DALARNA UNIVERSITY

Degree Thesis

Master's Level (Second cycle)

Body Rumen Fill Scoring of Dairy Cows Using Digital Images

Authors: Reza Derakhshan, Soroush Yousefzadeh Boroujeni
School: Dalarna University, Borlänge
Supervisor: Moudud Alam
Co-supervisors: Niclas Högberg, Louise Winblad von Walter
Examiner: Mia Xiaoyun Zhao
Course code: MI4001 Credits:
30
Date of examination: January 16, 2024

Abstract:

The research presented in this thesis focuses on an innovative use of digital imaging, and the machine learning techniques to assess the body rumen fill scoring in dairy cows. This study aims to enhance the efficiency of monitoring and managing dairy cow health, which is crucial for the dairy industry's productivity and sustainability.

The primary objective was to develop an automated annotation system for evaluating rumen fill status in dairy cows using digital images extracted from recorded videos. This system leverages advanced machine learning algorithms and neural networks, aiming to mimic manual assessments by veterinarians and specialists on farms. To achieve the above objectives, this thesis made use of already existing video records from a Swedish dairy farm hosting mainly the Swedish Red and the Swedish Holstein breeds. A subset of these images were then processed, manually classified using a modified rumen fill scoring system based on visual assessment, and supervised classification algorithms were trained on 277 manually annotated images.

The thesis explored various machine learning techniques for classifying these images, including Logistic Regression, Support Vector Machine (SVM), and a Deep Neural Network using the VGG16 architecture. These models were trained, validated, and tested with a dataset that included variations in cow color patterns, aiming to determine the most effective approach for automated rumen fill scoring. The results indicated that while each model had its strengths and weaknesses, the simple logistic model was performing the best in terms of test accuracy and F1 score.

This research contributes to the field of precision livestock farming, particularly in the context of dairy farming. By automating the process of rumen fill scoring, the study aims to provide dairy farmers with a reliable, efficient, and cost-effective tool for monitoring cow health. This tool has the potential to enhance dairy cow welfare, improve milk production, and support the sustainability of dairy farming operations. However, at the current state, the model accuracy of the best model was only moderate. There is a need for further improvement of the prediction performance possibly by adding more cow images, using improved image processing, and feature engineering.

Keywords: Dairy Cows, Image Processing, Machine Learning, Rumen Fill Scoring, Automated Annotation, Precision Agriculture, Dairy Farming, VGG16.

Table of Contents

1. Introduction and Literature Review

1.1. Background

The dairy industry plays a vital role in the economy, particularly in regions like the European Union where, in 2018, its output reached a value of approximately 52 billion euros, marking a significant contribution to the agricultural sector. This shift in consumer preferences is reshaping the dynamics of the market, prompting dairy farms to modify their production methods and product offerings to align with these new demands.

In the context of the Swedish dairy industry, a significant focus has been placed on sustainability and minimizing the environmental impact. A key initiative in this direction was launched in 2020 by key players in the Swedish beef and dairy sectors. Their objective was to explore and measure how the climate impact of these sectors could be reduced by the year 2050 [2]. The findings of this project [2] were promising, indicating that it's feasible for the Swedish beef and dairy industries to conform to the objectives of the Paris agreement. The target is to achieve climate-neutral farming by 2050, ensuring that this is done without sacrificing biodiversity, the health and welfare of animals, or the production of food and bioenergy. This approach is not only seen as a pathway to growth for the industry but also serves as a strategic plan for future development. It emphasizes the importance of balancing animal health, environmental conservation, productivity, and profitability in farming practices.

Dairy farming, as a crucial part of global agriculture, is gaining more attention due to the ethical treatment and welfare of dairy cows. Thompson [3] articulates the moral obligation to recognize animals' sentience and well-being, establishing a fundamental ethical duty within the broader context of sustainable agricultural practices.

Research by Smith et al. [4] underscores the interdependence of dairy cow welfare and productivity, emphasizing the necessity for optimal conditions to ensure sustained economic viability. Simultaneously, Brown and Jones [5] identify a shift in consumer preferences, signaling a rising demand for ethically produced dairy products. This trend not only reflects a growing societal concern for animal welfare but also has market implications for the dairy industry.

Global regulations and standards, as exemplified by the World Organization for Animal Health (OIE) [6], the Strasbourg Convention (1976), Council Directive 98/58/EC, and Regulation (EU) 2017/625, affirm the commitment to ethical treatment in dairy farming. These legal frameworks set stringent guidelines, establishing a foundation for maintaining public trust and credibility within the industry. In parallel, Johnson et al. [7] explore the intricate relationship between welfare and health in dairy cows, highlighting the significance of optimal welfare conditions in reducing diseases.

Sustainability in dairy farming demands a holistic approach, as argued by Green et al. [8]. They advocate for the integral role of dairy cow welfare in achieving a balance between economic, environmental, and ethical dimensions, critical for ensuring the long-term sustainability of the industry. Technological innovations, as

discussed by White and Black [9], offer promising avenues for enhancing dairy cow welfare through precision farming and sensor applications.

Global perspectives on dairy cow welfare, provided by the Global Dairy Platform [10], offer insights into diverse international approaches. Understanding these perspectives is vital for the development of comprehensive welfare standards adaptable to diverse cultural, economic, and geographical contexts. Despite these efforts, challenges persist, including economic constraints and entrenched cultural practices. Garcia and Patel [11] explore these challenges and discuss the potential for sustainable and welfare-centric practices to overcome them, pointing towards a more humane and economically viable future for the dairy industry.

1.2. Welfare of a Dairy Cow

Regular health check-ups and preventive measures are fundamental for maintaining the overall well-being of dairy cows. This involves routine veterinary care, vaccinations, and prompt treatment of any illnesses or injuries. One recent study has focused on assessing dairy cow welfare during different periods of farming, specifically grazing and housing. This research is crucial as it highlights the varying needs and welfare concerns of dairy cows depending on their environment and management practices. It underscores the importance of considering different farming stages and settings when evaluating and ensuring the overall welfare of dairy cows [12].

Bos et al. [13] emphasize the importance of nine essential factors for promoting the welfare of dairy cows in husbandry systems. First and foremost, adequate resting space is crucial, allowing cows to lie down and rest simultaneously comfortably. Equally important is the provision of nutritious feed and continuous access to fresh water, supporting essential bodily functions and milk production. Additionally, cows should have the freedom to engage in natural behaviors, necessitating enough space for movement within the herd and their environment. The handling of cows also plays a significant role, where calm and predictable human interactions are essential to minimize stress and fear, subsequently enhancing milk production.

Furthermore, the elimination of negative stimuli like current leakage and cow trainers is vital to prevent chronic stress, which can adversely affect cow health and welfare. The design of the environment should be obstacle-free, enabling cows to move, rest, and lie down without hindrance, while maintaining personal space. Climate control is another critical aspect, where the Temperature Humidity Index should be kept below 71 to prevent heat stress. Safe passageways and feeding areas, characterized by non-slip, dry, and clean floors, are necessary to prevent slipping or hoof damage. Lastly, adequate lighting is required to facilitate proper cow recognition, exploration, and social interaction, with recommended lighting levels of more than 200 lux during the day and darkness at night.

Implementing these factors contributes to a content and productive dairy herd, aligning with the interconnected nature of animal welfare, health, and productivity. In the context of this project, the focus has been on the second factor, "feed and water," and the methods employed to evaluate it. Examining this factor is crucial for understanding and enhancing the nutritional aspects of dairy cow husbandry.

1.3. Monitoring Cow Signals

Given the multitude of factors influencing the determination of a dairy cow's health, it is imperative to establish a comprehensive monitoring system for assessing various indicators. As highlighted by Hulsen [14], the complexity of factors affecting dairy cow well-being underscores the necessity of systematic monitoring. This approach involves the continuous assessment of behavioral, physiological, and other relevant indicators to ascertain the overall health and welfare of dairy herds. With this monitoring system in place, farm managers can make informed decisions, ensuring the proactive management of cow health and contributing to the overall productivity of the dairy farming operation.

Moran and R. Doyle [15] have listed nine distinct scoring systems, offering a nuanced understanding of dairy cow health and well-being. These scoring systems, outlined comprehensively in "Cow Signals Checkbook" by Hulsen [14], contribute invaluable numerical values to assess various crucial aspects of cow welfare.

Body Condition Score (BCS): BCS involves visually estimating the amount of muscle and fat covering the bones of an animal [14]. Maintaining an optimal BCS is widely recognized as crucial for reproductive performance, milk production, and overall cow health [16].

Locomotion/Lameness Score: This subjective assessment of how easily a cow walks on a level surface [14] is a key indicator of lameness, a prevalent issue affecting dairy herds. Lameness negatively impacts cow comfort, productivity, and overall herd welfare [17].

Hoof Score: This visual and descriptive assessment of hoof health [14] is instrumental in identifying issues like claw lesions and laminitis, both of which significantly impact cow mobility and comfort [18].

Leg Score: Assessing the stance of the hind legs [14], leg score contributes to lameness evaluation. Proper leg conformation is crucial for a cow's ability to move comfortably and avoid musculoskeletal issues [19].

Hygiene Score: Evaluating the contamination of manure and dirt on the udder and lower hind legs of cows [14], hygiene score is directly linked to udder health and the prevention of mastitis [20].

Rumen Score: This measure of rumen fill serves as an indicator of feed intake and the rate of feed passage [14]. Optimal rumen function is vital for nutrient utilization and overall cow health [21].

Manure Score: Examining the visual composition and physical consistency of feces [14], manure score provides insights into digestive health and nutrient absorption [22].

Teat Score: This evaluation of the impact of the milking system on teat health [14] is crucial for preventing mastitis, a common and costly udder infection [20].

Panting Score: Assessing the impact of heat stress on cow well-being [14], panting score is particularly relevant in regions experiencing high temperatures. Heat stress

negatively affects milk production, reproduction, and overall cow comfort [23].

These scoring systems collectively form a robust framework for assessing and monitoring various facets of cow health and welfare. They empower farmers with data-driven insights for informed decision-making and proactive management practices on the dairy farm. In the context of this project, our focus has been only on the second factor, "rumen score".

1.4. Digital Transformation in Dairy Industry

Digital transformation in the dairy industry has become a pivotal aspect of modern agricultural practices, significantly impacting the management, productivity, and welfare of dairy cows. The integration of advanced technologies has ushered in a new era of precision farming, offering farmers valuable insights into various aspects of herd health, behavior, and production. There are different dimensions of digital transformation in dairy farming including Environmental Monitoring, Blockchain for Automated Milking Systems, Automated Milking Systems and following two dimensions:

Precision Livestock Farming (PLF): Precision Livestock Farming involves the use of technology to monitor and manage individual animals within a herd. Sensor technologies, such as RFID tags, accelerometers, and pedometers, enable real-time tracking of cow behavior, activity levels, and health status [24]. This data-rich approach enhances early detection of health issues, leading to more proactive and targeted interventions.

Data Analytics for Health Monitoring: The integration of data analytics and machine learning has empowered farmers to make data-driven decisions for herd health management. By analyzing patterns in sensor data, predictive models can identify deviations in behavior or health parameters, enabling early detection of diseases, reproductive issues, or nutritional imbalances [25].

1.5. Rumen Fill Score

The rumen, a vital fermentation chamber in the stomach of ruminant animals like dairy cows (Figure 1), is integral for digesting fibrous feed materials and absorbing nutrients, contributing to the cow's energy and growth. Rumen health directly impacts feed efficiency, nutrient utilization, and, consequently, milk production, playing a pivotal role in the overall well-being and productivity of dairy cows. Rumen scoring, a valuable management practice, visually assesses rumen fill and function, offering insights into the cow's feed intake, digestion, and gastrointestinal health. This scoring system considers factors such as rumen size, shape, and contractions. Maintaining optimal rumen health is essential for maximizing milk production and ensuring the welfare of dairy cows. The evaluation focuses on scoring the left flank, with results detailed in Table 1 and visual representations in Figure 1, determined by observing the region between the ribs at the front, the vertebrae at the top, and the hook bone at the back.

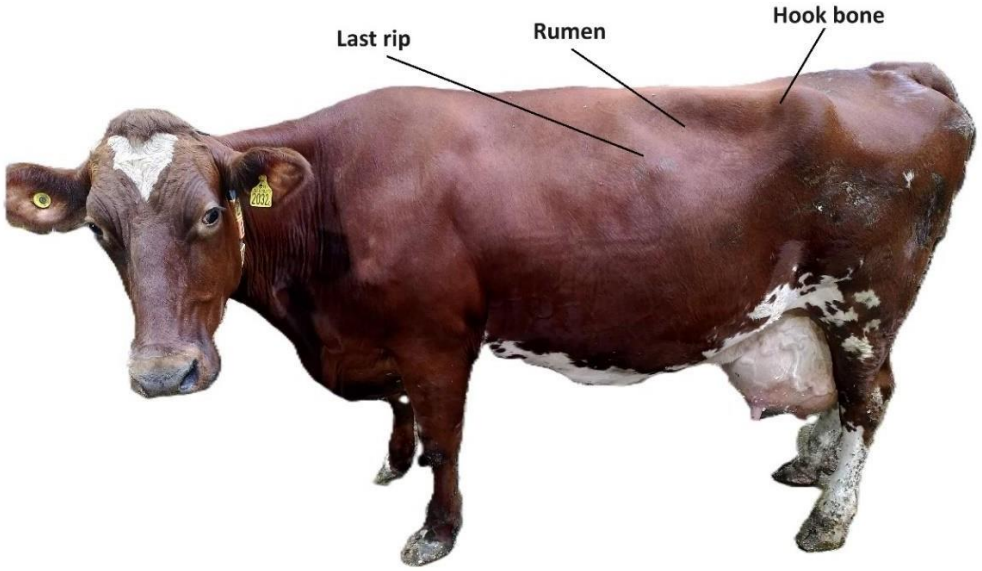

Figure 1. Cow's anatomy

Table 1. Descriptors used to quantify rumen scores and their diagnoses [15]

Score	Descriptor	Diagnosis/Status
1	• A deep dip in the left flank. • The skin under the lumbar vertebrae curves inwards. • The skin fold from the hook bone goes vertically downwards. • The para lumbar fossa behind the last rib is more than one hand width deep. • Viewed from the side, this part of the flank has a rectangular appearance.	The cow has eaten little or nothing which could be due to sudden illness or insufficient or unpalatable feed.
2	• The skin under the lumbar vertebrae curves inwards The skin fold from the hook bone runs diagonally forward towards the last rib. • The para lumbar fossa behind the last rib is one hand width deep. • Viewed from the side, this part of the flank has a triangular appearance.	This score is often seen in cows in the first week after calving. Later in lactation, this is a sign of insufficient feed intake or a too high rate of passage.
3	• The skin under the lumbar vertebrae goes vertically down for one hand width and then curves outward. • The skin fold from the hook bone is not visible • The para lumbar fossa behind the last rib is still just visible.	This is the right score for milking cows with a good feed intake and when the feed remains in the rumen for the optimal time.
4	• The skin under the lumbar vertebrae curves outwards. • No para lumber fossa is visible behind the last rib.	This is the correct score for cows nearing the end of lactation and for dry cows
5	• The lumbar vertebrae are not visible as the rumen is well filled. • The skin over the whole belly is quite tight There is no visible transition between the flank and ribs.	This is the correct score for dry cows.

Burfeind et al. [26] have discussed the development and assessment of a 5-point subjective scoring system for visually describing rumen fill in dairy cows. The

authors aimed to evaluate the system's performance as an indicator of changes in dry matter intake (DMI) and feed intake. The study found that rumen fill scores, ranging from 1 to 5, demonstrated substantial intra- and interobserver repeatability. Changes in visual rumen fill scores were correlated with changes in DMI, indicating its potential as an estimate of feed intake. However, caution is advised in clinical usage due to variability not associated with measured parameters such as physiological differences, rumen health, feed characteristics, hydration status, time of feeding, cow posture, method of scoring. The study suggests further research to refine cowside estimates of DMI and explore the system's utility in identifying cows at risk for disease.

In a comprehensive mapping study, Shine et al. provided an extensive overview of machine learning applications in dairy farming over the past two decades [27]. They present a thorough analysis of machine learning applications in dairy farming from 1999 to 2021. It reviews 129 publications, emphasizing the evolution of research areas, algorithmic approaches, and validation methods over time. The study highlights a significant increase in publications focused on dairy cow physiology and health, showing a shift from traditional animal husbandry issues to more nuanced health-related concerns. It also notes an increased utilization of neural network-based algorithms, indicating a trend away from statistical regression models towards more complex machine learning techniques. The paper underscores the importance of machine learning in advancing dairy farming practices, particularly in the realms of animal health and productivity.

Recent advancements in rumen fill analysis have leveraged neural networks and machine learning algorithms. A study [28] utilized ANN models, incorporating animal, feed, and environmental factors, to forecast rumen fill. These factors, extracted from various studies, formed the dataset for predicting rumen fill weight. The ANN model, using a three-layer Levenberg–Marquardt back-propagation network, showed a high prediction accuracy of 96% for rumen fill weight, with higher precision in cattle than sheep. Additionally, a random forest model, based on a binary tree algorithm, demonstrated an accuracy of 87% in predicting rumen fill, with varying accuracy in cattle and sheep during validation. The study concluded the superior performance of the ANN model over the random forest in rumen fill prediction for cattle and sheep.

In another study, Song et al. [29] developed a method for automatically assessing Reticulo-Ruminal Motility in Dairy Cows using 3-Dimensional Vision. This approach employs a 3D camera system, integrated into an automatic milking robot, to capture images of a cow's left paralumbar fossa. The system then processes these images to detect and analyze reticulo-ruminal contractions by identifying surface concavities. Demonstrating a high matching sensitivity, this technique emerges as a non-invasive and efficient method for real-time gastrointestinal health monitoring, greatly benefiting dairy farm management and animal welfare.

1.6. Objectives

The primary objective of this research is to develop an automated annotation system that evaluates rumen fill using digital images from recorded videos. This system will be based on the implementation of machine learning algorithms. The overarching

goal is to streamline and reduce the reliance on veterinarians and specialists on farms, potentially fostering a transformative shift towards digital solutions in the agricultural sector.

This project is specifically geared towards the estimation of rumen fill scoring (RFS) using 2D digital images. The proposed methodology involves the validation of automated scores by aligning them with established RFS computed manually, potentially incorporating subjective assessments based on image and video analyses. To achieve this, several machine learning algorithms including Deep Neural Network-based image classification methods were deployed for supervised classification. The utilization of such techniques holds promise for enhancing the efficiency of rumen health assessment, paving the way for a more digitized and automated approach in the context of livestock management on farms.

The structure of this paper is meticulously organized to facilitate a clear understanding of the research conducted. The introductory section sets the stage, outlining the significance and objectives of estimating rumen fill scoring (RFS) using 2D digital imaging. Following this, we delve into the literature review which provides context and highlights previous advancements in the field. The methodology section describes the detailed process of data collection, the algorithms employed, and the validation of automated RFS against manual assessments. Subsequent sections present the results of the machine learning models, offering a comparative analysis of their performance. Discussions interpret these findings, drawing connections with existing research and practical implications. The paper concludes with a summary of the insights gained, potential limitations of the study, and directions for future research. Appendices and references are provided for additional details and source verifications, ensuring the paper's completeness and utility as a resource for furthering automated livestock management practices.

2. Methodology

This chapter outlines the methodology which is focusing on the innovative use of digital imaging and machine learning to assess dairy cow welfare. The approach involves data collection through video capture, image processing for model input preparation, and the application of machine learning algorithms to evaluate rumen fill. This systematic process aims to improve the monitoring and management of dairy cow health, contributing to the field of precision agriculture.

2.1. Setup

This thesis uses video recording, from a Swedish farm (Swedish Livestock Research Centre, Uppsala, Sweden), which was conducted for some other purposes. The video recording was captured using the Ubiquiti UniFi G3-Flex Camera. The Ubiquiti UniFi G3-Flex Camera is well-suited for farm environments due to its high-resolution Full HD (1080p) video quality, essential for clear and detailed monitoring of agricultural activities. Its IPX4 weatherproof rating ensures durability under harsh outdoor conditions, a critical feature for farm settings. The camera's wide viewing angles, night vision capabilities with IR LED illumination, and Power Over Ethernet (PoE) support for simplified installation make it highly adaptable for various farm surveillance needs. Moreover, its versatile mounting options accommodate different monitoring scenarios across farm landscapes. This camera was strategically mounted on the left side of the passageway, a critical juncture that bifurcates the path leading cows either towards the milking machines or along a direct route intended for their passage. The placement of the camera was meticulously chosen to ensure optimal coverage of the cows' movement, capturing high-resolution video as they traversed this division in the corridor. The camera's positioning and the resulting footage were instrumental in monitoring and analyzing the cows' behavior (Figure 2).

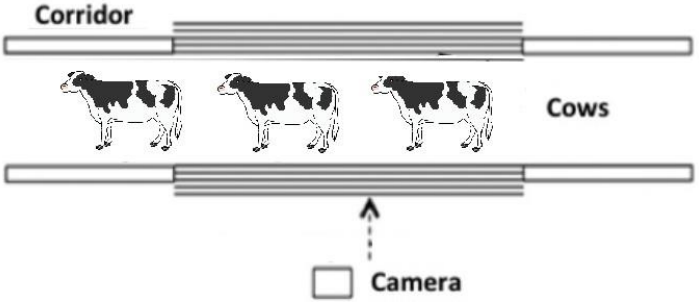

Figure 2. Schematic layout of passageway and installed camera

The farm used a guided cow traffic system with a smart gate system to steer cows towards the milking or feeding and resting areas. Only the cows sent to the latter destinations would be recorded. Every cow had an RFID tag in their right ear, which was read when a cow entered the passageway. The camera started a recording when the cow entered its field of view and ended the recording when it leaved it.

2.2. Data Acquisition

The dataset compiled for our analysis comprised imagery of two prominent dairy cattle breeds: the Swedish Red (SRB) and the Swedish Holstein, both of which are highly regarded for their milk production capabilities and adaptability to the Swedish climate. Data collection spanned an extensive period from 25 April to 11 May 2023, during which we meticulously captured and selected a total of 277 high-quality images from the video footage provided by the farm. Although our assessment was confined to a single farm, the dataset reflected a diverse array of five distinct color patterns exhibited by the cows.

In preparation for the application of machine learning techniques, each image was carefully cropped manually to a uniform size of 256 x 256 pixels. This standardization was not only critical for maintaining consistency across the dataset but also ensured compatibility with the input requirements of Convolutional Neural Networks (CNNs), along with other machine learning models under scrutiny in our study. The image size was specifically chosen to balance the need for sufficient detail with computational efficiency, thus optimizing the performance of the models in detecting and learning the intricate patterns necessary for our research objectives.

Figure 3 Capture the ROI

2.3. Classification

To accurately gauge the effectiveness of our models, we needed a reliable benchmark to quantify the rumen fill in cows, an indicator closely linked to their feed intake. Zaaijer and Noordhuizen [30] have pioneered a rumen fill scoring system that hinges on the visual assessment of the paralumbar fossa, offering a practical means to detect variations in feed consumption. Their scoring spectrum extends from 1 to 5, where a score of 1 corresponds to a minimal rumen fill, often signaling inadequate intake due to health issues, while a score of 5 denotes an optimally filled rumen, typical in non-lactating, dry cows. For lactating cows, a score of 3 is considered ideal, reflecting sufficient dry matter intake. This intuitive scoring framework was further substantiated by Burfiend et al. [26], who reported a moderate Spearman’s rank correlation (rs = 0.68) between the rumen fill scores and actual dry matter intake, attesting to the system's validity. Moreover, they provided a detailed summary of the criteria that defined each score as per Zaaijer and Noordhuizen's system, a synopsis of which is presented in Table 1 for reference.

To enhance the clarity and precision of our classification process for both the machine learning algorithms and our own analysis, we simplified the original five-class system by Zaaijer and Noordhuizen [30] into a three-class framework. Specifically, we merged classes 2 and 3 into a single class labeled as '2', and similarly combined classes 4 and 5 into a new class '3'. This restructuring resulted in a more streamlined classification system with three distinct classes, making it more manageable and suitable for our study's objectives.

In the initial phase of our classification, we undertook a comprehensive training process to ensure accurate and consistent image categorization. This involved closely following the guidelines and seeking expert advice from the professional supervisors to refine our skills in assessing the images based on the adapted scoring system (see also Section 2.5).

2.4. Data Processing

For a more comprehensive understanding of our data processing approach, we elucidated certain key terms that are frequently referenced throughout this report:

ROI (Region of Interest): This term refers to a specifically designated rectangular area within each image that contains the cow's rumen, which is the focus of our analysis.

RGB (Red, Green, and Blue): These are the primary colors of the light used in the digital representation of the images extracted from the videos, specifically within the ROI.

Our methodology for image extraction from the video footage was meticulously carried out by hand. Due to the camera's angular positioning relative to the cow passageway, automating this process proved challenging and prone to errors, particularly in terms of image quality and accurate positioning of the cows. Therefore, we adopted the following manual steps to ensure optimal data collection:

Optimal Cow Positioning: We identified that the most consistent and clearly visible images of the cows were captured when they were standing still with a straight spinal cord. Manual image extraction allowed us to select these optimal moments from the videos.

Image Cropping: The images were cropped to focus on the left side of the cow's rumen, using specific anatomical landmarks for consistency:

The left boundary was defined by the cow's rumen.
The right boundary was demarcated by the tail.
The upper boundary aligned with the cow's upper spine.
The lower boundary was just beneath the rumen.

Rumen Visibility: We ensured that the rumen was the most prominently visible part in each image.

Exclusion Criteria: Cows with unclear or dirty skin, which could potentially impede accurate assessment, were excluded from our analysis. Additionally, any

images that posed challenges in grading for any of us were removed from the dataset to maintain consistency and reliability.

Following the selection process, the approved images were automatically resized to the dimensions of 256 x 256 pixels, a size compatible with the input requirements for Convolutional Neural Networks (CNNs). We utilized the "ImageDataGenerator" library to effectively organize and randomly allocate these images into three distinct datasets: training, validation, and testing. This division facilitated a structured approach to training and evaluating the performance of our CNN models, ensuring a systematic and efficient analysis.

2.5. Manual Grading of the Images

Our grading methodology was based on Table 1 and also supported by expert guidance in the field. In this paper, the images were graded independently by two co-authors and these assessments were collectively reviewed to ensure consistency and accuracy, achieving a high agreement with a Kappa coefficient of 0.81. The grading methodology was refined into a three-tier system for more effective machine learning model training, prioritizing a balance between the detailed data needs and the practical aspects of machine learning. This simplification aimed to align with well-established methods, ensuring data clarity and consistency, crucial for accurate machine learning outcomes. It tried to address challenges like data sparsity and class imbalance, common in machine learning, by ensuring the best representation for themost categories. This approach thus facilitated more efficient model training and enhanced the potential for accurate, real-world applications.

The images that did not have agreement between two graders were resolved in the following way. If the scores in 5-score system were in the same class in three-tier system, for example classes 2 and 3 (in 5-score system), the true score graded 2 which had the same class (in three-tier system). If the graded were 4 or 5 (in 5-score system), this was graded 3 (in three-tier system). Finally, 13 images with higher difference were deleted from the dataset to bring the disagreements to the minimum. The total number of images in the dataset reduced to 276.

In our study, we divided the annotated images into three datasets: 216 for training, 31 for validation, and 30 for testing. This split was chosen to ensure a substantial training set for the VGG16 model to learn adequately, balanced against enough data to fine-tune and test the model without overfitting. The implementation was carried out in Python (3.9.7), leveraging its powerful libraries for efficient image processing and transfer learning. This setup was intended to provide a comprehensive learning base and rigorous evaluation framework for the model.

2.6. Transfer Learning

VGG16, a widely recognized convolutional neural network architecture, has demonstrated effectiveness in various computer vision tasks, including image classification. However, its performance on imbalanced datasets can be influenced by the inherent challenges associated with uneven class distributions. VGG16's deep architecture, consisting of multiple convolutional and fully connected layers, tends to learn hierarchical features that may prioritize prevalent classes, potentially

leading to suboptimal performance on minority classes in imbalanced datasets. Mitigating this challenge often involves implementing techniques such as class weights, data augmentation, or leveraging transfer learning with fine-tuning to

enhance the model's ability to generalize across all classes. As part of the methodology, addressing the imbalanced nature of the dataset when applying VGG16 is crucial for ensuring roust performance across diverse class distributions.

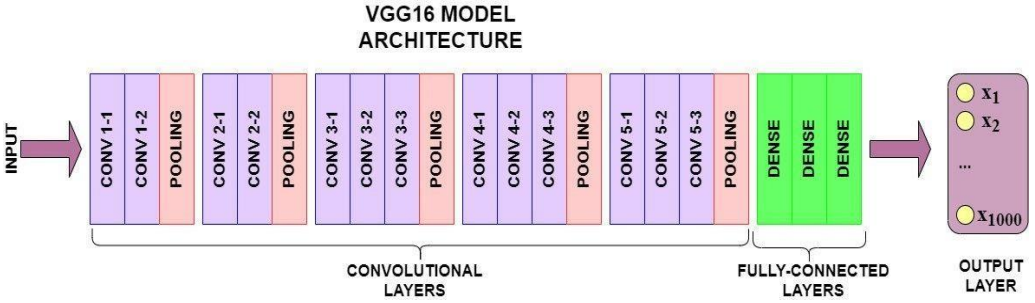

Figure 4 VGG architecture [32]

The choice of training the VGG16 model for 10 and 50 epochs was strategic. Training for 10 epochs allows us to quickly assess the model's initial learning curve and identify any early issues. Extending training to 50 epochs leverages the model's depth to enhance feature extraction and improve performance, balancing the need for thorough learning against computational efficiency. This approach ensures we optimize the model's learning potential without excessive resource expenditure. The learning process of the model was quantified using mathematical expressions to calculate the cross-entropy loss, which is common for classification problems. This loss function is expressed as Equation (1):

$$Cross - Entropy\ Loss = -\Sigma\ (y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)) \quad (1)$$

Here, yi represents the true label of the i-th sample, and pi is the predicted probability of the i-th sample according to the model. The training process was closely monitored to observe the improvement in performance as the epochs progressed. Validation accuracy and loss were also calculated following the same methodology to ensure consistency in performance evaluation across different data subsets.

Augmentation plays a pivotal role in optimizing the performance of VGG16, and as such, we conducted experiments both with and without augmentation. Given the specific anatomical characteristics, such as the left-side location of the cow's rumen, certain augmentation techniques, particularly those altering the image direction, were deemed impractical for effective model training. The comparative analysis between models trained with and without augmentation revealed that the incorporation of augmentation did not lead to an appreciable enhancement in the efficiency of training our neural network. This observation underscores the nuanced considerations required in selecting augmentation strategies tailored to the unique features of the dataset and the intricacies of the task at hand.

2.7. Other Models

In addition to employing convolutional neural networks (CNNs) as the primary model for this thesis, logistic regression, and support vector machine (SVM) were incorporated to diversify the analysis and evaluate their performance. These alternative models offer distinct approaches to classification tasks, allowing for a comprehensive comparison of results. Logistic regression, a linear model, provides simplicity and interpretability, while SVM, a non-linear model, excels in capturing complex decision boundaries. By incorporating multiple models, the research aims to assess the suitability of each approach and gain a holistic understanding of their respective strengths and weaknesses in addressing the research objectives.

A logistic regression model was utilized to analyze images of size 256x256 giving 256^2 features per image. Therefore, the so called "p>n" problem was inherent with the data set. We used the Principal Component Analysis (PCA) to reduce the dimensionality of the image data. PCA was applied to the images, reducing to four principal components. This dimensionality reduction was crucial to manage the computational complexity and to extract the most significant features from the images for the logistic regression analysis.

Once PCA was applied, the logistic regression (LR) model was trained on the transformed dataset. To ensure the robustness and generalizability of the model, a 10-fold cross-validation approach was employed. This technique involves dividing the dataset into ten subsets, using nine for training and one for validation, and iterating this process ten times with different subsets. This method helps in assessing the model's performance more accurately by reducing the variability that might arise from a single split of the data.

In Figure 5, the flowchart succinctly illustrates the comprehensive process from the initial acquisition of raw video data to the application of sophisticated machine learning algorithms.

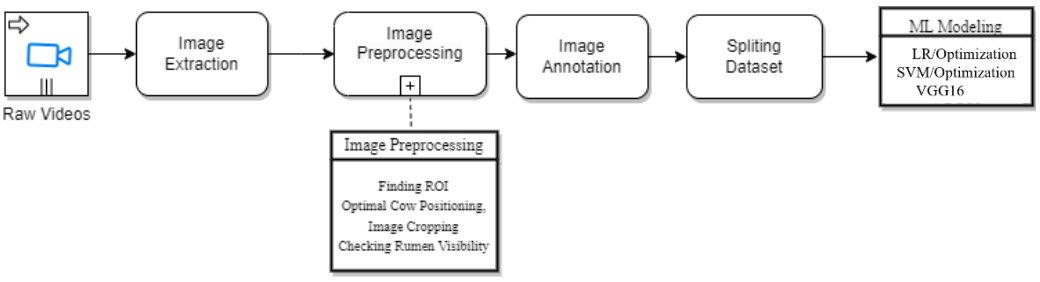

Figure 5 Flowchart of grading rumen fill

2.8. Implementation of Computational Models

In this section of the thesis, we delve into the detailed methodologies employed in the implementation of various computational models, highlighting the tools, software packages, and specific strategies used. The focus is on the comprehensive comparison of models such as Logistic Regression, SVM, and the VGG16 deep

learning architecture, particularly in the context of their application in image classification tasks.

1. **Data Preparation and Preprocessing**:

    - Utilization of the TensorFlow Keras library [33] for image data generation and augmentation.

    - Application of **ImageDataGenerator** from Keras for image rescaling, an essential step to normalize pixel values for effective model processing.

    - Organization of image data into distinct directories for training, validation, and testing, ensuring a structured approach to model training and evaluation.

2. **Model Implementation and Training**:

    - For Logistic Regression and SVM, scikit-learn [34] (a Python library) was employed, with the models being optimized through parameter tuning.

    - The VGG16 model [35], a pre-trained neural network from the Keras library, was employed for deep learning tasks. Its extensive use in image recognition tasks makes it a suitable choice for this study.

    - Training involved adjusting epochs and monitoring steps per epoch to assess model performance over time.

3. **Software and Packages**:

    - TensorFlow and Keras: Used for implementing and training the deep learning models, including the VGG16 architecture.

    - scikit-learn: Employed for more traditional machine learning models like Logistic Regression and SVM.

    - Python programming language served as the foundation for all implementations, known for its robust libraries and community support in machine learning and data science.

1. **Challenges and Solutions**:

    - Addressing overfitting, particularly in the context of deep learning models where extensive training can lead to models being overly tailored to the training data.

    - Balancing model complexity and computational efficiency, especially when dealing with large datasets and deep learning architectures.

- To combat overfitting: Implement dropout and early stopping in training.
- For complexity vs. efficiency: Use dimensionality reduction and lightweight neural networks

2.9. Performance Evaluation

The performance evaluation of various predictive models is reported in this section, utilizing a dataset with the goal of classifying data points accurately. The models were assessed based on four pivotal metrics: Accuracy, Precision, Recall, and F1 Score, each crucial for understanding different aspects of performance.

Evaluation Metrics Defined

- **Accuracy** measures the proportion of true results (both true positives and true negatives) among the total number of cases examined and is calculated as:

$$Accuracy = \frac{(True\ Positives\ (TP) + True\ Negatives\ (TN)}{TotalPopulation} \quad (2)$$

- **Precision** quantifies the number of positive class predictions that actuallybelong to the positive class, computed by:

$$Precision = \frac{TP}{TP\ +\ False\ Positives\ (FP)} \quad (3)$$

- **Recall** measures the proportion of actual positives that were identified correctly, given by:

$$Recall = \frac{TP}{TP\ +\ False\ Negatives\ (FN)} \quad (4)$$

- **F1 Score** is the harmonic mean of Precision and Recall, providing a balancebetween them, especially in uneven class distributions:

$$F1\ Score = 2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall} \quad (5)$$

3. Results

This chapter presents the results from different classification methods. We compare these methods to understand their strengths and weaknesses better, offering clear insights into how each one works.

3.1. VGG16 Neural Network

3.1.1. Model Performance (without augmentation)

We have performed transfer learning using the VGG16 pre-trained model on a custom dataset of images. First, the VGG16 model is loaded with weights pre-trained on the ImageNet dataset, excluding the top (fully connected) layers. The layers of the pre-trained model are then frozen to prevent their weights from being updated during training. On top of the frozen VGG16 model, a new sequential model is constructed. This model includes a flattening layer to convert the 3D output of the VGG16 base to a 1D feature vector, followed by a dense layer with 256 units and ReLU activation. A dropout layer with a dropout rate of 0.5 is added for regularization, and finally, a dense layer with 3 units (assuming three categories) and softmax activation is added to produce class probabilities. The model is compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as the metric.

3.1.1.1. Model Performance for 10 Epochs

The training process is executed using the **fit** method with **train_generator** for training data and **val_generator** for validation data. The training is performed for 10 epochs, and the results are printed for each epoch. The training loss, training accuracy, validation loss, and validation accuracy are displayed. The reported results suggest the model's performance on the given dataset, showing a progression of training and validation metrics across the epochs. In the results provided, the model achieves increasing accuracy on both training and validation sets, suggesting that the model is learning and generalizing well on the given data. The loss values are decreasing, which aligns with the goal of minimizing the loss during training. Table 2 presents the outcomes on the test dataset in the form of a confusion matrix table.

The transfer learning model exhibits commendable performance on the provided dataset, showcasing significant advancements in both loss reduction and accuracy enhancement throughout the training epochs. The initial training loss of 6.36 undergoes consistent reduction, reaching 0.21 by the tenth epoch. Concurrently, the training accuracy steadily climbs from 0.40 to an impressive 0.91, underscoring the model's proficiency in learning and adapting to the intricacies of the dataset.

Validation results align with positive trends, demonstrating a decline in validation loss from 4.08 to 0.97 and an increase in validation accuracy from 0.55 to 0.74 over the ten epochs. These findings affirm the model's capacity to generalize effectively to unseen data, showcasing its robustness. The convergence of training and

validation metrics suggests that the model successfully leverages meaningful representations from the pre-trained VGG16 base. The subsequent dense layers, configured for your specific dataset, contribute to the model's adaptability and overall efficacy in capturing the underlying patterns within the data. Figure 6 shows the progress of model accuracy and loss.

The training log provides a detailed overview of the model's progression, highlighting its ability to adapt to the dataset's complexities and achieve high accuracy. This suggests that the model has effectively learned discriminative features from the pre-trained base, enabling it to make accurate predictions on both the training and validation sets.

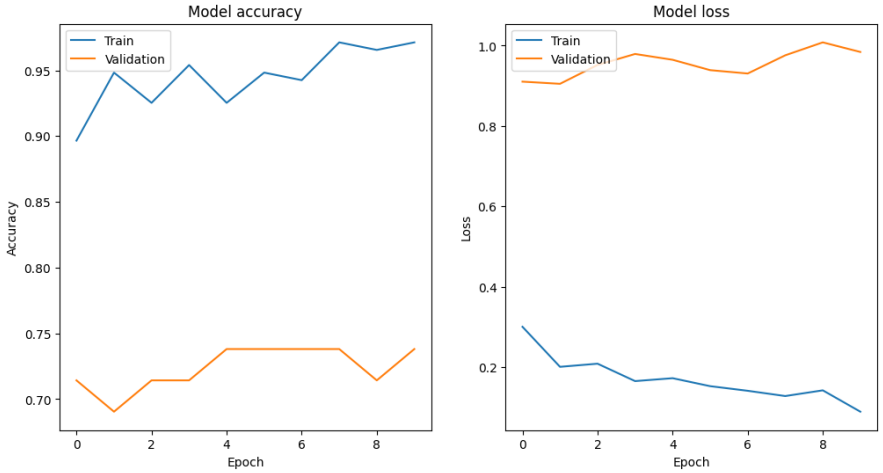

Figure 6 model accuracy and loss for VGG16- 10 epochs

Table 2 Confusion matrix –VGG16 without augmentation epochs 10

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	0	3	1
Actual Class 2	0	9	6
Actual Class 3	0	9	2

3.1.1.2. Model Performance for 50 Epochs

At the commencement of training, the model's initial training loss was 5.88, which notably improved to 0.5140 by the 5th epoch. This early stage also saw a training accuracy increase from 40.80% to 83.33%. By the 25th epoch, the model demonstrated a remarkable training accuracy of 98.28%, with a significantly reduced loss of 0.06. During validation, the highest accuracy reached was 76.19% at the 8th epoch, after which there was an observed increase in validation loss, suggesting a divergence between model performance on training and validation data, a potential indication of overfitting. We implemented early stopping to cease training before overfitting became more pronounced. Moreover, regularization techniques were used to discourage the complexity of the model that could lead to overfitting. Upon completion of training, the model underwent a final evaluation on a separate test set, resulting in a test loss of 2.09 andan accuracy of 53.33%. This highlighted a challenge in the model's ability to generalize to new, unseen data, which is a fundamental objective in the field of machine learning. Table 3 shows

the confusion matrix of the result on the test dataset. Figure 7 shows the accuracy and loss progress which is explained.

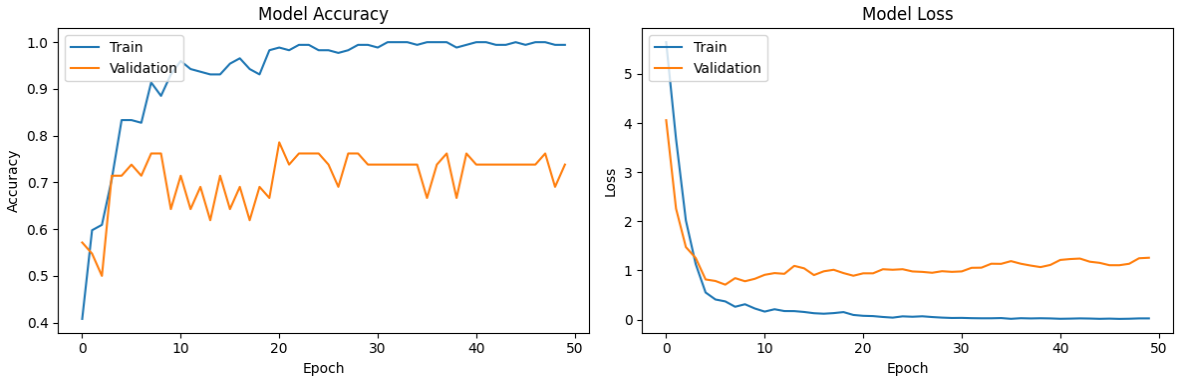

Figure 7 model accuracy and loss for VGG16- 50 epochs

Table 3 Confusion matrix – VGG16 without augmentation epochs 50

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	0	3	1
Actual Class 2	0	14	1
Actual Class 3	0	5	6

3.1.2. Model Performance (with augmentation)

In this thesis, data augmentation was employed as a technique to artificially expand the diversity of the training dataset by applying various transformations to the original images. Utilizing the **ImageDataGenerator** class, we systematically introduced random rotations up to 20 degrees, width and height shifts by 20%, shear transformations up to 20%, and zooming by 20%. Additionally, we included horizontal flips of the images. These specific augmentations were chosen to reflect potential variances encountered in real-world scenarios, with the aim of making the Convolutional Neural Network (CNN) model more robust to such variations and thus enhancing its generalization capabilities.

However, despite the theoretical benefits of these augmentations, the empirical results indicated a different outcome. While we observed a decrease in training loss from 4.96 to 0.77, suggesting the model was effectively learning to minimize errors, the training accuracy fluctuated, starting at 48.85% and demonstrating variance thereafter. These fluctuations could suggest difficulty in the model's ability to learn from the augmented images and could be a precursor to challenges in generalization to the validation set or unseen data. The results on the test dataset, yielding an accuracy of 50% and an F1 score of approximately 0.42, further corroborate the notion that the augmentation did not enhance the model's performance as anticipated. The confusion matrix, presented in Table 4, and the detailed metrics in the classification report elucidate the model's specific performance across different classes post-augmentation. Despite our efforts to create a model resilient to image variations through augmentation, the results necessitate a reevaluation of the augmentation strategy and perhaps a more tailored approach to improving model robustness and performance.

Table 4 Confusion matrix – transfer learning VGG16 with augmentation

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	0	3	1
Actual Class 2	0	13	2
Actual Class 3	0	9	2

3.2. Support Vector Machine (SVM)

In this section, we employ SVM with a linear kernel and a regularization parameter (C) set to 1. The SVM is implemented using the One- vs-Rest (OvR) strategy, allowing for the classification of multiple classes. This strategy trains a separate binary classifier for each class while treating the rest as a single class. This ensures that the model can handle multiclass classification tasks effectively.

In refining the SVM implementation, we utilized a linear kernel due to its effectiveness in high-dimensional spaces and its computational efficiency. The kernel's simplicity often leads to less overfitting when the number of features is large. To optimize the model's performance, a systematic grid search was conducted to tune the regularization parameter, C. The grid search evaluated a range of C values from 0.01 to 100 in logarithmic steps to determine the optimal balance between model complexity and training accuracy. This exhaustive search ensured the selection of a C value that minimizes the generalization error, which is set at 1 after observing the model's performance metrics. This process of parameter tuning via grid search is essential for the reproducibility of our results and allows for the SVM to be effectively applied in multiclass classification using the One-vs-Rest strategy.

The model is trained on a dataset comprising images from different categories. The dataset is divided into training and testing sets, with labels encoded using the LabelEncoder to transform categorical labels into numerical format. The SVM is then trained on the training set using the OvR approach. The model's performance is evaluated on the testing set, and metrics such as accuracy, precision, recall, and the confusion matrix are calculated to assess its effectiveness.

To further understand the model's performance, we visualize the Receiver Operating Characteristic (ROC) curve and Precision-Recall curve. The ROC curve illustrates the trade-off between true positive rate and false positive rate, while the Precision-Recall curve provides insights into the precision and recall relationship. These visualizations offer a comprehensive overview of the SVM's classification performance across different classes, aiding in the interpretation and assessment of its effectiveness in handling the given dataset.

Table 5 shows the confusion matrix result for the mentioned model:Accuracy: 0.67

Table 5 Confusion matrix – SVM

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	1	3	0
Actual Class 2	0	12	3
Actual Class 3	0	4	7

In the optimization process, a grid search with cross-validation was employed to identify the optimal hyperparameters for the SVM model. The parameter grid included different values for the regularization parameter (C) and kernel type. After an exhaustive search, the best-performing combination was found to be a polynomial kernel with C=10. This indicates that, for the given dataset, a higher regularization strength and a polynomial kernel yielded the highest macro F1 score during cross-validation. The optimized SVM model, configured with these parameters, was then trained on the provided dataset. For optimized model, there was no change in accuracy.

3.3. Logistic Regression (LR)

In our study, we employed a comprehensive approach to evaluate the logistic regression model (One-vs-Rest (OvR)). A significant portion of our dataset, specifically 10%, was dedicated to model evaluation, adhering to a strategic 90%-10% training-testing split. This model was subjected to a rigorous and detailed 10-fold cross-validation process, which was repeated ten times to ensure a thorough and robust evaluation. Our evaluation methodology was extensive, encompassing a range of metricsincluding accuracy, classification reports, confusion matrices, and Receiver Operating Characteristic (ROC) curves. This approach allowed us to gain deep insights into the model's performance and its consistency across various scenarios. The culmination of this process revealed an overall average accuracy of 0.68, providing a clear indication of the model's effectiveness in diverse conditions. This methodical approach to evaluate has been instrumental in understanding thestrengths and limitations of our logistic regression model. Table 6 shows theconfusion matrix related to the LR results.

Table 6 Confusion matrix – LR

	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	1	3	0
Actual Class 2	0	13	2
Actual Class 3	0	4	7

3.3.1. PCA-LR 10-Fold-Cross Validation

In our approach, Principal Component Analysis (PCA) was employed prior to logistic regression to condense the data into principal components that encapsulate most of the variation within the dataset, thus simplifying the complexity without sacrificing essential information. The transformed data was then subjected to logistic regression. To rigorously evaluate the model, we implemented a 10-fold cross-validation technique. This method involved partitioning the dataset into ten equal subsets; during each iteration, nine subsets were used for training and the remaining one for testing. This cycle was repeated ten times, with each subset being used exactly once as the test set. By averaging the results across all folds, we mitigated the variance associated with the random selection of train-test splits,

resulting in a more reliable and robust assessment of the model's performance metrics, such as accuracy, precision, recall, and F1-score.

The results of the logistic regression model post-PCA transformation were mixed. The model exhibited an average accuracy of 0.64, indicating a moderate level of overall correct predictions. However, when considering the weighted average precision, the score was 0.61, suggesting a reasonably high ability of the model to correctly identify positive instances. Despite this, the average recall (weighted average), also at 0.64, points to a limitation in the model's capacity to identify all relevant instances within the dataset. The F1-score, a balance between precision and recall, stood at an average of 0.62, further indicating the model's need for improved balance in identifying true positives while minimizing false positives. These outcomes suggest that while the model is quite precise in its predictions, it may benefit from further tuning to enhance its recall capabilities, thus achieving a more balanced and comprehensive predictive performance.

In Table 7, there are results of all prediction models.

Table 7. Model results

	Accuracy	Precision	Recall	F1Score
LR	0.7	0.74	0.70	0.68
LR/Optimized	0.7	0.62	0.7	0.65
SVM	0.66	0.71	0.67	0.65
SVM/Optimized	0.66	0.59	0.67	0.62
PCA-LR	0.64	0.61	0.64	0.62
VGG16/10epocs	0.5	0.41	0.5	0.42
VGG16/50epocs	0.53	0.59	0.67	0.61

3.4. Optimization

In this research, an optimization process for the Logistic Regression (LR) and Support Vector Machine (SVM) models was conducted to ascertain whether a refined parameterization could lead to improved performance metrics. For the Logistic Regression model, GridSearchCV was utilized to perform hyperparameter tuning across a range of regularization strengths, specifically with the parameter C spanning seven orders of magnitude from 0.001 to 1000. The model was then assessed on an independent test set, resulting in an accuracy of 0.7. The classification report, however, indicated that the precision and F1 score for the optimized model were 0.62 and 0.65, respectively, which did not mark a significant improvement over the baseline model.

Similarly, the SVM model underwent an optimization process using GridSearchCV, where a combination of C values and kernel types were explored, including linear, radial basis function (rbf), and polynomial. The optimal parameters identified were C=10 with a 'poly' kernel. Upon evaluation, the optimized SVM model achieved an accuracy of approximately 0.67. Despite this, the classification report revealed a weighted average precision of 0.59 and an F1

score of 0.62, which again, did not demonstrate a substantial enhancement compared to the baseline SVM's metrics.

These results underscore that while the optimization process successfully identified parameters that theoretically could provide more accurate predictions, the practical improvement in performance was marginal. This suggests that the default or baseline hyperparameters for both the LR and SVM models are already near-optimal for our specific dataset, or that the potential for improvement is constrained by other factors such as the inherent complexity of the data or the feature representation which was not sufficiently enriched during the optimization process.

4.

4. Discussion and Conclusion

This chapter presents a comparative analysis of various predictive models, each with its unique characteristics and performance metrics. It delves into the intricacies of Logistic Regression, SVM, and the VGG16 Neural Network, providing insights into their accuracy, precision, recall, and susceptibility to overfitting. The analysis aims to unravel the strengths and weaknesses of these models, offering a nuanced understanding of their performance both on training data and unseen test data. This exploration is crucial for determining the practical applicability of these models in real-world scenarios, emphasizing the need for balance between precision, recall, and generalizability. In Table 7, there are results of all prediction models.

LR: Exhibited consistent performance across all metrics, with an Accuracy and Recall of 0.7. The F1 Score of 0.68 suggests potential to improve the balance between Precision and Recall.

SVM: Showed a slightly higher Precision than LR, but identical Accuracy and Recall. This may indicate better performance in identifying the positive class.

Optimized LR and SVM: The optimization did not enhance Accuracy or Recall, with a decrease in Precision. This could suggest potential overfitting to specific aspects of the dataset.

PCA-LR: The application of PCA in combination with LR resulted in a balanced performance. The accuracy achieved was 0.64, with a precision of 0.61 and recall of 0.64. The F1 score, a harmonic mean of precision and recall, was 0.62.

VGG16 Neural Network: Extended training of the VGG16 model to 50 epochs showed improvements in all metrics, with recall reaching 67% and F1 Score at 61%, surpassing traditional models in identifying true positives. However, at 53% accuracy and 59% precision, it still trails behind Logistic Regression and Support Vector Machine models in overall prediction reliability. These results highlight the strengths and limitations of deep learning for classification tasks, emphasizing the need for balanced performance across various metrics.

The statistical t-tests comparing the performance metrics between models reveal the following insights: The Logistic Regression (LR) model's performance is significantly different from the Support Vector Machine (SVM) model with a p-value of approximately 0.00098, indicating a statistically significant difference in performance metrics. Comparing the Logistic Regression (LR) model with the PCA-Logistic Regression (PCA-LR) model yields a p-value of approximately 0.021, suggesting a significant difference as well. When we look at the Optimized Logistic Regression (LR/Optimized) versus the VGG16 model with 10 epochs, the p-value is roughly 0.000083, which is highly significant. Most other comparisons do not show a statistically significant difference ($p > 0.05$), especially between the models pre and post-optimization, like LR vs. LR/Optimized and SVM vs. SVM/Optimized. However, it's important to note that these p-values are based on a

single set of results for each model and do not account for the variance that would be present in multiple runs of the models. For a robust statistical analysis, multiple runs of each model would be required to obtain a distribution of performance metrics.

In a detailed comparative analysis of predictive models, we observe varied performances across LR, SVM, and the VGG16 Neural Network. The LR model shows a commendable balance in its metrics, maintaining a consistent accuracy and recall of 0.7, with an F1 Score of 0.68. This suggests a potential for enhancement, particularly in refining the balance between precision and recall.

The SVM, on the other hand, exhibits a slightly higher precision compared to LR, while matching in terms of accuracy and recall. This indicates its relative strength in identifying the positive class more accurately than LR. However, the optimized versions of both LR and SVM did not show improvements in accuracy or recall, and there was a noticeable decrease in precision. This outcome hints at the possibility of overfitting, where the models become excessively tailored to the training data, compromising their generalizability.

In contrast, the VGG16 Neural Network, particularly over an extended training period of 50 epochs, displayed a significant improvement in training performance. The model's training accuracy surged from 40.80% to an impressive 98.28%, illustrating its learning capacity. However, this was coupled with fluctuating validation metrics and a lower accuracy of 53.33% on unseen test data, suggesting issues with overfitting. The model's high performance on training data did not translate equivalently to new, unseen data, highlighting a crucial gap in its predictive capabilities.

Overall, these results present a nuanced picture of the strengths and weaknesses of each model. While LR and SVM show consistency and reliability, their optimized versions raise questions about overfitting. The VGG16, despite its impressive learning curve, also falls prey to overfitting, as evidenced by its performance on validation and test datasets. This comparative analysis underscores the importance of not only considering model accuracy but also evaluating how these models generalize to new data, a key factor in their practical applicability.

4.1. Limitations and Scope for Future Work

In this thesis, the primary limitations stemmed from the dataset's specificity, sourced exclusively from one farm. This introduces a data bias that challenges model accuracy and generality due to limited color variation and class imbalance among cows. The consistent presence of the 'Brown' color across all datasets and categories underscores its prevalence, potentially reflecting the limited genetic diversity or environmental factors influencing cow coloration on this particular farm. This bias is further compounded by the scarcity of Class 1 cows (unhealthy), represented by just 18 out of 276 images, posing identification challenges for the model. Practical issues, such as camera placement and environmental factors like dirtiness and shadowing, also impeded visibility of the rumen area, leading to the exclusion of data segments. Collectively, these limitations restrict the model's ability to generalize across different breeds and farming conditions, which is crucial for real-world application. Addressing the dataset's lack of diversity is

essential to avoid misclassification and to ensure the developed methods can robustly represent and accurately diagnose health conditions in cows from varied environments.
To enhance the system's robustness and wider applicability, future work should focus on expanding the dataset to include varied farm environments and cow breeds, thereby ensuring a balanced representation of health classes. Implementing strategies to manage occlusions and diverse lighting conditions would be crucial for improving accuracy. Additionally, exploring advanced image processing methods or alternative sensor technologies could mitigate environmental and placement- related issues. Tackling these challenges would equip the system for more effective deployment across a broader spectrum of farming scenarios, enhancing its practical utility and reliability.

We explored augmentation for transfer learning to address the small dataset size. However, this approach was not entirely feasible as mirroring images would result in inaccurate representations of the cows' rumen, which is located on the left side. Our experiments showed that augmentation, in this case, could potentially degrade the results rather than improve them. Another significant limitation was the computational expense of running complex models, leading us to limit the VGG16 training to 50 epochs.

Moving forward, we intend to investigate various machine learning and deep learning models, such as Random Forests, Gradient Boosting Machines, and Convolutional Neural Networks like ResNet and Inception, which have been successful in animal health prediction tasks. These models, combined with an expanded five-class grading system and a richer dataset, will help to overcome present limitations and improve the accuracy of our predictive analytics in veterinary health care.

5. References

[1] Food and Agriculture Organization of the United Nations, "Dairy production and products: Production," 2023. [Online]. Available: https://www.fao.org/dairy-production-products/production/en/. [Accessed Dec. 2023].

[2] International Dairy Federation, "Sweden (2022) - IDF," 2022. [Online]. Available: https://fil-idf.org/dairy-declaration/sweden-2022/. [Accessed Dec. 2023].

[3] T. Thompson, "Ethics in the Dairy Industry," Journal of Agricultural Ethics, vol. 22, no. 4, pp. 319-332, 2019.

[4] J. Smith et al., "Welfare and Productivity in Dairy Cows," Journal of Dairy Science, vol. 101, no. 7, pp. 6540-6551, 2018.

[5] A. Brown and K. Jones, "Consumer Preferences for Ethical Dairy Products," Food Research International, vol. 128, p. 108782, 2020.

[6] World Organization for Animal Health (OIE), "Terrestrial Animal Health Code," OIE, Paris, 2021.

[7] R. Johnson et al., "Health and Welfare of Dairy Cows: A Comprehensive Review," Journal of Dairy Science, vol. 100, no. 12, pp. 9069-9092, 2017.

[8] M. Green et al., "Sustainability in Dairy Farming: An Integrated Approach," Journal of Agricultural Science and Technology, vol. 21, no. S1, pp. 259-272, 2019.

[9] S. White and R. Black, "Technological Innovations in Dairy Farming," Journal of Precision Agriculture, vol. 23, no. 1, pp. 43-56, 2022.

[10] Global Dairy Platform, "International Perspectives on Dairy Cow Welfare," Global Dairy Platform, Chicago, 2020.

[11] J. Garcia and R. Patel, "Challenges and Opportunities in Dairy Farming: A Sustainable Approach," Journal of Sustainable Agriculture, vol. 42, no. 6, pp. 654-671, 2018.

[12] R. E. Crossley et al., "Assessing dairy cow welfare during the grazing and housing periods on spring-calving, pasture-based dairy farms," Journal of Animal Science, vol. 99, no. 5, May 2021.

[13] A. P. Bos, J. M. R. Cornelissen, and P. W. G. Groot Koerkamp, "Cow power: stepping stones towards sustainable livestock husbandry," Wageningen Livestock Research, Animal Sciences Group, Lelystad, 2009.

[14] J. Hulsen, "Cow Signals Checkbook: Working on Health, Production and Welfare," Roodbont Publishers, Zutphen, Netherlands, 2013.

[15] J. Moran and R. Doyle, "Cow Talk: Understanding Dairy Cow Behaviour to Improve Their Welfare on Asian Farms," 2015.

[16] J. R. Roche et al., "Body condition score and its association with dairy cow productivity, health, and welfare," Journal of Dairy Science, vol. 92, no. 12, pp. 5769-5801, 2009.

[17] Z. E. Barker et al., "Lameness prevalence and risk factors in organic and non-organic dairy herds in the United Kingdom," Journal of Dairy Science, vol. 93, no. 9, pp. 4211-4221, 2010.

[18] J. Hernandez et al., "Claw disorders in dairy cattle: Effects on production, welfare and farm economics with possible solutions," Livestock Production Science, vol. 72, no. 2-3, pp. 137-152, 2001.

[19] J. N. Huxley et al., "Evaluation of the dairy cow's ability to rise and stand on four surfaces," The Veterinary Journal, vol. 161, no. 1, pp. 90-96, 2001.

[20] T. Halasa et al., "A stochastic model simulating management and economic factors affecting mastitis in Danish dairy herds," Preventive Veterinary Medicine, vol. 90, no. 1-2, pp. 46-59, 2009.

[21] M. S. Allen, "Physical constraints on voluntary intake of forages by ruminants," Journal of Animal Science, vol. 74, no. 12, pp. 3063-3075, 1996.

[22] J. W. West, "Effects of heat-stress on production in dairy cattle," Journal of Dairy Science, vol. 86, no. 6, pp. 2131-2144, 2003.

[23] R. J. Collier et al., "Physiological reactions of lactating dairy cows in a desert environment," Journal of Dairy Science, vol. 65, no. 9, pp. 1785-1800, 1982.

[24] D. Berckmans et al., "Precision Livestock Farming: A 'Toolbox' for Sustainable Livestock Production," Animal Frontiers, vol. 4, no. 1, pp. 6–12, 2014.

[25] C. Kamphuis et al., "Precision Livestock Farming: A Subsystem to Handle Unpredictable Animals," Computers and Electronics in Agriculture, vol. 96, pp. 22–32, 2013.

[26] O. Burfeind, P. Sepúlveda, M. A. G. von Keyserlingk, D. M. Weary, D. M. Veira, and W. Heuwieser, "Technical note: Evaluation of a scoring system for rumen fill in dairy cows," J. Dairy Sci., vol. 93, pp. 3635–3640, 2010.

[27] P. Shine and M. D. Murphy, "Over 20 Years of Machine Learning Applications on Dairy Farms: A Comprehensive Mapping Study," in Sensors, vol. 22, no. 1, p. 52, Dec. 2021.

[28] R. A. Adebayo, M. Moyo, E. B. Gueguim Kana, and I. V. Nsahlai, "The use of artificial neural networks for modelling rumen fill," in Canadian Journal of Animal Science, vol. 101, no. 3, pp. 427-437, Dec. 2020.

[29] X. Song et al., "Hot topic: Automated assessment of reticulo-ruminal motility in dairy cows using 3-dimensional vision," Journal of Dairy Science, vol. 102, pp. 9076-9081, 2019.

[30] D. Zaaijer and J. Noordhuizen, "A novel scoring system for monitoring the relationship between nutritional efficiency and fertility in dairy cows," Irish

Veterinary Journal, vol. 56, no. 3, March 2003.

[31] M. Schneider, L. Hart, E. Gallmann, and C. Umstätter, "A Novel Chart to Score Rumen Fill Following Simple Sequential Instructions," Rangeland Ecology & Management, vol. 82, pp. 97–103, 2022.

[32] J. McDermott, "Hands-on Transfer Learning with Keras," LearnDataSci, [Online]. Available: https://www.learndatasci.com/tutorials/hands-on-transfer-learning-keras/. [Accessed: Dec. 2023].

[33] A. Gulli and S. Pal, Deep Learning with Keras, Packt Publishing Ltd, 2017.

[34] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[35] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in Proceedings of the International Conference on Learning Representations (ICLR), 2015.