



Early prediction of broiler gait scores using spatiotemporal kinematic tracking and explainable machine learning features

Hosna Mohammadialabadi^a, Guoming Li^{a,b,c,*}, Tongshuai Liu^{a,d},
Venkat Umesh Chandra Bodempudi^{a,b}, Oluwadamilola Oso^a, Mahtab Saeidifar^{a,b},
Sai Akshitha Reddy Kota^c, Ehsan Asali^{a,c}, Aravind Mandiga^{a,c}

^a Department of Poultry Science, College of Agricultural and Environmental Sciences, University of Georgia, Athens 30602, GA, USA

^b Institute for Artificial Intelligence, Franklin College of Arts and Sciences, University of Georgia, Athens 30602, GA, USA

^c School of Computing, Franklin College of Arts and Sciences, University of Georgia, Athens 30602, GA, USA

^d College of Animal Science & Technology, Henan University of Animal Husbandry and Economy, Zhengzhou 450046, Henan, China

ARTICLE INFO

Keywords:

Poultry
Machine learning
Computer vision
Gait scoring

ABSTRACT

Lameness is an important welfare issue for broilers, particularly in the later phases of production, and is commonly assessed using gait scores. Most existing automated gait scoring systems are designed for birds at 4 weeks of age and older, delaying opportunities for corrective interventions. The objective of this study was to predict broiler gait scores at 7 weeks of age by leveraging early-life (2 and 3 weeks of age) spatiotemporal kinematic features through interpretable and explainable machine learning models. Top-view videos of Cobb 500 male broilers were recorded in 48 experimental pens measuring 1.2 m wide $\times$ 3.0 m long. A total of 600 h of videos were randomly selected and analyzed from 480 color-marked birds. Using a YOLOv11 model with the ByteTrack tracking framework, individual birds were tracked and their trajectories were extracted. Based on these trajectories, three categories of features were computed: raw kinematic metrics, movement-threshold-filtered parameters, and pen-normalized values. Our optimized CatBoost model achieved 82.5% accuracy in classifying healthy (gait score 0) and impaired birds (scores 1 and 2). The most significant predictors were acceleration quantiles ($P = 0.017$), feeding time ratios ($P = 0.020$), and movement angle variability. The system maintained 99.2% mAP@0.5 detection accuracy on the test dataset in this study, even when the image resolution was downsampled to 960 $\times$ 540 pixels. As one of the first methods capable of predicting gait scores associated with lameness from movement patterns of birds at 2 and 3 weeks of age, this approach advances early gait score prediction compared to existing systems. This early-prediction framework leverages spatiotemporal kinematic features extracted from broilers at 2 and 3 weeks of age to predict gait scores at 7 weeks of age, providing a framework with the potential to support producers in identifying at-risk individual birds early and enabling timely welfare interventions.

1. Introduction

The United States is the largest broiler producer in the world, with 9.33 billion broilers and 27.7 million metric tons of harvest weight (USDA, 2025). The whole production system is valued at \$45.4 billion and the results of short production cycles, genetic improvements, and efficiency in feed-to-meat conversion. A broiler can reach a live weight of nearly 3 kg within a 49-day (7 weeks of age) production cycle (Council, 2024). However, this accelerated muscle growth often outpaces skeletal development, raising significant welfare concerns,

particularly in relation to leg health. More than 30% of fast-growing broilers suffers from leg weakness and skeletal deformities due to this physiological imbalance (Julian, 2005). Reports further indicate that 14–52% of commercial broilers suffer from moderate to severe leg issues (Bassler et al., 2013), especially during the late phase of a production cycle, when broilers spend 76–86% of their time lying down on the litter floor (Weeks et al., 2000). In a large-scale study, Knowles et al. (2008) observed that by day 40, 27.6% of broilers showed poor locomotion, while 3.3% were nearly unable to walk.

Gait scoring has been widely used to assess broiler leg health,

* Corresponding author at: Department of Poultry Science, College of Agricultural and Environmental Sciences, University of Georgia, Athens 30602, GA, USA.
E-mail address: gml@uga.edu (G. Li).

<https://doi.org/10.1016/j.compag.2026.111651>

Received 9 September 2025; Received in revised form 5 March 2026; Accepted 6 March 2026

Available online 29 March 2026

0168-1699/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

assigning numerical values to walking ability (Butterworth et al., 2009; Kestin et al., 1992). Scoring systems range from binary evaluations to ordinal scales such as three-, four-, and six-point methods (Cordeiro et al., 2009; Garner et al., 2002; Kestin et al., 1992; Webster et al., 2008). (Cordeiro et al., 2009; Garner et al., 2002; Kestin et al., 1992; Webster et al., 2008). Among these, the three-point system described by Webster et al. (2008) is one of the most widely applied approaches in the United States, assigning scores of 0 (normal), 1 (moderate difficulty), and 2 (severe lameness or immobility). This method is a standard in the U.S. broiler industry, where guidelines recommend that fewer than 1% of birds should be scored 2, and fewer than 30% should be scored 1 or 2 combined (Council, 2020). Despite its widespread use, traditional visual gait assessments are labor-intensive, time-consuming, and subject to observer variability and error (Webster et al., 2008; J. Yang et al. 2023a).

To tackle these limitations, recent studies have introduced image-based and machine learning approaches for automated gait evaluation (Li et al., 2023; Pereira et al., 2021; J. Yang et al., 2023). These include monitoring the trajectories of broilers, analyzing and assessing their posture, and obtaining spatiotemporal activity metrics using ceiling or side-mounted cameras (Aydin, 2017; Li et al., 2023; Van Hertem et al., 2018). Aydin (2017) monitored broilers walking through a test corridor using a digital camera and extracted gait-related kinematic features, reporting strong correlations between gait scores and speed ($r = 0.844$), step length ($r = 0.882$), and lateral body oscillation ($r = 0.861$). Van Hertem et al. (2018) applied a camera-based monitoring system to quantify flock activity and spatial distribution, reporting a linear relationship between flock activity and average gait score ($R^2 = 0.55$) and strong correlations with activity ($r = -0.741$) and distribution ($r = 0.705$). De Alencar Nääs et al. (2020) used walking speed, acceleration, strain, and sex as inputs to decision-tree models for gait classification, achieving 78% accuracy for a simplified three-point scale and 91% accuracy for a final binary decision tree primarily driven by walking speed. Yang et al. (2023b) employed top-view cameras to extract production and behavioral metrics and developed regression models for six-point gait scores, achieving R^2 values of 0.90–0.91 and RMSE values of 0.38–0.41 based on body weight, age, activity index, and distribution index. Fodor et al. (2025) collected top-view videos of broilers walking through a corridor and extracted four walking features (lateral body oscillation, step count, completion time, and length–width ratio), finding that these features reliably identified severely impaired birds with gait scores of over 3 but were less effective for distinguishing mild gait or leg health issues such as hock burn and footpad dermatitis. More broadly, Merenda et al. (2024) demonstrated the feasibility of computer vision combining YOLOv5l for detection, OSNet for tracking, and a Gradient Boosting Classifier for behavior recognition in group-housed broilers, achieving high accuracy for feeding (0.895) and drinking (0.900) behaviors but lower accuracy for active and inactive states.

While these advancements show promise, most studies focus on gait assessment during later stages of broiler growth (4 weeks of age or onward) or predicting gait scores of a flock of birds. Few studies address the early developmental phase (before 4 weeks of age) of individual broilers in group settings, even though early locomotor issues may be more responsive to intervention. Earlier detection of locomotor issues improves the response to intervention and increases the available time for management adjustments that is better for animal welfare and production optimization (Aydin, 2017), especially given that risk factors such as dark period, flock age, and litter quality substantially influence lameness outcomes (Bassler et al., 2013). However, previous computer vision research focused on group-level activity or behavioral indicator classification rather than extracting gait-relevant locomotor features of individual birds.

Trajectory extraction is fundamental to gait analysis and has significantly improved through advances in computer vision and machine learning. To calculate kinematic features, it is necessary to extract the trajectories of the selected broilers (Caplen et al., 2012; Li et al., 2023).

In this research, color markers were employed to differentiate individual birds within each pen, enabling human validation of certain birds. Trajectory extraction can be achieved through detection models, tracking algorithms, or a hybrid approach. Zhang et al. (2018) showed that combining detection and tracking models can improve both detection and tracking accuracy. For example, Campbell et al. (2024) developed a monocular camera system that detects and tracks broilers and applies unsupervised 2D trajectory clustering, reporting a detection precision of 0.98, recall of 0.90, and a multi-object tracking accuracy of 74.7%. Li et al. (2023) used synchronized top-view RGB and depth cameras along with deep-learning and image-processing models to extract bird trajectories and 16 spatiotemporal and 3D locomotive metrics, achieving 100% keypoint-tracking accuracy and a 3.62 mm RMSE for gait-related feature estimation. Collectively, these advances demonstrate the potential of computer vision to characterize broiler locomotion, reinforcing the need for more precise trajectory-based approaches for early gait evaluation of individual birds in group settings. However, there are certain challenges associated with this approach: When birds were near other broilers, the likelihood of ID loss in both tracking and detection increases significantly. In fact, approximately 56% of ID losses can be attributed to this proximity issue (Doornweerd et al., 2024).

Despite these advances, current computer vision systems have not yet progressed to deliver reliable early-stage gait score predictions for group-housed broilers. Some computer vision-based gait scoring platforms allow individual broiler chickens to walk through during assessment. However, those platforms cannot track multiple individual birds simultaneously, which leads to evaluation inefficiency and limits their practicality for monitoring gait scores in commercial houses containing tens of thousands of birds. Most of the deep learning related studies focus on model training and evaluation only, critical features supporting early gait score prediction remain unclear, reducing the reliability and understandability of artificial intelligence-based models used to support human decision-making. The objective of this study was to predict broiler gait scores at 7 weeks of age by leveraging early-life (2 and 3 weeks of age) spatiotemporal kinematic features through interpretable and explainable machine learning models.

2. Materials and methods

2.1. Data Acquisition

The live bird trial took place at the Poultry Research Center (PRC) of University of Georgia between April 18 to June 6, 2024. A total of 1,776 male Cobb 500 broilers were evenly distributed into 48 floor pens with each measured at 1.2 m wide and 3.0 m long, with each pen containing 37 birds. The overall layout of the 48 pens is presented in Fig. 1. The Cobb 500 management principles were followed for controlling the temperature and lighting (Cobb-Vantress, 2022). Feed and water were provided *ad libitum*. Broilers were raised up to 7 weeks of age. The light was set as full brightness (20–30 lx) from days 0–9 and as dim (~5 lx) from days 10–49. The temperature was set at 33.9 °C from days 0–6; 31.1 °C from days 7–13; 28.3 °C from days 14–139 20; 25.6 °C from days 21–27; 22.8 °C from days 28–34; and 21.1 °C from days 35–49. All procedures were approved by Institutional Animal Care and Use Committees at the University of Georgia (protocol number: A2023 07–016-Y1-A0).

A camera (NHD-887MSB, Swann Security, Santa Fe Springs, CA) was installed on the ceiling of each pen with distance of 3.05 m from the ground. The video recording was set at a sample rate of 15 images per second (fps), and a total of 24 one-hour video episodes were recorded daily for each pen. The data was analyzed from 00:00 to 14:00. The recording period also included the dark phase from 02:00 to 06:00. Although referred to as the dark phase, the period from 02:00 to 06:00 was maintained under dim lighting (~5 lx), ensuring birds remained visible. The dim phases could ensure young chicks to access the food as

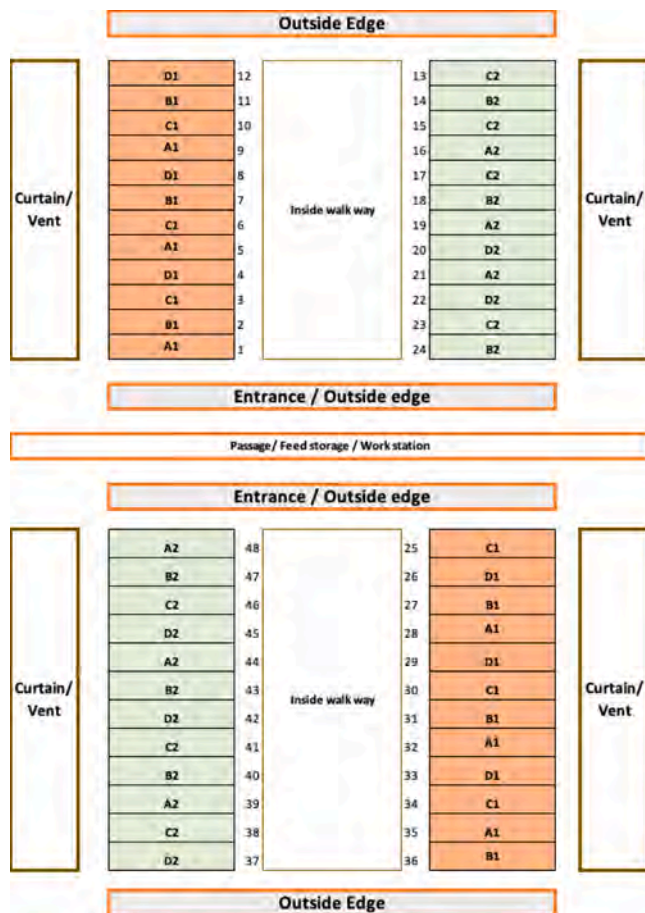


Fig. 1. Schematic drawing of the floor pen layout during the experiment, including 48 pens.

needed while allowing them to rest. The intentional inclusion of both bright and dim-light conditions aimed to enhance dataset variation and more accurately simulate the lighting transitions encountered in commercial farm environments. From 2 to 7 weeks of age, video recordings were conducted on four consecutive days per week to comply with farm biosecurity protocols and address practical data management constraints. Week 1 was skipped due to bird habituation and mortality replacement. A total of 600 h of videos, stored as MP4 files on a 20-terabyte external hard disk, from broilers at 2 and 3 weeks of age were randomly selected and analyzed from 480 color-marked individuals in this experiment. Fig. 2 shows a sample image of recorded videos extracted from a top-view camera.

The four-day recording schedule instead of seven days in a week was



Fig. 2. A sample image of recorded videos extracted from a top-view camera.

implemented due to a combination of practical data storage limitations and adherence to farm biosecurity protocols. First, the recording system generated approximately 256 GB of video data per pen per day, while the recording space was only 2 TB. To prevent data loss from being overwritten, manual backup via external storage drives were required. We needed three days to back up 4 days of data in a week. Meanwhile, to minimize disease risk and disturbance to the birds, farm rules strictly limited human entry into the facility to one visit per day. This single daily visit was used for essential husbandry tasks. The four-day recording window was strategically chosen to capture comprehensive behavioral data while complying with this access constraint.

Ten focal broilers were randomly selected from each of the 48 pens and marked with 10 different livestock safe spray colors, including green, purple, black, red, green-purple, green-black, green-red, purple-black, purple-red, and black-red. A subset of 480 broilers (10 birds from each of the 48 pens) was randomly selected for manual gait scoring on a three-point scale.

Following the guidelines of the three-point gait-scoring system from the creator (Webster et al., 2008) and the National Chicken Council (Council, 2022), each bird was assigned a score from 0 to 2: score 0 indicating the ability to walk 1.5 m with no signs of lameness, score 1 indicating the ability to walk 1.5 m with signs of lameness (obvious unevenness in steps or sitting down at least once), or score 2 indicating the inability to walk 1.5 m. For manual gait scoring, birds were placed on a custom-designed plywood platform (1.8 m long, 0.46 m wide, and 0.30 m high). The platform has 0.15 m start and finish sections on each end of a 1.5 m long walking space, with delineations marked every 0.30 m. Focal birds were individually placed on the platform start section and either walked 1.5 m independently or were encouraged to walk by a researcher waving behind and gently tapping the bird with a broom. A bird was encouraged to move when seeing another bird placed on the other side of the platform. There were no time-out criteria for gait scoring, and the time needed for gait-scoring one bird ranged from 3 to 20 min. Birds were considered to have completed the gait-scoring session when both feet had crossed into the finish section. All birds were scored by three independent observers, with the final score for each instance determined by majority vote.

2.2. Overall workflow

The overall workflow of early gait score prediction framework development is depicted in Fig. 3. A database of 600-hour videos of birds with 2 and 3 weeks of age was formed first, and each 1-hour-long video was split into three 20-minute-long videos to enhance computational efficiencies. Individual color-marked birds in the floor pens were tracked from the video database using custom-trained object detection and tracking models. Pixel-based measures were converted into real-world spatial units (centimeters) using a homography-based transformation derived from the known pen dimensions (1.2 m wide $\times$ 3.0 m long). Then a series of spatiotemporal and kinematic parameters (e.g., speed, traveling distance, and acceleration) were extracted from the calibrated trajectories of individual birds. Those features were fed into machine learning models for training to predict gait scores in 7 weeks of age. Detailed steps of the workflow are explained in the following sections. The framework was written in Python and developed on a 64-bit Windows computer equipped with an NVIDIA RTX 4000 GPU (20 GB VRAM) and 64 GB RAM.

2.3. Pixel to centimeter transformation

Converting pixel-based measurements to real-world units helps the interpretability of features and enables biological validation. While the extracted kinematic features remain an approximation of actual movement, converting pixel-based measurements to real-world units (e.g., cm/s) enhances the understanding of feature reasonability and enables comparison with previously reported bird movement values (e.g.,

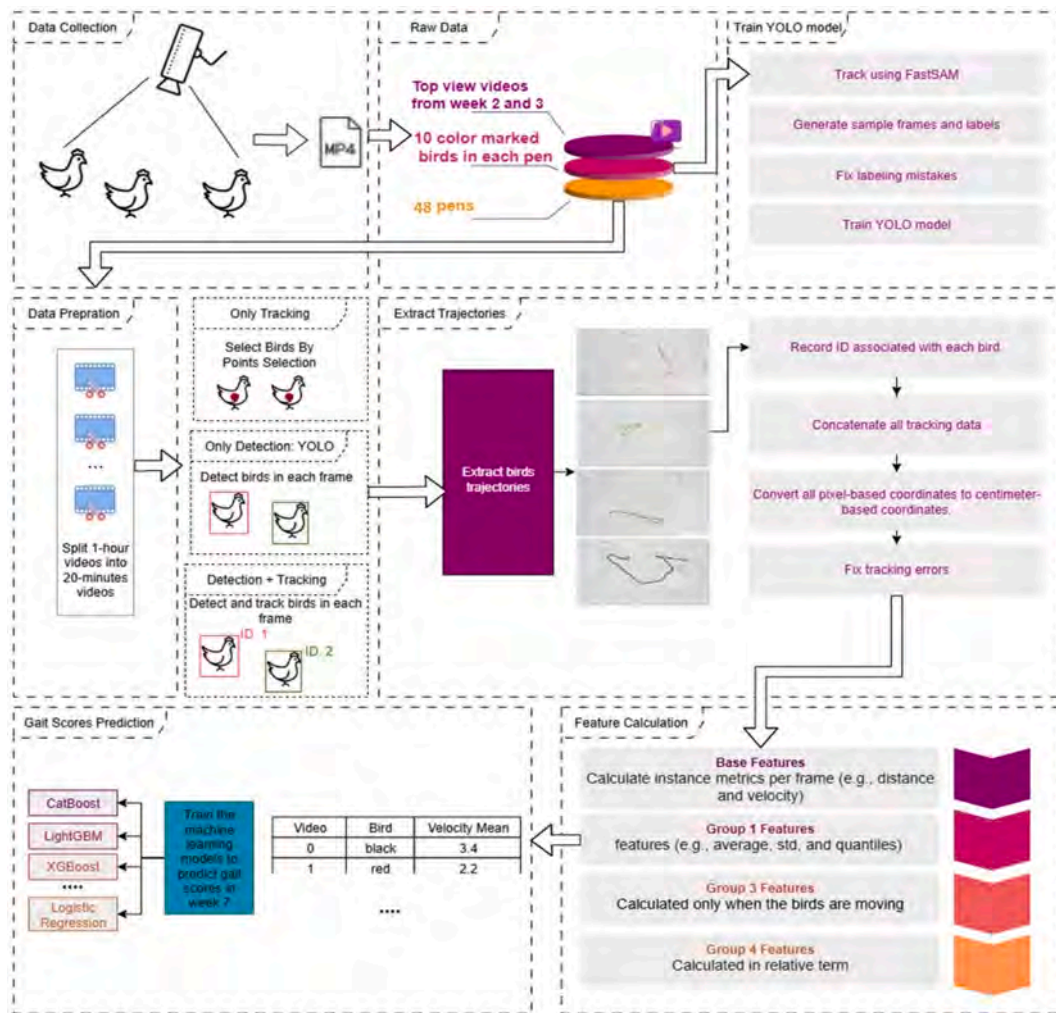


Fig. 3. Overall workflow of the early gait score prediction framework. Starting from raw video input, the workflow includes detection and tracking of individual focal birds at 2 and 3 weeks of age, computation and refinement of movement trajectories, extraction of spatiotemporal kinematic features, and training machine learning models to predict gait scores at 7 weeks of age.

walking speed) for biological validation, without affecting model performance or prediction precision. The stationary top-view cameras enabled perspective correction and scale normalization through RANSAC-based homography in OpenCV. Using this geometric method, pixel coordinates were mapped to a flat surface using known reference points in the real world. An example of the transformation process is indicated in Fig. 4. A standard coordinate system based on the known sizes of each pen: 1.2 m wide by 3.0 m long, which is the same as the frame of (0,0), (300,0), (0,120), and (300,120) in centimeters. The pixel coordinates were adjusted for each pen to ensure the spatial information consistency of each pen.

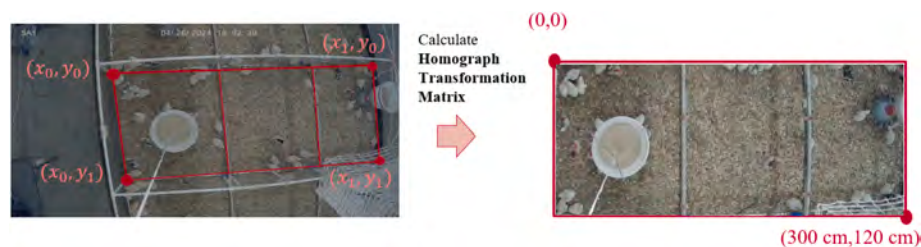


Fig. 4. Example illustration of converting trajectories from pixel based to centimeter based and fixing camera perspective view errors. The key points for the transformation are indicated as red dots.

2.4. Detection and tracking models

The color markers were randomly assigned to birds within each pen and were not specific to gait scores. Only one bird per color pattern was present in a given pen, this setup allows for the use of either a detection model capable of tracking 10 color patterns of birds or a combination of detection and tracking models to directly track multiple birds..

Detection-based methods like YOLO require supervised training, whereas general tracking algorithms such as the Track Anything Model (TAM) (J. Yang et al., 2023) do not require task-specific supervised training or fine-tuning on our dataset, as they are pre-trained on large general datasets (e.g., SA-1B, SA-V, SA-CO, and SA-FARI). The TAM

leverages the Segment Anything Model (Kirillov et al., 2023) for segmentation-based tracking. Fast-SAM (Zhao et al., 2023), a CNN-based alternative, was also evaluated. For this study, YOLOv11 was implemented in two configurations:

- Standalone detection: Direct extraction of trajectories from frame-wise detections.
- Detection-enhanced tracking: Integration with the ByteTrack algorithm to refine trajectories by mitigating ID switches and detection noise.

A comparative summary of these models is provided in Table 1.

Our pipeline began with an initial Fast-SAM fine-tuning on a small, custom dataset of 500 images (containing 5000 bird instances) annotated with a single “bird” class to improve general detection of broilers in our specific farm environment. Although the base Fast-SAM model is capable of performing detection and segmentation without any task-specific training, to optimize its performance for the specific broiler detection task in this study, the Fast-SAM model was subsequently fine-tuned using our annotated dataset.

This fine-tuned model was then used for large-scale tracking and manual ID labeling across 24 h of diverse videos sampled from multiple pens and different times of a day (morning, afternoon, and evening), producing tracking data with preliminary object IDs that were manually corrected and assigned to specific color-marked birds.

Incorrect detection that suddenly shifted a bird’s centroid by tens of centimeters between consecutive images or within 1/15 s resulted in unrealistically high velocity and acceleration values. These outliers substantially distorted the distributions of movement features, which were later summarized using statistical descriptors. In contrast, short periods of missed detections only reduced the number of samples but did not significantly alter the overall distribution shape of features such as velocity, acceleration, or movement entropy. Therefore, tracks with intermittent missing detections were retained, whereas tracking IDs associated with incorrect detections were removed to preserve feature reliability. From this meticulously labeled tracking data, we created a YOLO dataset by extracting 5000 images, with bounding boxes and individual class labels for each unique color, which was used to train YOLOv11 for individual re-identification; notably, while the initial Fast-SAM fine-tuning used a single “bird” class to boost detection recall, the final dataset annotated only the color-marked broilers with individual IDs, achieving the pipeline’s goal of individual identification for color-marked birds. The YOLOv11 was trained with 700 epochs and other default hyperparameters, and the best model was saved for further processing.

The trained YOLOv11 model was compared with all the models listed on Table 1 for individual bird tracking. Tracking accuracy metrics, including IDF1 and MOTA, were reported for YOLOv11 model with the ByteTrack tracking framework based on 4,500 images, to quantitatively evaluate tracking performance. To ensure consistency in comparative analyses (e.g., tracking performance, image resolution effects, and image rate impact), the same reference 5-minute video segments were used across all detection and tracking evaluations.

Ground-truth tracks were annotated at 5 FPS. Inference was performed at both 15 FPS and 5 FPS; however, for quantitative comparison

with the ground truth, predictions from the 15 FPS inference were temporally subsampled to 5 FPS. Subsampling the 15 FPS predictions did not necessarily yield the same results as inference performed natively at 5 FPS. Since the 15 FPS inference processes additional intermediate images, its predictions differ from those obtained via native 5 FPS inference.

Key metrics such as Intersection over Union (IoU), Precision, The Identification Metric (IDF1), Multiple Objects Tracking Accuracy (MOTA), and mean Average Precision (mAP) were used to quantify the classification and tracking performance of multiple birds (Equations (1)–(6)). Processing time (seconds per image) was computed by dividing the total processing time of the 5-minute video with the total number of images processed. The optimal model among the four major sets of models was determined with the higher tracking accuracy and less processing time per image.

Because a tracker could lose and re-detect the same bird, a single physical bird may be assigned multiple temporary IDs within one tracking session. ID switching was not a critical performance metric for our specific tracking problem. Since each bird was uniquely identifiable by its color, temporary tracking losses followed by re-identification did not compromise tracking integrity. The ByteTrack intentionally configured to prioritize new ID generation over ID switching during occlusion events. Specifically, the IoU matching threshold was increased to 0.9 and the track buffer duration was reduced to 15 images (or 3 s), ensuring that lost tracks were promptly terminated rather than incorrectly reassigned.

$$IoU = \frac{AreaofOverlap}{AreaofUnion} \quad (1)$$

where IoU denotes the intersection-over-union between predicted and ground-truth bounding boxes.

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)} \quad (2)$$

where TP and FP represent true positives and false positives, respectively.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (3)$$

where AP_i denotes the average precision for class i , and N is the number of classes.

$$AP = \int_0^1 P(R) dR \quad (4)$$

where $P(R)$ represents precision at a given recall level R .

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \quad (5)$$

where IDTP, IDFP, and IDFN denote identity true positives, identity false positives, and identity false negatives used for tracking evaluation.

$$MOTA = 1 - \frac{FN + FP + IDS_w}{GT} \quad (6)$$

where IDS indicates identity switches, GT denotes the number of ground-truth objects and FN and FP represent false negative and false positives, respectively.

The final detection and tracking model was selected based on a balanced assessment of MOTA and practical processing speed for large-scale video analysis. Specifically, a processing time of less than 1 s per image was established as a feasibility threshold to enable the analysis of long-duration videos required for gait score classification.

To enhance computational efficiency of tracking, the following two optimizations were implemented. The first one was about spatial

Table 1

Summary of tracking methodologies, noting the TAM and Fast-SAM in contrast to YOLO do not require training.

Model	Features
Track-Anything-Model	Track multiple objects
Fast-SAM	Track multiple objects
YOLOv11 model (only detection)	Detect unique objects
YOLOv11 model (combined with ByteTrack tracking model)	Detect unique objects and combine them with a tracking model to correct false detections

resolution downsampling. Resolution was reduced from $1,920 \times 1080$ pixels to 960×540 pixels. The second one was about temporal sub-sampling. The image rate decreased from 15 fps to 5 fps to reduce processing load.

To evaluate the impact of reduced image resolution on detection performance, the mAP metric was computed across 4,500 images extracted from a 5-minute video. Additionally, when reducing the image rate from 15 fps to 5 fps, tracking performance was assessed using MOTA and IDF1, while motion consistency was verified by calculating the average velocity and standard deviation of velocity across the same 5-minute video.

Algorithm 1 Enhancing Tracking Accuracy of YOLO and ByteTrack

Input:
 $T = \{\tau_i\}_{i=1}^n$: Dictionary of tracked objects where $\tau_i = (d_1, \dots, d_k)$ (sequence of detections)
 $t_{\min}$: Minimum frame threshold (default : $t_{\min} = 10$)

Output:
 T^* : Refined tracks with consistent class labels
 # Filter transient tracks:

```

1   $T_{\text{filtered}} \leftarrow \emptyset$ 
2  for each  $\tau_i$  in  $T$  do:
3    if  $|\tau_i| \geq t_{\min}$  :
4       $T_{\text{filtered}} \leftarrow T_{\text{filtered}} \cup \{\tau_i\}$ 
5  end
  #Enforce class consistency:
6   $T^* \leftarrow \emptyset$ 
7  for each  $\tau_i$  in  $T_{\text{filtered}}$  do:
8     $C \leftarrow \{c(d_j) | d_j \in \tau_i\}$  (extract class labels)
9     $f(c) \leftarrow \text{FrequencyDistribution}(C)$ 
10    $c^* \leftarrow c \in \text{argmax}_{c \in C} f(c)$ 
11   If  $f(c^*) > \max_{c \neq c^*} f(c)$  :
12     for  $d_j \in \tau_i$  :
13        $c(d_j) \leftarrow c^*$ 
14     End
15    $T^* \leftarrow T^* \cup \{\tau_i\}$ 
16   else:
17     discard  $\tau_i$ 
18 end
  return  $T^*$ 

```

2.5. Enhancing tracking accuracy

To improve tracking accuracy, YOLOv11 object detection was integrated with the ByteTrack algorithm based on the comparative results presented in Section 2.4. Previous studies have shown that combining detection and tracking models can enhance both detection and tracking performance (Zhang et al., 2018). As illustrated in Fig. 5, correct trajectories exhibit reasonable and continuous movement steps during tracking. However, as shown in Fig. 6, a bird may be correctly detected in one frame but misclassified in an intermediate frame (e.g., falsely

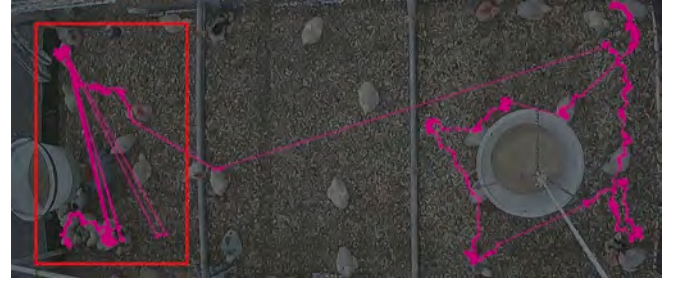


Fig. 6. An example of erroneous detections that cause back-and-forth jumps in the position of the broiler detected. These mistakes usually occurred in the corners where the algorithm did not detect the color markers with high accuracy due to challenging camera coverage.

identified as another bird), which can introduce substantial errors in the tracking process. If unaddressed, such misclassifications may result in unrealistic trajectory jumps or velocity spikes.

To address these tracking challenges, we implement the ByteTrack algorithm with post processing logics to improve robustness. The system handled two key issues. First, short-duration tracks (<10 images) were automatically discarded to eliminate temporary false detections (Algorithm 1, lines 2–5). Second, for each tracked bird, the class distribution of YOLO detections was analyzed, and outlier labels were replaced with the majority class. Tracks with uniform label distributions were removed (Algorithm 1, lines 6–18). This approach prioritized tracking precision over recall, ensuring only high-confidence trajectories to be retained.

2.6. Feature Calculation

To get useful features of locomotion quality from the bird trajectories, first the basic kinematic variables for each bird were calculated on a image-by-image basis. These variables were displacement (distance traveled), velocity, and acceleration which were calculated based on Equations (7) to (11).

$$\text{Distance}_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \quad (7)$$

$$\text{Velocity}_t = \frac{\text{Distance}_t - \text{Distance}_{t-1}}{(t_t - t_{t-1})} \quad (8)$$

$$\text{Acceleration}_t = \frac{\text{Velocity}_t - \text{Velocity}_{t-1}}{(t_t - t_{t-1})} \quad (9)$$

$$\text{Angle}_t = \tan^{-1} \left(\frac{y_t - y_{t-1}}{x_t - x_{t-1}} \right) \quad (10)$$

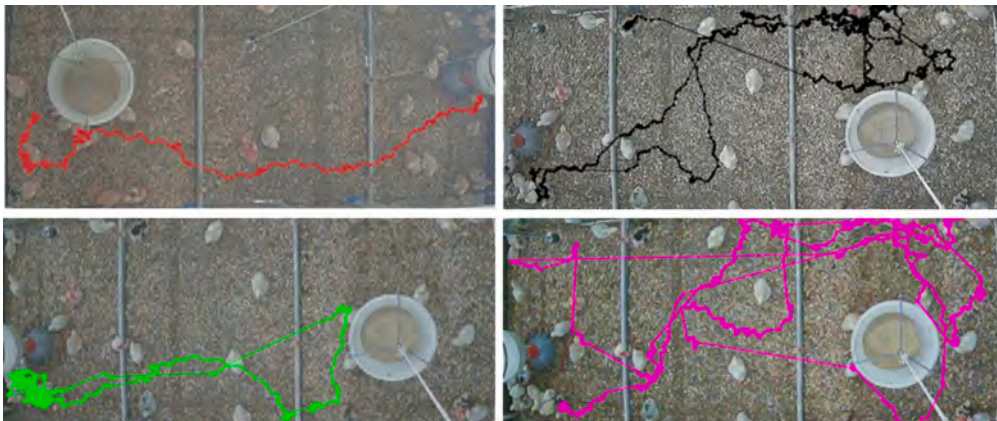


Fig. 5. Example of correct trajectory extractions from a 5-minute video segment. Each colored trajectory represents an individual broiler, and the absence of abrupt positional jumps indicates reliable tracking suitable for kinematic feature extraction.

$$\text{Path Prediction Error}_t = (x_t - \hat{x}_t)^2 + (y_t - \hat{y}_t)^2 \quad (11)$$

where x_t and y_t represent the coordinates of the bounding box center for the detected bird at time t , while $\hat{x}_t$ and $\hat{y}_t$ denote the expected coordinates of the bird at the same timestep.

The path prediction error feature demonstrated the likelihood of a bird walking in a straight line. Broiler chicks that are not lame usually move in straight lines, but those that have trouble with their legs tend to move in zig-zag ways or can't stay in one direction. This feature, then, measures how far off an ideal straight-line path is and can be used to tell if a broiler could have a problem with their gait. To calculate, first the next position of the bird was extrapolated by utilizing the bird's last two positions in consecutive images. Then, the difference between the expected and actual position is calculated using Equation (11).

These basic kinematic measurements were used to construct more advanced statistical features by aggregation functions including mean, maximum, minimum, median, standard deviation, 25th and 75th percentiles, skewness, and kurtosis. The base features were categorized as three groups.

- **Group (1) features:** These features were based on the raw kinematic calculations and were just basic statistical summaries of 1-hour videos.
- **Group (2) features:** This group of features was calculated when birds moved at least a predetermined distance. For this study, a threshold of 10 cm was applied.
- **Group (3) features:** These features were the same as Group (1) features but were calculated relative to the average values of other broilers within the same video.

These three groups of features covered most of the basic spatiotemporal kinematic features. In addition, the feeding time, distribution index, and entropy of movement were also computed per bird basis. In summary, a total of 276 columns of features were formed from this section.

2.6.1. Group (1) features

The basic set of kinematic features for Group I characteristics came from calculating bird trajectories image-by-image. Statistical aggregation methods were used on all images of each video to produce these features. In other words, kinematic features were first calculated on each image and then, statistical properties of these values were used as features. In total, this process generated 42 features.

2.6.2. Group (2) features

Even when the birds were not moving, the bounding box coordinates may change. Fig. 7 shows a group of three broilers, one of which is marked with a red identifier. The red bird does not move in images 1, 2, and 3, but the bounding box fluctuates. This results in computational errors of velocity, acceleration, and other characteristics that inadequately represent motion. To fix this, as Algorithm 2 described, a minimum movement threshold of 10 cm was used, and kinematic features

were calculated only when a bird's movement between two images goes over this threshold. This phase of filtering helps get rid of noise. Similar to Group (1) features, this group also generates 42 features.

Algorithm 2 Movement Threshold Filtering

Input:
 $P = [p_1, p_2, \dots, p_n]$: List of 2D position ($p_t = (x_t, y_t)$)
 $\delta_{\min}$: Movement threshold (e.g., $\delta_{\min}=10$ cm)
 $\Delta T = [\Delta t_1, \Delta t_2, \dots, \Delta t_n]$: List of time interval between images

Output:
 $F = [f_1, f_2, \dots, f_n]$: Feature values (where applicable)

#initialization
 $F \leftarrow \emptyset$

for each consecutive pair (p_t, p_{t-1}):
 #Displacement:
 $\Delta d_t = \|p_t - p_{t-1}\|$
 #Movement Classification:
 $s_t = \begin{cases} \text{moving,} & \text{if } \Delta d_t > \delta_{\min} \\ \text{stationary,} & \text{otherwise,} \end{cases}$
 # Calculate feature value and append results
 if $s_t == \text{moving}$:
 $f_t = \text{feature}_t(\text{parameters}_t)$
 $F \leftarrow F \cup \{f_t\}$

end
return F

2.6.3. Group (3) Features

It was also observed that during certain periods, all birds were active, whereas they may also be inactive at other time. Therefore, if only small periods were relied upon, generalization may not be achieved effectively. A set of relative features was also added here. These features were the same as Group (1) features but were calculated relative to the average values of other broilers within the same video. In addition to group (1) features, this post processing also applied to features related to distribution index, feeding, entropy, and fractal dimension and added 76 new features. To calculate this type of feature each feature was divided by the average of that feature in the same pen and same period among all birds (Equation (12)).

$$\text{Relative feature}_{t,i}^{ij} = \frac{\text{Feature}_{t,i}^{ij}}{\frac{1}{n} \sum_{j=0}^n \text{Feature}_{t,i}^{ij}} \quad (12)$$

where i is feature number; j is bird number; and t is time.

2.6.4. Other features

1) Feeding Time

In this research the feeding time was estimated using a spatial inference method that identifies bird presence around the feeder area. Due to the fixed location of the feeder, the feeder centroid was first found manually in one sample image from each camera, and then applied to other images.

A larger ellipse (Fig. 8) was selected because when broilers were feeding, the centroid of their bounding boxes was located within an ellipse larger than the feeder's dimensions. An ellipse with 10 percent

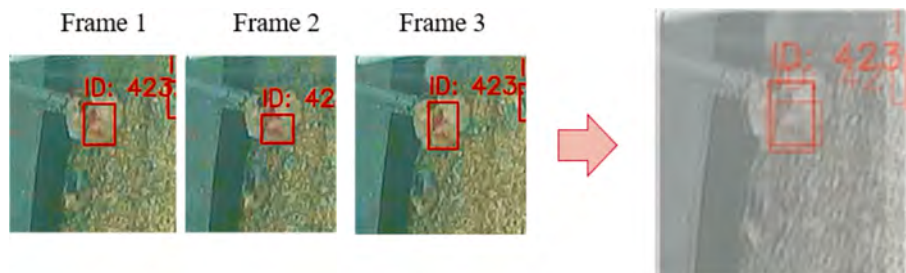


Fig. 7. A sample of bounding box fluctuation. The right image shows the overlaps of the images.

larger axis was visually determined to include broilers' body centroid when feeding.

Then in each image, if the centroid of a detected bird was located within the elliptical feeder area (Equation (13)), the bird was identified with feeding. This method was applied in (Li et al., 2020), which achieved the 89–93% accuracy for determining broilers at feeder. The feature of feeding time ratio was computed by dividing the number of images with the birds presented in feeder zone with total number of images (Equation (14)). This set of computations introduced 3 additional features for model training, including total feeding time ratio, feeding time ratio for right feeder, and feeding time ratio for left feeder.

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} \leq 1 \quad (13)$$

where (h, k) is the center of the feeder zone; a is the semi-major axis length of the fitted ellipse; b is the semi-minor axis length of the fitted ellipse; and (x, y) is the centroid coordinate of the bird in a image. This equation defines a bird locating within a fitted elliptical feeder zone.

$$\text{Feeding Time Ratio} = \frac{N_f}{N_T} \quad (14)$$

where N_f is the number of images with the bird in the defined feeder zone; and N_T is the total number of images within an examination period.

2) Distribution index

The Distribution Index (DI) measures uniformity of broilers position by comparing standard deviation of the position coordinates to the average position coordinates. Fig. 9 shows examples of different levels of DI across x axis.

A simplified distribution index (DI) was developed, retaining the key properties of the original index used by Van Hertem et al. (2018) but streamlining the calculation by directly applying it to the birds' x and y coordinates. Higher DI values indicate more widely dispersed broiler distributions. A higher DI value indicates that broiler positions were more widely dispersed rather than concentrated in a specific area. The index was calculated by comparing the standard deviation of the broilers' x and y coordinates to their average positions (Equations (15) and (16)). This process added six new features to the dataset: the standard deviation of x, the standard deviation of y, the average x position, the average y position, and the resulting distribution indices (DI_x and DI_y).

$$DI_x = \frac{\text{Std}(x)}{\text{Mean}(x)} \quad (15)$$

$$DI_y = \frac{\text{Std}(y)}{\text{Mean}(y)} \quad (16)$$

where DI_x and DI_y are the distribution index along x and y axis, respectively; *std* is standard deviation.

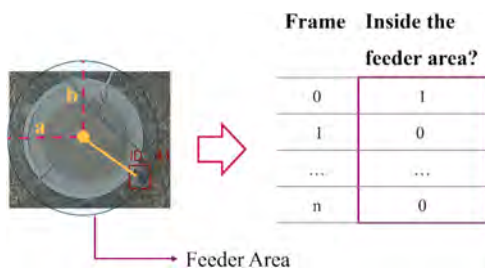


Fig. 8. Schematic drawing for the defined feeder zone and feeding time calculation.

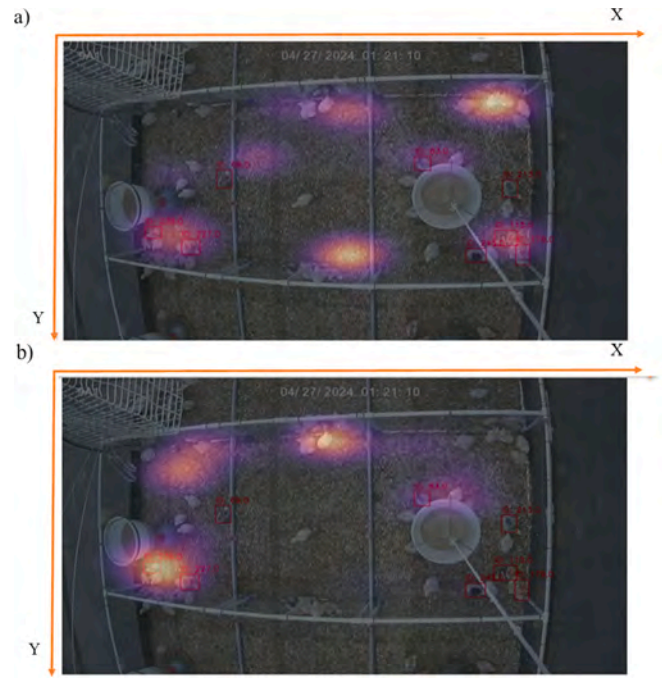


Fig. 9. Examples of different levels of distribution index across x axis. The image (a) belongs to a broiler with high distribution index across x axis whereas the image (b) belongs to a broiler with lower distribution index across x axis, as the former one had more pixel coordinate changes along x axis.

3) Entropy of movement

The entropy of broiler movement was represented by the Shannon entropy (Shannon, 1948) of a set of angles in a movement trajectory. The entropy was quantified with the histogram of movement angle using 20 bins (Equation (17)). The entropy value adds 1 column to the final feature data.

$$H(x) = - \sum p(x) \log(p(x)) \quad (17)$$

where $H(x)$ is the histogram of movement angle in a bin; and $p(x)$ is the probability of each bin. This equation quantifies the uncertainty associated with the distribution of movement direction.

4) Box counting fractal dimensions

The bounding box information of tracked individual broilers was further explored with the box counting fractal dimension. The indicator quantified the level of curly movement patterns of a bird. It also identified stability and randomness of a bird's movement over time, just like route prediction error and entropy. Lame birds could waddle toward feeders or drinkers, which resulted in high-level curly movement patterns instead of regular paths and further led to higher box counting fractal dimensions. A series of square boxes (or grid) were overlaid along tracked trajectories of individual birds. The density of the overlaid grid is determined by the box size value ϵ , with a smaller ϵ resulting in a denser grid (Fig. 10). The number of valid boxes N_ϵ that contained at least one segment of the trajectories was computed. The mapping and counting processes were repeated with 10 different box sizes, starting from $\epsilon = 50$ cm to $\epsilon = 10$ cm with a step size of 4 cm. The fitted curve was plotted with $\log(N_\epsilon)$ against $\log(1/N_\epsilon)$. The slope of the fitted curve represented the fractal dimension of curly bird movement patterns. Higher values of the slope suggest more intricate movement patterns of a bird across a 2D space. The set of computations added one column to the final feature data.

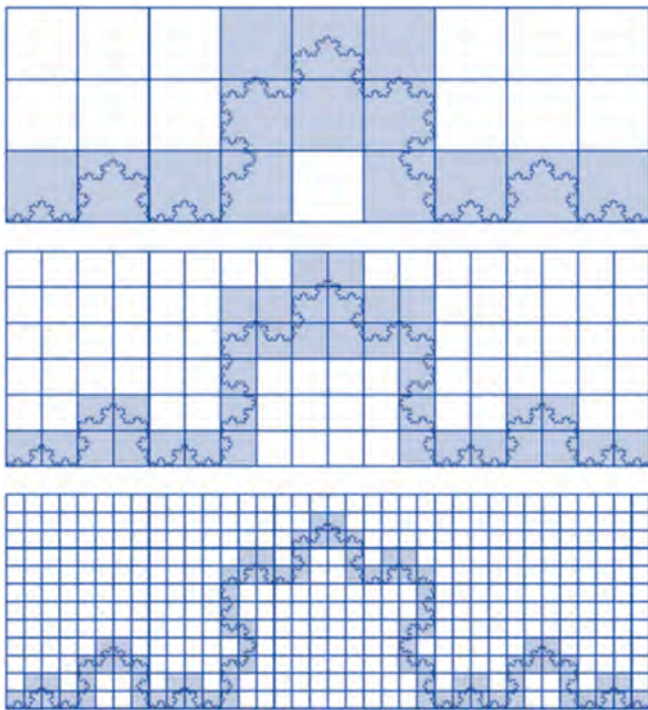


Fig. 10. Examples of box counting fractal dimensions for different ϵ values. The first image has a larger grid size ϵ compared to others. In this method we start from a high ϵ value and go toward the smaller values.

2.7. Machine learning model development for early gait score prediction

To classify gait scores, multiple machine learning algorithms were trained and evaluated (Table 2), including CatBoost, LightGBM, Extra-TreesClassifier, Quadratic Discriminant Analysis, XGBoost, Logistic Regression. The dataset consisted of 4,700 bird instances, each described by 276 features from the 48 pens. A total of 42 pens of data were used for training and validation, while the remaining unseen pens were reserved for testing to reduce data leakage and overfitting issues of early gait score prediction.

Precision (Equation (2)), Recall (Equation (18)), and F1 score (Equation (19)) were used to evaluate the classification accuracy of the final machine learning based gait score prediction models and the

Table 2

Machine learning classifiers information for the early gait score prediction. All unspecified hyperparameters were set to the default values defined in the Python package.

Model name	Python package name	Hyperparameters
LightGBM	lightgbm	gbdt boosting, max depth of 3, 35 leaves, L1 regularization of 2.91, L2 regularization of 4.0, and 1,000 estimators
CatBoost	catboost	depth of 3, 20,000 iterations, learning rate of 0.002, and l2_leaf_reg of 10
ExtraTreesClassifier	scikit-learn	100 estimators, gini criterion, min_samples_split of 2, min_samples_leaf of 1, max_features as 'sqrt'
XGBoost	Xgboost	gbtree booster, L1 regularization of 3, L2 regularization of 3, gamma equal to 0, and max_depth of 6
Quadratic Discriminant Analysis	scikit-learn	reg_param of 0.0, store_covariance equal to false, and tol equal to 0.0001
Logistic Regression	scikit-learn	a maximum of 1,000 iterations, penalty as 'l2', tol equal to 0.0001, C equal to 1.0, lbfgs as solver

optimal machine learning classifier was determined based on the highest F1-score. The optimal machine learning classifier was determined with the highest values of the three metrics. The feature importance analysis was conducted with the optimal model and SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), so that the most important features out of 276 ones can be sorted out for enhancing the mechanism of gait score prediction.

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (18)$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

For model inference, two methods were implemented. Method 1 generated predictions using a single one-hour video. Method 2 employed an ensemble technique to enhance prediction accuracy by aggregating results from multiple one-hour videos per pen. In this approach, the model outputs a probability for each gait score class, and the final predicted gait score for each pen was assigned to the class with the highest average probability across all videos for a given pen.

Fig. 11 shows the distribution of gait scores in the dataset. Before regrouping, most samples were in gait score 0 (about 60%), while gait scores 1 and 2 had fewer samples (31% and 9%, respectively). Due to the data imbalance, we tested two models. In first model, the three-point gait score system was used as target for prediction. In second model, gait scores 1 and 2 were combined into one impaired group. After combining gait scores 1 and 2 into one impaired group, the distribution became more balanced with 60% in group 0 and 40% in group (1). This regrouping reduced the effect of class imbalance and prepared the data for improved model training. Therefore, two classification strategies were evaluated in this study: a three-class classification (gait scores 0, 1, and 2) and a binary classification (gait score 0 versus combined gait scores 1 and 2).

2.8. Important features

To gain a comprehensive understanding of the model's predictions, we analyzed feature importance using both the inherent CatBoost scores and SHAP values. This dual approach combines a global perspective on overall feature utility with detailed insights into how each feature impacts individual predictions.

CatBoost Feature Importance measures how much a feature is used to make key decisions across all the trees in the model. A feature like 'acceleration mean' might be used very frequently for splitting data across many trees, even if those splits only cause small, incremental improvements in accuracy. This gives it a high 'weight' in the model's internal accounting. SHAP Values measure the average magnitude of the impact a feature has on the model's output predictions. A feature might be used often (high CatBoost importance) but have a consistently small and stable effect on the final score (low mean |SHAP| value). Conversely, a feature used only in a few critical decisions could have a massive impact when it is used, giving it a high SHAP value.

3. Results

3.1. Bird tracking performance

Table 3 shows the accuracy and processing speed of detection/tracking models. Overall, pre-trained TAM required 4 s to process one image with the 94% IDF1. The pre-trained Fast-SAM required 0.5 s for one image with the 66% IDF1. The custom-trained YOLOv11 model achieved a processing speed of 0.6–0.8 s per image and an IDF score ranging from 60% to 68%.

The top-down camera configuration provided a consistent and unchanging background with stable lighting and scale, which substantially reduced detection complexity. In addition, the use of color-marked

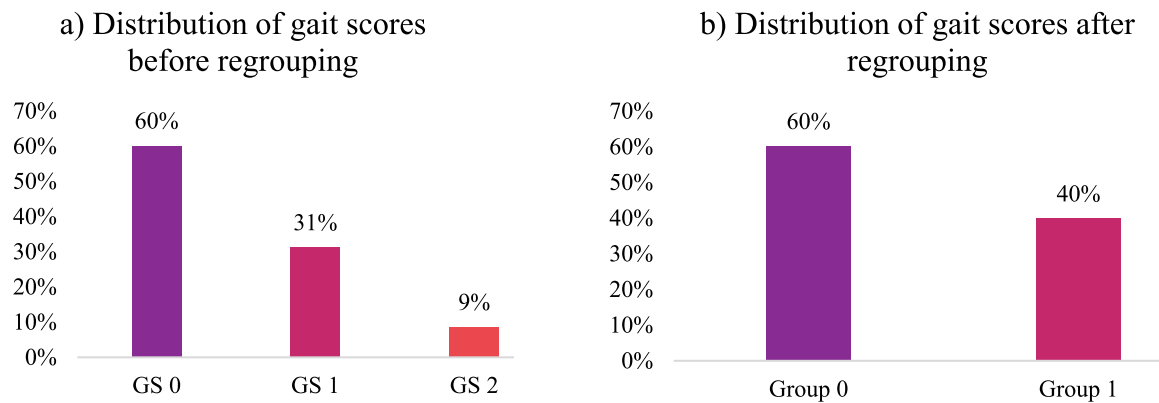


Fig. 11. The distribution of broilers gait score classes before and after combining gait scores 1 and 2 into one group. These gait scores distribution are based on gait scores of birds with 7 weeks of age.

Table 3

Performance of detection/tracking models in terms of processing time and Tracking performance. The mAP is not applicable to TAM and Fast-SAM. For the YOLOv11 model combined with ByteTrack (last row), the mAP measure is unreliable and is therefore omitted, as some detections are dropped to avoid velocity jumps. The image resolution for all models was 960×540 pixels.

Model	Training requirement	Training Time	Processing Speed (seconds per image)	Detection Performance mAP @0.5:0.95 (%)	Tracking performance IDF1 (%)	MOTA (%)
Track-Anything-Model	No training	0	4.0	—	94	81
Fast-SAM	No training	0	0.5	—	66	58
YOLOv11 model (only detection)	Train on 1,700 sample images	41 h (700 epochs)	0.6	90	60	55
YOLOv11 model (combine with ByteTrack)	Train on 1,700 sample images	41 h (700 epochs)	0.8	—	68	63

Note: In performance metrics mAP@0.5:0.95 is mean Average Precision at IoU from 0.5 to 0.95; IDF1 is Identity F1 Score; and MOTA is Multiple Object Tracking Accuracy.

broilers introduced strong visual contrast and consistency, significantly simplifying the detection task and reducing intra-class variability. Furthermore, the training and testing annotations were generated using Fast-SAM, which is built upon a YOLO-based architecture. This shared architectural foundation likely improved compatibility between the annotations and the YOLOv11 detector, contributing to higher detection accuracy.

Although the Track-Anything Model (TAM) showed competitive tracking performance (IDF1 = 94%), its computational intensity (4 s per image) made it impractical for processing the large volume of video data (at least multiple hours from each pen). Therefore, YOLOv11 with ByteTrack was selected as a more efficient alternative.

Table 4 shows the training results of YOLOv11 for different numbers of epochs. As the epochs increased from 50 to 700, the model became more stable and accurate. The precision values improved from 0.976 at 50 epochs to 0.998 at 700 epochs, showing that with more training the model could detect birds more correctly. Recall values changed in a smaller range, between 0.943 and 0.978, which means the model's

Table 4

Performance of the YOLOv11 training in different training epochs.

Epoch	Time (seconds)	Precision (B)	Recall (B)	mAP50 (B)	mAP50-95 (B)
50	10519	0.976	0.943	0.968	0.922
100	21079	0.983	0.956	0.975	0.930
200	42145	0.995	0.966	0.981	0.936
300	63154	0.996	0.972	0.986	0.944
400	84176	0.997	0.974	0.991	0.950
500	105214	0.997	0.975	0.994	0.951
600	126246	0.997	0.976	0.994	0.952
700	147240	0.998	0.978	0.995	0.954

ability to find all birds did not increase as much as precision. The mAP50 values were already high in all cases (0.968 to 0.995), but they also showed slight improvement with more epochs. In the case of mAP50-95, the value increased from 0.922 at 50 epochs to 0.954 at 700 epochs, which indicates better performance in detecting objects at different thresholds. Overall, longer training led to higher precision and accuracy, while recall remained relatively steady.

Fig. 12 shows the confusion matrix of the YOLOv11 model for detecting single- and two-color marked birds. Correct classifications appeared on the diagonal, while misclassifications were off-diagonal. The results indicate higher accuracy for two-color markers compared to single-color ones, confirming that marker design influences detection performance. An imbalance in the number of birds per class occurred in each pen despite the designated 10 color-marked birds, because occlusions (behind feeders, in corners, or from overlapping birds) prevented reliable detection in all cases.

Table 5 shows the detection results of YOLOv11 with two image sizes. The accuracy at $1,920 \times 1,080$ and 960×540 was almost the same (mAP@0.5: 0.994 vs. 0.992). The mAP@0.5:0.95 dropped slightly from 0.943 to 0.90, but the processing speed improved a lot (2.5 to 0.6 s per image). This means the smaller resolution gives much faster results with only a small loss in accuracy.

Table 6 compares the tracking results when using 15 FPS and 5 FPS videos. The error in average velocity was slightly higher at 15 FPS (8 to 24%) compared to 5 FPS (3.5 to 19%), while the standard deviation of velocity showed a larger range of errors at 5 FPS (5 to 28%) than at 15 FPS (13 to 19%). In terms of overall tracking performance, MOTA was 61% for inference at 15 FPS (subsampling to 5 FPS for evaluation) and 55% for native 5 FPS inference. Similarly, IDF1 was 67% and 60%, respectively. The higher image rate was therefore associated with improved tracking performance, but required approximately three times

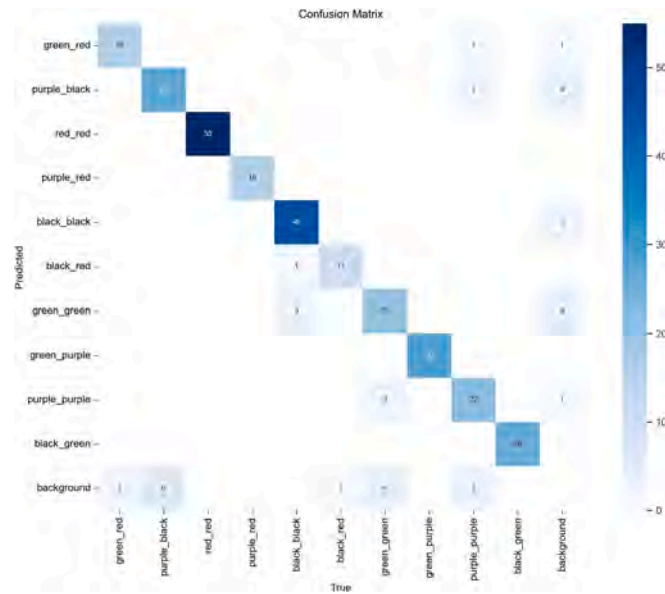


Fig. 12. Confusion matrix of YOLOv11 model in detecting color-marked birds. In this research single- and two-color marker methods are used.

Table 5
Detection performance with two image sizes for the YOLOv11 model.

Metric	Image size of 1,920 × 1,080 pixels	Image size of 960 × 540 pixels
mAP@0.5	0.994	0.992
mAP@0.5:0.95	0.943	0.90
Processing Speed (seconds per image)	2.5	0.6

Table 6
Effect of reducing images per second on tracking. The ground truth tracking data was based on 5 FPS video, and therefore, the 15 FPS tracking data was down-sampled to 5 FPS after tracking to make it possible to compare with the ground truth tracking data.

Metric	Model with 15 images per seconds error	Model with 5 images per seconds error
Average Velocity	8% to 24%	3.5% to 19%
Standard Deviation of Velocity	13% to 19%	5% to 28%
MOTA	61%	55%
IDF1	67%	60%

greater processing time.

3.2. Gait score prediction

Although the Track-Anything Model (TAM) showed competitive tracking performance (IDF1 = 94%), its computational intensity (4 s per image) made it impractical for processing the large volume of video data (at least multiple hours from each pen). Therefore, YOLOv11 with ByteTrack was selected as a more efficient alternative.

The machine learning models for gait score prediction were trained under two different configurations. First, they were trained to predict three distinct gait score classes (0, 1, and 2); the results for this three-class model are presented in Table 7. Subsequently, to address a high imbalance between classes, gait scores 1 and 2 were combined into a single “impaired” group, forming a binary classification task. The results for this binary model are presented in Tables 8 and 9.

Table 7 shows the performance of the CatBoost model across three

gait score classes. CatBoost was selected for detailed reporting of the three-class classification task (gait scores 0, 1, and 2) due to its superior and balanced performance across precision, recall, and F1-score compared to other evaluated models (see Tables 7 and 9). For gait score 0, the model achieved perfect precision (100%) with 72% recall, giving an 84% F1 score. For gait score 1, the precision dropped to 50% while recall was 75%, resulting in the lowest F1 score (60%). Gait score 2 showed higher balance, with 66% precision and 100% recall, producing an F1 score of 80%. Overall, the model reached 72% precision, 82% recall, and 77% F1 score, indicating stronger detection for extreme scores (0 and 2) compared to the intermediate score (1).

The features from the trajectories detected by both the Fast-SAM and YOLOv11 models were used to train separate machine learning models. This method showed how different tracking pipelines affected the accuracy of predicting gait scores. Table 8 compares the accuracy of three machine learning models (CatBoost, LightGBM, and XGBoost) using features from Fast-SAM and YOLOv11 trajectories. For Fast-SAM, CatBoost gave the highest accuracy (76.8%), while LightGBM had the lowest (68.7%). When using YOLOv11, all models performed better, especially with Method 2. CatBoost improved from 75.3% to 82%, LightGBM from 71.5% to 82.5%, and XGBoost from 68.6% to 77.5%. The improvements showed that YOLOv11 features were more effective than Fast-SAM features for gait score prediction. Method 1 only used one-hour videos to make predictions. Method 2, on the other hand, used an ensemble technique that aggregated the predictions of multiple 1-hour videos to make the predictions more accurate.

Similar to Table 8, Table 9 compares different machine learning models for gait score prediction when combining gait scores 1 and 2 into one class, but the focus is on method 2 and analyzing precision and recall in both classes. LightGBM reached the highest accuracy (82.5%) with strong precision and recall for the healthy group (gait score 0) (89.2% and 86.2%) and reasonable performance for the impaired group (gait score 1 and 2) (66.6% precision, 72.7% recall). CatBoost had a similar accuracy (82%), with slightly higher recall for the healthy group (89.6%) but lower recall for the impaired group (63%). Extra Trees and XGBoost both achieved 77.5% accuracy, though their precision and recall for the impaired group were lower (54 to 62%). Logistic Regression had the lowest overall accuracy (64.1%), while QDA showed high recall for the healthy group (96.4%) but completely failed on the impaired group (0%). Overall, boosting models, especially LightGBM and CatBoost, gave the best balance of accuracy and stability compared to the others.

3.3. Important features

As shown in Table 8, CatBoost achieved the highest accuracy for the three-class gait score classification. For the two-class classification in Table 9, although LightGBM achieved a slightly higher accuracy, the difference between the two models was marginal. Given CatBoost’s strong and consistent performance across both three-class and binary gait score classifications, it was selected for SHAP-based feature importance analysis to provide a reliable interpretation of the key features driving model predictions.

Fig. 13 depicts the primary contributing features, emphasizing those that significantly influenced the model’s output. Given that various

Table 7
Performance of CatBoost model classifying three gait scores. These results are based on 3 group classification and YOLO tracking model.

Models	Evaluation criteria (%)		
	Precision	Recall	F1 score
Gait Score 0	100	72	84
Gait Score 1	50	75	60
Gait Score 2	66	100	80
Overall Performance	72	82	77

Table 8

Classification accuracy of three machine learning models using trajectories based on Fast-SAM and YOLOv11 in a two gait score (binary) classification setting. For the YOLO model, results from method 1 and method 2 are presented in separate columns. In method 2, class probabilities predicted from multiple 1-hour videos are averaged to produce a single classification per video.

Models	Fast-SAM	YOLOv11		Difference
	Accuracy (%)	Accuracy (%) – Method 1	Accuracy (%) – Method 2	
CatBoost	76.8	75.3	82	6.7
LightGBM	68.7	71.5	82.5	11
XGBoost	70.2	68.6	77.5	8.9

Table 9

Precision and Recall metrics of different machine learning classification models. The classification task here was binary classification as gait scores 1 and 2 were combined to create one impaired group. The results are based on method 2. The binary classification is only tested on trajectories from the YOLO model. Besides base models such as LightGBM, CatBoost, and XGBoost, other models are tested including Extra Trees Classifier, Quadratic Discriminant Analysis, and Logistic Regression. This table is based on the two-gait score prediction.

	Accuracy (%)	Precision (% (%, Healthy))	Recall (% (%, Healthy))	Precision (% (%, Impaired))	Recall (% (%, Impaired))
Light Gradient Boosting (LightGBM)	82.5	89.2	86.2	66.6	72.7
Categorical Boosting Model (CatBoost)	82.0	86.6	89.6	70.0	63.0
Extra Trees Classifier	77.5	86.2	83.3	54.0	60.0
Extreme Gradient Boosting (XGBoost)	77.5	81.2	89.6	62.0	45.0
Quadratic Discriminant Analysis	69.2	71.0	96.4	0.0	0.0
Logistic Regression	64.1	80.0	61.5	47.0	69.0

machine learning models were assessed during this study, the SHAP analysis was exclusively utilized for the CatBoost model, one of the top-performing classifiers, offering comprehensive insights into the influence of each feature on the model's predictions. The movement angle features (variables beginning with movement_angle) and the distribution index along the y-axis (Y_di) were the most important features in predicting gait scores.

Table 10 lists the most important features identified by the CatBoost model. Among them, the mean acceleration was statistically significant (p -value = 0.008), with a gradual reduction from 0.5675 in gait score 0 to 0.4479 in gait score 2. In addition, standard deviation and maximum velocity of acceleration showed significant differences across gait scores, while feeding time ratio had smaller but still notable variation. These results indicate that kinematic acceleration features play the most important role in distinguishing gait scores.

Feature normality was assessed via Shapiro-Wilk tests ($\alpha = 0.05$), revealing non-normal distributions for more than 40% of kinematic features. We therefore employed the Kruskal-Wallis H test (a non-parametric alternative to one-way ANOVA) to evaluate median differences across gait score groups. We controlled multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure with $\alpha = 0.05$ to account for testing across important features.

These results suggest that normalizing features based on the average of the features among all broilers in that pen (group (3) features) when

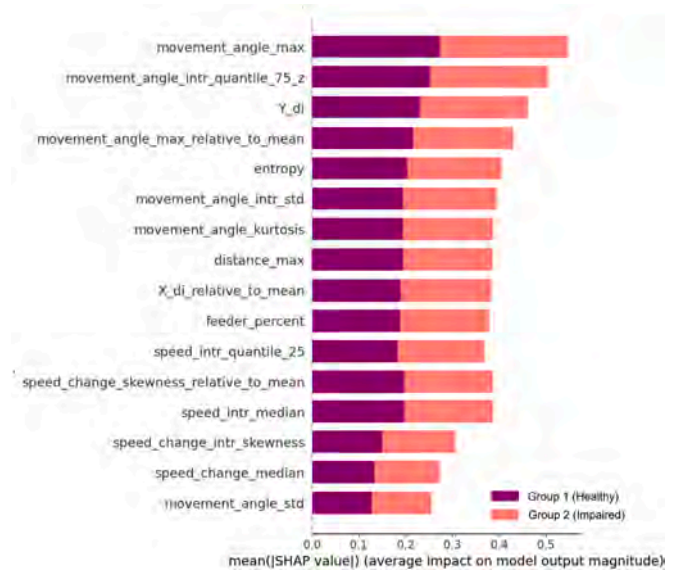


Fig. 13. Feature importance scores indicated by SHAP values with higher values being more important.

combined with other features can help to achieve better accuracies and generalize the model.

As threshold-based values were among the most important predictors, movement-threshold filtering was used as a noise-suppression step to enable reliable extraction of kinematic features for gait score prediction. The novelty of this work does not lie in the thresholding operation itself, but in the integrated early-prediction framework that combines threshold-filtered kinematic features with feeding-time inference, pen-normalized relative features, and explainable machine-learning models for early gait score prediction. Much of the previous research, including Li et al. (2023) used a 3 m long $\times$ 0.46 m wide platform. In this setup, birds typically move in only one direction, making the measurement of linear movement, velocity, and acceleration sufficient for analysis. In larger pens, however, birds can move in many directions. Due to bounding box fluctuations, measuring velocity image-by-image yields misleading values. Although thresholding is easy to implement, it is useful and was absent from prior research on gait score estimation or prediction. Fig. 14 compares the distributions of selected features across the three gait score classes. The boxplots illustrate differences in central tendency and variability, with several features (e.g., speed standard deviation, acceleration threshold, movement angle) showing clear separation between classes. These patterns highlight which kinematic variables carried the strongest discriminative power for predicting gait scores. Outliers are also shown, reflecting the variability within each class.

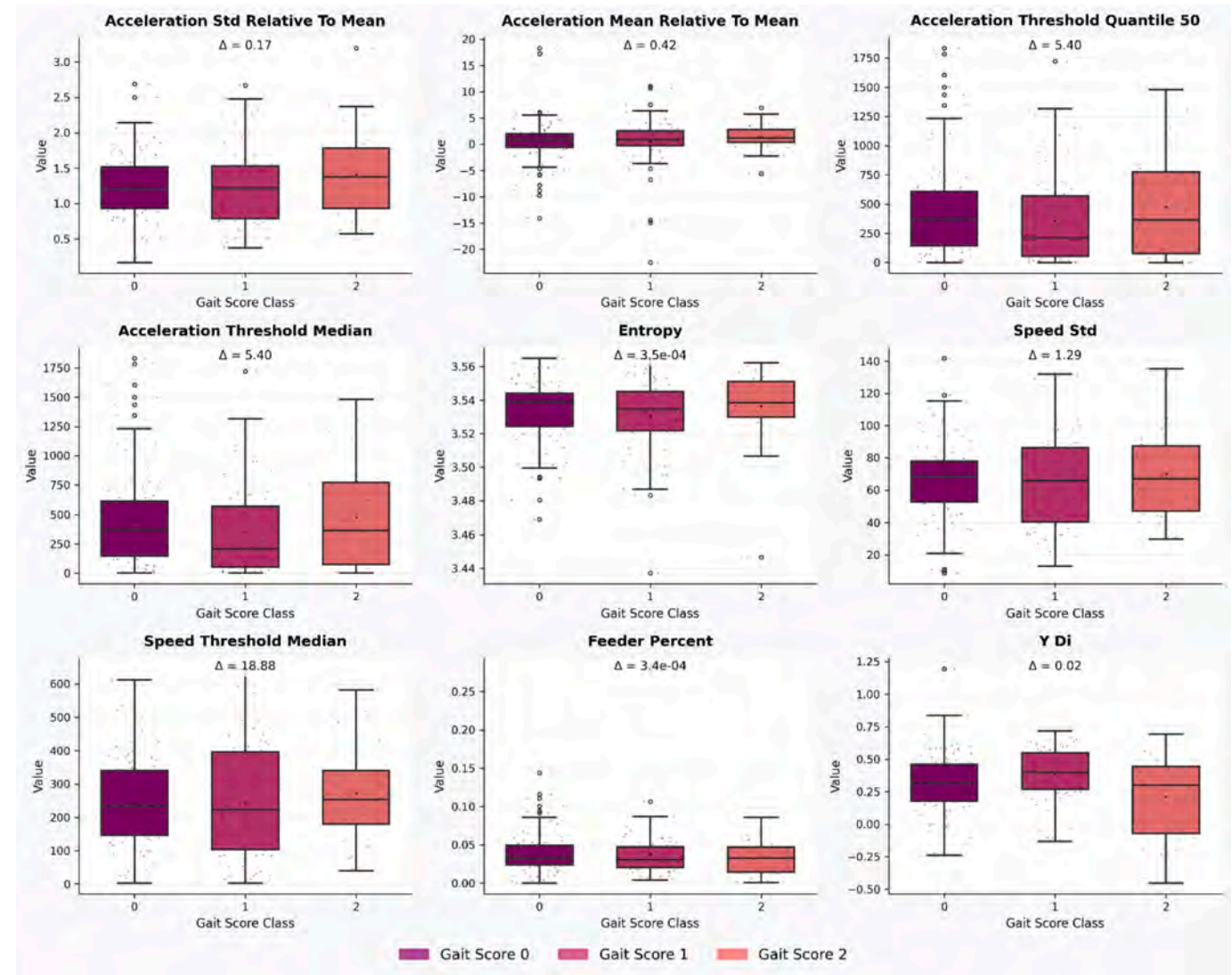
Healthy birds (gait score 0) generally walked longer distances, for example, about 12,456 cm in one hour on average for healthy birds compared with 12,139 cm for impaired ones. Speed threshold values were also consistently higher in gait score 0, with gaps of 5 to 16 units. Feeder use showed only small differences, while the distribution index was usually lower in impaired birds, suggesting a reduced spatial range of movement within the pen. Overall, gait score 0 birds were more active and stable in their locomotion across the day.

Fig. 15 shows temporal patterns of kinematic features for broilers with different gait scores. The curves show that measures such as feeder use, speed, and acceleration vary throughout the day. Notable differences emerged between healthy and impaired birds, with impaired groups generally showing higher variability in some features. These fluctuations suggest that time of day influences activity patterns and should be considered in gait analysis. The apparent discrepancy arises due to the fundamental difference in how these two metrics are

Table 10

Most important features from Catboost model with their average values in each gait score.

Feature name	Feature type	Significance (P-value)	FDR-adjusted P-value	Gait Score 0	Gait Score 1	Gait Score 2
Acceleration quantile 25	Threshold based	0.017	0.033	4.80	2.642	1.790
Acceleration mean	Relative to mean	0.008	0.033	0.568	0.540	0.448
Acceleration standard deviation	Relative to mean	0.042	0.042	2.500	2.900	4.500
Maximum velocity	Relative to mean	0.040	0.042	1.230	1.200	1.360
Feeding time ratio	–	0.020	0.033	0.050	0.040	0.030


Fig. 14. Distribution of features across gait score classes. Boxes show interquartile range (IQR), and dots denote outliers. Here only features with distinct characteristics are presented.

calculated. Normal distance (mean) includes raw per-image movement distances, including minor postural adjustments that are accumulated across all images, whereas the thresholded mean distance only displacements exceeding a predefined minimum (10 cm) were counted and smaller movements were treated as zero, leading to a substantially lower average value (Supplementary Table 1).

Fig. 16 illustrates sample bird trajectories in pens with two average gait scores. In Fig. 16(a), birds from a pen with a lower activity level (average gait score of 0.7) show shorter and less complex movement paths. In contrast, Fig. 16(b) depicts a pen with higher activity (average gait score of 0.3), where trajectories were longer and more widely distributed. These differences highlight the relationship between gait

condition and spatial movement patterns.

4. Discussion

4.1. Tracking pipeline Performance, Configuration, and limitations

Our tracking pipeline, which combined YOLOv11 model with the ByteTrack tracking framework, achieved an IDF1 score of 68% (Table 3). Downsampling the image resolution to 960×540 still maintained strong accuracy (99.2% mAP@0.5, Table 5), while processing time became much faster compared to the original resolution (0.6 vs. 2.5 s per image). In terms of the speed–accuracy trade-off

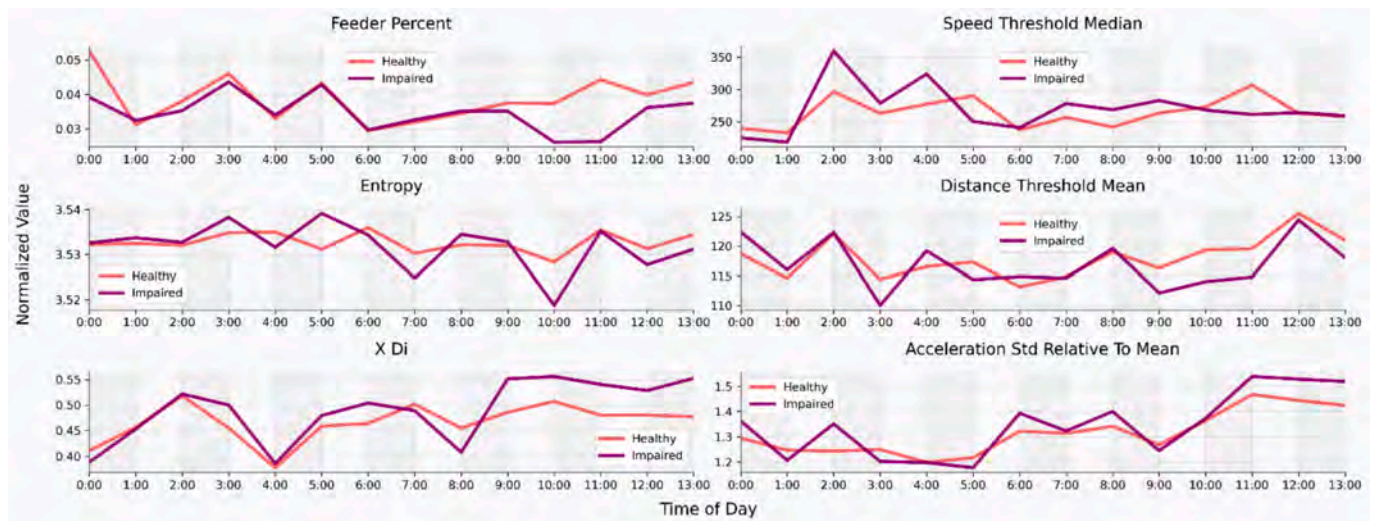


Fig. 15. Averages of feature values across color-marked birds during the day for different gait score groups (healthy and impaired). Distance-related features are reported in centimeters, and all derived features are calculated based on centimeter-scale distances. The data in 0:00 points for instance corresponds to the average of the feature value extracted from all trajectories between 0:00 AM and 01:00 AM. The data is from 0 h AM to 2 h PM which is around 13 h for each pen. Detailed values for these features are provided in Supplementary Table 1.

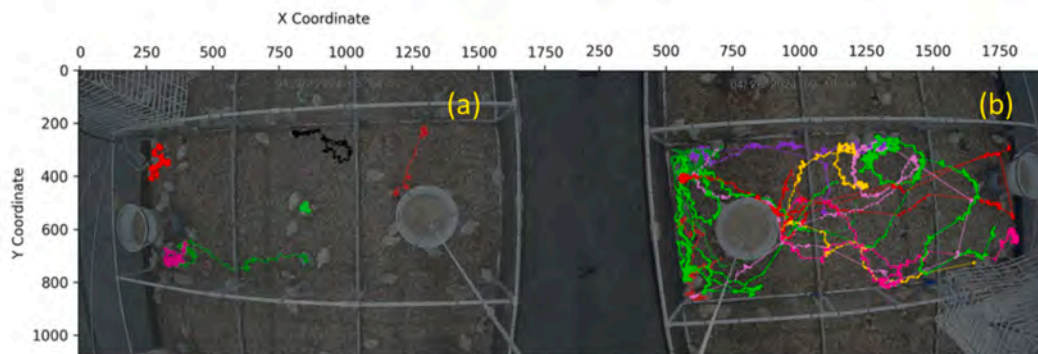


Fig. 16. Image (a) belongs to sample trajectories for a pen with the average gait score 0.7 and image (b) belongs to a sample trajectory for a pen with average gait score of 0.3. Here different colors represent different birds.

(Table 3), YOLOv11 achieved a favorable balance between efficiency and performance, with processing times of 0.6–0.8 s per image, outperforming Fast-SAM in tracking accuracy (IDF1: 68% vs. 66%) while being substantially faster than Track-Anything (4.0 s per image), supporting its selection for subsequent tracking analysis.

Our tracking pipeline reduced trajectory jumps as discussed in Section 2.5. “Trajectory jumps” refer to sudden, unrealistic displacements in an object’s tracked path, typically caused by detection errors or ID switches. These artifacts can produce spurious spikes in derived kinematic features (e.g., velocity) that do not reflect true animal movement. Extracted kinematic features such as velocity were highly sensitive to detection errors, and even minor inaccuracies can lead to significant discrepancies in the derived measurements.

Although this setup improved tracking accuracy, some errors still occurred when birds overlapped near feeders or drinkers. These challenges are similar to occlusion problems reported in previous studies (Doornweerd et al., 2024). Future improvements could come from integrating depth sensors or multi-camera fusion, as also suggested by Fodor et al. (2025).

4.2. Feature engineering and noise mitigation approach

The noise mitigation and predictive capability arise from the

subsequent feature engineering and modeling steps. Specifically, we (i) designed multiple explainable kinematic and behavioral features from early-life trajectories (2 and 3 weeks of age), (ii) applied movement-threshold filtering to suppress bounding-box-induced noise prior to feature aggregation, (iii) post processing trajectories to fix false detections and trajectory jumps explained in section 2.5, and (iv) integrated these features into an explainable machine learning framework for early gait score prediction. The feature engineering in this study integrates feeding-time inference, movement-threshold-filtered kinematics, pen-normalized relative metrics, and fractal-based trajectory descriptors derived from early-life (2 and 3 weeks of age) trajectories, which are subsequently used in an interpretable machine-learning framework to predict gait scores at 7 weeks of age. Feeding-time inference features were derived from trajectory data and designed specifically to capture behavior relevant to gait development. Movement-threshold-filtered kinematic features were applied to suppress bounding-box-induced noise in early-life tracking data and enable reliable aggregation of biologically meaningful locomotor descriptors. Pen-normalized relative features contextualized individual bird movement within group-housed environments. Fractal dimension analysis captured trajectory complexity and represented a novel feature within the context of broiler gait score prediction. While movement thresholds alone are not novel, their use here enables reliable extraction of biologically

meaningful features for early lameness prediction which is a novel work.

This study calculates velocity in a two-dimensional (2D) environment, unlike previous narrow-platform setups that constrained movement to one direction. In wider pens, birds move freely, requiring 2D velocity computation based on image-by-image displacement. The temporal interval for velocity calculation involves a trade-off: short intervals amplify noise, while long intervals underestimate movement due to path curvature. This approach created the problem addressed in section 2.6.2 and algorithm 2, which we solved using a thresholding method.

In this research, in addition to considering the average values of kinematic features, we also examined other statistical properties, such as quantiles, as well as distributional properties, including kurtosis and skewness. Furthermore, we analyzed feeding behaviors, trajectory path complexity, and normalized pen features, aspects that were either absent or not comprehensively addressed in previous research. This research expanded on and tested a richer set of kinematic features than those used in previous studies. For example, Aydin (2017) incorporated factors such as body weight, age, an activity index, and distribution index. De Alencar Nääs et al. (2020) examined mean velocity, mean acceleration, genetic factors, and sex. Li et al. (2023) included lateral body oscillation, linear velocity, linear acceleration, and forward-moving duration. Unlike prior gait-scoring studies that primarily rely on basic kinematics or activity metrics in flock level (Aydin, 2017), this study jointly integrates feeding-time inference, threshold-filtered kinematics, pen-normalized relative features, and interpretable machine learning for early-stage gait score prediction of individual broilers, which could help pinpoint abnormal birds for early intervention. Prior gait-scoring studies primarily relied on basic kinematics, posture metrics, or group-level activity indicators, rather than combining feeding-related, noise-robust features.

4.3. Novelty and comparative advantages of the experimental approach

This work is, to our knowledge, the first to demonstrate accurate early prediction of gait scores using data from birds with 2 and 3 weeks of age, whereas earlier studies mainly focused on later growth stages (Fodor et al., 2025; Pereira et al., 2021). Although models developed by De Alencar Nääs et al. (2020), Yang et al., 2023b, or Li et al. (2023) may achieve higher accuracy in tracking broilers and identifying features to differentiate gait scores, the proposed model has two distinct advantages. First, compared to prior studies using isolated birds or small platforms, our approach incorporates group-housed birds in pens that permit more natural movement and social interaction, offering improved ecological validity for studying gait in a flock setting, though still not fully replicating commercial stocking densities or farm layouts. Second, this research focuses on predicting later-stage gait scores using early-stage data. This time gap naturally increases the error rate; however, if the model's accuracy is sufficient, it can provide farmers with the significant advantage of taking preventive measures.

Unlike prior gait studies using narrow platforms (e.g., 300 cm × 40 cm corridor) or isolated bird setups, our experimental pens measured 170 cm × 300 cm and housed 37 birds, allowing more natural flock movement and interaction, though still below commercial stocking densities. Previous studies used constrained settings, such as narrow single-bird pens with a 300 cm × 40 cm corridor (Fodor et al., 2025), small multi-bird pens of 200 cm × 150 cm (Yang et al. 2023b), or side-view analyses of isolated birds in a 100 cm × 50 cm × 50 cm platform (De Alencar Nääs et al., 2020). Compared with those studies, our approach facilitates more natural flock behavior and movement.

4.4. Gait score prediction results and model performance

Our results showed that trajectory-based features extracted from Fast-SAM and YOLOv11 enabled reliable prediction of week gait scores of birds with 7 weeks of age using data from birds with 2 and 3 weeks of

age. The ensemble approach (method 2), which averaged the class probabilities across all video segments to come with one final prediction, achieved 82.5% accuracy (Table 8), which is higher than the 78% accuracy reported by Van Hertem et al. (2018) in older birds, and comparable to the performance reported by Yang et al. (2023b)) ($R^2 = 0.90$ to 0.91 for 3 to 9 weeks of age). Combining gait scores 1 and 2 into a single impaired class improved the overall and minimum precision and recall among gait scores, showing that regrouping helped reduce the effect of class imbalance. When comparing machine learning models (Table 9), boosting approaches such as CatBoost and LightGBM performed best, with accuracy above 82%, while XGBoost and Extra Trees remained slightly lower (77.5%). Logistic Regression had the weakest results, with an accuracy of 64.1% and poor precision for the impaired group. These differences highlight the advantage of boosting models in handling complex and non-linear patterns in the data.

4.5. Interpretation and importance of predictive features

The analysis highlighted acceleration quantiles (P-value = 0.017), acceleration mean (P-value = 0.008), feeding time ratios (P-value = 0.020), and movement angle variability as the most important predictors (Table 10, Fig. 13). Among these, acceleration quantiles and mean showed a strong difference between gait score groups, with lower values in impaired birds. This agrees with Aydin (2017), who reported that lame broilers often display irregular velocity patterns ($r = 0.844$). Feeding time also played a role, with impaired birds spending less time at feeders (impaired birds spent 3.4% of time feeding, while healthy ones spent 3.7% on average). When comparing feature groups, pen-normalized measures (Group (3) improved robustness by placing individual movement into the context of flock activity. For example, pens with generally lower movement had higher rates of impaired gait scores (Fig. 16). Threshold-based features (Group (2) also helped by reducing noise from bounding-box fluctuations, which strengthened their biological relevance. In contrast, trajectory complexity (e.g., fractal dimension) and path prediction errors were less useful, suggesting that simpler kinematic features are sufficient for early detection.

4.6. Limitations and challenges for real-world application

Despite these promising results, this study has several other limitations. In our experimental setup, color markers were used to distinguish individual birds. In a real-world farming environment, however, distinguishing, tracking, and calculating individual kinematic metrics is significantly more challenging. In a real farm setting, ID loss for tracked objects occurs more frequently than in a controlled pen measuring 1.2 m wide × 3.0 m long. For instance, Doornweerd et al. (2024) analyzed broiler tracking in a pen measuring 1.80 m × 2.61 m and found that broiler ID loss usually occurs (56% of the time) when broilers are in close proximity to each other. Rapid movement was another factor for ID loss. Notably, in that experiment, the camera setup had no blind spots. In top-view camera setups, the design of the pen can create blind spots. In this research, the feeders had a cone shape, which resulted in the loss of visibility for the birds when they were positioned underneath or behind the feeders, leading to difficulties in tracking and detection. On a real farm, it is difficult to create an environment without blind spots. Furthermore, the high number of broilers and greater stocking densities on a real farm make it more challenging to maintain tracking of an individual bird for an extended period. However, it is still possible to track birds for a few seconds until the system loses them. The key methodological distinction is that, on a real farm, the focus should be on tracking each unique ID separately and subsequently aggregating feature values across the entire farm population, rather than focusing on specific individual birds.

The brief tracking period for each ID is sufficient to calculate kinematic features such as distance, velocity, acceleration, movement angle, and path prediction error. The average feeding time ratio across the farm

can also be measured on a real farm and is similar to the work in this research. However, on real farms, it is difficult to calculate other metrics, such as the distribution index, entropy, or box-counting fractal dimensions. For the feasible metrics, we calculate distance, velocity, acceleration, movement angle, and path prediction error for each ID (not each physical bird). We can then aggregate these feature values across all IDs to compute summary statistics, including averages, standard deviations, maximum and minimum values, feeding time ratios, and quantiles. Unlike our controlled pens with fixed boundaries, commercial farms lack clearly defined spatial constraints, making it difficult to calculate distribution index, entropy, or box-counting fractal dimensions for individual birds. For future commercial implementation, three approaches could accommodate the tracking constraints of dense environments: aggregating data at the flock level, developing proxy metrics, or establishing well-defined monitoring zones through spatial segmentation to compute these features meaningfully.

A further limitation involves tracking accuracy, which can be compromised by issues such as occlusions during close interactions and camera blind spots. There are also inherent limitations in predicting gait scores. Environmental variables (e.g., litter quality) that affect birds between 2 and 7 weeks of age make it difficult to achieve very high prediction accuracy. Genetic variability is another contributing factor that is not included in our model. The remaining error rate ($\sim 17.5\%$) may be due to uncontrolled environmental factors (e.g., litter quality) or genetic variability in Cobb 500 broilers. Furthermore, any high accuracy that is achieved may be partially attributable to noise within the data.

While our experimental setup allows more natural flock movement than isolated bird platforms, we acknowledge it does not fully replicate commercial stocking densities ($\sim 1 \text{ ft}^2$ per bird). Scaling to commercial houses would require adaptation of the tracking pipeline to handle severe occlusion and high bird counts, for example through flock-level aggregation, representative sampling, or the use of multi-camera or depth-sensing systems to mitigate occlusion.

The subjectivity of manual gait scoring was present in the current dataset used for model training, as gait scores were assigned based on human observation, which is an inherent and unavoidable aspect of visual gait assessment. To minimize the impact of subjective judgment, a consensus-based scoring protocol was applied in the present study, in which at least three trained individuals independently scored the same birds and reached agreement before final gait scores were assigned. Previous studies have similarly identified inter- and intra-observer variation as a limitation of visual gait scoring (Pereira et al., 2021), with gait scores often assigned by human observers and inter-observer reliability not always quantified (van der Sluis et al., 2021). The reliability of subjective scoring methods like gait assessment is commonly quantified using Cohen's kappa (κ), a statistical measure that evaluates agreement between observers while accounting for chance (Landis & Koch, 1977). Values closer to 1 indicate stronger agreement. In poultry gait scoring literature, even under controlled conditions, inter-observer agreement is typically substantial but not perfect. For example, Tahamtani et al. (2021) reported an inter-observer κ of 0.70, while intra-observer agreement for experienced individuals ranged from $\kappa = 0.77$ to 0.90. This inherent variability in the manual scoring standard establishes a realistic ceiling for model accuracy. Other studies have addressed this limitation by using independent scoring followed by consensus to improve reliability, which is consistent with the approach used in the present study (Pereira et al., 2021). Therefore, the remaining prediction error should be interpreted in the context of the inherent subjectivity of manual gait scoring rather than as a limitation of the proposed modeling framework. The primary value of our system was not in achieving a perfect biological diagnosis, but in providing a reliable, high-throughput tool that applies a standardized welfare assessment consistently across thousands of birds with the least observation variabilities, enabling early flock-level risk detection that is impractical with manual scoring. An important consideration is that the model's predictions are ultimately anchored to the subjective nature of manual gait

scoring. The 'ground truth' used for training reflects human observer consensus, which includes inherent variability. Therefore, the model's objective is to automate and standardize the application of this established welfare protocol, not to define an objective biological ground truth. The achieved accuracy represents a high degree of alignment with human judgement, and the system's principal advantage lies in providing consistent, scalable application of the scoring criteria, thereby eliminating observer bias and enabling practical flock-level monitoring.

4.7. Future research directions

The data annotations of large dataset for gait score modeling are the major bottleneck. Future studies could focus on reducing the need for extensive manual annotation. For example, a loop that combines short periods of manual labeling, training a YOLO model, and using more advanced tracking algorithms could speed up the process and make large-scale gait scoring more practical. Another promising direction is the integration of 3D pose estimation, which could enable richer features and a more accurate interpretation of automated gait annotations.

Future research could also focus on predicting bone diseases or developing a prediction model in other gait-scoring methods. In addition, explainability tools like SHAP could be developed in a more farmer-friendly way, allowing insights to be translated into management decisions, for instance, targeting dietary interventions in pens where feeding time is low. Finally, analyzing variation in activity metrics across different hours (Fig. 15) may identify time windows with stronger predictive power, helping to refine monitoring schedules. To apply the results of this research on real farms, one potential solution is to focus on the average kinematic features of the entire flock to predict an average gait score, rather than attempting to score each bird individually. To do this, using a setup similar to this paper, the model can ignore losing a bird ID in the actual farm environment tracking. In fact, kinematic features of each tracking ID can be calculated separately and then should be aggregated over a time window. The weight of such gait score prediction model is not necessarily same as this paper, but the methodologies of this paper can be applied. In reality, there are no sufficient color markers covering every bird in a farm containing tens of thousands of birds. Perhaps, color-marking a small set of birds and using them to represent the flock could be an option.

5. Conclusion

In conclusion, this study successfully demonstrates the feasibility of early gait score prediction in broilers by leveraging spatiotemporal kinematic tracking and explainable machine learning. Robust individual bird tracking within group-housed settings was achieved using a YOLOv11 model with the ByteTrack tracking framework applied to color-marked broilers, enabling the extraction of reliable spatiotemporal trajectories for downstream analysis. The creation of new, explainable features such as pen-normalized and movement-threshold-filtered metrics successfully eliminated noise and facilitated model generalizability. The trained CatBoost model achieved 82.5% accuracy in healthy and impaired bird prediction weeks in advance, using predictive variables including acceleration statistics and feeding time ratios. The method provides a valuable platform for proactive intervention in welfare, providing a computer-vision-based, scalable solution to early lameness monitoring in commercial broiler production.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Claude in order to improve the readability and language of the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

CRediT authorship contribution statement

Hosna Mohammadiabadi: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Guoming Li:** Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Tongshuai Liu:** Writing – review & editing, Data curation. **Venkat Umesh Chandra Bodempudi:** Writing – review & editing, Data curation. **Oluwadamilola Oso:** Writing – review & editing, Data curation. **Mahtab Saeidifar:** Writing – review & editing, Methodology, Data curation. **Sai Akshitha Reddy Kota:** Writing – review & editing, Data curation. **Ehsan Asali:** Writing – review & editing, Data curation. **Aravind Mandiga:** Writing – review & editing, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported by Cobb Research Initiative.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.compag.2026.111651>.

Data availability

Data will be made available on request.

References

- Aydin, A., 2017. Development of an early detection system for lameness of broilers using computer vision. *Comput. Electron. Agric.* 136, 140–146.
- Bassler, A., Arnould, C., Butterworth, A., Colin, L., De Jong, I., Ferrante, V., Ferrari, P., Haslam, S., Wemelsfelder, F., Blokhuis, H., 2013. Potential risk factors associated with contact dermatitis, lameness, negative emotional state, and fear of humans in broiler chicken flocks. *Poult. Sci.* 92 (11), 2811–2826.
- Butterworth, A., Arnould, C., van Niekerk, T. F., Veissier, I., Keeling, L., van Overbeke, G., & Bedaux, V. (2009). Welfare quality® assessment for poultry (broilers, laying hens).
- Campbell, M., Miller, P., Díaz-Chito, K., Hong, X., McLaughlin, N., Parvinzamin, F., Del Rincon, J.M., O'Connell, N., 2024. A computer vision approach to monitor activity in commercial broiler chickens using trajectory-based clustering analysis. *Comput. Electron. Agric.* 217, 108591.
- Caplen, G., Hothersall, B., Murrell, J.C., Nicol, C.J., Waterman-Pearson, A.E., Weeks, C. A., Colborne, G.R., 2012. Kinematic analysis quantifies gait abnormalities associated with lameness in broiler chickens and identifies evolutionary gait differences. *PLoS One* 7 (7), e40800.
- Cobb-Vantress. (2022). *Cobb500 broiler performance & nutrition supplement*. <https://www.cobbgenetics.com/assets/Cobb-Files/2022-Cobb500-Broiler-Performance-Nutrition-Supplement.pdf>.
- Cordeiro, A., Nääs, I., Salgado, D., 2009. Field evaluation of broiler gait score using different sampling methods. *Brazilian J. Poultry Sci.* 11, 149–154.
- Council, N. C. (2020). *Animal Welfare Guidelines and Audit Checklist for Broilers*. N. C. Council. <https://www.nationalchickencouncil.org/wp-content/uploads/2017/07/NCC-Welfare-Guidelines-Broilers.pdf>.
- Council, N. C. (2022). *National Chicken Council Broiler Welfare Guidelines and Audit Checklist*. https://www.nationalchickencouncil.org/wp-content/uploads/2023/01/NCC-Broiler-Welfare-Guidelines_Final_Dec2022-1.pdf.
- Council, N. C. (2024). *U.S. broiler performance: 1925 to present*.
- De Alencar Nääs, I., da Silva Lima, N., Gonçalves, R., de Lima, L., Ungaro, H., Abe, J., 2020. Lameness prediction in broiler chicken using a machine learning technique. *Information Processing in Agriculture*. In.
- Doornweerd, J., Veerkamp, R., De Klerk, B., Van der Sluis, M., Bouwman, A., Ellen, E., Kootstra, G., 2024. Tracking individual broilers on video in terms of time and distance. *Poult. Sci.* 103 (1), 103185.
- Fodor, I., Taghavi, M., Ellen, E.D., van der Sluis, M., 2025. Top-view characterization of broiler walking ability and leg health using computer vision. *Poult. Sci.* 104 (2), 104724.
- Garner, J., Falcone, C., Wakenell, P., Martin, M., Mench, J., 2002. Reliability and validity of a modified gait scoring system and its use in assessing tibial dyschondroplasia in broilers. *Br. Poult. Sci.* 43 (3), 355–363.
- Julian, R.J., 2005. Production and growth related disorders and other metabolic diseases of poultry—a review. *Vet. J.* 169 (3), 350–369.
- Kestin, S., Knowles, T., Tinch, A., Gregory, N., 1992. Prevalence of leg weakness in broiler chickens and its relationship with genotype. *Vet. Rec.* 131 (9), 190–194.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., & Lo, W.-Y. (2023). Segment anything. *Proceedings of the IEEE/CVF international conference on computer vision*.
- Knowles, T.G., Kestin, S.C., Haslam, S.M., Brown, S.N., Green, L.E., Butterworth, A., Pope, S.J., Pfeiffer, D., Nicol, C.J., 2008. Leg disorders in broiler chickens: prevalence, risk factors and prevention. *PLoS One* 3 (2), e1545.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.
- Li, G., Gates, R.S., Meyer, M.M., Bobeck, E.A., 2023. Tracking and characterizing spatiotemporal and three-dimensional locomotive behaviors of individual broilers in the three-point gait-scoring system. *Animals* 13 (4), 717.
- Li, G., Zhao, Y., Purswell, J.L., Du, Q., Chesser Jr, G.D., Lowe, J.W., 2020. Analysis of feeding and drinking behaviors of group-reared broilers via image processing. *Comput. Electron. Agric.* 175, 105596.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Merenda, V.R., Bodempudi, V.U., Pairis-Garcia, M.D., Li, G., 2024. Development and validation of machine-learning models for monitoring individual behaviors in group-housed broiler chickens. *Poult. Sci.* 103 (12), 104374.
- Pereira, D. F., Nääs, I. d. A., & Lima, N. D. d. S. (2021). Movement analysis to associate broiler walking ability with gait scoring. *AgriEngineering*, 3(2), 394–402.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27 (3), 379–423.
- Tahamtani, F.M., Herskin, M.S., Foldager, L., Murrell, J., Sandercock, D.A., Riber, A.B., 2021. Assessment of mobility and pain in broiler chickens with identifiable gait defects. *Appl. Anim. Behav. Sci.* 234, 105183.
- USDA. (2025). *USDA Agricultural Projections to 2034*. W. A. O. B. Office of the Chief Economist, U.S. Department of Agriculture. <https://ers.usda.gov/sites/default/files/laserfiche/outlooks/110966/OCE-2025-1.pdf?v=25143>.
- van der Sluis, M., Ellen, E.D., de Klerk, B., Rodenburg, T.B., de Haas, Y., 2021. The relationship between gait and automated recordings of individual broiler activity levels. *Poult. Sci.* 100 (9), 101300.
- Van Hertem, T., Norton, T., Berckmans, D., Vranken, E., 2018. Predicting broiler gait scores from activity monitoring and flock data. *Biosyst. Eng.* 173, 93–102.
- Webster, A., Fairchild, B., Cummings, T., Stayer, P., 2008. Validation of a three-point gait-scoring system for field assessment of walking ability of commercial broilers. *J. Appl. Poult. Res.* 17 (4), 529–539.
- Weeks, C., Danbury, T., Davies, H., Hunt, P., Kestin, S., 2000. The behaviour of broiler chickens and its modification by lameness. *Appl. Animal Behav. Sci.* 67 (1–2), 111–125.
- Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., & Zheng, F. (2023). Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*.
- Yang, X., Zhao, Y., Gan, H., Hawkins, S., Eckelkamp, L., Prado, M., Burns, R., Purswell, J., Tabler, T., 2023b. Modeling gait score of broiler chicken via production and behavioral data. *Animal* 17 (1), 100692.
- Zhang, Z., Cheng, D., Zhu, X., Lin, S., & Dai, J. (2018). Integrated object detection and tracking with tracklet-conditioned detection. *arXiv 2018. arXiv preprint arXiv:1811.11167*.
- Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., & Wang, J. (2023). Fast segment anything. *arXiv preprint arXiv:2306.12156*.