Shahid Bahonar University of Kerman

Biomechanism and Bioenergy Research

Online ISSN: 2821-1855
Homepage: https://bbr.uk.ac.ir

Iranian Society of Agricultural Machinery Engineering and Mechanization

Exploring Key Visual Features for Early Lameness Detection: Toward Transparent Intelligence

Ali Khalili Tazehkandgheshlagha[1], Ali Jafaria[1]✉, Seyed Saeid Mohtasebia[1], Hossein Navid[2]

[1] Department of Agricultural Machinery Engineering, Faculty of Agriculture, University of Tehran, Karaj, Iran.
[2] Department of Biosystems Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran.

✉ Corresponding author: jafarya@ut.ac.ir

ARTICLE INFO

Article type:

Research Article

Article history:

Received 02 March 2025

Received in revised form 30 April 2025

Accepted 31 May 2025

Available Online 30 June 2025

Keywords:

Multivariate Time-Series Analysis, Feature Engineering, Deep Learning, Machine Learning, Animal Health and Welfare.

ABSTRACT

Lameness in cattle, characterized by abnormal stride and gait, poses significant economic and welfare challenges in agriculture. Traditional visual inspections lack accuracy and scalability, prompting the development of transparent computer vision-based detection systems. This study leverages a dataset of 170 cattle videos from public sources and the University of Tehran's Cattle Farm, preprocessed into 1226 one-second sub-clips (416×416 pixels, 25 FPS) to mitigate noise from unpredictable cattle behavior. Using the YOLOv7 model, we extracted 35 temporal features, including step sizes, speed, acceleration, and relative head-to-leg coordinates, focusing on the cattle's head, legs, and back. These features were further engineered using time-series characterization techniques and hypothesis testing, yielding 3773 features. A deep learning model, trained on these features, achieved 88.66% accuracy and 93.74% AUC, while a Light Gradient Boosting Machine model on engineered features reached 81.3% accuracy and 90.8% AUC. Sensitivity analysis highlighted leg and head-related features as critical for lameness detection. By emphasizing interpretable features and robust modeling, this approach enhances transparency, improving animal welfare and farm productivity under diverse conditions.

Cite this article: Khalili Tazehkandgheshlagha, A., Jafaria, A., Mohtasebia, S.S., & Navid, H (2025). Exploring Key Visual Features for Early Lameness Detection: Toward Transparent Intelligence. *Biomechanism and Bioenergy Research*, *4*(2), 55-73. https://doi.org/10.22103/bbr.2025.25007.1119

Publisher: Shahid Bahonar University of Kerman
DOI: https://doi.org/10.22103/bbr.2025.25007.1119

INTRODUCTION

Recent technological advancements have brought about a transformative shift in the agricultural sector, ushering in improvements in productivity, sustainability, and resilience (Ogundari & Bolarinwa, 2018; Peng et al., 2022). Embracing these innovations has become imperative to unlock agriculture's full potential (Ogundari & Bolarinwa, 2018). One notable area influenced by cutting-edge technologies is the timely and accurate diagnosis of animal diseases, crucial for effective treatment and economic loss prevention (Ryu & Tai, 2022).

The early detection and diagnosis of lameness in cattle is of particular significance. Lameness is characterized by abnormal stride and gait, often accompanied by pain (Schlageter-Tello et al., 2014). Neglecting lameness can result in substantial economic repercussions, such as premature culling (Barkema et al., 1994), diminished milk production (Alban et al., 1996; Green et al., 2002; Warnick et al., 2001), weight loss (Huxley, 2013), reduced fertility (Melendez et al., 2003), decreased slaughter value, and increased treatment costs(Ettema & Østergaard, 2006). Farmers and experts have traditionally relied on visual inspections to detect animal lameness. However, this method's accuracy and scalability are limited, prompting the development of intelligent detection systems (Flower & Weary, 2006). To address these limitations, a focus on non-intrusive methods has emerged as a valuable approach, with computer vision leading the way due to its minimal disruption of natural animal behavior, accuracy, accessibility, and cost-effectiveness (Alsaaod et al., 2015; Bagheri et al., 2024).

Research efforts have explored various approaches in this domain. One study employed a 3-dimensional depth camera to capture crucial aspects of cows' movement during walking, with the recorded key regions subsequently used to generate signals representing cow movement. This approach ultimately enabled the implementation of an early detection system for lame cows using a linear SVM[1] classifier, achieving a commendable accuracy rate of 95.7% (Abdul Jabbar et al., 2017). Another study gathered radar data from the front, back, and side of cows moving through the milking parlor corridor in three distinct directions. After denoising the radar data and extracting significant features, these were combined with veterinarian-conducted animal identification results. A SVM classifier was then employed for lameness detection, yielding an accuracy of 85% for dairy cows (Shrestha et al., 2018).

A lameness detection approach proposed by (Jia et al., 2025) integrated LCCCT[2] and DSKNN[3]. Their methodology involved decomposing videos, using LCCCT to extract upper contour pixel regions, and analyzing head and neck tilt data with DSKNN. By applying the KNN[4] algorithm to the processed data, they achieved a strong detection rate suitable for lameness classification. In a related approach, a detection method based on a double normal background statistical model was introduced. This model's structure and parameters were derived by analyzing and tracking pixel changes associated with dairy cows and their gait characteristics, enabling effective lameness detection (Jiang et al., 2019).

Further innovation came with the use of the YOLOv3[5] deep learning algorithm for intelligent lameness identification in cattle. The method analyzed cow videos, detected legs using YOLOv3, computed relative step size, and generated a feature vector. A classification model trained with LSTM[6] on this characteristic vector achieved an impressive accuracy of 98.79%, outperforming classifiers such as SVM, KNN, and DTC[7] (Wu et al., 2020).

Despite their high accuracy, a common drawback of many of these techniques is their reputation as "black boxes," mainly because they

[1] Support Vector Machine
[2] Local Circulation Center Compensation Tracking
[3] Distilling Data of K-Nearest Neighbor
[4] K-Nearest Neighbor
[5] You Only Look Once version 3
[6] Long Short-Term Memory
[7] Decision Tree Classification

often fail to provide transparent explanations for their underlying mechanisms. This limitation, in turn, hinders their adaptability and undermines their trustworthiness (Lecun et al., 2015). In an AI-dominated era marked by a lack of transparency in models, this research aims to demystify the opaque nature of intelligent lameness detection techniques, offering a clearer understanding of how they operate. Such transparency is crucial, as it is a pressing concern across various fields relying on machine learning (Abbasian Ardakani et al., 2023; Arrieta et al., 2020; Merkin et al., 2022; Oliveira et al., 2021; Schmid & Finzel, 2020). This research seeks to unravel the intricacies of lameness detection and has direct implications for animal health and welfare.

Further refinement of lameness detection methods can profoundly impact the diagnosis and treatment of lameness in animals, ultimately enhancing their overall well-being and productivity. However, it is crucial to acknowledge the inherent challenges in treating and preventing lameness disorders, with the prognosis not always consistently favorable. These difficulties can limit the potential benefits of automated lameness detection. Achieving a positive outcome relies heavily on our comprehension of lameness, which is nearly impossible without the necessary insights.

Therefore, our study investigates key visual features through advanced feature extraction techniques and employs multiple statistical models to enhance the transparency and interpretability of intelligent lameness detection systems. Unlike previous research, which often relies on opaque "black-box" models with limited explainability, our approach innovatively prioritizes transparency by dissecting the core components of lameness detection and elucidating how learning algorithms process visual data. A key advantage of this study lies in its novel modeling strategy, which uses minor sequences (one-second video sub-clips) to mitigate the noise introduced by unpredictable cattle behaviors, such as irregular step sizes or sudden stops. This method improves detection accuracy and robustness across diverse farm environments compared to existing approaches that struggle with such variability. Additionally, our comprehensive feature extraction, focusing on critical attributes like cattle step sizes, speed, acceleration, and relative head-to-leg coordinates, provides a more interpretable and reliable framework for lameness detection. By offering clear insights into the inner workings of these models, our research not only advances the accuracy and resilience of automated lameness detection but also sets a new standard for trustworthy AI systems in animal welfare applications.

MATERIALS AND METHODS

Study Population

The dataset used in this study consists entirely of video data. We collected this data from two primary sources: publicly available online platforms and a controlled cattle farm environment. Public data was gathered from resources such as YouTube[1], Kaggle[2], GitHub[3], and similar open-access platforms, providing a diverse range of cattle videos under various environmental and management conditions. Additionally, we obtained high-quality video recordings from the Cattle Farm of the College of Agriculture & Natural Resources at the University of Tehran[4], located in Karaj at the foothills of the central Alborz Mountains, Iran. These recordings focused exclusively on the Holstein Friesian breed. The final dataset integrates the publicly sourced videos and the farm-recorded samples, resulting in a comprehensive and diverse video dataset covering multiple cattle breeds and settings.

Pre-Processing

This study consists of two primary components, each necessitating distinct data types. The components include an Object

[1] YouTube
[2] Kaggle
[3] Github
[4] University of Tehran

Detection System and a Lameness Detection. Therefore, in this section, we will delve into the data preparation methods for each of these components.

Data for Object Detection System

We primarily collected data in video format. Our initial step to prepare this data for our Object Detection component involved extracting frames from a randomly chosen subset of these videos. During the frame extraction process, we utilized a Laplacian Filter, a commonly employed tool for detecting sharp edges in images. This filtering method helped us remove any blurry frames. This approach excluded frames with sharpness levels below a specific empirically determined threshold[1].

Furthermore, we implemented a Similarity Comparison Technique to remove redundant images that added little value. To accomplish this, we leveraged OpenAI's CLIP Model, specifically CLIP-ViT-B-32. This model encoded the images, enabling us to compare them against one another. This comparison yielded a list of pairs with the highest Cosine Similarity Scores (CSS). Images with a CSS score above our empirically established threshold of 0.93 were flagged as duplicates and promptly removed.

After pre-processing the frames, we constructed a dataset consisting of 428 images. We used an approximately 90/10 split, selecting 386 images for training and reserving the remaining 42 for testing. This ratio was chosen to maximize the amount of data available for training, which is often beneficial in computer vision tasks with limited datasets. While an 80/20 split is more commonly used in general machine learning, a 90/10 split is more appropriate in this context to help the model learn more effectively from a small dataset.

All images were resized uniformly to 416 × 416 pixels. This specific resolution was chosen to strike a balance between computational efficiency and sufficient spatial resolution for detecting relatively small body parts. Higher resolutions such as 640 or 1280 pixels were considered, but they would have significantly increased memory usage and processing time without a meaningful gain in detection performance for our use case.

The images were annotated using Roboflow's annotation tool. Each image contains a total of six bounding boxes corresponding to three object classes: four instances of "Leg," one instance of "Head," and one instance of "Back." The bounding boxes used were standard axis-aligned rectangles, which are compatible with most object detection models. The reasoning behind selecting these three classes is explained in the following sections. Figure 1 shows an example of these annotated images.

[1] For each image, the Laplacian variance is calculated and if the Laplacian variance is below the predefined threshold of 100, the corresponding image file is removed.

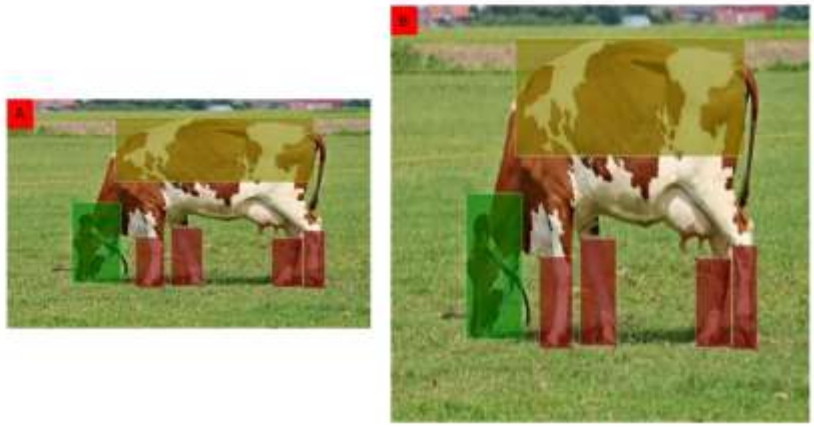

Figure 1. An annotated Cattle Image; featuring four annotation boxes for "Leg," one for "Head," and one for "Back." Image: A- displays the original size of the image, while Image B- depicts the image resized to dimensions of 416 x 416.

Data for Lameness Detection System

A Locomotion Score (LS) is a measure of lameness in cattle, and two commonly used LS systems are the three-point system and the five-point system (Sprecher et al., 1997). The five-point system can be transformed into a three-point system, where LS scores of 1 and 2 in the five-point system correspond to a score of 1 in the three-point system. A score of 3 in the five-point system is equivalent to a score of 2 in the three-point system, and scores of 4 and 5 in the five-point system correspond to a score of 3 in the three-point system.

In this study, each video collected featured a single cattle, so the number of cattle investigated in this research matched the number of videos. These videos were initially assessed using both the three-point and five-point LS systems. Videos initially measured using the five-point system were converted to the three-point system. In the three-point system, a score of 1 indicates healthy cattle, where lameness is not noticeable; hence, the cattle are considered healthy. Scores of 2 and 3 in the three-point system indicate lame cattle. After standardizing all the videos to the same LS system, we categorized them as either lame or healthy based on the scoring system. A score of 1 was classified as representing a healthy animal, while scores of 2 and 3 were classified as indicating lameness.

Subsequently, we resized all the videos to a uniform size of 416 × 416 pixels and ensured they played at 25 frames per second (FPS). Please refer to Table 1 for a comprehensive overview of the dataset, its divisions, and other relevant properties.

Table 1 - Lameness Detection Dataset Properties. This table shows the train/test split configurations, subsequently utilized in preparing the dataset for feature extraction.

Slicing	Count	Healthy	Lame
Source 1[1]	104	59	45
Source 2[2]	66	57	9
All	170	116	54
Training	150	106	44
Testing	20	10	10

Feature Extraction

We employed the YOLOv7 object detection model (Wang et al., 202۳) for our object detection algorithm, specifically leveraging the pre-trained weights provided by the YOLOv7 repository and explicitly using the YOLOv7-X variant. The training was performed on a Linux environment

[1] Source 1 refers to the part of the dataset that was collected from publicly available sources across the internet.

[2] Source 2 refers to the part of the dataset that was collected from the Cattle Farm associated with the College of Agriculture & Natural Resources at the University of Tehran.

using a Tesla T4 GPU[1], over 200 epochs with a batch size of 32. We utilized transfer learning, starting from the official pre-trained weights and fine-tuning on our custom data. The training was conducted using the default hyperparameters specified in the official YOLOv7 implementation, which includes the SGD[2] optimizer, a learning rate of 0.01, momentum of 0.937, and a weight decay of 0.0005.

Subsequently, we divided the videos into smaller sub-clips to prepare the lameness detection dataset for feature extraction. Given the unpredictable nature of cattle behavior while walking, for instance, variations in step size, speed, and occasional stops at irregular intervals, we created sub-clips with a duration of one second, equivalent to 25 frames or time steps. This decision was made to reduce data interference.

Figure 2 illustrates the irregularities in cattle movement patterns. While the seasonal patterns[3] in the data remain consistent over time, the variability in step size introduces noise that can cause the lameness detection algorithm to misclassify healthy cattle as lame, as lameness is often associated with shorter strides. By selecting one-second sub-clips for the videos, we aimed to minimize this noise, enabling the lameness detection algorithm to better capture the underlying seasonal correlations, even when dealing with shortened strides.

As a result of splitting the original videos into one-second sub-clips, we created 1226 video segments. Of these, 1074 were included in the training set, and 152 were allocated to the testing set. Table 2 displays the complete properties of the processed dataset utilized for feature extraction.

Table 2 - Feature Extraction Dataset Properties. The train/test split shown is identical to the configuration presented in Table 1.

Slicing	Count	Healthy	Lame
Source 1	857	369	488
Source 2	369	328	41
All	1226	697	529
Training	1074	616	458
Testing	152	81	71

We generated features by inputting the preprocessed video clips into the object detection system and recording the results as [x, y] coordinates representing the center points of the predicted bounding boxes. These features displayed temporal characteristics, as they were derived from sequential frames.

Following this, we designed a Python script to refine the acquired data, reshaping the structure of the recorded features. The resulting processed files were structured with 25 time steps and 12 distinct features. These features encompassed the x and y coordinates for the cattle's Head, Back, and four Legs, with each attribute stored in a separate column.

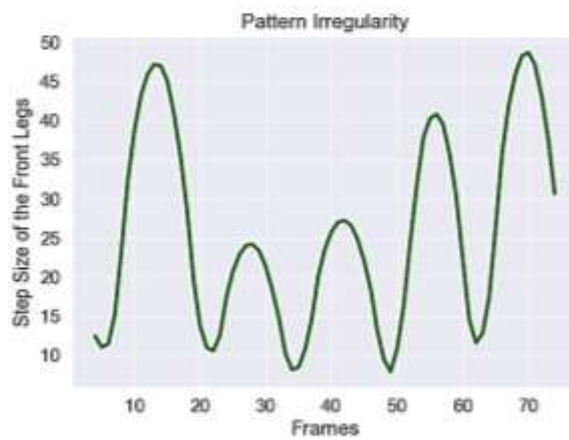

Figure 2. Changes in the step size of a single healthy cattle measured in pixels within a 3-second time frame.

In the case of identifying the front legs from the rear legs, we considered five possible scenarios. If no leg was detected or the number of detected legs was one or two, we set all corresponding features to zeroes. When three legs were detected,

[1] Graphics Processing Unit
[2] Stochastic Gradient Descent

[3] Seasonal patterns entail recurrent variations or trends in data that follow a consistent pattern over specific intervals, irrespective of the nature of those intervals.

we used the head position to distinguish the front legs from the rear legs, setting the two legs closest to each other while considering the other two as zeroes. If all four legs were detected, we again used the head coordinates to distinguish the front legs from the rear legs, adjusting all the values accordingly. This method effectively eliminated directional complications from the cattle's moving direction, whether moving from the right of the video to the left or vice versa.

We then derived additional features from the extracted [x, y] coordinates. The complete list of these features and their descriptions are shown in Table 3.

Table 3 - The Complete List of Features Extracted For Lameness Detection[1].

Feature Name	Description
xhead	x coordinate of the centroid of the cattle's head
yhead	y coordinate of the centroid of the cattle's head
xback	x coordinate of the centroid of the cattle's back
yback	y coordinate of the centroid of the cattle's back
xlegFrontFirst	x coordinate of the centroid of the cattle's first front leg
ylegFrontFrist	y coordinate of the centroid of the cattle's first front leg
xlegFrontSecond	x coordinate of the centroid of the cattle's second front leg
ylegFrontSecond	y coordinate of the centroid of the cattle's second front leg
xlegRearFirst	x coordinate of the centroid of the cattle's first rear leg
Feature Name	**Description**
ylegRearFirst	y coordinate of the centroid of the cattle's first rear leg
xlegRearSecond	x coordinate of the centroid of the cattle's second rear leg
ylegRearSecond	y coordinate of the centroid of the cattle's second rear leg
FRS[2]	The step size of the front legs
RRS[3]	The step size of the rear legs
HB[4]	Position of the head to the back
FFHead[5]	Position of the first front leg to the head
SFHead[6]	Position of the second front leg to the head
FRHead[7]	Position of the first rear leg to the head
SRHead[8]	Position of the second rear leg to the head
FFBack	Position of the first front leg to the back
SFBack	Position of the second front leg to the back
FRBack	Position of the first rear leg to the back
SRBack	Position of the second rear leg to the back
InsSpeed - FRS/Frame	Instantaneous speed of the front legs with FRS/Frame unit
InsSpeed - RRS/Frame	Instantaneous speed of the rear legs with RRS/Frame unit
InsSpeed - Overall	Overall instantaneous speed of the cattle
InsAcceleration - FRS/Frame	Instantaneous acceleration of the front legs
InsAcceleration - RRS/Frame	Instantaneous acceleration of the rear legs
InsAcceleration - Overall	Overall instantaneous acceleration of the cattle
AvgSpeed - FRS/Frame	The average speed of the front legs
AvgSpeed - RRS/Frame	The average speed of the rear legs
AvgSpeed - Overall	The overall average speed of the cattle
AvgAcceleration - FRS/Frame	The average acceleration of the front legs
AvgAcceleration - RRS/Frame	The average acceleration of the rear legs
AvgAcceleration - Overall	The overall average acceleration of the cattle

[1] The names are written as how they were used in the codes.
[2] Front Relative Step-Size
[3] Rear Relative Step-Size
[4] Head Back
[5] First Front Head
[6] Second Front Head
[7] First Rear Head
[8] Second Rear Head

Feature Engineering and Analysis

We developed a deliberately lightweight deep-learning model to isolate and quantify the contribution of each engineered feature to cattle lameness detection shown in Figure 3. The network begins with an LSTM layer that summarizes temporal dynamics into a compact hidden state. This s followed by a pooling operation that reduces dimensionality and noise, and then a single 1D convolutional layer that extracts the most salient temporal patterns. A second pooling stage further concentrates these features before they are passed to fully connected perceptron layers for final modeling and binary classification.

The reason for this simple design such as using only one convolutional layer—as opposed to stacking multiple layers—was intentional. Since the focus of this study is on feature analysis rather than optimizing predictive accuracy, we prioritized model interpretability and feature impact visibility. A deeper architecture with additional convolutional layers would have increased the model's capacity, making it harder to isolate the effect of individual features. In contrast, a smaller model ensures that each feature has a relatively larger influence on the model's behavior. This design choice helps us better understand the role and importance of each feature, which would be significantly diluted in a more complex, high-capacity model.

Before conducting the sensitivity analysis, we first trained the deep learning model using the full features. We trained the model for 150 epochs with 32 batches on a Core i7 CPU in a Linux environment, using the Adam optimizer with a learning rate of 0.001.

Then, we trained and tested each feature separately using the deep learning model and the same hardware and hyperparameters.

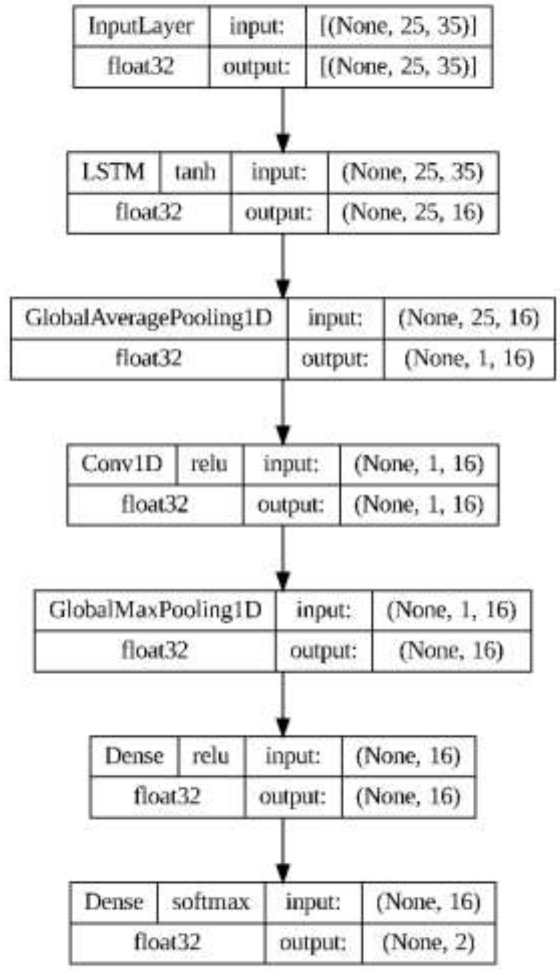

Figure 3. Deep Learning Model Architecture. The model was developed using Python, Tensorflow and Keras.

To further explore the impact of individual features, we employed a time-series feature extraction method using Hypothesis Testing, as outlined in (Christ et al., 2018). This method extracts a comprehensive set of statistical and mathematical features from time-series data based on significance tests to retain only the most relevant ones. Combining 63 distinct time-series characterization techniques, we transformed the initial lameness dataset—comprising 1226 samples and 35 manually selected features—into a structured dataset with 1226 rows and 27,406 columns, each representing a newly generated feature. These features encapsulated various aspects of the time-series sequences, including the number of peaks, average or maximal values, and more intricate characteristics like the timereversal symmetry statistic. The method we used for feature generation is detailed in (Christ et al., 2018).

Subsequently, we carried out a feature selection step to remove features that lacked a statistically significant relationship with the target variable. This process involved applying univariate hypothesis testing, where each extracted feature was independently evaluated for its correlation with the labels. Features that did not meet a predefined statistical significance threshold (p-value < 0.05) were discarded. The threshold was chosen based on standard statistical convention to control the false discovery rate while retaining informative features.

After feature selection, we applied additional preprocessing steps to prepare the data for analysis. These included standardizing the feature values to have zero mean and unit variance, and removing constant or near-constant features that provided no discriminative power. Standardization, in particular, was essential to ensure that features with different numerical ranges contributed equally to the learning process. These preprocessing steps helped improve the stability and performance of the models by reducing redundancy and ensuring the data was in a suitable format for training.

Following these preprocessing steps, we arrived at a dataset with dimensions 1226 × 3773 (Samples × Features), each feature bearing a prefix indicating its origin from the original feature. Using this feature-engineered dataset, we assessed the effectiveness of these engineered features by training and evaluating multiple models with identical hardware and train/test split configurations as those used in the deep learning model. Furthermore, in the final analysis, we quantified the contribution of each new feature to the model's overall predictive capability by counting how many newly derived features were associated with each original feature.

Evaluation Metrics[1]

Accuracy

The accuracy metric assesses how frequently predictions match the actual labels. Eq. (1) displays the formula for calculating accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

In Eq. (1),

TP (True Positives) shows an outcome where the algorithm correctly predicts the positive class.

TN (True Negatives) shows an outcome where the algorithm correctly predicts the negative class.

FP (False Positives) shows an outcome where the algorithm incorrectly predicts the positive class.

FN (False Negatives) shows an outcome where the algorithm incorrectly predicts the negative class.

Mean Average Precision

Mean Average Precision (mAP) is a metric used to measure the performance of a model for tasks such as object detection and is calculated across all classes, considering various Intersection over Union (IoU) thresholds. It is defined as:

$$mAP = \frac{1}{n}\sum_{k=1}^{n} AP_k \quad (1)$$

Where, n is the number of classes and AP_k is the AP of class k.

IoU, or Intersection over Union, is defined as:

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \quad (2)$$

Where

Area of Intersection refers to the shared area between the predicted and ground truth bounding boxes.

Area of Union represents the total area covered by both bounding boxes.

[1] This section excludes certain metrics as they were solely used for supplementary assessment.

AP, which stands for Average Precision, quantifies the area under the precision-recall curve and is defined as:

$$AP = \int_0^1 p(r)dr \tag{4}$$

Here, *p(r)* signifies precision at a given recall level *r*

And Precision is the proportion of true positive predictions among all positive predictions, while recall is the proportion of true positive predictions among all actual positives and are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

AUC

A ROC Curve[1] s a graphical representation that illustrates how well a classification algorithm performs across various classification thresholds. This curve is characterized by two key parameters:

1. True Positive Rate (Recall).
2. False Positive Rate, which is defined as:

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

AUC, or Area under the ROC Curve, quantifies the entire area beneath the ROC curve, spanning from the origin (0, 0) to the point (1, 1). This metric is particularly valuable when an imbalance exists between different classes within the dataset. Class imbalances in real-world scenarios are common, causing traditional accuracy metrics to exhibit bias towards the dominant class and yield unreliable results. The utilization of the AUC metric effectively mitigates this bias and renders the evaluation metric nearly immune to the effects of class imbalance.

RESULTS AND DISCUSSION

After training, the object detection model achieved a mean Average Precision (mAP) of 99.6% in accurately identifying various parts of the cattle's body. This score was calculated using the COCO[2] evaluation metric, mAP@0.5:0.95, which averages performance across multiple IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05. Unlike fixed-threshold evaluations, this metric provides a more comprehensive measure of both localization and classification accuracy by assessing model performance across a spectrum of overlap requirements.

This means the model was highly effective in both locating and correctly classifying each targeted body part across different samples, demonstrating strong performance in precision and recall. The training and evaluation[3] processes are shown in Figures 4 through 7.

---

[1] Receiver Operating Characteristic Curve

[2] COCO

[3] Due to the structure and constraints of the framework and pipeline used for training, the term testing refers to the validation phase — that is, evaluating the model on a held-out validation set. Since no hyperparameter tuning was performed or necessary in this case, the validation set effectively served as the testing set. The results shown here are therefore from the validation phase and are intended to visualize how the model's accuracy evolved during training. It's important to note that evaluation is used here in a broader sense, distinct from training, validation, or testing, and refers to the general process of assessing the model's performance — in line with the Cambridge Dictionary definition: "the process of judging or calculating the quality, importance, amount, or value of something."

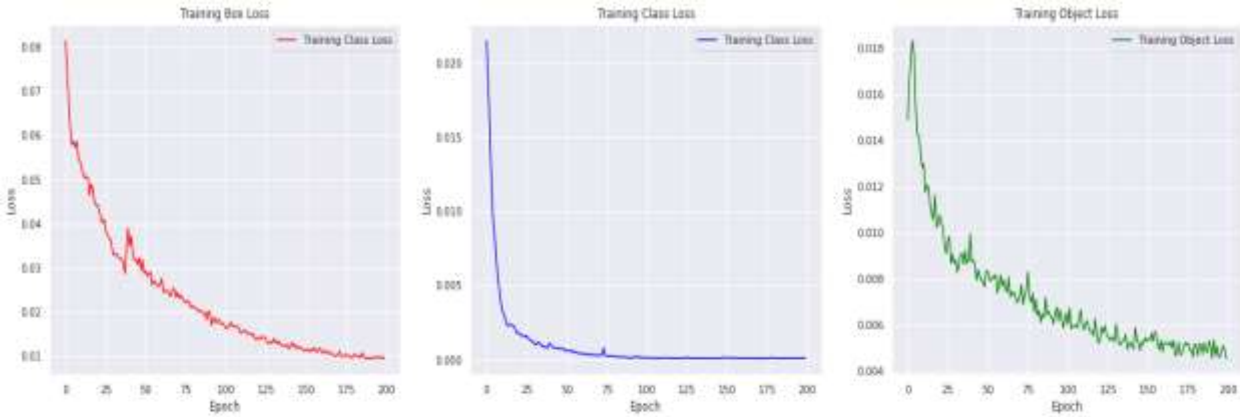

Figure 4. Training Losses for Object Detection Model. The left plot illustrates the box loss, while the center and right plots depict the class and object losses, respectively, for the Object Detection model.

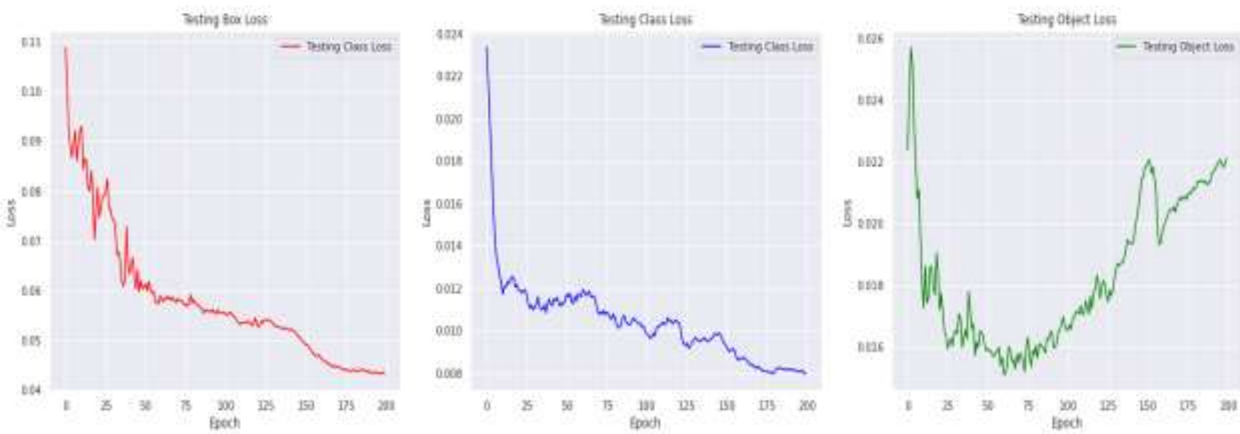

Figure 5. Evaluation Losses for Object Detection Model. The left plot illustrates the box loss, while the center and right plots depict the class and object losses, respectively, for the Object Detection model.

During the training process, the focus is directed towards reducing three key losses—box loss, class loss, and object loss— as depicted in Figures 4 and 5. The minimization of box loss enhances the model's precision in aligning predicted bounding boxes with the actual positions of objects in images. Simultaneously, diminishing class loss is crucial for accurately assigning class labels to detected objects. Additionally, reducing object loss helps the model adeptly distinguish between regions containing objects and those without. Due to the intricate characteristics of deep learning models, unraveling specific aspects of these models poses a significant challenge, which is a key focal point of our paper. Nonetheless, the notable uptick in the object loss in Figure 5 could suggest a subtle overfitting issue despite the decreasing trend in the other two losses. This phenomenon could be associated with hyperparameters, such as the learning rate and their values as they evolve throughout training.

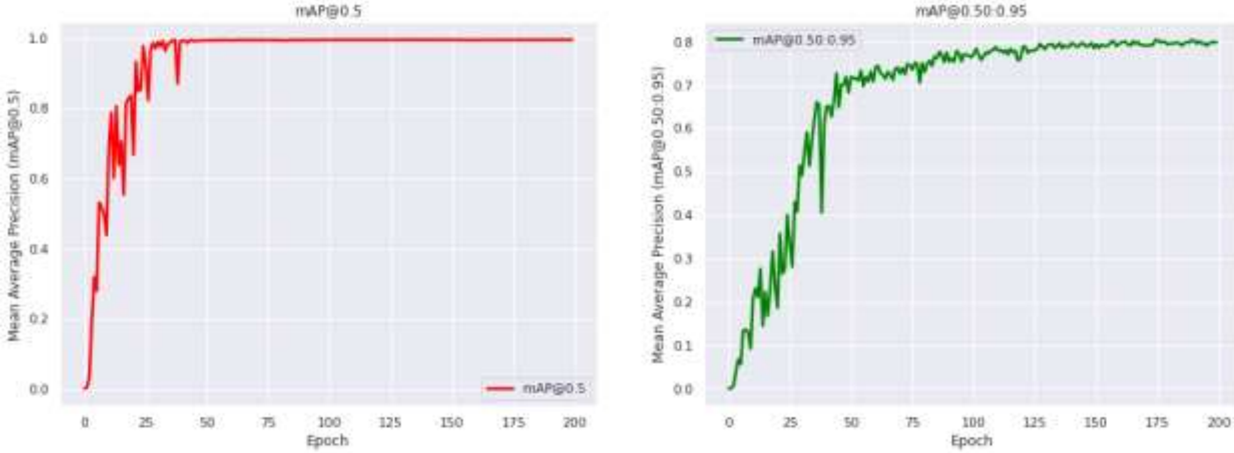

Figure 6. Object Detection Model Evaluation mAP.

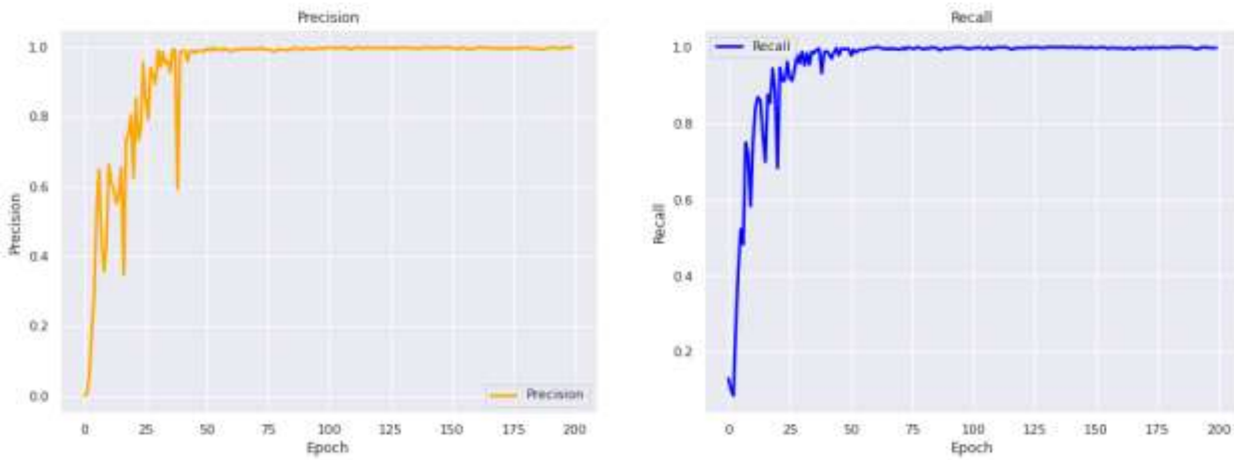

Figure 7. Object Detection Model Evaluation Precision, Recall.

Figures 6 and 7 depict the model's performance across various training epochs. The graphs commence with a low value in all the plots and exhibit a nearly consistent upward trend. The abrupt fluctuations observed in all plots may be attributed to the model making final adjustments to pre-trained weights just before reaching convergence. Figure 8 shows a few of the predicted bounding boxes.

Figure 8. Object Detection Model Predictions.

After training the deep learning model using all features, it achieved an 88.66% accuracy and a 93.74% AUC. Figure 9 shows visual representations of the model's training and evaluation performance, including loss, accuracy, and AUC. The sensitivity analysis results for the training and test sets are shown in Table 4, Figures 10 and 11.

Table 4 - Sensitivity Analysis Results Using the Deep Learning Model.

Feature Name	Train Loss	Train Accuracy	Train AUC	Test Loss	Test Accuracy	Test AUC
xhead	0.646	0.622	0.659	0.666	0.573	0.644
yhead	0.640	0.626	0.681	0.604	0.626	0.741
xback	0.666	0.595	0.623	0.672	0.560	0.569
yback	0.649	0.578	0.627	0.687	0.506	0.540
xlegFrontFirst	0.641	0.625	0.672	0.673	0.573	0.616
ylegFrontFrist	0.650	0.610	0.647	0.699	0.579	0.597
xlegFrontSecond	0.638	0.614	0.671	0.714	0.600	0.571
ylegFrontSecond	0.653	0.613	0.653	0.691	0.613	0.576
xlegRearFirst	0.558	0.669	0.761	0.662	0.560	0.623
ylegRearFirst	0.644	0.595	0.653	0.666	0.606	0.641
xlegRearSecond	0.507	0.694	0.802	0.678	0.553	0.590
ylegRearSecond	0.657	0.593	0.634	0.670	0.586	0.638
FRS	0.471	0.769	0.853	0.813	0.666	0.706
RRS	0.560	0.703	0.771	0.628	0.646	0.722
HB	0.651	0.589	0.631	0.691	0.519	0.515
Feature Name	**Train Loss**	**Train Accuracy**	**Train AUC**	**Test Loss**	**Test Accuracy**	**Test AUC**
FFHead	0.617	0.626	0.709	0.630	0.686	0.705
SFHead	0.621	0.644	0.703	0.633	0.660	0.693
FRHead	0.641	0.613	0.669	0.634	0.613	0.658
SRHead	0.642	0.613	0.657	0.642	0.600	0.645
FFBack	0.630	0.622	0.663	0.703	0.553	0.556
SFBack	0.630	0.630	0.704	0.687	0.540	0.564
FRBack	0.609	0.645	0.714	0.697	0.626	0.631
SRBack	0.626	0.632	0.687	0.690	0.573	0.613
InsSpeed - FRS/Frame	0.452	0.739	0.851	0.634	0.680	0.763
InsSpeed - RRS/Frame	0.540	0.735	0.804	0.560	0.713	0.782
InsSpeed - Overall	0.476	0.760	0.842	0.574	0.713	0.797
InsAcceleration - FRS/Frame	0.429	0.787	0.879	0.752	0.693	0.751
InsAcceleration - RRS/Frame	0.462	0.761	0.859	0.673	0.639	0.708
InsAcceleration - Overall	0.515	0.732	0.815	0.627	0.646	0.714
AvgSpeed - FRS/Frame	0.613	0.660	0.716	0.609	0.706	0.760
AvgSpeed - RRS/Frame	0.643	0.648	0.677	0.631	0.653	0.694
AvgSpeed - Overall	0.619	0.650	0.717	0.630	0.646	0.700
AvgAcceleration - FRS/Frame	0.565	0.717	0.777	0.553	0.713	0.788
AvgAcceleration - RRS/Frame	0.549	0.714	0.793	0.553	0.693	0.788
AvgAcceleration - Overall	0.509	0.760	0.826	0.559	0.720	0.782

The sensitivity analysis findings from the deep learning model revealed that features associated with the head and legs of cattle exerted the most substantial influence on the model's performance. These results imply that the condition of a cattle's head and legs plays a essential role in assessing lameness. Additionally, these particular features exhibited relatively minor losses compared to other features.

The comparison of results from models trained on the feature-engineered dataset revealed that the Light Gradient Boosting Machine model outperformed all other models, achieving the highest accuracy of 81.3% and an AUC of 90.8% subsequently confirming the validity of the engineered features. Table 5 displays the models' results on the feature-engineered features. The top 10 engineered features with the highest impact are also illustrated in Figure 10.

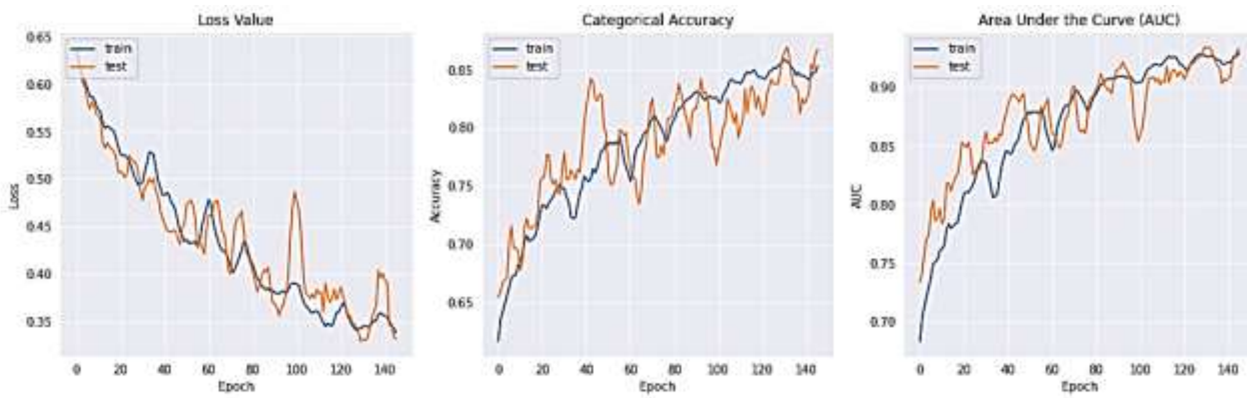

Figure 9. Deep Learning Model Results Using the Full Set of Features

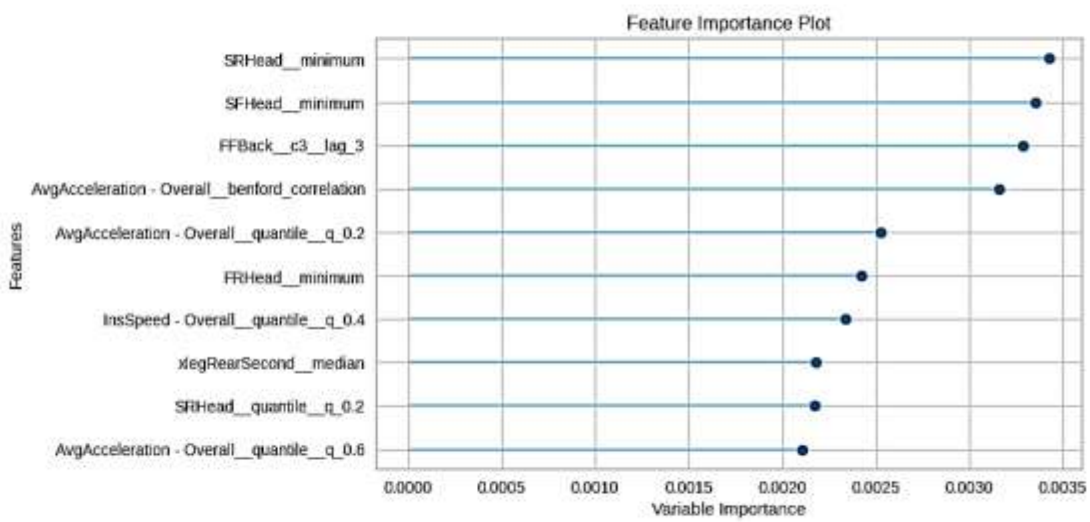

Figure 10. Top 10 Most Important Engineered Features.

Analyzing the feature count revealed that the original features with more filtered features (engineered features) had a more significant impact on the detection task. Figure 11 demonstrates these results.

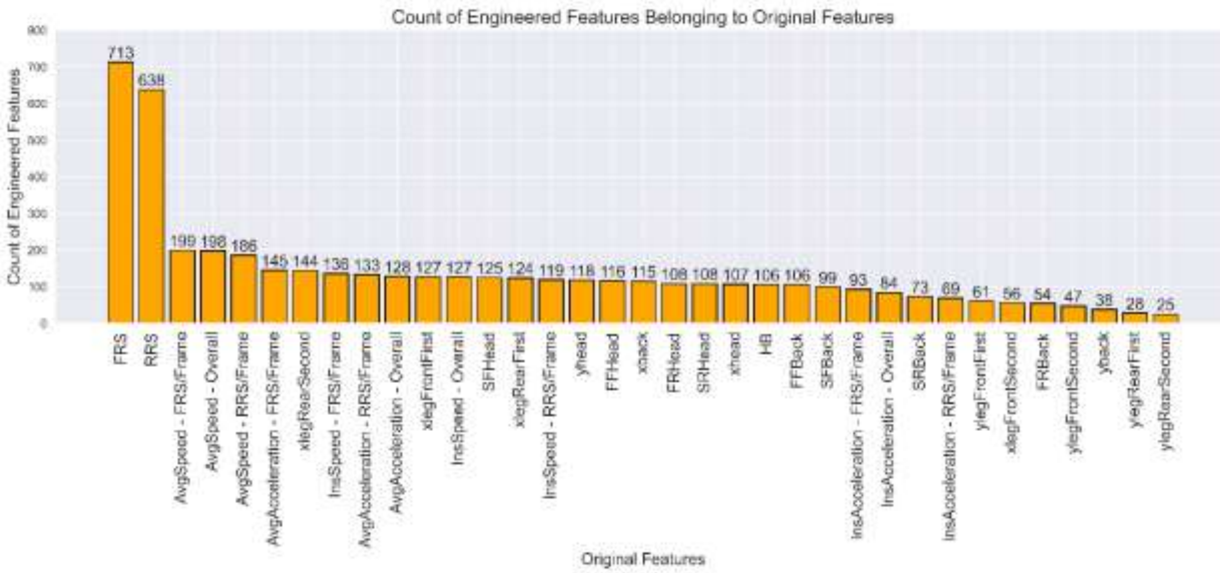

Figure 11. Contribution of Each Original Feature to Lameness Detection. The numbers indicate how many engineered features belong to the original features.

Table 5 - Results of the Models on the Feature-Engineered Test Set; Sorted by Accuracy.

Model Name	Accuracy	AUC
Light Gradient Boosting Machine	0.813	0.908
Extreme Gradient Boosting	0.806	0.904
Random Forest Classifier	0.793	0.885
Logistic Regression	0.773	0.877
Ada Boost Classifier	0.773	0.874
Extra Trees Classifier	0.766	0.916
Gradient Boosting Classifier	0.760	0.883
SVM - Linear Kernel	0.740	0.743
K Neighbors Classifier	0.660	0.730
Decision Tree Classifier	0.653	0.659
Ridge Classifier	0.640	0.626
Quadratic Discriminant Analysis	0.553	0.550
Dummy Classifier	0.533	0.500
Linear Discriminant Analysis	0.520	0.512
Naive Bayes	0.466	0.500

Examining the engineered features affirmed the significant influence of head and leg-related characteristics on accurately detecting cattle lameness. This discovery underscores the pivotal role of the head and legs in ascertaining lameness in cattle. Furthermore, the findings revealed that positional attributes such as the cattle's speed, acceleration, and body coordinates were the most vital factors for the detection process.

While the primary aim of this study was to enhance interpretability rather than outperform existing systems in terms of raw accuracy, it is still instructive to reflect on our results in light of prior work. Previous approaches such as (Abdul Jabbar et al., 2017) and (Wu et al., 2020) reported higher classification accuracies, reaching up to 95.7% and 98.79% respectively. However, these methods often rely on specialized hardware (e.g., depth cameras) or black-box deep learning models with limited insight into feature-level contributions. In contrast, our approach achieved a respectable accuracy of 88.66% and an AUC of 93.74%, while providing detailed sensitivity analyses and feature engineering that clarify the influence of specific visual features on detection performance. The relatively lower accuracy is an acceptable tradeoff for the transparency gained, particularly as our results consistently highlighted leg movement patterns and head–leg positional relationships as key indicators of lameness. This aligns with biomechanical expectations and existing veterinary understanding. Moreover, the engineered features derived from time-series analysis contributed significantly to model performance, reinforcing the biological relevance of temporal gait characteristics. The importance of features such as instantaneous acceleration and step size further supports the notion that early signs of lameness manifest through subtle movement deviations, which are effectively captured through our interpretable framework.

CONCLUSIONS

Based on our study, we found that unpredictable animal behaviors are the leading cause of noise that affects the performance of intelligent detection systems, making it difficult to make accurate predictions. To address this issue, we suggest using a modeling approach that involves minor sequences or features. This approach helps mitigate sudden cattle behaviors, improve accuracy, and make the model more resilient to environmental factors typically encountered on farms. By reducing the detection window, these systems can be trained to operate well under varying conditions and with diverse input data, given the unreliability of data due to the challenging farm environment.

Our research indicates that the most significant features for intelligent lameness detection are those related to the cattle's legs and head or a combination of both. While these features have traditionally been used individually to design intelligent lameness detection systems, our conclusion highlights the potential for greater accuracy and robustness. We have also identified that factors such as cattle step sizes (strides), speed, acceleration (calculated using Displacement per Frame), and the relative coordinates of the head to each leg significantly influence the detection task. Therefore, we can achieve superior and more resilient results by focusing on and incorporating these specific features into system design.

Our study thoroughly investigated the fundamental visual features necessary for advancing lameness detection systems, focusing on balancing robustness and accuracy. It prioritized transparency and reliability, recognizing the paramount importance of these factors, mainly when dealing with living beings such as animals.

Acknowledgments

The present research was carried out as an approved research study in the Faculty of Agriculture, University of Tehran (Grant No. 8987353). Therefore the financial support of the Research Assistant of the Faculty of Agriculture, University of Tehran (Grant No. 8993553), is gratefully acknowledged. We would also like to express our gratitude to all the people and organizations who helped in the present research, especially Dr. Ganj Khanlou (Faculty member at the faculty of agriculture) and the Animal Science Research Station at the University of Tehran.[30].

REFERENCES

Abbasian Ardakani, A., Mohammadi, A., Mirza-Aghazadeh-Attari, M., Faeghi, F., Vogl, T. J., & Acharya, U. R. (2023). Diagnosis of Metastatic Lymph Nodes in Patients With Papillary Thyroid Cancer: A Comparative Multi-Center Study of Semantic Features and Deep Learning-Based Models. *Journal of Ultrasound in Medicine*, *42*(6), 1211-1221. https://doi.org/10.1002/jum.16131

Abdul Jabbar, K., Hansen, M. F., Smith, M. L., & Smith, L. N. (2017). Early and non-intrusive lameness detection in dairy cows using 3-dimensional video. *Biosystems Engineering*, *153*, 63-69. https://doi.org/10.1016/j.biosystemseng.2016.09.017

Alban, L., Agger, J., & Lawson, L. (1996). Lameness in tied Danish dairy cattle: the possible influence of housing systems, management, milk yield, and prior incidents of lameness. *Preventive veterinary medicine*, *29*(2), 135-149. https://doi.org/10.1016/S0167-5877(96)01066-5

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. *Sensors*, *15*(6), 14513-14525. https://doi.org/10.3390/s150614513

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S.,

[30] University of Tehran

Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bagheri, N., Aghdam, M. J., & Ebrahimi, H. (2024). Estimating Nitrogen and Chlorophyll Content in Corn Using Spectral Vegetation Indices Derived From UAV Multispectral Imagery. *Biomechanism and Bioenergy Research*, *3*(1), 81-93. https://doi.org/10.22103/BBR.2024.23234.1082

Barkema, H. W., Westrik, J. D., van Keulen, K. A. S., Schukken, Y. H., & Brand, A. (1994). The effects of lameness on reproductive performance, milk production and culling in Dutch dairy farms. *Preventive veterinary medicine*, *20*(4), 249-259. https://doi.org/10.1016/0167-5877(94)90058-2

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, *307*, 72-77. https://doi.org/10.1016/j.neucom.2018.03.067

Ettema, J. F., & Østergaard, S. (2006). Economic decision making on prevention and control of clinical lameness in Danish dairy herds. *Livestock science*, *102*(1-2), 92-106. https://doi.org/10.1016/j.livprodsci.2005.11.021

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

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

Huxley, J. (2013). Impact of lameness and claw lesions in cows on health and production. *Livestock science*, *156*(1-3), 64-70. https://doi.org/10.1016/j.livsci.2013.06.012

Jia, Z., Zhao, Y., Mu, X., Liu, D., Wang, Z., Yao, J., & Yang, X. (2025). Intelligent Deep Learning and Keypoint Tracking-Based Detection of Lameness in Dairy Cows. *Veterinary Sciences*, *12*(3), 218. https://doi.org/10.3390/vetsci12030218

Jiang, B., Song, H., & He, D. (2019). Lameness detection of dairy cows based on a double normal background statistical model. *Computers and Electronics in Agriculture*, *158*, 140-149. https://doi.org/10.1016/j.compag.2019.01.025

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

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), 927-937. https://doi.org/10.1016/S0093-691X(02)01152-4

Merkin, A., Krishnamurthi, R., & Medvedev, O. N. (2022). Machine learning, artificial intelligence and the prediction of dementia. *Current Opinion in Psychiatry*, *35*(2), 123-129. https://doi.org/10.1097/YCO.0000000000000768

Ogundari, K., & Bolarinwa, O. D. (2018). Impact of agricultural innovation adoption: a meta-analysis. *Australian Journal of Agricultural and Resource Economics*, *62*(2), 217-236. https://doi.org/10.1111/1467-8489.12247

Oliveira, D. F., Vismari, L. F., Nascimento, A. M., de Almeida, J. R., Cugnasca, P. S., Camargo, J. B., Almeida, L., Gripp, R., & Neves, M. (2021). A new interpretable unsupervised anomaly detection method based on residual explanation. *IEEE Access*, *10*, 1401-1409. https://doi.org/10.1109/ACCESS.2021.3137633

Peng, J., Zhao, Z., & Liu, D. (2022). Impact of Agricultural Mechanization on Agricultural Production, Income, and Mechanism: Evidence From Hubei Province, China [Original Research]. *Frontiers in Environmental Science*, *10*, 838686. https://doi.org/10.3389/fenvs.2022.838686

Ryu, H. W., & Tai, J. H. (2022). Object detection and tracking using a high-performance artificial intelligence-based 3D depth camera: towards early detection of African swine fever. *Journal of Veterinary Science*, *23*(1), e17. https://doi.org/10.4142/jvs.21252

Schlageter-Tello, A., Bokkers, E. A., Koerkamp, P. W. G., Van Hertem, T., Viazzi, S., Romanini, C.

E., Halachmi, I., Bahr, C., Berckmans, D., & Lokhorst, K. (2014). Manual and automatic locomotion scoring systems in dairy cows: A review. *Preventive veterinary medicine*, *116*(1-2), 12-25. https://doi.org/10.1016/j.prevetmed.2014.06.006

Schmid, U., & Finzel, B. (2020). Mutual explanations for cooperative decision making in medicine. *KI-Künstliche Intelligenz*, *34*(2), 227-233. https://doi.org/10.1007/s13218-020-00633-2

Shrestha, A., Loukas, C., Kernec, J. L., 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

Sprecher, D., et al., Hostetler, D. E., & Kaneene, J. (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

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721

Warnick, L., Janssen, D., Guard, C., & Gröhn, Y. (2001). 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

Wu, D., Wu, Q., Yin, X., Jiang, B., Wang, H., He, D., & Song, H. (2020). Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector. *Biosystems Engineering*, *189*, 150-163. https://doi.org/10.1016/j.biosystemseng.2019.11.017