murray, R. D., Russell, W. B., Sutherst, J. E., & Ward, W. R. (1996b). Incidence and prevalence of lameness in dairy cattle. *Veterinary Record*, *138*(23), 563–567. https://doi.org/10.1136/vr.138.23.563

Coetzee, J. F., Shearer, J. K., Stock, M. L., Kleinhenz, M. D., & van Amstel, S. R. (2017). An Update on the Assessment and Management of Pain Associated with Lameness in Cattle. *Veterinary Clinics of North America - Food Animal Practice*, *33*(2), 389–411. https://doi.org/10.1016/j.cvfa.2017.02.009

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., Crooker, B. A., Van Tassell, C. P., Yang, J., Wang, S., Matukumalli, L. K., & Da, Y. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics*, *12*(1), 408. https://doi.org/10.1186/1471-2164-12-408

Collick, D., Ward, W., & Dobson, H. (1989). Associations between types of lameness and fertility. *Veterinary Record*, *125*(5), 103–106. https://doi.org/10.1136/vr.125.5.103

Cook, N. B. (2003). Prevalence of lameness among dairy cattle in Wisconsin as a function of housing type and stall surface. *Journal of the American Veterinary Medical Association*, *223*(9), 1324–1328. https://doi.org/10.2460/javma.2003.223.1324

Cook, N. B. (2014). *Environmental and Nutritional Causes of Lameness*.

Cook, N. B., Bennett, T. B., & Nordlund, K. V. (2004a). Effect of free stall surface on daily activity patterns in dairy cows with relevance to lameness prevalence. *Journal of Dairy Science*, *87*(9), 2912–2922. https://doi.org/10.3168/jds.S0022-0302(04)73422-0

Cook, N. B., Bennett, T. B., & Nordlund, K. V. (2004b). Effect of free stall surface on daily activity patterns in dairy cows with relevance to lameness prevalence. *Journal of Dairy Science*, *87*(9), 2912–2922. https://doi.org/10.3168/jds.S0022-0302(04)73422-0

Cook, N. B., & Nordlund, K. V. (2009). The influence of the environment on dairy cow behavior, claw health and herd lameness dynamics. *The Veterinary Journal*, *179*(3), 360–369. https://doi.org/10.1016/j.tvjl.2007.09.016

Cramer, G., Lissemore, K. D., Guard, C. L., Leslie, K. E., & Kelton, D. F. (2009). The association between foot lesions and culling risk in Ontario Holstein cows.

Journal of Dairy Science, *92*(6), 2572–2579. https://doi.org/10.3168/jds.2008-1532

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge University Press.

Croyle, S. L., Nash, C. G. R., Bauman, C., LeBlanc, S. J., Haley, D. B., Khosa, D. K., & Kelton, D. F. (2018). Training method for animal-based measures in dairy cattle welfare assessments. *Journal of Dairy Science*, *101*(10), 9463–9471. https://doi.org/10.3168/jds.2018-14469

Cruciani, F., Cleland, I., Nugent, C., McCullagh, P., Synnes, K., & Hallberg, J. (2018). Automatic annotation for human activity recognition in free living using a smartphone. *Sensors (Switzerland)*, *18*(7). https://doi.org/10.3390/s18072203

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised Learning. In *Machine Learning Techniques for Multimedia* (pp. 21–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

Dahl-Pedersen, K., Foldager, L., Herskin, M. S., Houe, H., & Thomsen, P. T. (2018). Lameness scoring and assessment of fitness for transport in dairy cows: Agreement among and between farmers, veterinarians and livestock drivers. *Research in Veterinary Science*, *119*(May), 162–166. https://doi.org/10.1016/j.rvsc.2018.06.017

Danscher, A. M., Toelboell, T. H., & Wattle, O. (2010). Biomechanics and histology of bovine claw suspensory tissue in early acute laminitis. *Journal of Dairy Science*, *93*(1), 53–62. https://doi.org/10.3168/jds.2009-2038

Daros, R. R., Eriksson, H. K., Weary, D. M., & von Keyserlingk, M. A. G. (2019). Lameness during the dry period: Epidemiology and associated factors. *Journal of Dairy Science*, *102*(12), 11414–11427. https://doi.org/10.3168/jds.2019-16741

Davis-Unger, J., Pajor, E. A., Schwartzkopf-Genswein, K., Marti, S., Dorin, C., Spackman, E., & Orsel, K. (2017). Economic impacts of lameness in feedlot

cattle. *Translational Animal Science*, *1*(4), 467–479. https://doi.org/10.2527/tas2017.0052

Delignette-Muller, M. L., & Dutang, C. (2015). **fitdistrplus** : An *R* Package for Fitting Distributions. *Journal of Statistical Software*, *64*(4). https://doi.org/10.18637/jss.v064.i04

Dembele, I., Špinka, M., Stěhulová, I., Panamá, J., & Firla, P. (2006). Factors contributing to the incidence and prevalence of lameness on Czech dairy farms. *Czech Journal of Animal Science*, *51*(3), 102–109. https://doi.org/10.17221/3916-cjas

Desrochers, A. (2017). Diagnosis and Prognosis of Common Disorders Involving the Proximal Limb. *Veterinary Clinics of North America: Food Animal Practice*, *33*(2), 251–270. https://doi.org/10.1016/j.cvfa.2017.03.002

Desrochers, A., Anderson, D. E., & St-Jean, G. (2001). Lameness Examination in Cattle. *Veterinary Clinics of North America: Food Animal Practice*, *17*(1), 39–51. https://doi.org/10.1016/S0749-0720(15)30053-0

DeVries, T. J., von Keyserlingk, M. A. G., & Weary, D. M. (2004). Effect of Feeding Space on the Inter-Cow Distance, Aggression, and Feeding Behavior of Free-Stall Housed Lactating Dairy Cows. *Journal of Dairy Science*, *87*(5), 1432–1438. https://doi.org/10.3168/JDS.S0022-0302(04)73293-2

Diaz, J., & Bodurov, N. (1986). Economic losses from hoof diseases in cows. *Veterinarno-meditsinski nauki*, *23*(6), 72–76.

Dimitrovski, I., Kocev, D., Kitanovski, I., Loskovska, S., & Džeroski, S. (2015). Improved medical image modality classification using a combination of visual and textual features. *Computerized Medical Imaging and Graphics*, *39*, 14–26. https://doi.org/10.1016/j.compmedimag.2014.06.005

Dohoo, I. R., & Martin, S. W. (1984). Disease, production and culling in Holstein-Friesian cows III. Disease and production as determinants of disease. *Preventive Veterinary Medicine*, *2*(5), 671–690. https://doi.org/10.1016/0167-5877(84)90013-8

Dolecheck, K., & Bewley, J. (2018). Animal board invited review: Dairy cow lameness expenditures, losses and total cost. *Animal*, *12*(7), 1462–1474. https://doi.org/10.1017/S1751731118000575

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Donovan, G. A., Risco, C. A., Temple, G. M. D., Tran, T. Q., & Horn, H. H. Van. (2004). *Influence of Transition Diets on Occurrence of Subclinical Laminitis in Holstein Dairy Cows* *. 73–84.

Donovan, T., & Litchfield, D. (2013). Looking for Cancer: Expertise Related Differences in Searching and Decision Making. *Applied Cognitive Psychology*, *27*(1), 43–49. https://doi.org/10.1002/acp.2869

Dunn, G. (2004). *Statistical evaluation of measurement errors: design and analysis of reliability studies* (2nd ed.). Arnold. https://go.exlibris.link/WF8qKLGx

EFSA. (2009). Scientific report on the effects of farming systems on dairy cow welfare and disease. *EFSA Journal*, *7*(7), 1143r. https://doi.org/10.2903/j.efsa.2009.1143r

Egger-Danner, C., Cole, J. B., Pryce, J. E., Gengler, N., Heringstad, B., Bradley, A., & Stock, K. F. (2015). Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. *Animal*, *9*(2), 191–207. https://doi.org/10.1017/S1751731114002614

Engel, B., Bruin, G., Andre, G., & Buist, W. (2003). Assessment of observer performance in a subjective scoring system: Visual classification of the gait of cows. *Journal of Agricultural Science*, *140*(3), 317–333. https://doi.org/10.1017/S0021859603002983

Enting, H., Kooij, D., Dijkhuizen, A. A., Huirne, R. B. M., & Noordhuizen-Stassen, E. N. (1997). Economic losses due to clinical lameness in dairy cattle. *Livestock Production Science*, *49*(3), 259–267. https://doi.org/10.1016/S0301-6226(97)00051-1

Eriksson, H. K., Daros, R. R., von Keyserlingk, M. A. G., & Weary, D. M. (2020). Effects of case definition and assessment frequency on lameness incidence estimates. *Journal of Dairy Science*, *103*(1), 638–648. https://doi.org/10.3168/jds.2019-16426

Evans, D. H., & McDicken, W. N. (2000). *Doppler ultrasound: physics, instrumentation and signal processing* (2nd ed.). Wiley.

Fabian, J., Laven, R. A., & Whay, H. R. (2014). The prevalence of lameness on New Zealand dairy farms: A comparison of farmer estimate and locomotion scoring. *Veterinary Journal*. https://doi.org/10.1016/j.tvjl.2014.05.011

Farm Animal Welfare Council. (2009). *Farm Animal Welfare in Great Britain: Past, Present and Future*.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Fieguth, P. (2022). *An Introduction to Pattern Recognition and Machine Learning* (1st 2022.). Springer International Publishing. https://go.exlibris.link/T7C7fnnR

FileConverto Network. (n.d.). *mp4compress mute-video*. 2019-2020. https://www.mp4compress.com/mute-video/#

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The Affect Heuristic in Judgments of Risks and Benefits. *Journal of Behavioral Decision Making*.

Fioranelli, F., Ritchie, M., & Griffiths, H. (2015). Classification of Unarmed/Armed Personnel Using the NetRAD Multistatic Radar for Micro-Doppler and Singular Value Decomposition Features. *IEEE Geoscience and Remote Sensing Letters*, *12*(9), 1933–1937. https://doi.org/10.1109/LGRS.2015.2439393

Fioranelli, F., Shah, S. A., Haobo, L., Shrestha, A., Yang, S., & Le Kernec, J. (2019). *Radar sensing for healthcare*. *55*(19), 2019–2021. https://doi.org/10.1049/el.2019.2378

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Fitzsimmons, C. J., Thompson, C. A., & Sidney, P. G. (2020). Confident or familiar? The role of familiarity ratings in adults' confidence judgments when estimating fraction magnitudes. *Metacognition and Learning*, *15*(2), 215–231. https://doi.org/10.1007/s11409-020-09225-9

Fleiss, L. J. (1971). MEASURING NOMINAL SCALE AGREEMENT AMONG MANY RATERS. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/https://doi.org/10.1037/h0031619

Flower, F. C., de Passille, A. M., Weary, D. M., Sanderson, D. J., & Rushen, J. (2007). Softer, higher-friction flooring improves gait of cows with and without sole ulcers. *Journal of Dairy Science*, *90*(3), 1235–1242. https://doi.org/10.3168/jds.S0022-0302(07)71612-0

Flower, F. C., & Weary, D. M. (2006). Effect of hoof pathologies on subjective assessments of dairy cow gait. *Journal of Dairy Science*, *89*(1), 139–146. https://doi.org/10.3168/jds.S0022-0302(06)72077-X

Foditsch, C., Oikonomou, G., Machado, V. S., Bicalho, M. L., Ganda, E. K., Lima, S. F., Rossi, R., Ribeiro, B. L., Kussler, A., & Bicalho, R. C. (2016). Lameness prevalence and risk factors in large dairy farms in upstate New York. Model development for the prediction of claw horn disruption lesions. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0146718

Folger, P. (2014). *Basic Radar Principles and General Characteristics*. 1–34.

Ford, J. C., O'Rourke, K., Veinot, J. P., & Walley, V. M. (2000). Histologic estimation of coronary artery stenoses: Reproducibility and the effect of training. *Cardiovascular Pathology*, *9*(5), 251–255. https://doi.org/10.1016/S1054-8807(00)00044-2

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Sage. https://go.exlibris.link/wFgSX8sM

Frankena, K., Somers, J. G. C. J., Schouten, W. G. P., Stek, J. V. Van, Metz, J. H. M., Stassen, E. N., & Graat, E. A. M. (2009). The effect of digital lesions and floor type on locomotion score in Dutch dairy cows. *Preventive Veterinary Medicine*, *88*, 150–157. https://doi.org/10.1016/j.prevetmed.2008.08.004

Fraser, D. (2013). *Understanding Animal Welfare The Science in its Cultural Context* (1st ed.). Wiley.

Galagedara, L. W., Parkin, G. W., & Redman, J. D. (2003). An analysis of the ground-penetrating radar direct ground wave method for soil water content measurement. *Hydrological Processes*, *17*(18), 3615–3628. https://doi.org/10.1002/hyp.1351

Galati, F., Ourselin, S., & Zuluaga, M. A. (2022). From Accuracy to Reliability and Robustness in Cardiac Magnetic Resonance Image Segmentation: A Review. *Applied Sciences (Switzerland)*, *12*(8). https://doi.org/10.3390/app12083936

Galindo, F., & Broom, D. M. (2002). The effects of lameness on social and individual behavior of dairy cows. *Journal of Applied Animal Welfare Science*, *5*(3), 193–201. https://doi.org/10.1207/S15327604JAWS0503_03

Gamer, M., Lemon, J., & Fellows, I. P. S. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (R package version 0.84.1). https://cran.r-project.org/package=irr

Gamer, M., Lemon, J., Fellows, I., & Puspendra, S. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*.

Garbarino, E. J., Hernandez, J. A., Shearer, J. K., Risco, C. A., & Thatcher, W. W. (2004). Effect of lameness on ovarian activity in postpartum Holstein cows. *Journal of Dairy Science*, *87*(12), 4123–4131. https://doi.org/10.3168/jds.S0022-0302(04)73555-9

Garcia, E., König, K., Allesen-Holm, B. H., Klaas, I. C., Amigo, J. M., Bro, R., & Enevoldsen, C. (2015a). Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy

cows. *Journal of Dairy Science*, *98*(7), 4560–4571. https://doi.org/10.3168/jds.2014-9266

Garcia, E., König, K., Allesen-Holm, B. H., Klaas, I. C., Amigo, J. M., Bro, R., & Enevoldsen, C. (2015b). Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy cows. *Journal of Dairy Science*, *98*(7), 4560–4571. https://doi.org/10.3168/jds.2014-9266

Gauthreaux, S. A. (2003). Radar Ornithology and Biological Conservation. *The Auk*, *120*(2), 266–277. https://doi.org/10.2307/4090179

Gibbons, J., Vasseur, E., Rushen, J., & De Passillé, A. M. (2012). A training programme to ensure high repeatability of injury scoring of dairy cows. *Animal Welfare*, *21*(3), 379–388. https://doi.org/10.7120/09627286.21.3.379

Gong, J., Yan, J., Li, D., & Kong, D. (2022). Detection of Micro-Doppler Signals of Drones Using Radar Systems with Different Radar Dwell Times. *Drones*, *6*(9). https://doi.org/10.3390/drones6090262

Gonzalez, R. C., & Woods, R. E. (2006). *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://sdeuoc.ac.in/sites/default/files/sde_videos/Digital%20Image%20Processing%203rd%20ed.%20-%20R.%20Gonzalez%2C%20R.%20Woods-ilovepdf-compressed.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Graham, L. C. (1974). Synthetic Interferometer Radar For Topographic Mapping. *Proceedings of the IEEE*, *62*(6), 763–768. https://doi.org/10.1109/PROC.1974.9516

Green, L. E., Hedges, V. J., Schukken, Y. H., Blowey, R. W., & Packington, A. J. (2002). The impact of clinical lameness on the milk yield of dairy cows. *Journal of Dairy Science*, *85*(9), 2250–2256. https://doi.org/10.3168/jds.S0022-0302(02)74304-X

Greenough, P. R. (2007). *Bovine Laminitis and Lameness A hands-on approach*. Saunders/Elsevier. https://eleanor.lib.gla.ac.uk/record=b2619598

Gregory, N., Craggs, L., Hobson, N., & Krogh, C. (2006). Softening of cattle hoof soles and swelling of heel horn by environmental agents. *Food and Chemical Toxicology*, *44*(8), 1223–1227. https://doi.org/10.1016/j.fct.2006.01.018

Griffiths, B. E., White, D. G., & Oikonomou, G. (2018). A cross-sectional study into the prevalence of dairy cattle lameness and associated herd-level risk factors in England and Wales. *Frontiers in Veterinary Science*, *5*(APR), 1–8. https://doi.org/10.3389/fvets.2018.00065

Grimaldi, P., Lau, H., & Basso, M. A. (2015). There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. In *Neuroscience and Biobehavioral Reviews* (Vol. 55, pp. 88–97). Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2015.04.006

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test part 3:* https://doi.org/10.6028/NIST.IR.8280

Gu, X., Deligianni, F., Lo, B., Chen, W., & Yang, G. Z. (2018). Markerless gait analysis based on a single RGB camera. *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks, BSN 2018*, *2018-Janua*(March), 42–45. https://doi.org/10.1109/BSN.2018.8329654

Gucht, T. Van De, Saeys, W., Meensel, J. Van, Nuffel, A. Van, & Vangeyte, J. (2018). Farm-specific economic value of automatic lameness detection systems in dairy cattle : From concepts to operational simulations. *Journal of Dairy Science*, *101*(1), 637–648. https://doi.org/10.3168/jds.2017-12867

Gudaj, R. T., Brydl, E., Lehoczky, J., & Komlósi, I. (2012). Analysis of lameness traits and type traits in Hungarian Holstein-Friesian cattle. *Biotechnology in Animal HusbandryBiotehnologija u Stocarstvu*, *28*(2), 195–204. https://doi.org/10.2298/bah1202195g

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy

in retinal fundus photographs. *JAMA - Journal of the American Medical Association*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Gwet, K. L. (2008). Intrarater Reliability. In *Wiley Encyclopedia of Clinical Trials* (Vol. 1, Issue 1). John Wiley & Sons, Inc. https://doi.org/10.1002/9780471462422.eoct631

Hadley Wickham. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Häggman, J., Juga, J., Sillanpää, M. J., & Thompson, R. (2013). Genetic parameters for claw health and feet and leg conformation traits in Finnish Ayrshire cows. *Journal of Animal Breeding and Genetics*, *130*(2), 89–97. https://doi.org/10.1111/j.1439-0388.2012.01007.x

Haley, D. B., Rushen, J., & Passillé, A. M. De. (2000). *Behavioural indicators of cow comfort : activity and resting behaviour of dairy cows in two types of housing*. *August 1999*.

Hall, K. H. (2002). Reviewing intuitive decision-making and uncertainty: the implications for medical education. *Medical Education*, *36*(3), 216–224. https://doi.org/10.1046/j.1365-2923.2002.01140.x

Hammond, J. S., Keeney, R. L., & Raiffa, H. (1998). *The Hidden Traps in Decision Making Harvard Business Review*.

Hao, D., Zhang, L., Sumkin, J., Mohamed, A., & Wu, S. (2020). Inaccurate Labels in Weakly-Supervised Deep Learning: Automatic Identification and Correction and Their Impact on Classification Performance. *IEEE Journal of Biomedical and Health Informatics*, *24*(9), 2701–2710. https://doi.org/10.1109/JBHI.2020.2974425

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Haskell, M. J., Rennie, L. J., Bowell, V. A., Bell, M. J., & Lawrence, A. B. (2006). Housing system, milk production, and zero-grazing effects on lameness and leg injury in dairy cows. *Journal of Dairy Science*, *89*(11), 4259–4266. https://doi.org/10.3168/jds.S0022-0302(06)72472-9

Hasselmann, K., & Hasselmann, S. (1991). On the nonlinear mapping of an ocean wave spectrum into a synthetic aperture radar image spectrum and its inversion. *Journal of Geophysical Research*, *96*(C6), 10713. https://doi.org/10.1029/91jc00302

Hayashi, S., Saho, K., Shioiri, K., Fujimoto, M., & Masugi, M. (2021). Utilization of micro-doppler radar to classify gait patterns of young and elderly adults: An approach using a long short-term memory network. *Sensors*, *21*(11). https://doi.org/10.3390/s21113643

He, X., Yang, Z., & Tsien, J. Z. (2011). A hierarchical probabilistic model for rapid object categorization in natural scenes. *PLoS ONE*, *6*(5). https://doi.org/10.1371/journal.pone.0020002

Hedgepeth, J., Fuhriman, D., Acker, W., & McFadden, B. (1999). *Fish Behavior Measured by a Tracking Radar-Type Acoustic Transducer near Hydroelectric Dams*.

Hedges, J., Blowey, R. W. ;, Packington, A. J. ;, Callaghan, C. J. O., & Green, L. E. (2001). A Longitudinal Field Trial of the Effect of Biotin on Lameness in Dairy Cows. *Journal of Dairy Science*, *84*(9), 1969–1975. https://doi.org/10.3168/jds.S0022-0302(01)74639-5

Hemsworth, P. H., Coleman, G. J., Barnett, J. L., & Borg, S. (2000). Relationships between human-animal interactions and productivity of commercial dairy cows. In *J. Anim. Sci* (Vol. 78). https://academic.oup.com/jas/article/78/11/2821/4625626

Hendrycks, D., Mazeika, M., Wilson, D., & Gimpel, K. (2018). *Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise*. http://arxiv.org/abs/1802.05300

Henry Kamulegeya, L., Okello, M., Mark Bwanika, J., Musinguzi, D., Lubega, W., Rusoke, D., Nassiwa, F., & Börve, A. (2019). *Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning*. https://doi.org/https://doi.org/10.1101/826057

Heringstad, B., Egger-Danner, C., Charfeddine, N., Pryce, J. E., Stock, K. F., Kofler, J., Sogstad, A. M., Holzhauer, M., Fiedler, A., Müller, K., Nielsen, P., Thomas, G., Gengler, N., de Jong, G., Ødegård, C., Malchiodi, F., Miglior, F., Alsaaod, M., & Cole, J. B. (2018). Invited review: Genetics and claw health: Opportunities to enhance claw health by genetic selection. *Journal of Dairy Science*, *101*(6), 4801–4821. https://doi.org/10.3168/jds.2017-13531

Herjanic, B., & Reich, W. (1997). Development of a structured psychiatric interview for children : Agreement between child and parent on individual symptoms. *Journal of Abnormal Child Psychology*, *25*(1), 21–31. https://doi.org/10.1023/A:1025703323438

Hernandez, J. A., Garbarino, E. J., Shearer, J. K., Risco, C. A., & Thatcher, W. W. (2007). Evaluation of the efficacy of prophylactic hoof health examination and trimming during midlactation in reducing the incidence of lameness during late lactation in dairy cows. *Journal of the American Veterinary Medical Association*, *230*(1), 89–93. https://doi.org/10.2460/javma.230.1.89

Hernandez, J., Shearer, J. K., & Webb, D. W. (2002). Effect of lameness on milk yield in dairy cows. *Journal of the American Veterinary Medical Association*, *220*(5), 640–644. https://doi.org/10.2460/javma.2002.220.640

Hernandez-Mendo, O., Von Keyserlingk, M. A. G., Veira, D. M., & Weary, D. M. (2007). Effects of pasture on lameness in dairy cows. *Journal of Dairy Science*, *90*(3), 1209–1214. https://doi.org/10.3168/jds.S0022-0302(07)71608-9

Hlawatsch, F. (Franz), & Auger, F. (2008). *Time-frequency analysis : concepts and methods*. ISTE.

Holzhauer, M., Middelesch, H., Bartels, C. J. M., Frankena, K., Verhoeff, J., Noordhuizen-Stassen, E. N., & Noordhuizen, J. P. T. M. (2005). Assessing the repeatability and reproducibility of the Leg Score: A Dutch Claw Health Scoring System for dairy cattle. *Tijdschrift Voor Diergeneeskunde*, *130*(14–15), 440–443.

Howell, J. M. (1983). Toxicity and Environmental Problems TOXICITY PROBLEMS ASSOCIATED WITH TRACE ELEMENTS IN DOMESTIC ANIMALS. *British Society of Animal Production 1983*.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2008). *A Practical Guide to Support Vector Classication*.

Huang, Y., & Boyle, K. (2008). Antenna Basics. In *Antennas* (pp. 107–127). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470772911.ch4

Huang, Y. C., & Shanks, R. D. (1995a). Within herd estimates of heritabilities for six hoof characteristics and impact of dispersion of discrete severity scores on estimates. *Livestock Production Science*, *44*(2), 107–114. https://doi.org/10.1016/0301-6226(95)00061-3

Huang, Y. C., & Shanks, R. D. (1995b). Within herd estimates of heritabilities for six hoof characteristics and impact of dispersion of discrete severity scores on estimates. *Livestock Production Science*, *44*(2), 107–114. https://doi.org/10.1016/0301-6226(95)00061-3

Huang, Y. C., Shanks, R. D., & McCoy, G. C. (1995). Evaluation of fixed factors affecting hoof health. *Livestock Production Science*, *44*(2), 115–124. https://doi.org/https://doi.org/10.1016/0301-6226(95)00062-5

Hubbard, R. A., Johnson, E., Chubak, J., Wernli, K. J., Kamineni, A., Bogart, A., & Rutter, C. M. (2017). Accounting for misclassification in electronic health records-derived exposures using generalized linear finite mixture models. *Health Services and Outcomes Research Methodology*, *17*(2), 101–112. https://doi.org/10.1007/s10742-016-0149-5

Hudson, C., Whay, H., & Huxley, J. (2008). Recognition and management of pain in cattle. *In Practice*, *30*(3), 126–134. https://doi.org/10.1136/inpract.30.3.126

Hultgren, J. (2002). *Foot / leg and udder health in relation to housing changes in Swedish dairy herds*. *53*, 167–189.

Hunting, W. (1895). *The Art of Horse-Shoeing: A Manual for Farriers*. Hansebooks.

Huxley, J. N., & Whay, H. R. (2006). Current attitudes of cattle practitioners to pain and the use of analgesics in cattle. *Veterinary Record*, *159*(20), 662–668. https://doi.org/10.1136/vr.159.20.662

Huzzey, J. M., DeVries, T. J., Valois, P., & Von Keyserlingk, M. A. G. (2006). Stocking density and feed barrier design affect the feeding and social behavior of dairy cattle. *Journal of Dairy Science*, *89*(1), 126–133. https://doi.org/10.3168/jds.S0022-0302(06)72075-6

Ito, K., von Keyserlingk, M. A. G., LeBlanc, S. J., & Weary, D. M. (2010). Lying behavior as an indicator of lameness in dairy cows. *Journal of Dairy Science*, *93*(8), 3553–3560. https://doi.org/10.3168/jds.2009-2951

Ito, K., Weary, D. M., & von Keyserlingk, M. A. G. (2009). Lying behavior: Assessing within- and between- herd variation in free-stall-housed dairy cows. *Journal of Dairy Science*, *92*(9), 4412–4420. https://doi.org/10.3168/jds.2009-2235

Jaarsma, T., Jarodzka, H., Nap, M., Van Merrienboer, J. J. G., & Boshuizen, H. P. A. (2014). Expertise under the microscope: Processing histopathological slides. *Medical Education*, *48*(3), 292–300. https://doi.org/10.1111/medu.12385

Jacquin, L., Imoussaten, A., Trousset, F., Montmain, J., & Perrin, D. (2019). Evidential Classification of Incomplete Data via Imprecise Relabelling: Application to Plastic Sorting. In *Proceedings* (pp. 122–135). https://doi.org/10.1007/978-3-030-35514-2_10

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Japkowicz, N., & Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.*, *6*, 429–449.

Jensen, M. B., Pedersen, L. J., & Munksgaard, L. (2005). The effect of reward duration on demand functions for rest in dairy heifers and lying requirements as measured by demand functions. *Applied Animal Behaviour Science*, *90*(3–4), 207–217. https://doi.org/10.1016/j.applanim.2004.08.006

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and

future. In *Stroke and Vascular Neurology* (Vol. 2, Issue 4, pp. 230–243). BMJ Publishing Group. https://doi.org/10.1136/svn-2017-000101

Johnson, J. M., & Khoshgoftaar, T. M. (2022). A Survey on Classifying Big Data with Label Noise. *Journal of Data and Information Quality*, *14*(4). https://doi.org/10.1145/3492546

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Juliusson, E. Á., Karlsson, N., & Gärling, T. (2005). Weighing the past and the future in decision making. *European Journal of Cognitive Psychology*, *17*(4), 561–575. https://doi.org/10.1080/09541440440000159

Jun, Y., Eo, T., Kim, T., Shin, H., Hwang, D., Bae, S. H., Park, Y. W., Lee, H. J., Choi, B. W., & Ahn, S. S. (2018). Deep-learned 3D black-blood imaging using automatic labelling technique and 3D convolutional neural networks for detecting metastatic brain tumors. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-27742-1

Jung, A. (2022). *Machine learning: the basics*. Springer.

Jung, I., Kulldorff, M., & Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine*, *26*(7), 1594–1607. https://doi.org/10.1002/sim.2607

Jungbluth, T., Benz, B., & Wandel, H. (2003). SOFT WALKING AREAS IN LOOSE HOUSING SYSTEMS FOR DAIRY COWS. *Fifth International Dairy Housing Conference for 2003*. https://doi.org/10.13031/2013.11618

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, *4*, 237–285. https://doi.org/10.1613/jair.301

Kagiyama, N., Piccirilli, M., Yanamala, N., Shrestha, S., Farjo, P. D., Casaclang-Verzosa, G., Tarhuni, W. M., Nezarat, N., Budoff, M. J., Narula, J., & Sengupta, P. P. (2020). Machine Learning Assessment of Left Ventricular Diastolic Function Based on Electrocardiographic Features. *Journal of the*

American College of Cardiology, *76*(8), 930–941. https://doi.org/10.1016/j.jacc.2020.06.061

Kahneman, D. (2011). Thinking, fast and slow. In *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*.

Kang, X., Zhang, X. D., & Liu, G. (2021). A review: Development of computer vision-based lameness detection for dairy cows and discussion of the practical applications. In *Sensors (Switzerland)* (Vol. 21, Issue 3, pp. 1–24). MDPI AG. https://doi.org/10.3390/s21030753

Kao, T. Y. J., Yan, Y., Shen, T. M., Chen, A. Y. K., & Lin, J. (2013). Design and analysis of a 60-GHz CMOS doppler micro-radar system-in-package for vital-sign and vibration detection. *IEEE Transactions on Microwave Theory and Techniques*, *61*(4), 1649–1659. https://doi.org/10.1109/TMTT.2013.2247620

Karabulut, E. M., Özel, S. A., & İbrikçi, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, *1*, 323–327. https://doi.org/10.1016/j.protcy.2012.02.068

Katzenberger, K., Rauch, E., Erhard, M., Reese, S., & Gauly, M. (2020). Inter-rater reliability of welfare outcome assessment by an expert and farmers of South Tyrolean dairy farming. *Italian Journal of Animal Science*, *19*(1), 1079–1090. https://doi.org/10.1080/1828051X.2020.1816509

Kendal, B. (2011). The beginnings of air radio navigation and communication. *Journal of Navigation*, *64*(1), 157–167. https://doi.org/10.1017/S0373463310000251

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*(1/2), 81. https://doi.org/10.2307/2332226

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014*, 372–378. https://doi.org/10.1109/SAI.2014.6918213

Khetan, A., Lipton, Z. C., & Anandkumar, A. (2017). *Learning From Noisy Singly-labeled Data*. http://arxiv.org/abs/1712.04577

Kielland, C., Ruud, L. E., Zanella, A. J., & Østerås, O. (2009). Prevalence and risk factors for skin lesions on legs of dairy cattle housed in freestalls in Norway. *Journal of Dairy Science*, *92*(11), 5487–5496. https://doi.org/10.3168/jds.2009-2293

Kim, J. W., Lee, B. H., Shaw, M. J., Chang, H. L., & Nelson, M. (2001). Application of decision-tree induction techniques to personalized advertisements on internet storefronts. *International Journal of Electronic Commerce*, *5*(3), 45–62. https://doi.org/10.1080/10864415.2001.11044215

King, M. T. M., LeBlanc, S. J., Pajor, E. A., & DeVries, T. J. (2017). Cow-level associations of lameness, behavior, and milk yield of cows milked in automated systems. *Journal of Dairy Science*, *100*(6), 4818–4828. https://doi.org/10.3168/jds.2016-12281

Kingsley, S., & Quegan, S. (1999). *Understanding Radar Systems*.

Klapproth, F. (2008). Time and decision making in humans. *Cognitive, Affective and Behavioral Neuroscience*, *8*(4), 509–524. https://doi.org/10.3758/CABN.8.4.509

Knierim, U., & Winckler, C. (2009). On-farm welfare assessment in cattle: Validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach. *Animal Welfare*, *18*(4), 451–458.

Koenig, S., Sharifi, A. R., Wentrot, H., Landmann, D., Eise, M., & Simianer, H. (2005). Genetic Parameters of Claw and Foot Disorders Estimated with Logistic Models. *Journal of Dairy Science*, *88*(9), 3316–3325. https://doi.org/10.3168/jds.S0022-0302(05)73015-0

Kofler, J., Hangl, A., Pesenhofer, R., & Landl, G. (2011). Evaluation of claw health in heifers in seven dairy farms using a digital claw trimming protocol and claw data analysis system Evaluierung der Klauengesundheit von Färsen in 7 Milchvieh-betrieben mittels digitaler Klauendatendokumentation und

Klauendatenanalyse. *Berl Münch Tierärztl Wochenschr*, *124*(8), 10–19. https://doi.org/10.2376/0005-9366-124-10

Kolawole, M. O. (2002). *Radar systems, peak detection and tracking*. Newnes. https://doi.org/10.1016/B978-0-7506-5773-0.X5000-6

Kolbehdari, D., Wang, Z., Grant, J. R., Murdoch, B., Prasad, A., Xiu, Z., Marques, E., Stothard, P., & Moore, S. S. (2008). A Whole-Genome Scan to Map Quantitative Trait Loci for Conformation and Functional Traits in Canadian Holstein Bulls. *Journal of Dairy Science*, *91*(7), 2844–2856. https://doi.org/10.3168/jds.2007-0585

König, S., Wu, X. L., Gianola, D., Heringstad, B., & Simianer, H. (2008). Exploration of relationships between claw disorders and milk yield in Holstein cows via recursive linear and threshold models. *Journal of Dairy Science*, *91*(1), 395–406. https://doi.org/10.3168/jds.2007-0170

Koo, V. C., Chan, Y. K., Gobi, V., Chua, M. Y., Lim, C. H., Lim, C. S., Thum, C. C., Lim, T. S., Ahmad, Z., Mahmood, K. A., Shahid, M. H., Ang, C. Y., Tan, W. Q., Tan, P. N., Yee, K. S., Cheaw, W. G., Boey, H. S., Choo, A. L., & Sew, B. C. (2012). A new unmanned aerial vehicle synthetic aperture radar for environmental monitoring. *Progress in Electromagnetics Research*, *122*(September 2011), 245–268. https://doi.org/10.2528/pier11092604

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *INFORMATICA-JOURNAL OF COMPUTING AND INFORMATICS*, *31*(3), 249–268.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine*, *21*(14), 2109–2129. https://doi.org/10.1002/sim.1180

Kristensen, E., Dueholm, L., Vink, D., Andersen, J. E., Jakobsen, E. B., Illum-Nielsen, S., Petersen, F. A., & Enevoldsen, C. (2006). Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. *Journal of Dairy Science*, *89*(9), 3721–3728. https://doi.org/10.3168/jds.S0022-0302(06)72413-4

Kull, J. A., Proudfoot, K. L., Pighetti, G. M., Bewley, J. M., O'Hara, B. F., Donohue, K. D., & Krawczel, P. D. (2019). Effects of acute lying and sleep deprivation on the behavior of lactating dairy cows. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0212823

Lai, E., Danner, A. L., Famula, T. R., & Oberbauer, A. M. (2021). Genome-Wide Association Studies Reveal Susceptibility Loci for Noninfectious Claw Lesions in Holstein Dairy Cattle. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.657375

Lallich, S., Muhlenbach, F., & Zighed, D. A. (2002). *Improving Classification by Removing or Relabeling Mislabeled Instances*.

Lam, C., Yi, D., Guo, M., & Lindsey, T. (2018). Automated Detection of Diabetic Retinopathy using Deep Learning. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2017*, 147–155. http://www.ncbi.nlm.nih.gov/pubmed/29888061

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lange, A., Waiblinger, S., Heinke, A., Barth, K., Futschik, A., & Lürzel, S. (2020). Gentle interactions with restrained and freemoving cows: Effects on the improvement of the animal-human relationship. *PLoS ONE*, *15*(11 November). https://doi.org/10.1371/journal.pone.0242873

Lange, S., & Hammer, A. (1978). Thermal Noise Analysis in Conical-Scan Radars. *IEEE Transactions on Aerospace and Electronic Systems*, *AES-14*(2), 400–403. https://doi.org/10.1109/TAES.1978.308670

Laven, R. (2006). *Nutrition and lameness in dairy cattle - a different perspective ? 11*(2), 6–8.

Laven, R., Huxley, J., Whay, H., & Stafford, K. (2009). Results of a survey of attitudes of dairy veterinarians in New Zealand regarding painful procedures and conditions in cattle. *New Zealand Veterinary Journal*, *57*(4), 215–220. https://doi.org/10.1080/00480169.2009.36904

Leach, K. A., Whay, H. R., Maggs, C. M., Barker, Z. E., Paul, E. S., Bell, A. K., & Main, D. C. J. (2010a). Working towards a reduction in cattle lameness: 1. Understanding barriers to lameness control on dairy farms. *Research in Veterinary Science*, *89*(2), 311–317. https://doi.org/10.1016/j.rvsc.2010.02.014

Leach, K. A., Whay, H. R., Maggs, C. M., Barker, Z. E., Paul, E. S., Bell, A. K., & Main, D. C. J. (2010b). Working towards a reduction in cattle lameness: 1. Understanding barriers to lameness control on dairy farms. *Research in Veterinary Science*, *89*(2), 311–317. https://doi.org/10.1016/j.rvsc.2010.02.014

Leach, K. A., Whay, H. R., Maggs, C. M., Barker, Z. E., Paul, E. S., Bell, A. K., & Main, D. C. J. (2010c). Working towards a reduction in cattle lameness: 2. Understanding dairy farmers' motivations. *Research in Veterinary Science*, *89*(2), 318–323. https://doi.org/10.1016/j.rvsc.2010.02.017

Lean, I. J., Westwood, C. T., Golder, H. M., & Vermunt, J. J. (2013). Impact of nutrition on lameness and claw health in cattle. *Livestock Science*, *156*(1–3), 71–87. https://doi.org/10.1016/j.livsci.2013.06.006

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly: Management Information Systems*, *45*(3), 1501–1525. https://doi.org/10.25300/MISQ/2021/16564

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*, 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Li, C., & Lin, J. (2008). Random body movement cancellation in doppler radar vital sign detection. *IEEE Transactions on Microwave Theory and Techniques*, *56*(12), 3143–3152. https://doi.org/10.1109/TMTT.2008.2007139

Li, H., Mehul, A., Le Kernec, J., Gurbuz, S. Z., & Fioranelli, F. (2021). Sequential Human Gait Classification with Distributed Radar Sensor Fusion. *IEEE Sensors Journal*, *21*(6), 7590–7603. https://doi.org/10.1109/JSEN.2020.3046991

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. In *ACM Computing Surveys* (Vol. 50, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3136625

Li, X., Fioranelli, F., Yang, S., Romain, O., & Le Kernec, J. (2021). *Radar-based hierarchical human activity classification*. *January*, 1373–1379. https://doi.org/10.1049/icp.2021.0566

Li, X., Li, Z., Fioranelli, F., Yang, S., Romain, O., & Le Kernec, J. (2020). Hierarchical radar data analysis for activity and personnel recognition. *Remote Sensing*, *12*(14), 1–22. https://doi.org/10.3390/rs12142237

Li, X. R., & Bar-Shalom, Y. (1993). Design of interacting multiple model algorithm for tracking in air traffic control systems. *Proceedings of the IEEE Conference on Decision and Control*, *1*(3), 906–911. https://doi.org/10.1109/cdc.1993.325013

Li, Z., Le Kernec, J., Abbasi, Q., Fioranelli, F., Yang, S., & Romain, O. (2023a). Radar-based human activity recognition with adaptive thresholding towards resource constrained platforms. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-30631-x

Li, Z., Le Kernec, J., Abbasi, Q., Fioranelli, F., Yang, S., & Romain, O. (2023b). Radar-based human activity recognition with adaptive thresholding towards resource constrained platforms. *Scientific Reports*, *13*(1), 3473. https://doi.org/10.1038/s41598-023-30631-x

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. In *Medical Image Analysis* (Vol. 42, pp. 60–88). Elsevier B.V. https://doi.org/10.1016/j.media.2017.07.005

Liu, X., Gao, K., Liu, B., Pan, C., Liang, K., Yan, L., Ma, J., He, F., Zhang, S., Pan, S., & Yu, Y. (2021). Advances in Deep Learning-Based Medical Image Analysis. *Health Data Science*, *2021*, 1–14. https://doi.org/10.34133/2021/8786793

Livesey, C. T., Marsh, C., Metcalf, J. A., & Laven, R. A. (2002). Hock injuries in cattle kept in straw yards or cubicles with rubber mats or mattresses. *Veterinary Record*, *150*, 677–679. https://doi.org/10.1136/vr.150.22.677

Logue, D. N., Offer, J. E., & Hyslop, J. J. (1994). Relationship of diet, hoof type and locomotion score with lesions of the sole and white line in dairy cattle. *Animal Science*, *59*(2), 173–181. https://doi.org/10.1017/S0003356100007650

Lorenzini, I., Grimm, K., Haidn, B., & Misha, E. (2018). Development of a prediction model for automatic lameness detection in dairy cows. *Arbeitswissenschaftliches Kolloquium 2018*, *March*.

Lorenzini, I., Grimm, K., Haidn, B., & Misha, E. (2019, March). *Advancements in the analysis of behavioural and performance data for early lameness detection in dairy cows*.

Lorenzini, I., Grimm, K., Misha, E., & Haidn, B. (2017). *Using a three point lameness scoring system combined with a clinical examination to increase the reliability of locomotion scoring*. *September*, 1–2.

Lunenburg, F. C. (2010). THE DECISION MAKING PROCESS. In *NATIONAL FORUM OF EDUCATIONAL ADMINISTRATION AND SUPERVISION JOURNAL* (Vol. 27).

Lv, Q., Dong, Y., Sun, Y., Li, C., & Ran, L. (2015). Detection of bio-signals from body movement based on high-dynamic-range Doppler radar sensor (Invited). *2015 IEEE MTT-S International Microwave Workshop Series on RF and Wireless Technologies for Biomedical and Healthcare Applications, IMWS-BIO 2015 - Proceedings*, *2*, 88–89. https://doi.org/10.1109/IMWS-BIO.2015.7303791

Madsen, S. N., Zebker, H. A., & Martin, J. (1993). Topographic Mapping Using Radar Interferometry: Processing Techniques. *IEEE Transactions on Geoscience and Remote Sensing*, *31*(1), 246–256. https://doi.org/10.1109/36.210464

Maertens, W., Baert, J., Van Nuffel, A., Vangeyte, J., Song, X., Berckmans, D., & Sonck, B. (2012). *Spatiotemporal Quadruped Gait Analysis based on Pressure Mat Data-Preliminary Results*.

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

Mahafza, B. R. (2016). *Radar Signal Analysis and Processing Using MATLAB*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420066449

Main, D. C. J., Leach, K. A., Barker, Z. E., Sedgwick, A. K., Maggs, C. M., Bell, N. J., & Whay, H. R. (2012). Evaluating an intervention to reduce lameness in dairy cattle. *Journal of Dairy Science*, *95*(6), 2946–2954. https://doi.org/10.3168/jds.2011-4678

Malchiodi, F., Koeck, A., Mason, S., Christen, A. M., Kelton, D. F., Schenkel, F. S., & Miglior, F. (2017). Genetic parameters for hoof health traits estimated with linear and threshold models using alternative cohorts. *Journal of Dairy Science*, *100*(4), 2828–2836. https://doi.org/10.3168/jds.2016-11558

Manske, T., Bergsten, C., & Hultgren, J. (2003). The Effect of Maintenance Claw Trimming on the Prevalence of Claw Lesions and the Need for Therapeutic Claw Trimming. *Acta Veterinaria Scandinavica*, *44*(Suppl 1), P59. https://doi.org/10.1186/1751-0147-44-s1-p59

Manske, T., Hultgren, J., & Bergsten, C. (2002a). Prevalence and interrelationships of hoof lesions and lameness in Swedish dairy cows. *Preventive Veterinary Medicine*, *54*, 247–263.

Manske, T., Hultgren, J., & Bergsten, C. (2002b). *The effect of claw trimming on the hoof health of Swedish dairy cattle*. *54*, 113–129.

Manson, F. J., & Leaver, J. D. (1988a). The influence of concentrate amount on locomotion and clinical lameness in dairy cattle. *Animal Science*, *47*(2), 185–190. https://doi.org/10.1017/S0003356100003251

Manson, F. J., & Leaver, J. D. (1988b). The influence of concentrate amount on locomotion and clinical lameness in dairy cattle. *Animal P*, *47*, 185–190. https://doi.org/10.1017/S0003356100003251

Manson, F. J., & Leaver, J. D. (1988c). The influence of dietary protein intake and of hoof trimming on lameness in dairy cattle. *Animal Production*, *47*(02), 191–199. https://doi.org/10.1017/S0003356100003263

Manson, F. J., & Leaver, J. D. (1989). The effect of concentrate: silage ratio and of hoof trimming on lameness in dairy cattle. *Animal Science*, *49*(1), 15–22. https://doi.org/10.1017/S0003356100004207

Manteuffel, C. (2019). Parturition detection in sows as test case for measuring activity behaviour in farm animals by means of radar sensors. *Biosystems Engineering*, *184*, 200–206. https://doi.org/10.1016/j.biosystemseng.2019.06.018

Marcatili, P., Nielsen, M. W., Sicheritz-Pontén, T., Jensen, T. K., Schafer-Nielsen, C., Boye, M., Nielsen, M., & Klitgaard, K. (2016). A novel approach to probe host-pathogen interactions of bovine digital dermatitis, a model of a complex polymicrobial infection. *BMC Genomics*, *17*(1). https://doi.org/10.1186/s12864-016-3341-7

March, S., Brinkmann, J., & Winkler, C. (2007). Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. *Animal Welfare*, *16*(2), 131–133.

Marchionatti, E., Fecteau, G., & Desrochers, A. (2014). Traumatic conditions of the coxofemoral joint: Luxation, femoral head-neck fracture, acetabular fracture. In *Veterinary Clinics of North America - Food Animal Practice* (Vol. 30, Issue 1, pp. 247–264). W.B. Saunders. https://doi.org/10.1016/j.cvfa.2013.11.001

Martín-Herrero, J. (2007). Hybrid object labelling in digital images. *Machine Vision and Applications*, *18*(1), 1–15. https://doi.org/10.1007/s00138-006-0041-3

Mason, C., & Offer, J. (2007). *Preventing lameness in dairy cows: Hoof lesions; their identification, treatment, management and prevention*.

MATLAB R2022b. (2022). The MathWorks Inc.

Matsumoto, T., Okumura, S., & Hirata, S. (2022). Non-contact respiratory measurement in a horse in standing position using millimeter-wave array radar. In *Journal of Veterinary Medical Science* (Vol. 84, Issue 10, pp. 1340–1344). Japanese Society of Veterinary Science. https://doi.org/10.1292/jvms.22-0238

Mattia, F., Le Toan, T., Picard, G., Posa, F. I., D'Alessio, A., Notarnicola, C., Gatti, A. M., Rinaldi, M., Satalino, G., & Pasquariello, G. (2003). Multitemporal C-band radar measurements on wheat fields. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(7 PART I), 1551–1560. https://doi.org/10.1109/TGRS.2003.813531

Maule, A. J., Hockey, G. R. J., & Bdzola, L. (2000). Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica*, *104*(3), 283–301. https://doi.org/10.1016/S0001-6918(00)00033-0

Maxwell, O. J. R., Hudson, C. D., & Huxley, J. N. (2015). Effect of early lactation foot trimming in lame and non-lame dairy heifers: A randomised controlled trial. *Veterinary Record*, *177*(4), 100. https://doi.org/10.1136/vr.103155

McCabe, A. (2007). *A Byzantine Encyclopaedia of Horse Medicine : The Sources, Compilation, and Transmission of the Hippiatrica*. Oxford University Press. http://ebookcentral.proquest.com/lib/gla/detail.action?docID=415029

McDaniel, B. T. (1997). *Breeding Programs to Reduce Foot and Leg Problems*.

McGinn, T., Wyer, P. C., Newman, T. B., Keitz, S., Leipzig, R., & Guyatt, G. (2004). Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). In *CMAJ. Canadian Medical Association Journal* (Vol. 171, Issue 11, pp. 1369–1373). Canadian Medical Association. https://doi.org/10.1503/cmaj.1031981

Mcgraw, K. O., & Wong, S. P. (1996). Forming Inferences About Some Intraclass Correlation Coefficients. In *Psychological Methods: Vol. l* (Issue 1).

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://www.ncbi.nlm.nih.gov/pubmed/23092060

McNally, J. C., Crowe, M. A., Roche, J. F., & Beltman, M. E. (2014). Effects of physiological and/or disease status on the response of postpartum dairy cows to synchronization ofestrus using an intravaginal progesterone device. *Theriogenology*, *82*(9), 1263–1272. https://doi.org/10.1016/j.theriogenology.2014.08.006

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). *A Survey on Bias and Fairness in Machine Learning*. http://arxiv.org/abs/1908.09635

Melendez, P., Bartolome, J., Archbald, L. F., & Donovan, A. (2003). The association between lameness, ovarian cysts and fertility in lactating dairy cows. *Theriogenology*, *59*(3–4), 927–937. https://doi.org/10.1016/S0093-691X(02)01152-4

Melendez, P., Gomez, V., Bothe, H., Rodriguez, F., Velez, J., Lopez, H., Bartolome, J., & Archbald, L. (2018). Ultrasonographic ovarian dynamic, plasma progesterone, and non-esterified fatty acids in lame postpartum dairy cows. *Journal of Veterinary Science*, *19*(3), 462. https://doi.org/10.4142/jvs.2018.19.3.462

Mellett, J. S. (1995). Ground penetrating radar applications in engineering, environmental management, and geology. *Journal of Applied Geophysics*, *33*(1–3), 157–166. https://doi.org/10.1016/0926-9851(95)90038-1

Melvin, W. L., & Scheer, J. (2010). *Principles of Modern Radar: Basic principles* (M. A. Richards, J. A. Scheer, & W. A. Holm, Eds.). Institution of Engineering and Technology. https://doi.org/10.1049/SBRA021E

Menke, C., Waiblinger, S., Fölsch, D. W., & Wiepkema, P. R. (1999). Social Behaviour and Injuries of Horned Cows in Loose Housing Systems. *Animal Welfare*, *8*(3), 243–258. https://doi.org/10.1017/S0962728600021734

Merriam-Webster. (n.d.-a). *"Assessor". In Merriam-Webster.com dictionary. Retrieved April 6, 2023, from https://www.merriam-webster.com/dictionary/assessor*.

Merriam-Webster. (n.d.-b). *"Binary", retrieved April 6, 2023 https://www.merriam-webster.com/dictionary/binary*. Merriam-Webster.Com Dictionary.

Merriam-Webster. (n.d.-c). *"Dichotomize", retrieved April 6 2023, from https://www.merriam-webster.com/dictionary/dichotomize*.

Merriam-Webster. (n.d.-d). *"Evaluator". In Merriam-Webster.com dictionary. Retrieved April 6, 2023, from https://www.merriam-webster.com/dictionary/evaluate*.

Merriam-Webster. (n.d.-e). *"Rater". In Merriam-Webster.com dictionary. Retrieved April 6, 2023, from https://www.merriam-webster.com/dictionary/rater*.

Milian-Suazo, F., Erb, H. N., & David Smith, R. (1989). Risk Factors for Reason-specific Culling of Dairy Cows. In *Preventive Veterinary Medicine* (Vol. 7).

Mitev, J., Gergovska, Z., Miteva, T., & Penev, T. (2011). Influence of lameness on daily milk yield, lactation curve and body condition score during lactation in Black-and White cows. *Bulgarian Journal of Agricultural Science*, *17*(5), 704–711.

Mokaram Ghotoorlar, S., Mehdi Ghamsari, S., Nowrouzian, I., & Shiry Ghidary, S. (2012). Lameness scoring system for dairy cows using force plates and artificial intelligence. *Veterinary Record*, *170*(5), 126. https://doi.org/10.1136/vr.100429

Monrad, J., Kassuku, A. A., Nansen, P., & Willeberg, P. (1983). AN EPIDEMIOLOGICAL STUDY OF FOOT ROT IN PASTURED CATTLE. In *Acta vet. scand* (Vol. 24).

Morris, M. J., Kaneko, K., Walker, S. L., Jones, D. N., Routly, J. E., Smith, R. F., & Dobson, H. (2011). Influence of lameness on follicular growth, ovulation, reproductive hormone concentrations and estrus behavior in dairy cows.

Theriogenology, *76*(4), 658–668. https://doi.org/10.1016/j.theriogenology.2011.03.019

Morton, D. B., & Griffiths, P. H. (1985). Guidelines on the recognition of pain, distress and discomfort in experimental animals and an hypothesis for assessment. *The Veterinary Record*, *116*(16), 431–436. https://doi.org/10.1136/vr.116.16.431

Mukaka, M. M. (2012). Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. In *Malawi Medical Journal* (Vol. 24, Issue 3). www.mmj.medcol.mw

Mülleder, C., Waiblinger, S., & Troxler, J. (2004). *Analyse der Einflussfaktoren auf Tiergerechtheit, Tiergesundheit & Leistung von Milchkühen im Boxenlaufstall*.

Mülling, C. K. W., Green, L., Barker, Z., Scaife, J., Amory, J., & Speijers, M. (2006). *Risk factors associated with foot lameness in dairy cattle and a suggested approach for lameness reduction. Christoph K.W. Mülling et al. - Proceedings of World Buiatrics Congress - Nice 2006*. http://www.ivis.org

Mulsant, B. H., Kastango, K. B., Rosen, J., Stone, R. A., Mazumdar, S., & Pollock, B. G. (2002). Interrater reliability in clinical trials of depressive disorders. *The American Journal of Psychiatry*, *159*(9), 1598. https://doi.org/10.1176/appi.ajp.159.9.1598

Muñoz-Ferreras, J. M., Pérez-Martínez, F., Calvo-Gallego, J., Asensio-López, A., Dorta-Naranjo, B. P., & Blanco-del-Campo, A. (2008). Traffic surveillance system based on a high-resolution radar. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(6), 1624–1633. https://doi.org/10.1109/TGRS.2008.916465

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press. https://go.exlibris.link/gSy2Mxkh

Murray, R. D., Downham, D. Y., Clarkson, M. J., Faull, W. B., Hughes, J. W., Manson, F. J., Merritt, J. B., Russell, W. B., Sutherst, J. E., & Ward, W. R.

(1996). *Epidemiology of lameness in dairy cattle : description and analysis of foot lesions. 1978*, 586–591.

Murray, R. D., Russell, W. B., Sutherst, J. E., & Ward, W. R. (1996). Epidemiology of lameness in dairy cattle : the influence of cubicles and indoor and outdoor walking surfaces. *Veterinary Record*, 130–136.

Neichev, O., Bodurov, N., Binev, K., Petrov, M., & Filipov, Z. (1980). Economic effect on the milk yield in cows after trimming and treating their hooves. *Veterinarno-meditsinski nauki*, *17*(5), 38–44.

Nguyen, D., Yamada, S., Park, B.-K., Lubecke, V., Boric-Lubecke, O., & Host-Madsen, A. (2007). Noise considerations for remote detection of life signs with microwave Doppler radar. *Conference Proceedings (IEEE Engineering in Medicine and Biology Society. Conf.)*, *2007*, 1667. https://doi.org/10.1109/IEMBS.2007.4352628

Nisbet, R., Miner, G., & Yale, K. (2018). *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier. https://doi.org/10.1016/C2012-0-06451-4

Nocek, J. E. (1997). Bovine Acidosis: Implications on Laminitis. *Journal of Dairy Science*, *80*(5), 1005–1028. https://doi.org/10.3168/jds.S0022-0302(97)76026-0

Oakden-Rayner, L. (2019). The rebirth of cad: How is modern ai different from the cad we know? *Radiology: Artificial Intelligence*, *1*(3). https://doi.org/10.1148/ryai.2019180089

Ødegård, C., Svendsen, M., & Heringstad, B. (2013). Genetic analyses of claw health in Norwegian Red cows. *Journal of Dairy Science*, *96*(11), 7274–7283. https://doi.org/10.3168/jds.2012-6509

Olechnowicz, J., & Jaśkowski, J. M. (2010). Risk factors influencing lameness and key areas in reduction of lameness in dairy cows. *Medycyna Weterynaryjna*, *66*(8), 507–511.

Olechnowicz, J., & Jaskowski, J. M. (2011). Behaviour of lame cows: A review. *Veterinarni Medicina*, *56*(12), 581–588. https://doi.org/10.17221/4435-VETMED

Olofsson, J. (1999). Competition for total mixed diets fed for ad libitum intake using one or four cows per feeding station. *Journal of Dairy Science*, *82*(1), 69–79. https://doi.org/10.3168/jds.S0022-0302(99)75210-0

Olsen, K. E., & Asen, W. (2017). Bridging the gap between civilian and military passive radar. *IEEE Aerospace and Electronic Systems Magazine*, *32*(2), 4–12. https://doi.org/10.1109/MAES.2017.160030

Omontese, B. O., Bellet-Elias, R., Molinero, A., Catandi, G. D., Casagrande, R., Rodriguez, Z., Bisinotto, R. S., & Cramer, G. (2020). Association between hoof lesions and fertility in lactating Jersey cows. *Journal of Dairy Science*, *103*(4), 3401–3413. https://doi.org/10.3168/jds.2019-17252

Onyiro, O. M., Andrews, L. J., & Brotherstone, S. (2008). Genetic Parameters for Digital Dermatitis and Correlations with Locomotion, Production, Fertility Traits, and Longevity in Holstein-Friesian Dairy Cows. *Journal of Dairy Science*, *91*(10), 4037–4046. https://doi.org/10.3168/jds.2008-1190

Ossent, P., & Lischer, C. (1998). Bovine laminitis: The lesions and their pathogenesis. *In Practice*, *20*(8), 415–427. https://doi.org/10.1136/inpract.20.8.415

P. G. Rajkondawar, A. M. Lefcourt, N. K. Neerchal, R. M. Dyer, M. A. Varner, B. Erez, & U. Tasch. (2002). THE DEVELOPMENT OF AN OBJECTIVE LAMENESS SCORING SYSTEM FOR DAIRY HERDS: PILOT STUDY. *Transactions of the ASAE*, *45*(4). https://doi.org/10.13031/2013.9941

Palmer, M. A., & O'Connell, N. E. (2015). Digital dermatitis in dairy cows: A review of risk factors and potential sources of between-animal variation in susceptibility. In *Animals* (Vol. 5, Issue 3, pp. 512–535). MDPI AG. https://doi.org/10.3390/ani5030369

Pandey, S., & Singh, J. (2015). A 0.6 V, low-power and high-gain ultra-wideband low-noise amplifier with forward-body-bias technique for low-voltage operations. *IET Microwaves, Antennas and Propagation*, *9*(8), 728–734. https://doi.org/10.1049/iet-map.2014.0581

Parsons, S. (2010). Introduction to Machine Learning, Second Editon by Ethem Alpaydin, MIT Press, 584 pp., $55.00. ISBN 978-0-262-01243-0. *The Knowledge Engineering Review*, *25*(3), 353–353. https://doi.org/10.1017/S0269888910000056

Pastell, M. E., & Kujalaf, M. (2007). A probabilistic neural network model for lameness detection. *Journal of Dairy Science*, *90*(5), 2283–2292. https://doi.org/10.3168/jds.2006-267

Pastell, M., Takko, H., Gröhn, H., Hautala, M., Poikalainen, V., Praks, J., Veermäe, I., Kujala, M., & Ahokas, J. (2006). Assessing cows' welfare: Weighing the cow in a milking robot. *Biosystems Engineering*, *93*(1), 81–87. https://doi.org/10.1016/j.biosystemseng.2005.09.009

Pastell, M., Tiusanen, J., Hakojärvi, M., & Hänninen, L. (2009). A wireless accelerometer system with wavelet analysis for assessing lameness in cattle. *Biosystems Engineering*, *104*(4), 545–551. https://doi.org/10.1016/j.biosystemseng.2009.09.007

Patel, H., Singh Rajput, D., Thippa Reddy, G., Iwendi, C., Kashif Bashir, A., & Jo, O. (2020). A review on classification of imbalanced data for wireless sensor networks. In *International Journal of Distributed Sensor Networks* (Vol. 16, Issue 4). SAGE Publications Ltd. https://doi.org/10.1177/1550147720916404

Pedersen, S. I. L., Huxley, J. N., Hudson, C. D., Green, M. J., & Bell, N. J. (2022). Preventive hoof trimming in dairy cattle: Determining current practices and identifying future research areas. *Veterinary Record*, *190*(5). https://doi.org/10.1002/vetr.1267

Pennisi, E. (2011). Researchers use weather radar to track bat movements. In *Science* (Vol. 331, Issue 6020, p. 998). https://doi.org/10.1126/science.331.6020.998

Pérez, J. M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., & Martín, J. I. (2005). *Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance* (pp. 381–389). https://doi.org/10.1007/11551188_41

Pérez-Cabal, M. A., & Alenda, R. (2014). Clinical lameness and risk factors in a Spanish Holstein population. *Livestock Science*, *164*(1), 168–174. https://doi.org/10.1016/j.livsci.2014.03.012

Pérez-Cabal, M. A., García, C., González-Recio, O., & Alenda, R. (2006). Genetic and phenotypic relationships among locomotion type traits, profit, production, longevity, and fertility in Spanish dairy cows. *Journal of Dairy Science*, *89*(5), 1776–1783. https://doi.org/10.3168/jds.S0022-0302(06)72246-9

Pesce, E., Joseph Withey, S., Ypsilantis, P. P., Bakewell, R., Goh, V., & Montana, G. (2019). Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Medical Image Analysis*, *53*, 26–38. https://doi.org/10.1016/j.media.2018.12.007

Phillips, C. J. C., Chiy, P. C., Bucktrout, M. J., Collins, S. M., Gasson, C. J., Jenkins, A. C., & Costa, M. J. R. P. D. A. (2000). Frictional properties of cattle hooves and their conformation after trimming. *Veterinary Re*, *146*, 607–609.

Phillips, C. J. C., & Morris, I. D. (2000). The Locomotion of Dairy Cows on Concrete Floors That are Dry , Wet , or Covered with a Slurry of Excreta. *Journal of Dairy Science*, *83*(8), 1767–1772. https://doi.org/10.3168/jds.S0022-0302(00)75047-8

Phillips, C. J. C., & Morris, I. D. (2001). The Locomotion of Dairy Cows on Floor Surfaces with Different Frictional Properties. *Journal of Dairy Science*, *84*(3), 623–628. https://doi.org/10.3168/jds.S0022-0302(01)74517-1

Pitkethly, M. J. (1992). Radar Absorbing Materials and their Potential use in Aircraft Structures. *Low Profile Absorbers and Scatterers*, 7–10.

Plaizier, J. C., Krause, D. O., Gozho, G. N., & McBride, B. W. (2008). Subacute ruminal acidosis in dairy cows: The physiological causes, incidence and consequences. *Veterinary Journal*, *176*(1), 21–31. https://doi.org/10.1016/j.tvjl.2007.12.016

Polderman, K. H., Jorna, E. M., & Girbes, A. R. (2001). Inter-observer variability in APACHE II scoring: Effect of strict guidelines and training. *Intensive Care Medicine*, *27*(8), 1365–1369. https://doi.org/10.1007/s001340101012

Pötzsch, C. J., Collis, V. J. ;, Blowey, R. W. ;, Packington, A. J. ;, & Green, L. E. (2003). The Impact of Parity and Duration of Biotin Supplementation on White Line Disease Lameness in Dairy Cattle ABSTRACT. *Journal of Dairy Science*, *86*(8), 2577–2582. https://doi.org/10.3168/jds.S0022-0302(03)73852-1

Poursaberi, A., Bahr, C., Pluk, A., Van Nuffel, A., & Berckmans, D. (2010). Real-time automatic lameness detection based on back posture extraction in dairy cattle: Shape analysis of cow with image processing techniques. *Computers and Electronics in Agriculture*, *74*(1), 110–119. https://doi.org/10.1016/j.compag.2010.07.004

Proakis, J. G., & Salehi, M. (2002). *Communication systems engineering* (2nd ed.). Prentice Hall. https://go.exlibris.link/lNRBvGGj

Proudfoot, K. L., Veira, D. M., Weary, D. M., & Keyserlingk, M. A. G. Von. (2009). Competition at the feed bunk changes the feeding , standing , and social behavior of transition dairy cows. *Journal of Dairy Science*, *92*(7), 3116–3123. https://doi.org/10.3168/jds.2008-1718

R Core Team. (2020). *R: A Language and Environment for Statistical Computing* (Version 1.2.1335). R Foundation for Statistical Computing. http://www.r-project.org/

Rahman, G., Islam, Z., & Rahman, M. G. (2013). *Data Quality Improvement by Imputation of Missing Values*. https://www.researchgate.net/publication/236635604

Rajala-Schultz, P. J., & Gro Èhn, Y. T. (1999). *Culling of dairy cows. Part II. Effects of diseases and reproductive performance on culling in Finnish Ayrshire cows*.

Rajkondawar, P. G., Liu, M., Dyer, R. M., Neerchal, N. K., Tasch, U., Lefcourt, A. M., Erez, B., & Varner, M. A. (2006). Comparison of Models to Identify Lame Cows Based on Gait and Lesion Scores , and Limb Movement Variables. *Journal*

of Dairy Science, *89*(11), 4267–4275. https://doi.org/10.3168/jds.S0022-0302(06)72473-0

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., … Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, *15*(11). https://doi.org/10.1371/journal.pmed.1002686

Randall, L., Green, M. J., Chagunda, M. G. G., Mason, C., Archer, S. C., Green, L. E., & Huxley, J. N. (2015). Low body condition predisposes cattle to lameness: An 8-year study of one dairy herd. *Journal of Dairy Science*, *98*(6), 3766–3777. https://doi.org/10.3168/jds.2014-8863

Randall, L. V., Green, M. J., Chagunda, M. G. G., Mason, C., Green, L. E., & Huxley, J. N. (2016). Lameness in dairy heifers; impacts of hoof lesions present around first calving on future lameness, milk yield and culling risk. *Preventive Veterinary Medicine*, *133*, 52–63. https://doi.org/10.1016/j.prevetmed.2016.09.006

Randhawa, S. S., Dua, K., Dhaliwal, P. S., Uppal, S. K., & Singh, S. T. (2008). Effect of formalin footbathing on the prevalence of foot lesions and conformational indices in dairy cattle. *Veterinary Record*, *163*(11), 335–337. https://doi.org/10.1136/vr.163.11.335

Rashad, A. M. A., Kohla, A. A., Aziz, M. A., & El-Hedainy, D. K. A. (2022). Prevalence and risk factors of lameness in dairy cattle in Alexandria, Egypt. *Spanish Journal of Agricultural Research*, *20*(1), 1–8. https://doi.org/10.5424/sjar/2022201-18245

Ratzinger, F., Haslacher, H., Perkmann, T., Pinzan, M., Anner, P., Makristathis, A., Burgmann, H., Heinze, G., & Dorffner, G. (2018). Machine learning for fast identification of bacteraemia in SIRS patients treated on standard care wards: a cohort study. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-30236-9

Raven, T. E. (2003). *CATTLE FOOTCARE AND CLAW TRIMMING* (Aloys Lurvink, Ed.; 3rd, illustr ed.). Farming Press, 1985.

Reamaroon, N., Sjoding, M. W., Lin, K., Iwashyna, T. J., & Najarian, K. (2019). Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome. *IEEE Journal of Biomedical and Health Informatics*, *23*(1), 407–415. https://doi.org/10.1109/JBHI.2018.2810820

Reicher, R. (1985). [Serious mistakes in hoof trimming of cattle by the layman]. *Tierarztliche Praxis*, *13*(3), 291–294.

Reinemann, D. (2007). *Stray voltage field guide*.

Remnant, J. G., Tremlett, A., Huxley, J. N., & Hudson, C. D. (2017). Clinician attitudes to pain and use of analgesia in cattle: where are we 10 years on? *Veterinary Record*, *181*(15), 400–400. https://doi.org/10.1136/vr.104428

Ren, Y., Zhang, X., Ma, Y., Yang, Q., Wang, C., Liu, H., & Qi, Q. (2020). Full Convolutional Neural Network Based on Multi-Scale Feature Fusion for the Class Imbalance Remote Sensing Image Classification. *Remote Sensing*, *12*(21), 3547. https://doi.org/10.3390/rs12213547

Reynolds, M. H., Cooper, B. A., & Day, R. H. (1997). Radar study of seabirds and bats on windward Hawai'i. *Pacific Science*, *51*(1), 97–106.

Robcis, R., Ferchiou, A., Berrada, M., Ndiaye, Y., Herman, N., Lhermie, G., & Raboisson, D. (2023). Cost of lameness in dairy herds: An integrated bioeconomic modeling approach. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2022-22446

Ross, M. W. ;, & Dyson, S. J. (2010). *Diagnosis and Management of Lameness in the Horse* (2nd ed.).

Rowlands, G. J., Russell, A. M., & Williams, L. A. (1983). Effects of season, herd size, management system and veterinary practice on the lameness incidence in dairy cattle. *The Veterinary Record*, *113*(19), 441–445. https://doi.org/10.1136/vr.113.19.441

Ruan, W. J., Goldstein, R. B., Chou, S. P., Smith, S. M., Saha, T. D., Pickering, R. P., Dawson, D. A., Huang, B., Stinson, F. S., & Grant, B. F. (2008). The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): Reliability of new psychiatric diagnostic modules and risk factors in a general population sample. *Drug and Alcohol Dependence*, *92*(1–3), 27–36. https://doi.org/10.1016/j.drugalcdep.2007.06.001

Rudge, A. W., & Adatia, N. A. (1978). Offset-Parabolic-Reflector Antennas: A Review. *Proceedings of the IEEE*, *66*(12), 1592–1618. https://doi.org/10.1109/PROC.1978.11170

Rushen, J., Haley, D., & De Passille, A. M. (2007). Effect of softer flooring in tie stalls on resting behavior and leg injuries of lactating cows. *Journal of Dairy Science*, *90*(8), 3647–3651. https://doi.org/10.3168/jds.2006-463

Russell, A. M., Rowlands, G. J., Shaw, S. R., & Weaver, A. D. (1982). Survey of lameness in British dairy cattle. *The Veterinary Record*, *111*(8), 155–160. https://doi.org/10.1136/vr.111.8.155

Russello, H., van der Tol, R., & Kootstra, G. (2022). T-LEAP: Occlusion-robust pose estimation of walking cows using temporal information. *Computers and Electronics in Agriculture*, *192*. https://doi.org/10.1016/j.compag.2021.106559

Russo, J. E., & Dosher, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 676–696. https://doi.org/10.1037/0278-7393.9.4.676

Rutherford, K. M. D., Langford, F. M., Jack, M. C., Sherwood, L., Lawrence, A. B., & Haskell, M. J. (2009). Lameness prevalence and risk factors in organic and non-organic dairy herds in the United Kingdom. *Veterinary Journal*, *180*(1), 95–105. https://doi.org/10.1016/j.tvjl.2008.03.015

Sadiq, M. B., Ramanoon, S. Z., Mansor, R., Syed-Hussain, S. S., & Mossadeq, W. M. S. (2020). Claw trimming as a lameness management practice and the association with welfare and production in dairy cows. In *Animals* (Vol. 10, Issue 9, pp. 1–18). MDPI AG. https://doi.org/10.3390/ani10091515

Salau, A. O., & Jain, S. (2019). Feature Extraction: A Survey of the Types, Techniques, Applications. *2019 International Conference on Signal Processing and Communication (ICSC)*, 158–164. https://doi.org/10.1109/ICSC45622.2019.8938371

Salleh, M. S., Mazzoni, G., Höglund, J. K., Olijhoek, D. W., Lund, P., Løvendahl, P., & Kadarmideen, H. N. (2017). RNA-Seq transcriptomics and pathway analyses reveal potential regulatory genes and molecular mechanisms in high- and low-residual feed intake in Nordic dairy cattle. *BMC Genomics*, *18*(1), 1–17. https://doi.org/10.1186/s12864-017-3622-9

Sánchez-Molano, E., Bay, V., Smith, R. F., Oikonomou, G., & Banos, G. (2019). Quantitative Trait Loci Mapping for Lameness Associated Phenotypes in Holstein–Friesian Dairy Cattle. *Frontiers in Genetics*, *10*(October), 1–9. https://doi.org/10.3389/fgene.2019.00926

Sarjokari, K., Kaustell, K. O., Hurme, T., Kivinen, T., Peltoniemi, O. A. T., Saloniemi, H., & Rajala-Schultz, P. J. (2013). Prevalence and risk factors for lameness in insulated free stall barns in Finland. *Livestock Science*, *156*(1–3), 44–52. https://doi.org/10.1016/j.livsci.2013.06.010

Šárová, R., Stěhulová, I., Kratinová, P., Firla, P., & Špinka, M. (2011a). Farm managers underestimate lameness prevalence in Czech dairy herds. *Animal Welfare*, *20*(2), 201–204.

Šárová, R., Stěhulová, I., Kratinová, P., Firla, P., & Špinka, M. (2011b). Farm managers underestimate lameness prevalence in Czech dairy herds. *Animal Welfare*, *20*(2), 201–204.

Schlageter-Tello, A., Bokkers, E. A. M., Groot Koerkamp, P. W. G., Van Hertem, T., Viazzi, S., Romanini, C. E. B., Halachmi, I., Bahr, C., Berckmans, D., & Lokhorst, K. (2014). Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. *Journal of Dairy Science*, *97*(9), 5533–5542. https://doi.org/10.3168/jds.2014-8129

Schlageter-Tello, A., Bokkers, E. A. M., Groot Koerkamp, P. W. G., Van Hertem, T., Viazzi, S., Romanini, C. E. B., Halachmi, I., Bahr, C., Berckmans, D., & Lokhorst, K. (2015a). Comparison of locomotion scoring for dairy cows by

experienced and inexperienced raters using live or video observation methods. *Animal Welfare*, *24*(1), 69–79. https://doi.org/10.7120/09627286.24.1.069

Schlageter-Tello, A., Bokkers, E. A. M., Groot Koerkamp, P. W. G., Van Hertem, T., Viazzi, S., Romanini, C. E. B., Halachmi, I., Bahr, C., Berckmans, D., & Lokhorst, K. (2015b). Relation between observed locomotion traits and locomotion score in dairy cows. *Journal of Dairy Science*, *98*(12), 8623–8633. https://doi.org/10.3168/jds.2014-9059

Schleicher, B., Nasr, I., Trasser, A., & Schumacher, H. (2013). IR-UWB radar demonstrator for ultra-fine movement detection and vital-sign monitoring. *IEEE Transactions on Microwave Theory and Techniques*, *61*(5), 2076–2085. https://doi.org/10.1109/TMTT.2013.2252185

Schoenfeld, A. H. (2011). *How We Think: A Theory of Goal-oriented Decision Making and Its Educational Applications*.

Schröder, M., Yordanova, K., Bader, S., & Kirste, T. (2016). Tool support for the online annotation of sensor data. *ACM International Conference Proceeding Series*, *23-24-June-2016*. https://doi.org/10.1145/2948963.2948972

Schwarz, S., & Chaslus-Dancla, E. (2001). Use of antimicrobials in veterinary medicine and mechanisms of resistance. In *Vet. Res* (Vol. 32).

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

Semmel, A. K. (1979). Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment. By Irving L. Janis and Leon Mann. (New York: Free Press, 1977. Pp. xx + 488. $15.95.). *American Political Science Review*, *73*(1), 213–214. https://doi.org/10.2307/1954755

Sen, T., & Boe, W. (1991). Confidence and accuracy in judgements using computer displayed information. *Behaviour and Information Technology*, *10*(1), 53–64. https://doi.org/10.1080/01449299108924271

Shahinfar, S., Khansefid, M., Haile-Mariam, M., & Pryce, J. E. (2021). Machine learning approaches for the prediction of lameness in dairy cows. *Animal*, *15*(11). https://doi.org/10.1016/j.animal.2021.100391

Sharma, M., Purohit, G. N., & Mukherjee, S. (2018). Information retrieves from brain MRI images for tumor detection using hybrid technique k-means and artificial neural network (KMANN). In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 4, pp. 145–157). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-981-10-4600-1_14

Shearer, J. K., & Amstel, S. R. Van. (2001). FUNCTIONAL AND CORRECTIVE CLAW TRIMMING. *Veterinary Clinics of North America: Food Animal Practice*, *17*(1), 53–72. https://doi.org/10.1016/S0749-0720(15)30054-2

Shearer, J. K., Stock, M. L., Amstel, S. R. Van, & Vetc, M. M. E. D. (2013). *Assessment and Management of Pain Associated with Lameness in Cattle*. *29*, 135–156. https://doi.org/10.1016/j.cvfa.2012.11.012

Shearer, J. K., & Van Amstel, S. R. (2001). *FUNCTIONAL AND CORRECTIVE CLAW TRIMMING*.

Shearer, J. K., & van Amstel, S. R. (2017a). Pathogenesis and Treatment of Sole Ulcers and White Line Disease. In *Veterinary Clinics of North America - Food Animal Practice* (Vol. 33, Issue 2, pp. 283–300). W.B. Saunders. https://doi.org/10.1016/j.cvfa.2017.03.001

Shearer, J. K., & van Amstel, S. R. (2017b). Pathogenesis and Treatment of Sole Ulcers and White Line Disease. *Veterinary Clinics of North America - Food Animal Practice*, *33*(2), 283–300. https://doi.org/10.1016/j.cvfa.2017.03.001

Shearer, J. K., Van Amstel, S. R., & Brodersen, B. W. (2012). Clinical Diagnosis of Foot and Leg Lameness in Cattle. In *Veterinary Clinics of North America - Food Animal Practice* (Vol. 28, Issue 3, pp. 535–556). https://doi.org/10.1016/j.cvfa.2012.07.003

Sheng, Y., Dong, D., He, G., & Zhang, J. (2022). How Noise Can Influence Experience-Based Decision-Making under Different Types of the Provided

Information. *International Journal of Environmental Research and Public Health*, *19*(16). https://doi.org/10.3390/ijerph191610445

Shepard, E. L. C., Wilson, R. P., Quintana, F., Laich, A. G., Liebsch, N., Albareda, D. A., Halsey, L. G., Gleiss, A., Morgan, D. T., Myers, A. E., Newman, C., & Macdonald, D. W. (2010). Identification of animal movement patterns using tri-axial accelerometry. *Endangered Species Research*, *10*(1), 47–60. https://doi.org/10.3354/esr00084

Shi, J., & Wu, J. (2021). *Distilling effective supervision for robust medical image segmentation with noisy labels*. http://arxiv.org/abs/2106.11099

Shin, D. K., Ahmed, M. U., & Rhee, P. K. (2018). Incremental deep learning for robust object detection in unknown cluttered environments. *IEEE Access*, *6*, 61748–61760. https://doi.org/10.1109/ACCESS.2018.2875720

Shrestha, A., Le Kernec, J., Fioranelli, F., Marshall, J. F., & Voute, L. (2017, May 17). Gait analysis of horses for lameness detection with radar sensors. *International Conference on Radar Systems (Radar 2017)*. https://doi.org/10.1049/cp.2017.0427

Shrestha, A., Loukas, C., le Kernec, J., Fioranelli, F., Busin, V., Jonsson, N., King, G., Tomlinson, M., Viora, L., & Voute, L. (2018). Animal Lameness Detection With Radar Sensing. *IEEE Geoscience and Remote Sensing Letters*, *15*(8), 1189–1193. https://doi.org/10.1109/LGRS.2018.2832650

Skolnik, M. I. (2001). *Introduction to radar systems* (Third). McGraw-Hill. https://go.exlibris.link/c3DyqkgP

Smart, M. E. (1985). Nutritional Factors of Lameness and Metabolic Bone Disease in Cattle. *Veterinary Clinics of North America: Food Animal Practic*, *1*(1). https://doi.org/10.1016/S0749-0720(15)31347-5

Sogstad, Å. M., Fjeldaas, T., & Østerås, O. (2005). Lameness and Claw Lesions of the Norwegian Red Dairy Cattle Housed in Free Stalls in Relation to Environment, Parity and Stage of Lactation. In *Acta vet. scand* (Vol. 46, Issue 4).

Sogstad, Å. M., Østerås, O., Fjeldaas, T., & Nafstad, O. (2007). Bovine claw and limb disorders related to culling and carcass characteristics. *Livestock Science*, *106*(1), 87–95. https://doi.org/10.1016/j.livsci.2006.07.003

Solano, L., Barkema, H. W., Pajor, E. A., Mason, S., LeBlanc, S. J., Zaffino Heyerhoff, J. C., Nash, C. G. R., Haley, D. B., Vasseur, E., Pellerin, D., Rushen, J., de Passillé, A. M., & Orsel, K. (2015a). Prevalence of lameness and associated risk factors in Canadian Holstein-Friesian cows housed in freestall barns. *Journal of Dairy Science*, *98*(10), 6978–6991. https://doi.org/10.3168/jds.2015-9652

Solano, L., Barkema, H. W., Pajor, E. A., Mason, S., LeBlanc, S. J., Zaffino Heyerhoff, J. C., Nash, C. G. R., Haley, D. B., Vasseur, E., Pellerin, D., Rushen, J., de Passillé, A. M., & Orsel, K. (2015b). Prevalence of lameness and associated risk factors in Canadian Holstein-Friesian cows housed in freestall barns. *Journal of Dairy Science*, *98*(10), 6978–6991. https://doi.org/10.3168/jds.2015-9652

Somers, J. G. C. J., Frankena, K., Noordhuizen-Stassen, E. N., & Metz, J. H. M. (2003). Prevalence of claw disorders in Dutch dairy cows exposed to several floor systems. *Journal of Dairy Science*, *86*(6), 2082–2093. https://doi.org/10.3168/jds.S0022-0302(03)73797-7

Sonka, M., Hlavac, V., & Boyle Roger. (2008). *Image Processing, Analysis, and Machine Vision*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://kgut.ac.ir/useruploads/1550563201478ety.pdf

Soumekh, M. (1996). Slant Plane Circular SAR Imaging. *IEEE Transactions on Image Processing*, *5*(8), 1252–1265.

Sprecher, D. J., Hostetler, D. E., & Kaneene, J. B. (1997). A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. *Theriogenology*, *47*(6), 1179–1187. https://doi.org/10.1016/S0093-691X(97)00098-8

Stashak, T. S. (2008). Adams and Stashak's lameness in horses. In *Adams' Lahmheit bei Pferden*. Verlag M. & H. Schaper.

Stoddard, G. C., & Cramer, G. (2017). A Review of the Relationship Between Hoof Trimming and Dairy Cattle Welfare. *Veterinary Clinics of North America - Food Animal Practice*, *33*(2), 365–375. https://doi.org/10.1016/j.cvfa.2017.02.012

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812–820. https://doi.org/10.1109/TII.2014.2349359

Szeghalmy, S., & Fazekas, A. (2023). A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. *Sensors (Basel, Switzerland)*, *23*(4). https://doi.org/10.3390/s23042333

Szklo, M., Nieto, F. J., & (Firm), P. (2019). *Epidemiology: beyond the basics* (Fourth). Jones & Bartlett Learning. https://go.exlibris.link/NBLqN8xC

Tadich, N., Flor, E., & Green, L. (2010a). Associations between hoof lesions and locomotion score in 1098 unsound dairy cows. *Veterinary Journal*, *184*(1), 60–65. https://doi.org/10.1016/j.tvjl.2009.01.005

Tadich, N., Flor, E., & Green, L. (2010b). Associations between hoof lesions and locomotion score in 1098 unsound dairy cows. *Veterinary Journal*, *184*(1), 60–65. https://doi.org/10.1016/j.tvjl.2009.01.005

Taneja, M., Byabazaire, J., Jalodia, N., Davy, A., Olariu, C., & Malone, P. (2020). Machine learning based fog computing assisted data-driven approach for early lameness detection in dairy cattle. *Computers and Electronics in Agriculture*, *171*(March), 105286. https://doi.org/10.1016/j.compag.2020.105286

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00349-y

Teed, N., Lund, A. K., & Knoblauch, R. (1993). The duration of speed reductions attributable to radar detectors*. *Accident Analysis and Prevention*, *25*(2), 131–137. https://doi.org/10.1016/0001-4575(93)90052-X

Telezhenko, E., & Bergsten, C. (2005). Influence of floor type on the locomotion of dairy cows. *Applied Animal Behaviour Science*, *93*(3–4), 183–197. https://doi.org/10.1016/j.applanim.2004.11.021

Telezhenko, E., Bergsten, C., Magnusson, M., & Nilsson, C. (2009). Effect of different flooring systems on claw conformation of dairy cows. *Journal of Dairy Science*, *92*(6), 2625–2633. https://doi.org/10.3168/jds.2008-1798

Ternman, E., Hänninen, L., Pastell, M., Agenäs, S., & Nielsen, P. P. (2012). Sleep in dairy cows recorded with a non-invasive EEG technique. *Applied Animal Behaviour Science*, *140*(1–2), 25–32. https://doi.org/10.1016/j.applanim.2012.05.005

Tesco Animal Health and Welfare Reporting 2020/21. (n.d.). Retrieved January 31, 2023, from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.tescoplc.com/media/757848/tesco-animal-health-and-welfare-reporting-202021.pdf

Thompson, Thomas S. ; Baron, R. O. Sr. (1998). *SYSTEM AND METHOD PROVIDING FOR REAL-TIME WEATHER TRACKING AND STORM MOVEMENT PREDCTION*.

Thomsen, P. T., Anneberg, I., & Herskin, M. S. (2012). Differences in attitudes of farmers and veterinarians towards pain in dairy cows. *Veterinary Journal*, *194*(1), 94–97. https://doi.org/10.1016/j.tvjl.2012.02.025

Thomsen, P. T., & Baadsgaard, N. P. (2006). Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows. *Preventive Veterinary Medicine*, *75*(1–2), 133–139. https://doi.org/10.1016/j.prevetmed.2006.02.004

Thomsen, P. T., Foldager, L., Raundal, P., & Capion, N. (2019). Lower odds of sole ulcers in the following lactation in dairy cows that received hoof trimming around drying off. *Veterinary Journal*, *254*. https://doi.org/10.1016/j.tvjl.2019.105408

Thomsen, P. T., Munksgaard, L., & Togersen, F. A. (2008). Evaluation of a lameness scoring system for dairy cows. *Journal of Dairy Science*, *91*(1), 119–126. https://doi.org/10.3168/jds.2007-0496

Tijssen, M., Serra Bragança, F. M., Ask, K., Rhodin, M., Andersen, P. H., Telezhenko, E., Bergsten, C., Nielen, M., & Hernlund, E. (2021). Kinematic gait characteristics of straight line walk in clinically sound dairy cows. *PLoS ONE*, *16*(July), 1–20. https://doi.org/10.1371/journal.pone.0253479

Titterington, F. M., Knox, R., Buijs, S., Lowe, D. E., Morrison, S. J., Lively, F. O., & Shirali, M. (2022). Human–Animal Interactions with Bos taurus Cattle and Their Impacts on On-Farm Safety: A Systematic Review. In *Animals* (Vol. 12, Issue 6). MDPI. https://doi.org/10.3390/ani12060776

Toomay, J. C., & Hannen, P. J. (2004). *Radar Principles for the Non-Specialist*. Institution of Engineering and Technology. https://doi.org/10.1049/SBRA032E

Tougui, I., Jilbab, A., & Mhamdi, J. El. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare Informatics Research*, *27*(3), 189–199. https://doi.org/10.4258/HIR.2021.27.3.189

Tran, N. K., Sen, S., Palmieri, T. L., Lima, K., Falwell, S., Wajda, J., & Rashidi, H. H. (2019). Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. *Burns*, *45*(6), 1350–1358. https://doi.org/10.1016/j.burns.2019.03.021

Tranter, W. P., & Morris, R. S. (1991). A case study of lameness in three dairy herds. *New Zealand Veterinary Journal*, *39*(3), 88–96. https://doi.org/10.1080/00480169.1991.35668

Tupin, J. P., & Couse, J. M. (2016). *ANMAL, HEALTH AND WELLNESS MONITORING USING UWB RADAR*.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. In *New Series* (Vol. 185, Issue 4157).

Ulaby, F., Li, R., & Shanmugan, K. (1982). Crop Classification Using Airborne Radar and Landsat Data. *Geoscience and Remote Sensing, IEEE Transactions On*, *GE-20*, 42–51. https://doi.org/10.1109/TGRS.1982.4307519

Underwood, E. J. (1971). *Trace elements in human and animal nutrition. Ed. 3.* New York, USA, Academic Press, Inc.

Van De Gucht, T., Saeys, W., Van Nuffel, A., Pluym, L., Piccart, K., Lauwers, L., Vangeyte, J., & Van Weyenberg, S. (2017). Farmers' preferences for automatic lameness-detection systems in dairy cattle. *Journal of Dairy Science*, *100*(7). https://doi.org/10.3168/jds.2016-12285

van der Spek, D., van Arendonk, J. A. M., Vallée, A. A. A., & Bovenhuis, H. (2013). Genetic parameters for claw disorders and the effect of preselecting cows for trimming. *Journal of Dairy Science*, *96*(9), 6070–6078. https://doi.org/10.3168/jds.2013-6833

van der Tol, P. P. J., Metz, J. H. M., Noordhuizen-Stassen, E. N., Back, W., Braam, C. R., & Weijs, W. A. (2002). The Pressure Distribution Under the Bovine Claw During Square Standing on a Flat Substrate. *Journal of Dairy Science*, *85*(6), 1476–1481. https://doi.org/10.3168/JDS.S0022-0302(02)74216-1

Van Der Tol, P. P. J., Metz, J. H. M., Noordhuizen-Stassen, E. N., Back, W., Braam, C. R., & Weijs, W. A. (2003). The vertical ground reaction force and the pressure distribution on the claws of dairy cows while walking on a flat substrate. *Journal of Dairy Science*, *86*(9), 2875–2883. https://doi.org/10.3168/jds.S0022-0302(03)73884-3

Van Marle-Köster, E., & Visser, C. (2021). Unintended consequences of selection for increased production on the health and welfare of livestock. In *Archives Animal Breeding* (Vol. 64, Issue 1, pp. 177–185). Copernicus GmbH. https://doi.org/10.5194/aab-64-177-2021

Van Metre, D. C. (2017). Pathogenesis and Treatment of Bovine Foot Rot. In *Veterinary Clinics of North America - Food Animal Practice* (Vol. 33, Issue 2, pp. 183–194). W.B. Saunders. https://doi.org/10.1016/j.cvfa.2017.02.003

Van Nuffel, A., Vangeyte, J., Mertens, K. C., Pluym, L., De Campeneere, S., Saeys, W., Opsomer, G., & Van Weyenberg, S. (2013). Exploration of measurement variation of gait variables for early lameness detection in cattle using the GAITWISE. *Livestock Science*, *156*(1–3), 88–95. https://doi.org/10.1016/j.livsci.2013.06.013

Van Nuffel, A., Zwertvaegher, I., Pluym, L., Van Weyenberg, S., Thorup, V. M., Pastell, M., Sonck, B., & Saeys, W. (2015). Lameness detection in dairy cows: Part 1. How to distinguish between non-lame and lame cows based on differences in locomotion or behavior. In *Animals* (Vol. 5, Issue 3, pp. 838–860). MDPI AG. https://doi.org/10.3390/ani5030387

van Tubergen, V., Heuft-Dorenbosch, L., Schulpen, G., Landewé, R., Wijers, R., van der Heijde, D., van Engelshoven, J., & van der Linden, S. (2003). *Radiographic assessment of sacroiliitis by radiologists and rheumatologists: does training improve quality?* 519–525. https://doi.org/10.1136/ard.62.6.519

Vanasse, A., Courteau, J., Fleury, M. J., Grégoire, J. P., Lesage, A., & Moisan, J. (2012). Treatment prevalence and incidence of schizophrenia in Quebec using a population health services perspective: Different algorithms, different estimates. *Social Psychiatry and Psychiatric Epidemiology*, *47*(4), 533–543. https://doi.org/10.1007/s00127-011-0371-y

Vanhoudt, A., Yang, D. A., Armstrong, T., Huxley, J. N., Laven, R. A., Manning, A. D., Newsome, R. F., Nielen, M., van Werven, T., & Bell, N. J. (2019). Interobserver agreement of digital dermatitis M-scores for photographs of the hind feet of standing dairy cattle. *Journal of Dairy Science*, *102*(6), 5466–5474. https://doi.org/10.3168/jds.2018-15644

Vermunt, J. J., & Greenough, P. R. (1995). Structural characteristics of the bovine claw: Horn growth and wear, horn hardness and claw conformation. *British Veterinary Journal*, *151*(2), 157–180. https://doi.org/10.1016/S0007-1935(95)80007-7

Viswanathan, G., Bhatia, R. C., Kamble, V. P., & Rao, S. R. (1997). Indian Doppler weather radar system - an overview. *International Geoscience and Remote*

Sensing Symposium (IGARSS), *3*, 1129–1131. https://doi.org/10.1109/igarss.1997.606373

Von Keyserlingk, M. A. G., Barrientos, A., Ito, K., Galo, E., & Weary, D. M. (2012). Benchmarking cow comfort on North American freestall dairies: Lameness, leg injuries, lying time, facility design, and management for high-producing Holstein dairy cows. *Journal of Dairy Science*, *95*(12), 7399–7408. https://doi.org/10.3168/jds.2012-5807

Vuttipittayamongkol, P., Elyan, E., & Petrovski, A. (2021). On the class overlap problem in imbalanced data classification. *Knowledge-Based Systems*, *212*. https://doi.org/10.1016/j.knosys.2020.106631

Walker, K. E., Middleton, J. R., Gull, T., Payne, C. A., & Adkins, P. R. F. (2023). Bacterial culture and susceptibility of samples taken from septic foot lesions of adult beef cattle. *Journal of Veterinary Internal Medicine*. https://doi.org/10.1111/jvim.16645

Walker, S. L., Smith, R. F., Jones, D. N., Routly, J. E., Morris, M. J., & Dobson, H. (2010). The effect of a chronic stressor, lameness, on detailed sexual behaviour and hormonal profiles in milk and plasma of dairy cattle. *Reproduction in Domestic Animals*, *45*(1), 109–117. https://doi.org/10.1111/j.1439-0531.2008.01263.x

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*, 64606–64628. https://doi.org/10.1109/ACCESS.2021.3074243

Wang, P., Ma, Y., Liang, F., Zhang, Y., Yu, X., Li, Z., An, Q., Lv, H., & Wang, J. (2020). Non-contact vital signs monitoring of dog and cat using a UWB radar. *Animals*, *10*(2). https://doi.org/10.3390/ani10020205

Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., Liu, X., Chen, J., Zhou, H., Ben Ayed, I., & Zheng, H. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-26216-9

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*. https://doi.org/10.1109/CVPR.2017.369

Wang, Z., Song, H., Wang, Y., Hua, Z., Li, R., & Xu, X. (2022). Research Progress and Technology Trend of Intelligent Morning of Dairy Cow Motion Behavior. *Smart Agriculture*, *4*(2), 36–52. https://doi.org/10.12133/j.smartag.SA202203011

Warnick, L. D., Janssen, D., Guard, C. L., & Gröhn, Y. T. (2010). The Effect of Lameness on Milk Production in Dairy Cows. *Journal of Dairy Science*, *84*(9), 1988–1997. https://doi.org/10.3168/jds.s0022-0302(01)74642-5

Wassink, G. J., Grogono-Thomas, R., Moore, L. J., & Green, L. E. (2003). Risk factors associated with the prevalence of footrot in sheep from 1999 to 2000. *The Veterinary Record*, *152*, 351–358.

Weary, D. M., Huzzey, J. M., & Von Keyserlingk, M. A. G. (2009). Board-invited Review: Using behavior to predict and identify ill health in animals. *Journal of Animal Science*, *87*(2), 770–777. https://doi.org/10.2527/jas.2008-1297

Webster, A. J. F. (2001). *Effects of Housing and Two Forage Diets on the Development of Claw Horn Lesions in Dairy Cows at*. *44*(0), 56–65. https://doi.org/10.1053/tvjl.2001.0569

Weigele, H. C., Gygax, L., Steiner, A., Wechsler, B., & Burla, J. B. (2018). Moderate lameness leads to marked behavioral changes in dairy cows. *Journal of Dairy Science*, *101*(3), 2370–2382. https://doi.org/10.3168/jds.2017-13120

Weishaupt, M. A. (2008). Adaptation strategies of horses with lameness. *The Veterinary Clinics of North America. Equine Practice*, *24*(1), 79–100. https://doi.org/10.1016/j.cveq.2007.11.010

Wells, S. J., Trent, A. M., Marsh, W. E., McGovern, P. G., & Robinson, R. A. (1993). Individual cow risk factors for clinical lameness in lactating dairy cows. *Preventive Veterinary Medicine*, *17*(1–2), 95–109. https://doi.org/10.1016/0167-5877(93)90059-3

Wells, S. J., Trent, A. M., Marsh, W. E., & Robinson, R. A. (1993). Prevalence and severity of lameness in lactating dairy cows in a sample of Minnesota and Wisconsin herds. *Journal of the American Veterinary Medical Association*, *202*(1), 78–82.

Westin, R., Vaughan, A., de Passillé, A. M., DeVries, T. J., Pajor, E. A., Pellerin, D., Siegford, J. M., Vasseur, E., & Rushen, J. (2016). Lying times of lactating cows on dairy farms with automatic milking systems and the relation to lameness, leg lesions, and body condition score. *Journal of Dairy Science*, *99*(1), 551–561. https://doi.org/10.3168/jds.2015-9737

Westwood, C. T., Bramley, E., & Lean, I. J. (2003). Review of the relationship between nutrition and lameness in pasture-fed dairy cattle Review of the relationship between nutrition and lameness. *New Zealand Veterinary Journa*, *51*(5), 208–218. https://doi.org/10.1080/00480169.2003.36369

Whay, H. (2002). Locomotion scoring and lameness detection in dairy cattle. *In Practice*, *September*.

Whay, H., & Huxley, J. (2005). Pain relief in cattle: A practitioners perspective. *Cattle Practice*, *13*, 81–85.

Whay, H. R., Waterman, A. E., & Webster, A. J. F. (1997). Associations between locomotion, claw lesions andnociceptive threshold in dairy heifers during the peri-partum period. *The Veterinary Journal*, *154*(2), 155–161. https://doi.org/10.1016/S1090-0233(97)80053-6

Wilson, T. R., LeBlanc, S. J., DeVries, T. J., & Haley, D. B. (2018). Effect of stall design on dairy calf transition to voluntary feeding on an automatic milk feeder after introduction to group housing. *Journal of Dairy Science*, *101*(6), 5307–5316. https://doi.org/10.3168/jds.2017-14011

Wilson-Welder, J. H., Alt, D. P., & Nally, J. E. (2015). Digital dermatitis in cattle: Current bacterial and immunological findings. In *Animals* (Vol. 5, Issue 4, pp. 1114–1135). MDPI AG. https://doi.org/10.3390/ani5040400

Winckler, C., & Willen, S. (2001). The Reliability and Repeatability of a Lameness Scoring System for Use as an Indicator of Welfare in Dairy Cattle. *Acta*

Agriculturae Scandinavica A: Animal Sciences, *51*(777257720), 103–107. https://doi.org/10.1080/090647001316923162

Witthoft, N., Sha, L., Winawer, J., & Kiani, R. (2018). Sensory and decision-making processes underlying perceptual adaptation. *Journal of Vision*, *18*(8), 1–20. https://doi.org/10.1167/18.8.10

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). *A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples*. http://www.biomedcentral.com/1471-2288/13/61

Wyn–Jones, G. (1988). *Equine Lameness*. Wiley-Blackwell.

Yang, H., Fong, S., Wong, R., & Sun, G. (2013). Optimizing classification decision trees by using weighted naïve bayes predictors to reduce the imbalanced class problem in wireless sensor network. *International Journal of Distributed Sensor Networks*, *2013*. https://doi.org/10.1155/2013/460641

Zebeli, Q., Dijkstra, J., Tafaj, M., Steingass, H., Ametaj, B. N., & Drochner, W. (2008). Modeling the adequacy of dietary fiber in dairy cows based on the responses of ruminal pH and milk fat production to composition of the diet. *Journal of Dairy Science*, *91*(5), 2046–2066. https://doi.org/10.3168/jds.2007-0572

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High Agreement and High Prevalence: The Paradox of Cohen's Kappa. *The Open Nursing Journal*, *11*(1), 211–218. https://doi.org/10.2174/1874434601711010211

Zhang, Z., & Sejdić, E. (2019). Radiological images and machine learning: Trends, perspectives, and prospects. In *Computers in Biology and Medicine* (Vol. 108, pp. 354–370). Elsevier Ltd. https://doi.org/10.1016/j.compbiomed.2019.02.017

Zheng, E., Yu, Q., Li, R., Shi, P., & Haake, A. (2021). *A Continual Learning Framework for Uncertainty-Aware Interactive Image Segmentation*. www.aaai.org