



Full-Length Article

Machine learning classification of dilated cardiomyopathy severity in commercial broiler chickens

Ümit Bilginer^{a,b,*} , Arthur F.A. Fernandes^b , Brenda Flack^b, Nicholas Anthony^b, Vivian Likness^b

^a Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI, 53706, USA

^b Cobb Vantress, Inc, 4703 US HWY 412 E, Siloam Springs, 72761, AR, USA

ARTICLE INFO

Keywords:

Broiler
Cardiac morphometry
Dilated cardiomyopathy
Feature importance
Machine learning

ABSTRACT

Dilated cardiomyopathy (DCM) is a significant cardiovascular disorder in commercial broiler chickens associated with congestive heart failure, ascites, and sudden death, leading to substantial economic losses in poultry production. Objective and accurate identification of DCM severity remains challenging due to reliance on subjective postmortem scoring systems. This study evaluated seven machine learning algorithms for classifying DCM severity into three ordinal categories (Score 1 = normal, Score 2 = moderate, Score 3 = severe) using 15 features comprising 6 cardiac morphometric measurements and 9 physiological and production parameters from 430 male Cobb 500 broiler chickens. Models were trained using 5-fold cross-validation with hyperparameter tuning via grid search, and stability was assessed through repeated 10 × 5-fold stratified cross-validation. XGBoost, Ridge regression, and Elastic Net achieved the highest test accuracy of 74.7% (95% CI: 64.2 to 83.4%), with macro AUC values of 0.85 for all three models. No statistically significant differences were observed between model performances (McNemar's test, all $P > 0.05$). Morphometric features, particularly heart width (w1) and the width-to-length ratio (w1l1), contributed approximately 62% of total feature importance, while physiological parameters contributed 38%. SHapley Additive exPlanations (SHAP) analysis confirmed that w1 was the most influential predictor across all severity classes. Score 3 (severe DCM) classification was limited by small sample size ($n = 29$, 6.7%). On the held-out test set, five of seven models achieved Score 3 sensitivity of 33.3% (2 of 6 cases); however, bootstrap analysis on cross-validated predictions yielded a lower mean sensitivity of 6.9% (95% CI: 0.0 to 17.2%), reflecting the high instability of this estimate. A supplementary binary classification (Normal vs. DCM) using repeated cross-validation yielded mean accuracies up to 79.0% with AUC up to 0.846. These findings demonstrate that machine learning approaches can provide objective and repeatable support for DCM severity assessment, with cardiac morphometric measurements providing the strongest discriminative signal. However, predictive performance was weakest and least stable for the severe class (Score 3), owing to its limited representation (6.7% of birds). Future work should focus on increasing severe DCM sample representation and integrating imaging-based features for improved classification.

Introduction

In poultry, dilated cardiomyopathy (DCM) is a progressive cardiovascular disorder characterized by ventricular dilation, impaired systolic function, and myocardial wall thinning, and leading to congestive heart failure and sudden death (Crespo and Shivaprasad, 2013). In commercial broiler chickens, DCM has become increasingly prevalent due to intensive genetic selection for faster growth rates and increased breast muscle yield, which imposes substantial metabolic and

cardiovascular demands on the developing heart (Julian, 2005; Olkowski, 2007). Reported flock mortality attributable to cardiovascular causes, including DCM, ranges from 2% to 8% in commercial operations (Olkowski, 2007), though prevalence varies with genetics, management practices, and growth rate. The condition results in significant economic losses through increased mortality, decreased feed efficiency, and carcass condemnations (Bordbar et al., 2025). Currently, DCM severity in broiler chickens is mostly evaluated after death through visual scoring. Trained personnel examine the heart and assign ordinal

* Corresponding author at: Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI, 53706, USA.

E-mail address: bilginer@wisc.edu (Ü. Bilginer).

<https://doi.org/10.1016/j.psj.2026.107473>

Received 23 April 2026; Accepted 21 July 2026

Available online 21 July 2026

0032-5791/© 2026 The Authors. Published by Elsevier Inc. on behalf of Poultry Science Association Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

severity scores based on gross cardiac morphological changes including ventricular dilation, wall thinning, and myocardial pallor. Although these scoring systems are widely used in research and industry settings, they are inherently limited by inter-observer variability and lack quantitative standardization (Crespo and Shivaprasad, 2013). Because scoring relies on subjective visual judgment, the same heart may be graded differently by different observers, or by the same observer on different occasions, and hearts falling between adjacent grades are particularly prone to disagreement. This limited repeatability reduces the consistency and comparability of severity data across studies, and underlines the need for objective and reproducible assessment tools.

Machine learning (ML) approaches have emerged as promising tools for disease classification in veterinary medicine, offering the potential for objective, repeatable, and quantitative assessment of complex multivariate phenotypes (Liakos et al., 2018; Alves et al., 2024; Chidziwisano et al., 2025). Recent applications in poultry health include automated detection of infectious diseases from fecal images (Machuve et al., 2022), lameness scoring, and body condition assessment using image analysis and sensor data (Okinda et al., 2020). However, the application of ML methods to cardiac disease classification in broiler chickens has not been reported. To our knowledge, no previous study has used ML algorithms for DCM severity classification in poultry. In addition, most previous poultry machine learning studies have relied on a single algorithm or a narrow feature set. As a result, it remains unclear which modelling approach is most suitable for graded cardiac traits, and how much cardiac morphometry contributes beyond physiological measures. A systematic comparison of several algorithm families under a common evaluation protocol is therefore needed to establish a reliable baseline for DCM severity classification.

Cardiac morphometric measurements, including heart dimensions, ventricular wall thickness, and derived ratios, provide objective quantitative descriptors of cardiac remodeling associated with DCM (Cohn et al., 2000; Olkowski et al., 2020). Physiological blood gas parameters, such as blood pH, partial pressures of oxygen and carbon dioxide, and lactate concentrations, reflect the metabolic consequences of cardiac dysfunction and may provide complementary diagnostic information (Wideman et al., 2013). The integration of both feature domains through multivariate ML algorithms may capture complex nonlinear relationships between cardiac structure, physiological status, and disease severity that univariate approaches cannot detect. In human cardiology, ML approaches have been applied to cardiomyopathy classification with promising results (Alimadadi et al., 2020; Accardo et al., 2022; Zhou et al., 2023).

The objectives of this study were to: (1) evaluate and compare seven ML algorithms for three-class DCM severity classification using cardiac morphometric and physiological features from commercial broiler chickens; (2) identify the most informative features and their relative contributions through importance ranking and SHapley Additive exPlanations (SHAP) analysis; and (3) assess classification performance for the underrepresented severe DCM category. DCM severity is an ordinal trait, so classification errors are not equally costly, misclassifying a severe heart as normal is more consequential than confusing two adjacent grades. We nonetheless treated the outcome as a nominal three-class problem, so that a broad range of standard classifiers could be compared under identical conditions. The pattern of misclassifications was then examined to recover the ordinal structure that overall accuracy alone does not capture.

Material and methods

Animals and data collection

Data were collected from 430 male Cobb 500 broiler chickens

originating from a single pure line maintained at a Cobb-Vantress pedigree facility. Birds were managed according to the Cobb Broiler Management Guide (Cobb-Vantress Inc., 2021) and processed at approximately 7 weeks of age. These birds had been held for one additional week after selection for the purposes of cardiac evaluation. Prior to processing, venous blood samples were collected from the wing vein of each bird and analyzed on-site using an E poc point-of-care blood gas analyzer (Siemens Healthineers). Samples were processed within minutes of collection to minimize pre-analytical variation. Cardiac evaluations were performed postmortem. Each heart was scored for DCM severity independently by five trained observers, and the only the final score assigned by majority consensus was retained and used in this study: Score 1 (normal, no gross cardiac abnormalities; $n = 246$, 57.2%), Score 2 (moderate DCM, mild to moderate ventricular dilation; $n = 155$, 36.0%), and Score 3 (severe DCM, marked ventricular dilation with wall thinning; $n = 29$, 6.7%).

Feature description

Fifteen features were measured or derived for each bird across two domains.

Six cardiac morphometric features were obtained from standardized dorsal-view photographs of excised hearts. The atria were removed prior to photography, and a ruler was placed alongside each heart for pixel-to-centimeter calibration. Total ventricle length ($l1$, cm) was measured from base to apex. Three widths were measured at equidistant points along this length: $w1$ (proximal width, cm), $w2$ (mid-ventricle width, cm), and $w3$ (distal width, cm). Three ratio features were derived: width-to-length ratio ($w1l1 = w1/l1$), mid-width-to-length ratio ($w2l1 = w2/l1$), and width ratio ($w1w2 = w1/w2$). Heart length ($l1$) was used to derive the ratio features but was not included as a separate model input because the information it provides is already captured by, and therefore redundant with, the derived ratios. Following photography, the total ventricle was weighed, the left ventricular free wall was dissected away, and the right ventricle was weighed separately to calculate the right ventricle to total ventricle weight ratio (RVTVRatio) (Fig. 1).

Nine physiological and production features comprised blood pH (pH), partial pressure of oxygen (pO_2 , mmHg), partial pressure of carbon dioxide (pCO_2 , mmHg), potassium concentration (K , mmol/L), blood lactate concentration (Lac , mmol/L), hematocrit (Hct , %), body weight gain ($Gain$, g), live body weight ($Weight$, g), and RVTVRatio.

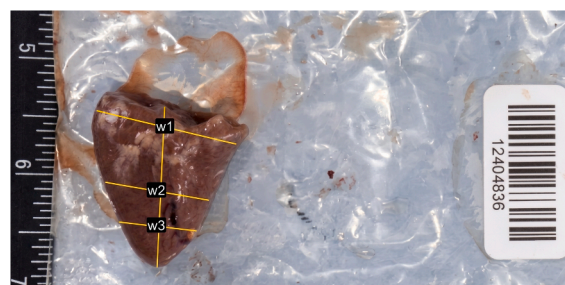


Fig. 1. Dorsal-view photograph of an excised broiler heart illustrating the morphometric measurement protocol. Cardiac dimensions ($w1$, $w2$, $w3$, $l1$) were measured from standardized photographs with pixel-to-centimeter calibration.

Machine learning models

Although regularized regression methods (Ridge, Lasso, Elastic Net) originate from classical statistics, they are widely adopted in the machine learning framework for supervised classification tasks due to their capacity for automated regularization, built-in feature selection, and compatibility with cross-validation-based tuning pipelines (Jiang et al., 2020). The inclusion of these algorithms alongside tree-based and kernel-based methods allows for a comprehensive comparison across fundamentally different modeling strategies, ranging from linear to nonlinear decision boundaries. Seven ML classification algorithms were evaluated, representing a range of model architectures. Tree-based ensemble methods (Random Forest (RF); Breiman, 2001) and Extreme Gradient Boosting (XGBoost; Chen and Guestrin, 2016)), kernel-based method (Support Vector Machine with radial basis function kernel (SVM-RBF); Cortes and Vapnik, 1995)), instance-based method (K-Nearest Neighbors (KNN); Cover and Hart, 1967)) and regularized multinomial logistic regression (Ridge (L2 penalty, $\alpha = 0$; Hoerl and Kennard, 1970), Lasso (L1 penalty, $\alpha = 1$; Tibshirani, 1996), and Elastic Net (combined L1/L2 penalty, $\alpha = 0.5$; Zou and Hastie, 2005)).

Hyperparameter tuning

Hyperparameters were optimized via grid search with 5-fold cross-validation using the R caret package (Kuhn, 2008). Full hyperparameter search spaces for all seven algorithms are provided in Supplementary Table S1. The same grid-search procedure, cross-validation folds, and data partitions were applied to all seven algorithms to ensure a fair comparison, and no algorithm was evaluated with untuned default settings. A fixed random seed was used for all data partitioning and resampling, and identical training and validation splits were shared across algorithms to ensure reproducibility.

Training and evaluation protocol

Data were split into training (80%, $n = 343$) and test (20%, $n = 87$) sets using stratified random sampling to preserve class proportions. All features were centered and scaled prior to model fitting. The centering and scaling parameters were estimated from the training data only and then applied to the test set. Within cross-validation, this standardization, together with hyperparameter tuning, was carried out separately within each training fold, so that model performance was never informed by validation or test observations. Model training employed 5-fold stratified cross-validation on the training set for hyperparameter selection, with the best hyperparameter configuration selected based on cross-validated accuracy. Final model performance was evaluated on the held-out test set.

To assess performance stability, repeated 10×5 -fold stratified cross-validation (50 total folds) was conducted on the full dataset using the optimal hyperparameters identified during initial training. This provides distributions of accuracy, Cohen's kappa, and macro F1-score, enabling assessment of model consistency across different data partitions.

Class imbalance consideration

The dataset exhibited strong class imbalance, with Score 3 representing only 6.7% of observations. The primary analysis was conducted without oversampling, as Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) applied during a sensitivity analysis did not consistently improve test set performance. Specifically, SMOTE-augmented models showed improved Score 3 sensitivity in some cases but at the cost of reduced overall accuracy; for instance, RF accuracy decreased from 72.4% to 69.8% with SMOTE. These results are consistent with prior findings that oversampling methods may not benefit all model architectures equally, particularly when the minority

class is extremely small and substantial overlap exists between adjacent class distributions (Batista et al., 2004; He and Garcia, 2009). Results without SMOTE are reported as the primary analysis.

Statistical analysis and performance metrics

Model performance was assessed using overall accuracy, Cohen's kappa coefficient (κ ; Cohen, 1960), macro-averaged area under the receiver operating characteristic curve (macro AUC), and class-specific sensitivity, specificity, precision, and F1-score. Exact binomial (Clopper-Pearson) 95% confidence intervals were computed for test set accuracy. Bootstrap confidence intervals (2,000 resamples) were computed for Score 3 sensitivity to quantify uncertainty associated with the small sample size. McNemar's test was used for pairwise model comparison on the test set. Model calibration was assessed using multi-class Brier scores and log-loss computed via 5-fold cross-validation (Supplementary Table S2). SHAP (Lundberg and Lee, 2017) values were computed for XGBoost and RF to provide interpretable, class-specific feature importance estimates.

Kruskal-Wallis rank sum tests (Kruskal and Wallis, 1952) were used to assess univariate differences in feature distributions across DCM severity groups. A supplementary binary classification analysis (Normal [Score 1] vs. DCM [Score 2 + 3]) was conducted using repeated 10×5 -fold cross-validation to evaluate diagnostic performance when collapsing severity grades (Supplementary Table S3). An additional supplementary analysis evaluated model performance using only morphometric features and body weight (7 features) to assess the contribution of physiological parameters (Supplementary Table S4). All analyses were implemented in R version 4.3.1 (R Core Team, 2023) using the caret (Kuhn, 2008), pROC (Robin et al., 2011), and

Table 1

Descriptive statistics (mean $\pm$ SD) of 15 features across DCM severity groups in 430 male Cobb 500 broiler chickens.

Feature	Score 1 (n = 246)	Score 2 (n = 155)	Score 3 (n = 29)	P-value
pH	7.475 $\pm$ 0.079	7.440 $\pm$ 0.083	7.427 $\pm$ 0.088	1.42e-05
pO2	32.938 $\pm$ 7.193	28.392 $\pm$ 7.518	29.269 $\pm$ 6.456	7.28e-08
pCO2	38.767 $\pm$ 9.504	42.533 $\pm$ 9.893	43.652 $\pm$ 9.852	1.53e-04
K	5.443 $\pm$ 0.449	5.559 $\pm$ 0.513	5.593 $\pm$ 0.503	0.073
Lac	5.094 $\pm$ 2.491	6.454 $\pm$ 2.883	6.942 $\pm$ 3.189	4.56e-07
Hct	23.447 $\pm$ 2.141	23.355 $\pm$ 2.218	23.276 $\pm$ 1.750	0.856
Gain	2652.5 $\pm$ 222.5	2613.4 $\pm$ 239.9	2637.7 $\pm$ 209.3	0.309
Weight	4011.7 $\pm$ 367.1	4001.3 $\pm$ 291.6	4005.3 $\pm$ 256.3	0.859
RVTVRatio	0.201 $\pm$ 0.031	0.213 $\pm$ 0.036	0.222 $\pm$ 0.031	3.25e-05
w1	1.256 $\pm$ 0.098	1.358 $\pm$ 0.102	1.437 $\pm$ 0.112	3.09e-25
w2	1.021 $\pm$ 0.075	1.055 $\pm$ 0.092	1.144 $\pm$ 0.100	5.30e-10
w3	0.762 $\pm$ 0.085	0.750 $\pm$ 0.081	0.746 $\pm$ 0.118	0.206
w1l1	0.853 $\pm$ 0.092	0.950 $\pm$ 0.104	1.024 $\pm$ 0.101	5.01e-25
w2l1	0.694 $\pm$ 0.078	0.738 $\pm$ 0.089	0.816 $\pm$ 0.093	1.44e-12
w1w2	1.233 $\pm$ 0.090	1.292 $\pm$ 0.107	1.260 $\pm$ 0.099	3.41e-07

P-values from Kruskal-Wallis rank sum test. Morphometric features: w1 = proximal heart width; w2 = mid-ventricle width; w3 = distal width; w1l1 = w1-to-length ratio; w2l1 = w2-to-length ratio; w1w2 = w1-to-w2 ratio. Physiological/production features: pH = blood pH; pO2 = partial pressure of O₂; pCO2 = partial pressure of CO₂; K = potassium; Lac = blood lactate; Hct = hematocrit; Gain = body weight gain; Weight = live body weight; RVTVRatio = right ventricle to total ventricle weight ratio. All morphometric measurements in cm; pO2 and pCO2 in mmHg; K in mmol/L; Lac in mmol/L; Hct in %; Gain and Weight in g.

randomForest (Liaw and Wiener, 2002) packages. Supplementary analyses (repeated CV, SHAP, binary classification, calibration) were conducted in Python 3.11 using scikit-learn (Pedregosa et al., 2011), XGBoost (Chen and Guestrin, 2016), and SHAP (Lundberg and Lee, 2017).

Results

Descriptive statistics

Descriptive statistics and Kruskal-Wallis test results for all 15 features are presented in Table 1. Of the 15 features, 10 showed statistically significant differences across DCM severity groups ($P < 0.05$). Five of the 6 morphometric features (w1, w2, w1l1, w2l1, w1w2) and some physiological blood gas (pH, pO₂, pCO₂, Lac) and production parameters (RVTVRatio) differed significantly among groups, while w3 and 4 other features (K, Hct, Gain, Weight) did not ($P > 0.05$).

Heart width (w1) showed the most pronounced differences across groups, with mean values of 1.256 ± 0.098 cm (Score 1), 1.358 ± 0.102 cm (Score 2), and 1.437 ± 0.112 cm (Score 3) ($P < 0.001$). This increasing trend in heart width with increasing severity is consistent with the progressive ventricular dilation characteristic of DCM.

Model performance

Overall model performance metrics are presented in Table 2 and Fig. 2. Three models (XGBoost, Ridge, and Elastic Net) achieved the highest test accuracy of 74.7% (95% CI: 64.2 to 83.4%; 65 of 87 correct classifications), corresponding to Cohen's kappa values of 0.52, 0.52, and 0.52, respectively. RF, SVM, and Lasso achieved test accuracies of 72.4% (95% CI: 61.8 to 81.5%), with kappa values ranging from 0.47 to 0.50. KNN had the lowest performance (accuracy = 63.2%, 95% CI: 52.2 to 73.3%; $\kappa = 0.33$). McNemar's pairwise tests indicated no statistically significant differences between any model pair (all $P > 0.05$), suggesting comparable predictive performance across algorithms on this dataset.

Cross-validated macro AUC values ranged from 0.76 (KNN) to 0.85 (Elastic Net), indicating moderate to good discriminative ability across all severity classes (Table 2). The regularized logistic regression models (Ridge, Lasso, Elastic Net) achieved AUC values comparable to the tree-based methods despite lower cross-validation accuracies, indicating comparable discriminative ability across models.

Table 2

Overall performance of seven machine learning models for three-class DCM severity classification (test set, $n = 87$).

Model	CV Acc	Test Acc	95% CI	Kappa	Macro AUC	Macro F1
Random Forest	0.774	72.4%	[61.8-81.5]	0.470	0.838	0.624
XGBoost	0.761	74.7%	[64.3-83.4]	0.523	0.850	0.646
SVM	0.684	72.4%	[61.8-81.5]	0.478	0.840	0.616
KNN	0.630	63.2%	[52.2-73.3]	0.333	0.755	0.509
Ridge	0.666	74.7%	[64.3-83.4]	0.523	0.850	0.729
Lasso	0.666	72.4%	[61.8-81.5]	0.498	0.843	0.712
Elastic Net	0.673	74.7%	[64.3-83.4]	0.519	0.852	0.706

CV Acc = 5-fold cross-validation accuracy; 95% CI = Clopper-Pearson exact binomial confidence interval for test accuracy.

Repeated cross-validation

Repeated 10 $\times$ 5-fold stratified cross-validation results are presented in Table 3. Mean accuracies across 50 folds ranged from 67.3% ($\pm 4.3\%$; XGBoost) to 72.8% ($\pm 4.1\%$; SVM).

Standard deviations of accuracy across folds were generally about 4%, indicating moderate but expected variability given the sample size and class imbalance. The 95% percentile intervals for accuracy showed considerable overlap across all models, which aligns with the McNemar's test results that found no significant pairwise differences.

Class-specific performance

Class-specific performance metrics are presented in Table 4. Classification performance varied substantially across severity classes, reflecting the ordinal nature and imbalanced distribution of the outcome variable.

Score 1 (normal) was classified with the highest sensitivity across all models, ranging from 76.0% (KNN) to 86.0% (Elastic Net), with specificity of 67.6 to 78.4%. Score 2 (moderate DCM) showed intermediate sensitivity of 48.4% (KNN) to 71.0% (XGBoost), with specificity consistently above 76.8%. The low Score 2 sensitivity suggests substantial overlap in feature distributions between adjacent severity categories.

Score 3 (severe DCM) showed the most variable performance. Sensitivity was 33.3% for 5 of 7 models (RF, XGBoost, SVM, KNN, Elastic Net), while Lasso achieved 83.3% and Ridge 66.7%. However, these estimates must be interpreted with extreme caution due to the very small test set representation ($n = 6$ for Score 3 in the test set, from a total of 29 in the full dataset). Bootstrap analysis (2,000 resamples) on cross-validated predictions yielded a Score 3 sensitivity of 6.9% (95% CI: 0.0 to 17.2%), confirming the wide uncertainty around these estimates. Specificity for Score 3 was high ($\geq 90.1\%$) across all models, indicating that severe DCM cases, when correctly identified, were rarely confused with less severe categories.

The confusion matrix for XGBoost (Table 5) reveals that misclassifications occurred exclusively between adjacent severity categories. No Score 1 bird was misclassified as Score 3, and no Score 3 bird was misclassified as Score 1. This pattern indicates that while the model struggles with the boundaries between neighboring categories, it reliably separates the extremes of the severity spectrum.

Feature importance

Feature importance rankings from RF (Gini impurity-based) and XGBoost (gain-based) are presented in Table 6 and corroborated by SHAP analysis (Fig. 3). Both methods identified heart width (w1) as the most important predictor (RF: 68.6, XGBoost: 100.0 [normalized]), followed by the width-to-length ratio (w1l1; RF: 50.9, XGBoost: 52.3). Collectively, morphometric features accounted for approximately 62% of total feature importance in both RF and XGBoost, while physiological features contributed approximately 38%. These proportions should be interpreted as approximate rather than exact, model-independent values. The morphometric domain contains several strongly correlated ratios derived from shared base measurements, whereas the physiological domain comprises a larger number of features (nine versus six) both factors influence how importance is distributed between the two groups, as discussed below.

Among physiological features, blood lactate (Lac; RF rank 5, XGBoost rank 5), partial pressure of oxygen (pO₂; RF rank 7, XGBoost rank 7), and RVTV ratio (RVTVRatio; RF rank 9, XGBoost rank 6) were the most informative. SHAP analysis revealed class-specific patterns: w1 showed the strongest positive association with Score 3 classification, while pH and pO₂ had differential effects across severity classes (Fig. 3).

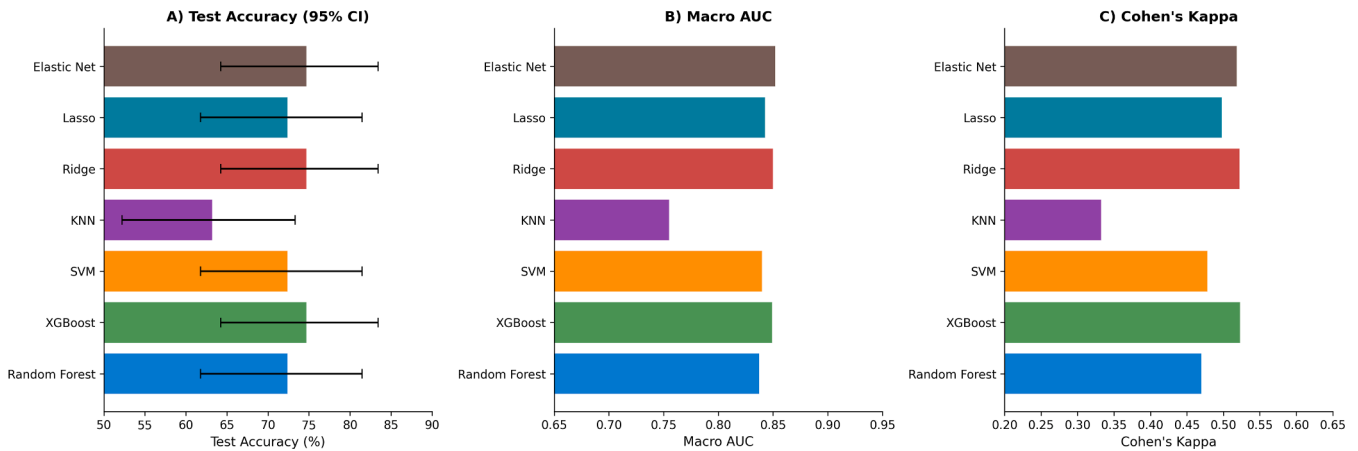


Fig. 2. Model performance comparison for three-class DCM severity classification. (A) Test accuracy with exact binomial 95% confidence intervals, (B) Macro AUC, (C) Cohen's kappa coefficient. $n = 87$ test observations.

Table 3

Repeated 10×5 -fold stratified cross-validation results (50 folds per model).

Model	Mean Acc (%)	SD Acc (%)	95% PI	Mean Kappa	Mean F1
Random Forest	72.1	3.9	[65.1-79.1]	0.448	0.496
XGBoost	67.3	4.3	[57.2-75.3]	0.372	0.490
SVM	72.8	4.1	[64.2-80.9]	0.462	0.486
KNN	67.7	4.2	[61.6-74.4]	0.349	0.445
Ridge	72.0	4.4	[64.0-80.0]	0.463	0.575
Lasso	72.1	4.4	[64.2-80.0]	0.463	0.576
Elastic Net	72.1	4.4	[64.2-80.0]	0.463	0.576

PI = 95% percentile interval from fold-level accuracy distributions.

Table 4

Class-specific performance metrics for seven ML models (test set, $n = 87$).

Model	Class	Sensitivity	Specificity	Precision	F1
Random Forest	Score 1	84.0%	70.3%	79.2%	0.816
Random Forest	Score 2	61.3%	78.6%	61.3%	0.613
Random Forest	Score 3	33.3%	98.8%	66.7%	0.444
XGBoost	Score 1	82.0%	78.4%	83.7%	0.828
XGBoost	Score 2	71.0%	76.8%	62.9%	0.667
XGBoost	Score 3	33.3%	98.8%	66.7%	0.444
SVM	Score 1	82.0%	73.0%	80.4%	0.812
SVM	Score 2	64.5%	78.6%	62.5%	0.635
SVM	Score 3	33.3%	97.5%	50.0%	0.400
KNN	Score 1	76.0%	67.6%	76.0%	0.760
KNN	Score 2	48.4%	78.6%	55.6%	0.517
KNN	Score 3	33.3%	90.1%	20.0%	0.250
Ridge	Score 1	82.0%	70.3%	78.8%	0.804
Ridge	Score 2	64.5%	82.1%	66.7%	0.656
Ridge	Score 3	66.7%	98.8%	80.0%	0.727
Lasso	Score 1	76.0%	73.0%	79.2%	0.775
Lasso	Score 2	64.5%	80.4%	64.5%	0.645
Lasso	Score 3	83.3%	96.3%	62.5%	0.714
Elastic Net	Score 1	86.0%	67.6%	78.2%	0.819
Elastic Net	Score 2	58.1%	85.7%	69.2%	0.632
Elastic Net	Score 3	66.7%	97.5%	66.7%	0.667

Score 3 test set $n = 6$. Estimates for this class carry wide uncertainty.

Calibration

Multi-class Brier scores ranged from 0.129 (SVM) to 0.164 (XGBoost), with log-loss values of 0.651 to 1.009 (Supplementary

Table 5

Confusion matrix for the best-performing model (XGBoost) on the held-out test set ($n = 87$).

Predicted \ Actual	Score 1	Score 2	Score 3
Score 1	41	8	0
Score 2	9	22	4
Score 3	0	1	2

Rows = predicted class; columns = actual class.

Table 6

Feature importance rankings from random forest (Gini impurity) and XGBoost (gain-based, normalized to 100).

Rank	Feature (RF)	Importance (RF)	Feature (XGB)	Importance (XGB)
1	w1	68.6	w1	100.0
2	w1l1	50.9	w1l1	52.3
3	w1w2	35.8	w2	38.4
4	w2l1	30.6	w1w2	30.6
5	Lac	29.0	Lac	27.0
6	w2	28.4	RVTVRatio	24.3
7	pO2	27.3	pO2	23.7
8	pH	22.9	w3	23.5
9	RVTVRatio	22.2	w2l1	21.6
10	pCO2	22.2	pH	21.4
11	w3	21.0	pCO2	16.1
12	K	16.0	Gain	10.4
13	Gain	13.7	Weight	8.2
14	Weight	7.7	K	2.1
15	Hct	7.5	Hct	0.0

RF importance = mean decrease in Gini impurity. XGBoost importance = total gain, normalized so that the top feature = 100.

Table S2). The regularized logistic regression models generally showed the best calibration, consistent with their probabilistic formulation.

Binary classification

A supplementary binary classification (Normal vs. DCM) using repeated 10×5 -fold cross-validation yielded mean accuracies of up to 79.0% with AUC values of up to 0.846 (Supplementary Table S3). The binary formulation improved classification performance compared to three-class models, as expected when collapsing ordinal categories reduces overlap between adjacent classes.

SHAP Feature Importance by DCM Severity Class (XGBoost)

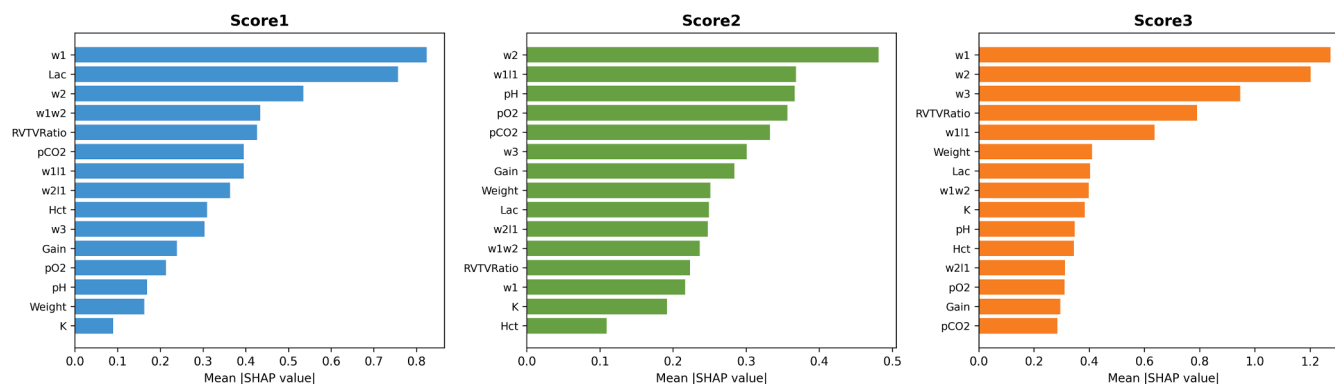


Fig. 3. SHAP (SHapley Additive exPlanations) feature importance by DCM severity class from XGBoost. Bars represent mean absolute SHAP values, indicating the average impact of each feature on model output for each class.

Discussion

This study demonstrated that machine learning algorithms can achieve moderate classification accuracy up to 74.7% for three-class DCM severity assessment in broiler chickens using a combination of cardiac morphometric, physiological and production features. The fact that several different algorithms performed similarly, with no significant pairwise differences according to McNemar's test, suggests that the main limitation on classification accuracy comes from the natural overlap between neighboring severity categories rather than the choice of model. This reflects the fundamental challenge of ordinal classification when class boundaries are subjectively defined (Gutiérrez et al., 2016). Similar accuracy levels have been reported in human cardiomyopathy classification. For example, Accardo et al. (2022) achieved 73.3% accuracy when differentiating ischemic heart disease from DCM using HRV features, and Zhou et al. (2023) reported 75% accuracy with XGBoost applied to echocardiographic data for ischemic cardiomyopathy versus DCM classification. In that study, segmental wall motion anomalies were identified as the most important predictor, which aligns with our finding that heart width was the dominant feature (Zhou et al., 2023). Across these studies, classification accuracy consistently falls within the 70 to 80% range despite differences in species, feature types, and algorithms, suggesting that phenotypic overlap between disease grades limits what can be achieved with gross morphological and physiological features.

It is not surprising that morphometric features, particularly w1 and the w1l1 ratio, were the most important predictors, since DCM involves progressive ventricular dilation that directly affects heart dimensions (Olkowski, 2007). The w1l1 ratio captures the shape change from an elongated normal heart to a more globular dilated morphology, explaining its high discriminative value. This is consistent with the pathological anatomy illustrated by Olkowski et al. (2020), where ventricular dilation in DCM-affected broilers is most pronounced in the proximal region of the heart, which corresponds to the area captured by the w1 measurement. The SHAP analysis further confirmed that w1 contributes the most to classification decisions across all severity classes, with increasing SHAP values corresponding to higher severity scores. It should be emphasized that these importance values reflect each feature's contribution to the model's predictions rather than a causal relationship with disease severity. A high SHAP value for w1 indicates that the model relies heavily on heart width to separate classes, not that heart width causes DCM. Feature importance is therefore interpreted here in a strictly predictive, rather than causal, sense. These findings support the utility of simple cardiac dimensional measurements as objective markers for DCM severity stratification. Furthermore, an important observation from the repeated cross-validation analysis is the discrepancy between

single split and averaged performance estimates. XGBoost achieved the highest test set accuracy (74.7%) but the lowest mean accuracy under repeated 10×5 -fold cross-validation ($67.3 \pm 4.3\%$), on the other hand SVM showed the opposite pattern with the highest repeated CV accuracy ($72.8 \pm 4.1\%$) but a lower test set accuracy of 72.4%. This reversal suggests that the specific 80/20 data partition used for the primary evaluation may have favored certain algorithms over others. Such instability in model ranking across different data splits has been documented in cardiovascular imaging studies, where Singh et al. (2021) demonstrated that single split validation can produce significantly variable AUC estimates depending on the random seed, and recommended repeated stratified cross-validation or bootstrap validation as more reliable alternatives. In the present study, the overlapping 95% percentile intervals across all models under repeated CV, combined with the non-significant McNemar's pairwise comparisons, confirm that no single algorithm can be declared superior for this classification task. We therefore recommend that future studies in this domain prioritize repeated cross-validation performance over single-split estimates when reporting and comparing model performance.

Physiological blood gas parameters presented a meaningful but secondary role (approximately 38% of total importance). Among these, blood lactate concentration emerged as the most informative physiological predictor, which may be attributed to its role as a marker of tissue hypoxia and anaerobic metabolism in cardiac dysfunction (Wideman et al., 2013; Mohammed et al., 2017; Barrett et al., 2021). The contribution of pO2 and pH further reflects the gas exchange impairment associated with congestive heart failure and pulmonary congestion in DCM-affected birds. Although pO2 was a significant predictor overall, its group means did not increase monotonically with severity (Score 1 = 32.9, Score 2 = 28.4, Score 3 = 29.3 mmHg; Table 1), indicating a nonlinear rather than a simple graded association. The w1w2 ratio showed a comparable non-monotonic pattern despite being statistically significant, whereas w3 did not differ significantly across severity groups ($P = 0.21$). These patterns indicate that not every feature contributes a clear directional signal, consistent with the substantial overlap between adjacent severity classes.

The relative importance assigned to each feature domain should be interpreted in context. Variable importance is conditional on the features already in the model: when heart shape measurements are included, they absorb much of the variance that physiological measures would otherwise explain, which reduces the apparent contribution of the latter. This does not mean that physiological parameters lack biological relevance. Their marginal contribution is simply smaller once cardiac geometry is accounted for. It should be noted that this study developed a postmortem phenotypic classification model rather than a genetic evaluation or breeding-value prediction model. Applying these

traits to genetic improvement would require their integration into an appropriate genetic evaluation framework, for example through sib or progeny testing, which is beyond the scope of the present work. The following considerations therefore concern the potential phenotyping role of these traits rather than direct genetic prediction. From a breeding perspective, the practical value of a trait depends not only on its predictive power but also on whether it can be collected at scale on live animals. Heart morphometric measurements require postmortem dissection and are therefore limited to progeny testing or sib testing schemes. If the goal is to develop a screening tool for use in broiler production settings, physiological parameters measurable on live birds may carry greater applied value despite their lower model importance, and this tradeoff warrants further investigation.

To directly test whether cardiac morphometric measurements alone could support severity classification without blood gas data, a supplementary analysis was conducted using only the six morphometric features and body weight (Supplementary Table S4). All seven algorithms showed reduced performance: the best-performing model (XGBoost) achieved 68.6% test accuracy, a 6 percentage-point drop from the full 15-feature model. The remaining models showed declines of 8 to 12 points. More telling was the complete failure to detect Score 3 cases: all seven morphometric-only models produced 0% Score 3 sensitivity on the test set. This result makes clear that while morphometric features carry the strongest individual signal, physiological parameters such as blood lactate and pO₂ are essential for identifying the most severe cases. In practice, both measurement domains are needed for meaningful severity assessment.

The low sensitivity for Score 3 (severe DCM) classification represents the primary limitation of the three-class models. With only 29 severe DCM cases in the full dataset (6.7%) and 6 in the test set, the models had limited exposure to this class during training. The bootstrap analysis confirmed wide uncertainty (95% CI: 0.0 to 17.2%) around Score 3 sensitivity estimates. Notably, SMOTE oversampling did not consistently improve Score 3 detection and in some cases reduced overall accuracy. This outcome may be explained by the fact that SMOTE-generated samples do not necessarily conform to the true minority class distribution, which can degrade rather than improve classification performance (Elreedy et al., 2024). The supplementary binary classification (Normal vs. DCM) demonstrated improved performance when severity grades were collapsed (Supplementary Table S3), supporting the use of simplified classifications in screening applications where detecting the presence of any DCM is prioritized over severity grading.

It should also be noted that the boundary between Score 2 (moderate) and clinically significant cardiac pathology is not sharply defined. Whether moderate ventricular dilation represents early-stage disease or a benign adaptation to rapid growth remains an open question. A reduced dataset comparing only Score 1 and Score 3 birds could help clarify model performance for the most diagnostically distinct groups, though the already limited Score 3 sample size would further constrain such an analysis.

Several methodological considerations should be noted. The ground truth labels were based on postmortem gross cardiac morphology, with hearts classified according to the degree of ventricular dilation and wall thinning, features widely recognized as hallmarks of DCM in broiler chickens (Olkowski, 2007; Olkowski et al., 2020). Severity scores were determined by majority consensus among five trained observers, which reduces but does not eliminate the influence of individual subjectivity on class assignments. Variability in severity assignment, particularly along the boundary between adjacent categories, represents an inherent ceiling on achievable classification accuracy. The use of a single 80/20 train-test split, while standard in many ML studies, may yield performance estimates sensitive to the particular data partition. The repeated 10 × 5-fold cross-validation analysis was conducted to address this concern and confirmed that performance distributions were consistent with the held-out test results. The moderate standard deviations observed across folds (approximately 4% for accuracy) reflect expected

variability for this sample size and highlight the importance of reporting performance distributions rather than single-point estimates.

Some morphometric features were derived as ratios from shared base measurements, resulting in multicollinearity within the feature set. This is a common characteristic of ratio-based morphometric data and does not compromise the ML approaches used here, as tree-based methods partition feature space through sequential splits and their predictive accuracy is not degraded by collinearity, although importance estimates may be redistributed among correlated features (Hastie et al., 2009; Dormann et al., 2013). The consistent feature importance rankings across RF, XGBoost, and SHAP analyses support the robustness of the identified important features despite correlations.

Several further limitations should be acknowledged. First, although 430 birds is a reasonable sample for this type of study, only 29 belonged to the severe class. This, however, was expected since the birds were raised following standard practices, and, even though DCM is an important condition, it is not expected to severely affect a high percentage of the animals. The imbalance could have increased the risk of overfitting for the most flexible algorithms, but overfitting was not observed. This imbalance was mitigated through cross-validation, regularization, and repeated resampling, but cannot be fully eliminated at the present sample size. Second, all models were developed and evaluated on a single dataset from one pure line at one facility, and no external validation on an independent population was performed, so generalizability to other genetic backgrounds, management systems, and production environments remains to be established. Third, the reference labels derive from subjective postmortem scoring; although the majority consensus among five trained observers reduces individual bias, individual observer scores were not retained, so an inter-observer agreement statistic (such as Fleiss' kappa) could not be computed. Future studies should record individual scores to quantify labelling reliability and should pursue external validation across multiple genetic lines and production settings.

Brier scores across models ranged from 0.129 (SVM) to 0.164 (XGBoost), indicating generally comparable but not identical calibration (Supplementary Table S2). The regularized logistic regression models tended to show slightly more favorable calibration, which is expected given their inherently probabilistic formulation. Although these differences are modest in the present study, calibration is an often-overlooked aspect of prediction model evaluation that can have meaningful consequences when models are used for risk-based decision-making (Van Calster et al., 2019).

The present findings have practical implications for poultry breeding and management. Cardiac morphometric measurements were the strongest individual predictors of DCM severity, but the supplementary morphometric-only analysis showed that these measurements alone are insufficient for reliable classification, particularly for severe cases. Both morphometric and physiological features are needed for meaningful severity assessment. The moderate overall accuracy (74.7%) indicates that current ML models should be viewed as complementary tools to, rather than replacements for, expert pathological assessment.

Conclusions

This study demonstrated that machine learning algorithms can provide objective, repeatable classification of DCM severity in commercial broiler chickens, with test accuracies up to 74.7% for three-class classification and up to 79.0% for binary (Normal vs. DCM) classification. Cardiac morphometric features, particularly heart width and the width-to-length ratio, were the most discriminative predictors, contributing approximately 62% of total feature importance. However, a supplementary analysis using only morphometric features and body weight confirmed that physiological blood gas parameters are essential for detecting severe DCM cases, as models without these features showed 0% Score 3 sensitivity. The underrepresentation of severe DCM cases (6.7% of observations) limited sensitivity for this class and represents

the primary target for improvement in future work. Because these models were developed and validated on a single dataset from one genetic line, broader claims regarding their application to commercial poultry production or breeding programs should be regarded as preliminary until confirmed by external validation on independent populations. Expanded severe DCM sampling, integration with advanced imaging modalities, and end-to-end deep learning approaches that classify severity directly from cardiac photographs without manual feature extraction represent promising directions for improving classification performance. These findings establish a foundation for objective DCM phenotyping in broiler chickens and support further development of ML-assisted cardiac health assessment tools for the poultry industry.

CRediT authorship contribution statement

Ümit Bilginer: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Arthur F.A. Fernandes:** Writing – review & editing, Writing – original draft, Validation, Software, Formal analysis, Data curation, Conceptualization. **Brenda Flack:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Nicholas Anthony:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Vivian Likness:** Writing – review & editing, Validation, Supervision, Resources, Formal analysis, Data curation, Conceptualization.

Disclosures

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.

Acknowledgements

The authors thank Cobb-Vantress Inc. for providing the dataset and supporting this research. We also thank the farm personnel who conducted the cardiac evaluations and data collection.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.psj.2026.107473](https://doi.org/10.1016/j.psj.2026.107473).

Data availability

The data that support the findings of this study are available from Cobb-Vantress Inc. upon reasonable request.

References

- Accardo, A., Restivo, L., Ajčević, M., Miladinović, A., Iskra, K., Silveri, G., Sinagra, G., 2022. Toward a diagnostic CART model for ischemic heart disease and idiopathic dilated cardiomyopathy based on heart rate total variability. *Med. Biol. Eng. Comput.* 60, 2655–2663.
- Alimadadi, A., Manandhar, I., Aryal, S., Munroe, P.B., Joe, B., Cheng, X., 2020. Machine learning-based classification and diagnosis of clinical cardiomyopathies. *Physiol. Genomics* 52, 391–400.
- Alves, A.A., Fernandes, A.F., Breen, V., Hawken, R., Rosa, G.J., 2024. Monitoring mortality events in floor-raised broilers using machine learning algorithms trained with feeding behavior time-series data. *Comput. Electron. Agric.* 224, 109124.
- Barrett, S.R.W., Seymour, R.S., 2021. Extreme hypoxia and high lactate concentrations in early chicken embryos show that cutaneous oxygen uptake is limited by diffusion and metabolism is partially anaerobic. *J. Comp. Physiol. B* 191, 1007–1016.
- Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* 6, 20–29.
- Bordbar, F., Guo, W., Wadood, A.A., Nie, Q., 2025. Skeletal muscle myopathies and cardiovascular diseases in poultry: challenges for meat production. *Worlds Poult. Sci. J.* 1–25.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357.
- Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: *Pages 785–794 in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY.
- Chidziwisano, G., Samikwa, E., Daka, C., 2025. Deep learning methods for poultry disease prediction using images. *Comput. Electron. Agric.* 230, 109765.
- Cobb-Vantress Inc., 2021. *Cobb Broiler Management Guide*. Cobb-Vantress Inc, Siloam Springs, AR. Accessed Feb. 2024. <https://www.cobb-vantress.com/>.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.
- Cohn, J.N., Ferrari, R., Sharpe, N., an International Forum on Cardiac Remodeling, 2000. Cardiac remodeling—concepts and clinical implications. *J. Am. Coll. Cardiol.* 35, 569–582.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20, 273–297.
- Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27.
- Crespo, R., Shivaprasad, H.L., 2013. In: Swayne, D.E. (Ed.), *Developmental, Metabolic, and Other Noninfectious Disorders*. Pages 1233–1270 in *Diseases of Poultry*, 13th ed. Wiley-Blackwell, Ames, IA.
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46.
- Elreedy, D., Atiya, A.F., Kamalov, F., 2024. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Mach. Learn.* 113, 4903–4923.
- Gutiérrez, P.A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F., Hervás-Martínez, C., 2016. Ordinal regression methods: survey and experimental study. *IEEE Trans. Knowl. Data Eng.* 28, 127–146.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning*, 2nd ed. Springer, New York, NY.
- He, H., García, E.A., 2009. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284.
- Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- Jiang, T., Gradus, J.L., Rosellini, A.J., 2020. Supervised machine learning: a brief primer. *Behav. Ther.* 51, 675–687.
- Julian, R.J., 2005. Production and growth related disorders and other metabolic diseases of poultry—a review. *Vet. J.* 169, 350–369.
- Kruskal, W.H., Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47, 583–621.
- Kuhn, M., 2008. Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26.
- Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. *Sensors* 18, 2674.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R News* 2, 18–22.
- Lundberg, S. M., and S. I. Lee. 2017. A unified approach to interpreting model predictions. Pages 4766–4777 in *advances in neural information processing systems* 30.
- Machuve, D., Nwankwo, E., Mduma, N., Mbelwa, J., 2022. Poultry diseases diagnostics models using deep learning. *Front. Artif. Intell.* 5, 733345.
- Mohammed, R., Salinas, C.E., Giussani, D.A., Blanco, C.E., Cogolludo, A.L., Villamor, E., 2017. Acute hypoxia-reoxygenation and vascular oxygen sensing in the chicken embryo. *Physiol. Rep.* 5, e13501.
- Okinda, C., Lu, M., Liu, L., Tongse, S., Nyalala, T.M., 2020. A review on computer vision systems in monitoring of poultry: a welfare perspective. *Artif. Intell. Agric.* 4, 184–208.
- Olkowski, A.A., 2007. Pathophysiology of heart failure in broiler chickens: structural, biochemical, and molecular characteristics. *Poult. Sci.* 86, 999–1005.
- Olkowski, A.A., Wojnarowicz, C., Laarveld, B., 2020. Pathophysiology and pathological remodelling associated with dilated cardiomyopathy in broiler chickens predisposed to heart pump failure. *Avian Pathol.* 49, 428–439.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- R Core Team, 2023. *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Muller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* 12, 77.
- Singh, V., Pencina, M., Einstein, A.J., Liang, J.X., Berman, D.S., Slomka, P., 2021. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Sci. Rep.* 11, 14490.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* 58, 267–288.
- Van Calster, B., McLernon, D.J., Van Smeden, M., Wynants, L., Steyerberg, E.W., 2019. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 17, 230.
- Wideman, R.F., Rhoads, D.D., Erf, G.F., Anthony, N.B., 2013. Pulmonary arterial hypertension (ascites syndrome) in broilers: a review. *Poult. Sci.* 92, 64–83.
- Zhou, M., Deng, Y., Liu, Y., Su, X., Zeng, X., 2023. Echocardiography-based machine learning algorithm for distinguishing ischemic cardiomyopathy from dilated cardiomyopathy. *BMC Cardiovasc. Disord.* 23, 476.
- Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* 67, 301–320.