Computers and Electronics in Agriculture 224 (2024) 109124

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Monitoring mortality events in floor-raised broilers using machine learning algorithms trained with feeding behavior time-series data

Anderson A.C. Alves [a], Arthur F.A. Fernandes [b], Vivian Breen [b], Rachel Hawken [b], Guilherme J. M. Rosa [c,*]

[a] *Department of Animal and Dairy Science, University of Georgia, 425 River Rd, Athens, GA 30602, USA*
[b] *Cobb-Vantress Inc., 4703 US Highway 412 East, Siloam Springs 72761, USA*
[c] *Department of Animal and Dairy Sciences, University of Wisconsin-Madison, 1675 Observatory Dr., Madison, WI 53706, USA*

ARTICLE INFO

Keywords:
Animal welfare
Classification algorithms
Mortality rate
RFID

ABSTRACT

In this study, we explored the integration of machine learning (ML) techniques with feeding behavior (FB) time series data to predict mortality events (animals culled or found dead) in floor-raised broilers. Our dataset included 2,667,617 daily observations for eight FB traits from 95,711 birds across 146 feeding trials. After data cleaning, the class distribution was 93.7 % healthy birds and 6.3 % withdrawn birds (culled or found dead), coded as 0 and 1 respectively. Mortality predictions were made one or three days before the observed events. Time series data for different FB traits were utilized to extract 22 time series features per trait, creating a structured feature dataset (days in the feeding trial + 128 time series features). We compared different ML algorithms: gradient boosting machine (GBM), multilayer perceptron neural network (MLP), logistic regression (LR), random forest (RF), and support vector machine (SVM). Due to the imbalanced nature of the data, we evaluated two sampling strategies: a random under-sampling technique (RUS) and a combined strategy (RUS + SMOTE). Models were assessed using 20-fold cross-validation and an independent test set. Statistical tests indicated consistent differences in most FB traits between control and withdrawn birds at least 7 days before the event. Features derived from traits like daily feed intake, number of visited feeders, visiting activity interval, and number of meals presented high predictive importance for mortality monitoring in broilers. In the cross-validation, classifiers achieved an average (standard deviation) of up to 0.87 (0.02) for the area under the ROC curve (AUC) and 0.55 (0.03) for the area under the precision-recall curve (AUPRC). This demonstrated a significant increase in classification performance compared to a no-skill classifier. However, performance dropped notably when extending the prediction window from one to three days in advance. The performance observed in the independent set was similar to that observed during cross-validation, indicating the robustness of our approach. The RUS + SMOTE strategy slightly outperformed RUS across all methods. GBM and SVM algorithms performed best, with no significant differences between them. Additionally, comparable results could be obtained by utilizing a reduced set of features with high predictive importance in comparison with models trained on the full feature set. In summary, Our findings indicate that large-scale feeding behavior data collected from electronic feeders offer valuable insights for predicting illness-related mortality events in floor-raised broilers using machine learning methods. Further research is needed to investigate the feasibility and cost-effectiveness of such monitoring systems in commercial settings.

1. Introduction

In commercial broiler production, the removal of birds from housing pens due to mortality or other illness-related conditions poses a substantial challenge for large-scale systems. These occurrences can be attributed to various factors, including infectious diseases, leg-associated problems, and suboptimal management conditions. Addressing these issues is paramount, as they lead to significant economic losses annually for the poultry industry (Sullivan, 1994; Spackman et al., 2016; Astill et al., 2018).

Preventing losses due to health-related issues requires the adoption of optimized management strategies, which in turn involves the constant

* Corresponding author.
E-mail address: grosa@wisc.edu (G.J.M. Rosa).

https://doi.org/10.1016/j.compag.2024.109124
Received 9 November 2023; Received in revised form 7 May 2024; Accepted 1 June 2024
Available online 8 June 2024
0168-1699/© 2024 Published by Elsevier B.V.

monitoring of the flock's health status for improving disease outbreak detection. Observational data collected from measurement routines and designed experiments provide an important source to understand better the factors underlying broiler chickens' mortality (Fossum et al., 2009; Schwean-Lardner et al., 2013; Zhang et al., 2018). Nonetheless, continuous manual observation of the birds' health by trained personnel becomes unfeasible due to logistical and welfare reasons, as it is both time-consuming and labor-intensive and may cause unnecessary stress to the animals.

The increasing availability of different sensor technologies in livestock farming offers opportunities to build systems for automated and non-invasive surveillance of animal performance, welfare, and health status (Astill et al., 2018; Brito et al., 2020; Ventura et al., 2020; Pérez-Enciso and Steibel, 2021; Rosa, 2021). For instance, in poultry species, audio sensors have been used for the automated detection of respiratory diseases (Carpentier et al., 2019; Cuan et al., 2022). Similarly, data generated by wearable sensors may help monitor the general health status of broiler chickens and other poultry species (Sassi et al., 2016). Additionally, the use of digital image processing has been increasingly suggested as an alternative for health monitoring in the poultry industry (Aydin et al., 2010; Zhuang et al., 2018; Zhuang and Zhang, 2019; Liu et al., 2021). Nevertheless, since commercial poultry species are densely housed in the same pen, large-scale applications of computer vision systems for image segmentation, bird individual identification, and posterior health monitoring are challenging (Zhuang et al., 2018).

The onset of diseases in animals is generally followed by typical behavioral changes in feeding, social interaction, and general activity, a suite of signals commonly termed sickness behavior (Millman, 2007). Hence, tracking subtle alterations in individual behavior and activity patterns may provide useful information to classify the animal health status and for early detection of diseases, although such a task would require means to automate the collection of individual behavioral responses.

Electronic feeders equipped with radio-frequency identification systems are becoming a feasible solution for measuring feed intake and other feeding behavior traits continuously in a large number of group-housed animals (Howie et al., 2011; Mendes et al., 2011; Lu et al., 2017). The adoption of such technology becomes particularly important in the poultry breeding industry as it dismisses the use of cage-based trials for measuring individual feed intake (Alves et al., 2024; Bley and Bessei, 2008; Howie et al., 2011; Yan et al., 2019). Beyond providing useful information on feed efficiency for breeding purposes, the use of electronic feeders unveils opportunities to track patterns in social and feeding behaviors, potentially in near real-time, which may contribute to enhancing farm-level management decisions (Pérez-Enciso and Steibel, 2021).

Different studies have pointed out feeding behavior traits collected in electronic feeding stations as potential health indicators in cattle. Such studies have shown that the onset of different illnesses such as bovine respiratory disease, ketosis, lameness, and other health disorders is accompanied by early changes in feed intake, intake per meal, number of visits, feeding time, and feeding rate (Gonzalez et al., 2008; Wolfger et al., 2015; Sutherland et al., 2017; Duthie et al., 2021). Conversely, limited research has been done to investigate the role of visit-based feeding behavior traits as health status indicators in the poultry industry.

Due to the amount of data potentially generated daily by large poultry flocks monitored with electronic feeders, and the complexity of such information at the individual level, the use of robust analytical tools such as machine learning (ML) methods might contribute to the process of pattern learning. Hence, this study aims to explore strategies to integrate ML methods and feeding behavior patterns measured through electronic feeders for automated and non-invasive prediction of illness-related mortality events in floor-raised birds.

2. Material and methods

2.1. Animals, housing, and data editing

The data used in this study were recorded from 95,711 pure-line broiler chickens of both sexes during 146 consecutive feeding trials of 28-day length, occurring between the years 2017 and 2022. The birds were housed in experimental pens equipped with an electronic feeding system developed by the Cobb Vantress, Inc. (Siloam Springs, AR) engineering team. The facilities were located in Oklahoma, each feeding trial consisted of up to three individual pens (typically of 56 ft length × 13 ft width) per house coupled with at least 6 feeding units containing either 8 or 16 stations (electronic feeders) of 1 ft^2 each and water lines equipped with nipple drinkers. The pens were housed with an average density of 1.5 $birds/ft^2$. All management procedures were performed according to the standard recommendations described in the Cobb Broiler management guide (Cobb-Vantress Inc., 2021).

Feeding behavior was continuously monitored during visits to the feeder using a passive radio-frequency identification (RFID) system. These electronic feeders were equipped with antennas that received the signal of individual low-frequency transponders attached to the birds' wings (Fig. 1). The entrance of the feeder was narrowed to allow individual access to the birds. For each access, the feeders recorded the individual codes of the animal, house, pen, and feeding station, as well as the visit date, time of entrance into and exit from the feeding station, total time spent at the feeder (seconds), and amount of food consumed (g). The visit data were automatically transferred and stored on a local server. Fig. 1 shows a schematic representation of the raw data acquisition process.

In total, 99,472,151 visit logs were obtained throughout the feeding trials, these visit logs were summarized in 2,667,617 daily observations for different feeding behavior traits. The following FB traits were considered as potential health status indicators: daily feed intake (DFI, g/day), daily number of visits (NVIS, n/day), time spent at the feeders (TSF, h/day), number of visited feeders (NVF, n/day), visiting activity interval (VAI, h/day), feeding rate (FR, g/hour), daily number of meals (NMEAL, n/day), average intake per meal (INTMEAL, g/meal), and meal length (MLEN, min). DFI was computed as the sum of all intakes recorded per visit during a day for an animal. For the NVIS, all individual visits were counted, regardless of the intake amount. TSF was computed as the total time the animal spent in the feeders during the day. The NVF considered only the number of unique feeders visited during a day by the same bird (i.e., multiple visits to the same feeder accounted for one single observation or count). The VAI considered the interval between the first and last visits of a bird during the day, whereas FR was calculated as the ratio between DFI and TSF. The NMEAL was computed by clustering visits occurring within a defined meal criterion interval as part of the same meal (Howie et al., 2009). INTMEAL was considered as the total feed intake within each meal, averaged over all meals during the day. Lastly, MLEN was computed as the difference in minutes between the last and first visit within the same meal, averaged over all meals within a day.

The identification of animals that did not complete the feeding tests was annotated by farm personnel and categorized as illness-related mortality events, totaling 8.76 % of all observations in the raw data. The main causes of removing events are birds found dead and welfare culling (V. Breen, Cobb-Vantress, Inc., Siloam Springs, AR, personal communication).

Since the main interest was in the patterns extracted from the time series of the FB traits, only animals that stayed more than 6 days in the feeding trial were maintained in the dataset. Additionally, animals with missing records for any day were removed from the data. After data editing, the overall class distribution was 6.3 % for withdrawn birds (*e.g.*, culled or found dead) and 93.7 % for healthy birds (completed feeding trial). Birds that finished the feeding trial were coded as 0 whereas those withdrawn were coded as 1.

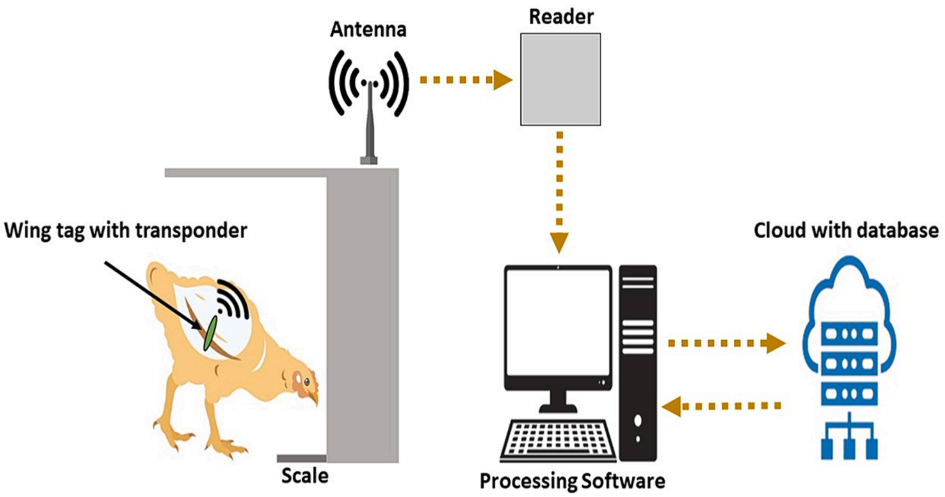

Fig. 1. Schematic representation for the acquisition process of the raw visit data. During the feeder visit the transponder attached to the bird's wing is activated by the antenna and transmits back the data which is decoded by the reader and sent to the local server. The processed information is stored in a cloud database.

2.2. Time-series feature extraction

All FB traits were considered as repeated measures over the feeding trial period, generating a time series for each animal and trait analyzed. As a starting point, the main interest in this work was to assess if feeding behavior could be used to predict mortality events one day in advance of their occurrence. For this purpose, the healthy birds were randomly assigned to control groups, with each control group being associated with a specific day on which mortality events occurred. For example, the control group for the mortality events that occurred on the ninth day of the feeding trial consisted of randomly chosen healthy birds whose feeding behavior was considered only up to their eighth day. In this way, both groups of healthy (control group) and withdrawn birds had time series of similar length. We also repeated the process described above considering a three-day prediction window. The R package *tsfeatures* (Hyndman et al., 2022) and other general-use R functions (R Core Team, 2022) were used for extracting different features from the individual time series (TS). A set of 22 features was generated per trait, totaling 198 features per animal. Additionally, the time series length (***day_ID***) for each animal was added as a numeric covariable (ranging from 6 to 27) in all classification methods.

Shortly, the following time series features were extracted: ***f.mean*** (the time-series global average); ***f.sd*** (the time-series global standard deviation); ***f.range*** (the overall time-series range); ***mean_diff*** (the difference between the averages of the first and second halves of the time series); ***linearity*** and ***curvature*** (linear and quadratic coefficients of the time series orthogonal quadratic regression); ***entropy*** (the time series spectral entropy); ***trev_num*** (the numerator of a normalized nonlinear autocorrelation function); ***nonlinearity*** (a coefficient based on a modification of the Terasvirta's nonlinearity test; Terasvirta *et al.*, 1993; Hyndman et al., 2022); ***ncross*** (measures how often a time series cross its median line); ***npeak*** (the number of peaks of a time series); ***flat_spots*** (computed by dividing the sample space of a time series into equal-sized intervals, and computing the number of datapoints for which the time-series maintains the values at the same level); ***motiftwo*** (returns an entropy of words in the binary alphabet built for the time series); ***embed_incircle*** (the proportion of points inside a given circular boundary space); ***trend*** (the strength of a time series trend); ***spike*** (computed as the variance of leave-one-out variances in the time series); ***std1st_der*** (the standard deviation of the time series first-derivative); ***acf1, acf2,*** and ***acf3*** (the first, second and third autocorrelation coefficients of the time series); ***nacf*** (the index for the first negative autocorrelation coefficient in the time series); ***e_acf1*** (the first autocorrelation coefficient of the time series decomposition residual).

It must be highlighted that the same time-series feature was extracted according to the different feeding behavior traits studied. For instance, the feature *acf1* (first autocorrelation coefficient) was extracted for all feeding behavior time series, generating nine different values (i.e., *acf1_DFI, acf1_NVIS, acf1_TSF, acf1_NVF, acf1_VAI, acf1_FR, acf1_NMEAL, acf1_INTMEAL, acf1_MLEN*). Table 1 shows the average values for the 22 time-series features extracted according to the feeding behavior traits studied. Providing a comprehensive theoretical foundation for these computed time-series features is beyond the scope of this study, for more information, the interested reader is referred to Wang et al. (2006), Fulcher (2017), Kang et al. (2017), and references therein.

2.3. Machine learning methods

We compared the classification performance of five different Machine Learning (ML) methods: logistic regression (LR; Walker and Duncan, 1967), gradient boosting machine (GBM; Friedman, 2001), multilayer perceptron neural network (MLP; Haykin, 1998), random forest (RF; Breiman, 2001), and support vector machine (SVM; Vapnik, 1995). All methods were fitted using the scikit-learn library (Pedregosa et al., 2011), available for Python 3 (Van Rossum and Drake, 2009). The hyperparameter fine-tuning process of all ML methods considered was performed using a custom genetic algorithm (GA) implementation (https://github.com/alvesand/pyga) written in Python programming language (Python Software Foundation, 2022). A brief description of each investigated method is provided below.

In the LR method, the optimization problem can be written in matrix form as follows:

$$\text{argmin} - \frac{1}{n}[\mathbf{y}\ln p(\mathbf{y}|\mathbf{X}) + (1_{\text{nx1}} - \mathbf{y})\ln(1_{\text{nx1}} - p(\mathbf{y}|\mathbf{X}))]^{T}1_{\text{nx1}} + \alpha\|\beta\|_2^2 \tag{1}$$

where n represents the number of training observations; $p(\mathbf{y}|\mathbf{X}) = \frac{1}{1+e^{-(\mathbf{X}\beta)}}$ is an n x 1 vector of probability estimates for positive cases according to the logistic function, with **X** representing an $n \times m$ matrix of time series features, and β representing a $m \times 1$ vector of linear regression coefficients; 1_{nx1} is a vector of ones of dimension $n \times 1$; $\alpha\|\beta\|_2^2$ is the *L2* regularization norm, with α representing a user-defined constant that controls the penalty magnitude.

For the MLP we used a single-hidden layer architecture trained with backpropagation and stochastic gradient descent algorithms. The general classifier has the following form:

$$p(\mathbf{y}|\mathbf{X}) = f_o\left\{\left[g^{(h)}(\mathbf{X}\mathbf{W}^{(h)} + \mathbf{B}^{(h)})\right]\mathbf{w}_o + \mathbf{b}_o\right\} \tag{2}$$

Table 1

Average values for features extracted from time series of different feeding behavior traits (daily feed intake – DFI, number of visits –NVIS, time spent at the feeder – TSF, number of visited feeders – NVF, visiting activity interval – VAI, feeding rate – FR, number of meals – NMEAL, intake per meal – INTMEAL, and length of meal – MLEN) measured in group-housed broilers using electronic feeders.

Time-series features[1]	**Feeding Behavior Traits**[2]								
	DFI (g)	**NVIS (count)**	**TSF (min)**	**NVF (count)**	**VAI (h)**	**FR (g/h)**	**NMEAL (count)**	**INTMEAL (g)**	**MLEN (min)**
f.mean	176.98	41.79	1.62	17.29	19.39	126.02	13.48	15.17	8.07
f.sd	38.84	13.81	0.43	5.56	3.01	27.53	3.68	5.33	2.87
f.range	145.11	47.59	1.56	19.17	11.22	97.59	12.87	18.54	10.25
mean_diff	36.54	−5.92	0.19	−3.44	0.68	11.96	−2.19	6.23	2.40
linearity	95.98	−15.03	0.49	−8.59	2.21	31.85	−5.25	16.37	6.17
curvature	−34.72	−6.68	−0.34	−1.16	−4.40	−0.32	−2.25	0.85	0.51
entropy	0.067	0.061	0.046	0.041	0.22	0.049	0.014	0.066	0.062
trev_num	0.014	0.009	0.015	0.005	0.049	0.0025	0.021	0.005	0.001
nonlinearity	3.35	NA*	3.33	NA*	3.39	3.44	NA*	3.82	3.74
ncross	5.19	5.05	5.96	5.65	7.78	5.35	5.36	4.23	4.97
npeak	5.28	5.03	5.26	5.19	5.48	5.27	4.90	5.48	5.47
flat_spots	2.53	2.68	2.31	2.38	3.66	2.40	2.44	2.77	2.66
motiftwo	1.58	1.59	1.68	1.67	1.46	1.64	1.65	1.52	1.59
embed_incircle	0.27	0.44	0.36	0.46	0.067	0.45	0.39	0.50	0.51
trend	0.62	0.54	0.49	0.46	0.38	0.53	0.44	0.65	0.57
spike	3.5e-05	6.1e-05	5.2e-05	7.1e-05	6.9e-05	5.7e-05	7.5e-05	4.8e-05	5.4e-05
std1st_der	0.26	0.30	0.31	0.33	0.324	0.30	0.33	0.27	0.29
acf1	0.32	0.36	0.26	0.28	0.004	0.34	0.28	0.46	0.38
acf2	0.21	0.18	0.12	0.12	−0.028	0.15	0.11	0.30	0.21
acf3	0.12	0.073	0.033	0.038	−0.0438	0.041	0.033	0.18	0.099
nacf	5.84	5.07	4.55	4.72	2.77	4.87	4.74	6.25	5.43
e_acf1	−0.18	−0.11	−0.15	−0.15	−0.19	−0.10	−0.12	−0.17	−0.14

[1] *f.mean*, *f.sd* and *f.range*: the time-series average, standard deviation and range; *mean_diff*: the difference between the averages of the first and second halves of the time series; *linearity* and *curvature*: linear and quadratic coefficients of the orthogonal regression; *entropy*: the spectral entropy; *trev_num*: the numerator of a normalized nonlinear autocorrelation function; *ncross:* how often a time series cross its median line; *npeak*: the number of peaks; *flat_spots*: the maximum run length across equal-sized intervals; *motiftwo*: an entropy of words in the binary alphabet built for the time series; *embed_incircle*: the proportion of points inside a given circular boundary space; *trend*: the strength of a time series trend; *spike*: the variance of leave-one-out variances in the time series; *std1st_der*: the standard deviation of the time series first-derivative; *acf1*, *acf2*, and *acf3*: the first, second and third autocorrelation coefficients of the time series; *nacf*: the index for the first negative autocorrelation coefficient in the time series; *e_acf1*: the first autocorrelation coefficient of the time series decomposition residual; *information excluded due to lack of variability.

in which $\mathbf{W}^{(h)}$ is the hidden-layer weight matrix with dimension $m \times k$ (where m represents the number of time series features and k stands for the number of hidden neurons), $\mathbf{B}^{(h)}$ is a $n \times k$ matrix of neuron biases, $g^{(h)}$ is the hidden-layer activation function, $\mathbf{w}_o$ and $\mathbf{b}_o$ are $k \times 1$ vector of weights and biases in the output layer, f_o is the logistic function that activates the raw outputs to lie between 0 and 1. The weight gradients in Equation (2) are updated by minimizing the cross-entropy loss function (having a similar form as described in Equation (1). We used our GA implementation to find the MLP hyperparameters with the best classification performance in the training set. The following intervals were assumed for the hyperparameters: *learning rate* $\in$ [0.001, 0.1], $k \in$ [16, 264], $\alpha \in$ [0.0001, 0.1], *batch size* $\in$ [16, 264], $g^{(h)} =$ {*logistic, tanh, relu, identity*}.

The RF is an ensemble learning algorithm that fits several decision trees (DT) classifiers on different bootstrap subsets of the training dataset and combines all DT predictions to improve its overall classification performance. This technique is termed bootstrap aggregating or bagging for short. In the RF, each DT node is built using a randomly selected subset of the feature space, which helps to decorrelate the tree ensemble. The RF algorithm grows every DT according to recursive binary splitting rules, by selecting the best feature and cut-off threshold for each node according to a given loss criterion until achieving homogeneous or near homogeneous classes in the terminal nodes. Each DT classifies unobserved data by attributing the classes with the greatest frequency in the terminal leaves. During the GA-based fine-tuning, the following intervals were assumed for the RF hyperparameters: *n_estimators* $\in$ [50, 500], *max_features* $\in$ [5,36], *min_samples_split* $\in$ [0.01, 0.15], *min_samples_leaf* $\in$ [0.01, 0.15], *max_depth* = {None, 2, 5, 10, 15, 20, 25, 30}, and *criterion* = {*gini, entropy*}. A detailed description of these hyperparameters can be found in the scikit-learn online documentation (Scikit-learn Developers, 2007).

Similar to RF, GBM also uses an ensemble of decision trees as weak learners to build a model with better predictive performance. Nonetheless, instead of building the trees independently, GBM uses a boosting technique for sequentially growing the trees by improving the prediction or classification performance of previous trees (James et al., 2013). At each iteration, the GBM algorithm fits a DT iteratively using the residuals of the previous model as the response. For binary classification problems, the objective is to minimize the log-loss function for updating the gradients. At every iteration, the predictions are updated by the following rule:

$$\widehat{f}^{t}(\mathrm{X}) = \widehat{f}^{t-1}(\mathrm{X}) + \lambda\psi^{t}(e_{b-1}; x^{t}) \tag{3}$$

where t represents the current iteration (for $t = 1, 2, \ldots$ T), $\psi^{t}(\mathbf{e}_{b-1}; x^{t})$ is the prediction based on the current base learner $\psi^{t}(.)$, $\mathbf{e}_{b-1}$ is a vector of residuals from the previous learner, x^{t} is a subset of the features input matrix, and $\lambda \in (0,1)$ is some shrinkage factor. GBM's final predictions are given by the weighted sum of the predictions of all base learners in the ensemble. The hyperparameters and intervals considered during the fine-tuning optimization were: *learning_rate* $\in$ [0.001, 0.15], *n_estimators* $\in$ [50, 200], max_features $\in$ [5,36], min_samples_split $\in$ [0.01, 0.15], min_samples_leaf $\in$ [0.01, 0.15], max_depth = {None, 2, 5, 10, 15, 20, 25, 30}, and criterion = {friedman_mse, squared_error}.

The SVM algorithm optimizes a constrained maximum margin problem by projecting the dataset features into a higher dimensional space, in which complex non-linear patterns present in the dataset might be linearly separable in the feature space mapping. In the dual formulation, the SVM can be optimized by maximizing the following Lagrangian function:

$$\widetilde{L}(a) = -\frac{1}{2}\sum_{i}\sum_{j} a_i a_j y_i y_j \mathrm{K}\langle \mathbf{x}_i \mathbf{x}_j \rangle + \sum_{i} a_i \tag{4}$$

subject to $\sum_i a_i y_i = 0$ and $0 \le a_i \le C$. In Eq. (4), a_i and a_j are Lagrange multipliers associated with the observations i and j, C is a positive regularization parameter and $K\langle \mathbf{x}_i\mathbf{x}_j \rangle$ is the Kernel function that defines the inner product in the feature space. The SVM hyperparameters were optimized within the following intervals: $C \in [0.1, 3]$, $gamma \in [0.001, 0.5]$, and $kernel = \{rbf, linear, poly, sigmoid\}$.

2.4. Sampling techniques

Since monitoring mortality events is an imbalanced classification problem, we assessed the impact of two data sampling techniques for balancing the label distributions during the training of the classification algorithms. These sampling approaches prevent the ML methods from putting too much weight on the majority class during the learning process, leading to misleading classifications in the validation and testing sets. The first sampling strategy investigated was the random under-sampling (RUS) method. For the RUS, we down-sampled the majority class by randomly selecting a subset of examples from it until achieving a nearly even class distribution in the training set (as reviewed by Ali et al., 2019).

In the second sampling strategy, we used a combination of under-sampling and oversampling techniques. Specifically, we first used RUS to down-sample the majority class until achieving a near 0.7/0.3 ratio (70 % of healthy birds and 30 % of withdrawn birds) in the training set. This process was followed by a full oversampling of the minority class with the synthetic minority oversampling technique (SMOTE; Chawla et al., 2022) until achieving a 0.5/0.5 ratio. This strategy is hereinafter termed RUS + SMOTE. Please note that the sampling strategies discussed here are applied only to the data used for training the models, with the actual class distribution being preserved in the validation and testing sets.

2.5. Validation strategies, feature selection, and comparison metrics

The full dataset comprised information on birds from 32 consecutive generations (G1 to G32). Based on this information, we divided the data into training (animals from G1 to G30, N = 66,866) and testing sets (animals from G31 to G32, N = 7011). During hyperparameter fine-tuning, we further divided the training set into two groups. Approximately 67 % of the training data was used for fitting the methods considering different hyperparameter combinations, while the remaining ~33 % of observations were used to monitor the methods' classification performance within the genetic algorithm (GA) implementation. The best hyperparameter configuration was defined according to the method and sampling strategy (Random Under Sampling – RUS or RUS combined with Synthetic Minority Over-sampling Technique – SMOTE).

After hyperparameter fine-tuning, we performed a 20-fold cross-validation (CV) scheme in the training set to check for the generalization capability of the final tuned models and to allow fairer comparisons across them. Note that for every iteration in the 20-fold CV, the sampling process was repeated, generating a training set according to the fold and observations sampled with RUS and RUS + SMOTE techniques. Furthermore, for each iteration in the cross-validation process we stored the features' predictive importance according to the regression coefficients of the Logistic Regression (LR) model and the impurity-based importance estimated with the tree-based approaches (Random Forest – RF and Gradient Boosting Machine – GBM). The impurity-based importance is computed by minimizing a splitting criterion (Gini impurity in RF and loss function in GBM) that reflects how well observations are separated in the output space. This impurity reduction process is then averaged across all trees for each feature (Zhang and Ma, 2012).

During the cross-validation, 20 variable importance (VIM) scores were generated per feature according to the three different models (LR, RF, and GBM). We then averaged the VIM scores to rank and select the top 20 % features with the highest predictive importance according to the sampling technique (RUS and RUS + SMOTE) and the feature selection method (LR, RF, and GBM). This information was used to retrain all classification algorithms (LR, MLP, GBM, RF, SVM) according to six reduced training datasets generated by the combination of the sampling techniques and feature selection methods. Finally, all models trained with the full (all features) and reduced (top 20 % features) training sets were evaluated according to their classification performance in the independent testing set (animals from G31 to G32). The objectives of this procedure were twofold: a) to assess if the results found with the full variable set in the CV scheme would replicate well in an independent dataset that did not contribute to the model's training and fine-tuning processes (i.e., check for possible model overfitting), and b) to avoid over-optimistic results due to the inclusion of privileged information for the models trained with the feature selection step. Fig. 2 summarizes in a flowchart all the main steps taken for data preparation and analysis and how they interact.

The classifiers' performance was assessed based on different metrics that take into account the incidence of the four possible prediction outcomes: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The metrics used for comparing the classification methods were $specificity = TN/(TN+FP)$, $sensitivity = TP/(TP+FN)$, $precision = TP/(TP+FP)$, and $F1 = 2 \times \frac{precision \times sensitivity}{precision + sensitivity}$. We also compared the models based on the area under the Receiver Operating Characteristic (ROC) curve (AUC), and the area under the precision-recall curve (AUPRC). In this study, sensitivity is also referred to as recall or true positive rate (TPR) interchangeably.

2.6. Statistical analysis

For comparing the FB averages between the two groups (alive and withdrawn), we used a matched pair design using information from the training data set. For each withdrawn bird, we randomly selected a healthy bird and assigned it to a control group. Only the FB data until one day before the mortality event were retained for animals from the control group, according to their matched pairs in the withdrawn group. We analyzed this dataset using a multivariate linear model for testing the effect of health status for each FB trait on a particular day before the mortality events. The model included the effects of health status (healthy or withdrawn), sex, hatch, and contemporary groups (birds raised in the same house and pen, during a specific feeding trial). The *p-values* computed according to the day were used to identify the relative period that each trait became consistently divergent before the mortality event (i.e., the *p-values* were always smaller than 0.05 from that day on).

T-statistics and their *p-values* were computed to test the overall effect of the sampling strategy in the classification metrics assessed by the cross-validation scheme, assuming the following model:

$$m_{ijkl} = \mu + ML_i + ST_j + Fold_k + e_{ijkl},$$

where m_{ijkl} is the value observed for the metric (specificity, sensitivity, precision, F1, AUC, or AUPRC) considering the classifier i, sampling strategy j, validation fold k, and randomly resampled training subset l, ML_i is the effect of the machine learning algorithm (GBM, LR, MLP, RF, and SVM), ST_j is the effect of the sampling technique (RUS or RUS + SMOTE), $Fold_k$ is the effect of the k^{th} validation fold ($k = 1, 2, \ldots, 20$), and e_{ijkl} is the random residual term. Additionally, multiple pairwise t-tests were performed to compare the classification performance across methods within the sampling strategies.

3. Results

3.1. Preliminary analysis

As an illustrative example, Fig. 3 shows the average trend of different feeding behavior (FB) traits according to the survival status on the 21st day of the feeding tests (control group 21). Visual inspection of the line

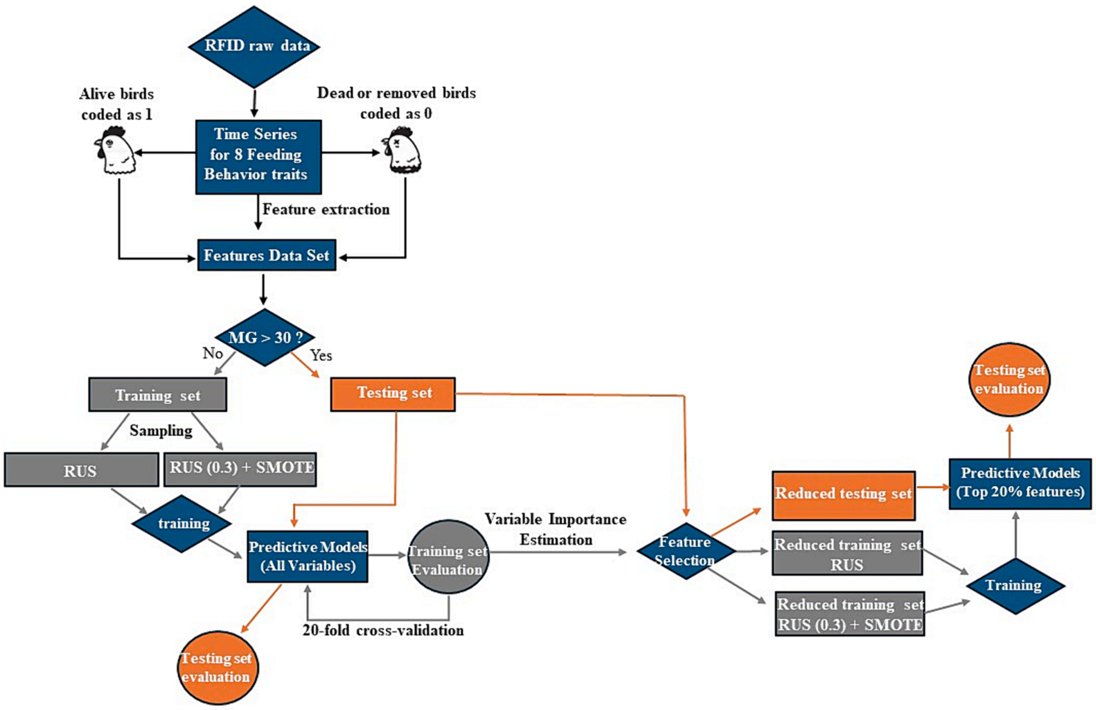

Fig. 2. Schematic representation of the main data preparation and analysis steps performed to assess the performance of different machine learning models for monitoring mortality in broiler chickens based on time series features extracted for different feeding behavior traits measured with a radio-frequency identification system. MG stands for mating generation. RUS is the random under-sampling technique. RUS + SMOTE is a combination of the RUS and synthetic minority oversampling technique (SMOTE).

charts indicates that mortality occurrences on day 21 were associated to some extent with prior changes in FB (Fig. 3). At the population level, the *p-values* for the statistical tests suggested that the average for the FB traits became consistently divergent a few days relative to the mortality events; more specifically, 7 days for TSF (*p-value* $\leq$ 0.011), 8 days for MLEN (*p-value* $\leq$ 0.0078), 10 days for INTMEAL, NV, and FR (*p-values* $\leq$ 0.0126, 0.009, and 0.0078, respectively), 15 days for VAI (*p-value* $\leq$ 0.0055), and 17 days for DFI and NVF (*p-values* $\leq$ 0.0028 and 0.01). Therefore, statistical tests indicated consistent differences in most FB traits between healthy and withdrawn birds at least 7 days before the event. On average, the sick birds were observed to feed less, reduced their feeding rate, visited fewer feeders, presented less visiting activity, and had a smaller intake per meal as the mortality event approached (Fig. 3). Nonetheless, shaded colors in Fig. 3 show wide and overlapping variation for FB trends according to the group (healthy or withdrawn), suggesting that monitoring the mortality risk at the individual level based on FB is a challenging task.

3.2. Cross-validation results

Table 2 shows the classification performance (assessed in a 20-fold cross-validation scheme) of different ML algorithms for predicting one day in advance the illness-related mortality events in floor-raised broilers based on feeding behavior traits. The classification ability depended heavily on the classifier (GBM, LR, MLP, RF, and SVM) and the sampling strategy (RUS or RUS + SMOTE) adopted, with averages (standard deviation) ranging from 0.84 (0.10) to 0.95 (0.00) for specificity, from 0.64 (0.03) to 0.74 (0.03) for sensitivity, between 0.27 (0.02) and 0.47 (0.02) for precision, and from 0.38 (0.10) to 0.54 (0.02) for F1 score (Table 2).

The combination of under and oversampling strategies (RUS + SMOTE) delivered slightly better performance than the RUS training strategy across the classification methods, considering the specificity (*p-value* $= 0.001$), precision (*p-value* $= 1.8 \times 10^{-6}$), and F1 (*p-value* $= 4.19 \times 10^{-7}$), and slightly worse sensitivity (*p-value* $= 0.0004$). The GBM algorithm, combined with the RUS + SMOTE strategy, achieved the best overall performance, with the highest averages for specificity, precision, and F1 score, although this model/sampling strategy combination had the lowest sensitivity average (Table 2). As expected, a sensitivity-precision trade-off across methods was observed. In other words, model/sampling strategy combinations that improved precision typically presented lower sensitivity and vice versa. As suggested by the F1 scores, the best balance between false positives and false negatives was achieved by the GBM, followed by the SVM algorithm, both trained under the RUS + SMOTE strategy (Table 2).

We assessed the AUC and AUPRC metrics to investigate which model/training strategy aggregated the highest performance considering the trade-off between true and false positive rates and between precision and sensitivity (recall) at different threshold values used for assigning the predicted probability values as positive classes. Fig. 4 depicts the boxplots for AUC and AUPRC values assessed in a 20-fold cross-validation according to the classification method and the sampling strategy, considering the predictions one day in advance.

The average (standard deviation) for the AUC values ranged between 0.846 (0.02) and 0.871 (0.02), values considered far superior to the performance expected for a random classifier (0.50). Conversely, the AUPRC ranged from 0.495 (0.03) to 0.549 (0.03) according to the classifier/sampling strategy combination, while an AUPRC of around 0.063 (6.3 % of positive outcomes) would be expected for a random estimator. When we increased the prediction window to three days before the observed events, there was observed a performance reduction for all classification models (Fig. 5). Still, the observed values for AUC were consistently higher than the performance expected for a random classifier, with averages (standard deviation) ranging between 0.77 (0.03) and 0.80 (0.02). Similarly, the observed values for the AUPRC lay consistently above the 0.06 value expected for a random estimator (Fig. 5).

Overall, visual inspection of the boxplots suggests no major

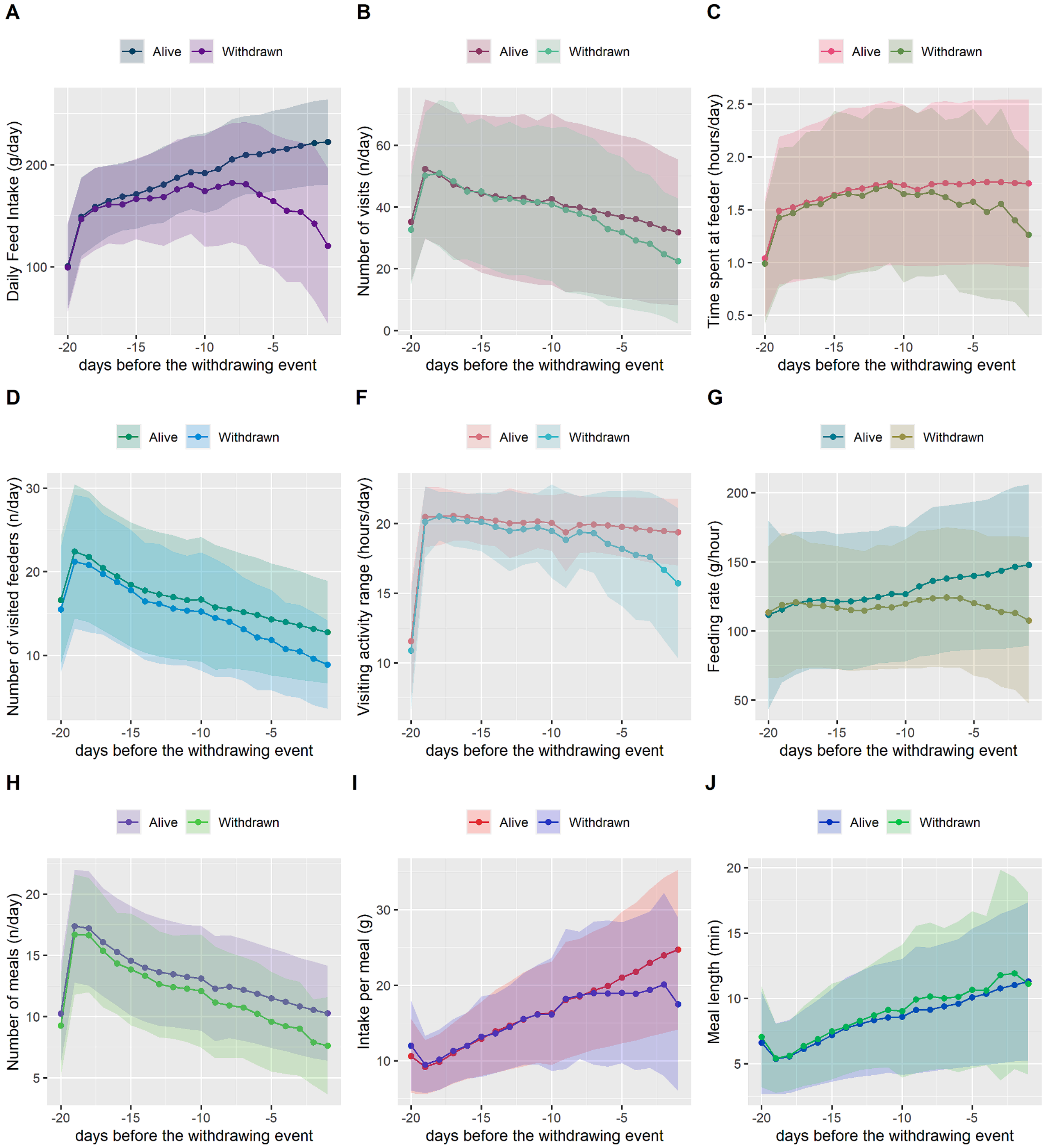

Fig. 3. Time series of different feeding behavior traits according to the permanence status on the 21st day of the feeding trial. The line charts show the average data from 20 days to 1 day relative to the event of interest. Shaded colors represent the observations' standard deviation for each day during the feeding test interval.

differences in the AUC and AUPRC values between the sampling techniques within the same classification model (Figs. 4 and 5). Nevertheless, there was a small advantage of the RUS + SMOTE over the RUS strategy across models considering the AUPRC metric (*p-value* < 0.01) in both prediction intervals (1 or 3 days in advance). This advantage can be better observed for GBM, considering the three-day prediction window (Fig. 5). Regardless of the sampling technique employed, the classification algorithms with the highest AUC values were the GBM and SVM, with no statistical evidence of performance differences observed between them (Figs. 4 and 5). The same pattern was observed for AUPRC, with these methods presenting very similar performance in both training strategies. In turn, the smallest AUC and AUPRC were achieved by the MLP, regardless of the sampling strategy (RUS or RUS + SMOTE) and prediction interval (1 or 3 days in advance).

3.3. *Performance on the testing data based on the full-feature set*

Fig. 6 shows the ROC and precision-recall (PR) curves along with the

Table 2
Average (standard deviation) for classification metrics observed in a 20-cross validation scheme performed to assess the performance of different machine learning methods trained for predicting illness-related mortality events (death or welfare culling) in broiler chickens based on time series features extracted from feeding behavior trends measured with a real-time radio-frequency system. Predictions were performed one day in advance of the observed events.

Algorithm[1]	Sampling[2]	Specificity	Sensitivity	Precision	F1
GBM	RUS	0.879 (0.01)	**0.745** (0.03)	0.288 (0.02)	0.415 (0.02)
	RUS + SMOTE	**0.952** **(0.00)**	0.643 (0.03)	**0.469** (0.02)	**0.542** (0.02)
LR	RUS	0.871 (0.01)	0.727 (0.03)	0.270 (0.02)	0.393 (0.02)
	RUS + SMOTE	0.876 (0.01)	0.728 (0.03)	0.277 (0.02)	0.401 (0.02)
MLP	RUS	0.842 (0.10)	0.707 (0.11)	0.292 (0.15)	0.380 (0.10)
	RUS + SMOTE	0.844 (0.10)	0.722 (0.08)	0.295 (0.11)	0.389 (0.10)
RF	RUS	0.893 (0.01)	0.720 (0.03)	0.306 (0.02)	0.429 (0.02)
	RUS + SMOTE	0.921 (0.01)	0.676 (0.03)	0.358 (0.02)	0.467 (0.02)
SVM	RUS	0.929 (0.01)	0.688 (0.04)	0.389 (0.02)	0.496 (0.03)
	RUS + SMOTE	0.932 (0.00)	0.691 (0.04)	0.399 (0.03)	0.505 (0.02)

[1] GBM: Gradient Boosting Machine; MLP: Multilayer Perceptron; NB: Naïve Bayes Classifier; RF: Random Forest; SVM: Support Vector Machine.

[2] RUS: random under-sampling on the training set; RUS + SMOTE: the majority class is under-sampled randomly until the ratio (0.70:0.30) followed by a full oversampling of the minority class with the SMOTE (synthetic minority oversampling Technique).

respective areas under these curves (AUC and AUPRC) for the independent data according to the classification method and training strategy, considering the one-day prediction interval and all available features for training the classifiers. The ROC and PR curves highlight the overall good performance of the classification models compared to a no-skill classifier (red lines) at distinguishing between withdrawn and healthy animals. For instance, Fig. 6A suggests that the SVM could identify within one day in advance 75 % of all animals that were removed from the feeding tests (TPR) at the expense of classifying erroneously 12.5 % of healthy animals as positive cases (FPR). By setting different thresholds for the models, it would be possible to increase the TPR beyond 75 % at an approximately linear cost for the FPR (Fig. 6A and 6B). Conversely, this relatively high TPR (75 %) would produce low precision values of around 30 % at the best scenarios, which roughly implies that for every 3 animals classified as non-healthy, two would be false positives (Fig. 6C and 6D). Fortunately, PR curves (Fig. 6C and 6D) clearly show that it is possible to adjust the models' threshold to achieve much higher precision at a cost of reducing (at a slower pace) the classification sensitivity (Recall).

AUC values observed in the testing set ranged from 0.84 to 0.87, whereas the interval observed for AUPRC lay between 0.51 and 0.60 (Fig. 6). Overall, these values are within the intervals spanned by the cross-validation scheme (Fig. 4). Nonetheless, unlike the CV-based results, the AUC differences observed across models in the testing set were much smaller, with the SVM achieving the best overall performance in both training strategies. Furthermore, SVM also aggregated the highest AUPRC regardless of the sampling strategy (0.60 and 0.59), followed by GBM (0.58) and RF (0.57), trained with RUS + SMOTE and RUS strategies, respectively (Fig. 6C and 6D). As observed in the 20-fold CV, the AUPRC values obtained in the independent data were far superior to that expected for a random classifier (0.08 for the independent dataset). In general, these results provide evidence for the absence of overfitting problems in all classifiers.

3.4. Variable importance and feature selection

Impurity-based variable importance (VI) computed with the GBM and RF, as well as the LR coefficients were used to rank the time series features based on their predictive contribution. The 15 top-ranked features according to the method (GBM, RF, or LR) and sampling strategy (RUS or RUS + SMOTE) are presented in Fig. 7. Four specific DFI-derived features (*linearity*, *mean_diff*, *f.mean,* and *curvature*) had the highest VI in GBM and RF, irrespective of the sampling strategy used (Fig. 7A, 7B, 7C, and 7D). Additionally, other time series features played relatively higher or lower importance in the tree-based methods (GBM and RF) depending on the sampling strategy considered. This is the case for features derived for VAI (*e.g., linearity, f.mean, nafc, embed_incircle, mean_difference*), DFI (*e.g., trev_num, ncross*), INTMEAL (*e.g., trev_num, ncross,* and *trend*), and NMEAL (*flat_spots*). In both sampling strategies,

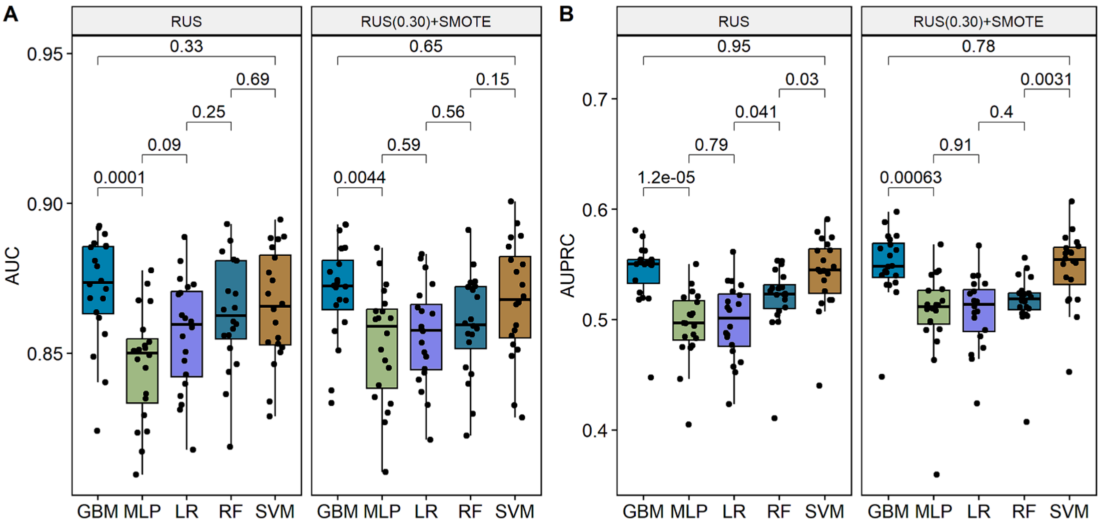

Fig. 4. Classification performance of different machine learning methods in predicting illness-related mortality events in broiler chickens one day in advance. The features for prediction were extracted from feeding behavior trends. The classification performance was assessed using the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPRC) in a 20-fold cross-validation scheme. Two different resampling techniques were employed during algorithm training: random under-sampling (RUS) and a combination of RUS with the synthetic minority oversampling technique (SMOTE). P-values were obtained by conducting multiple pairwise t-tests across classification methods within the sampling strategies.

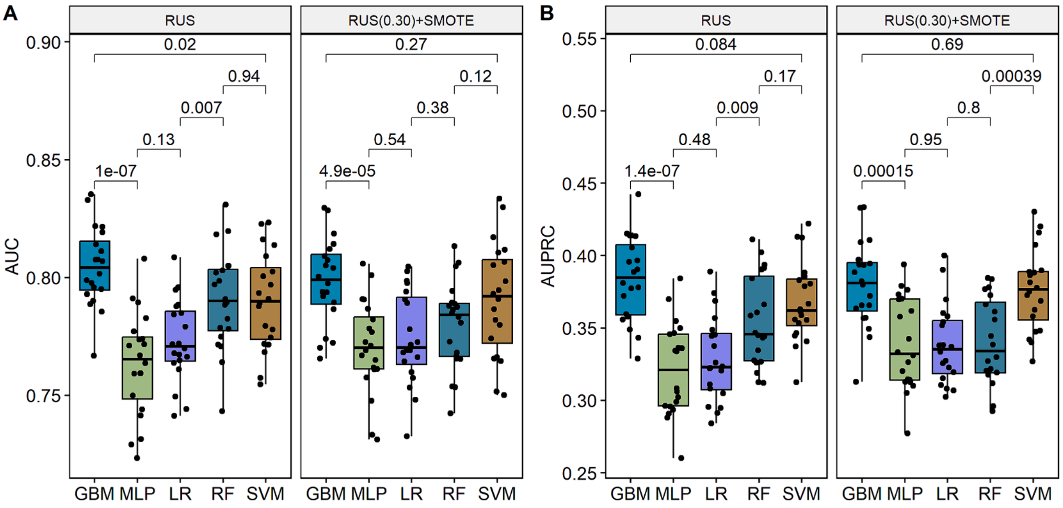

Fig. 5. Classification performance of different machine learning methods in predicting illness-related mortality events in broiler chickens three days in advance. The features for prediction were extracted from feeding behavior trends. The classification performance was assessed using the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPRC) in a 20-fold cross-validation scheme. Two different resampling techniques were employed during algorithm training: random under-sampling (RUS) and a combination of RUS with the synthetic minority oversampling technique (SMOTE). P-values were obtained by conducting multiple pairwise t-tests across classification methods within the sampling strategies.

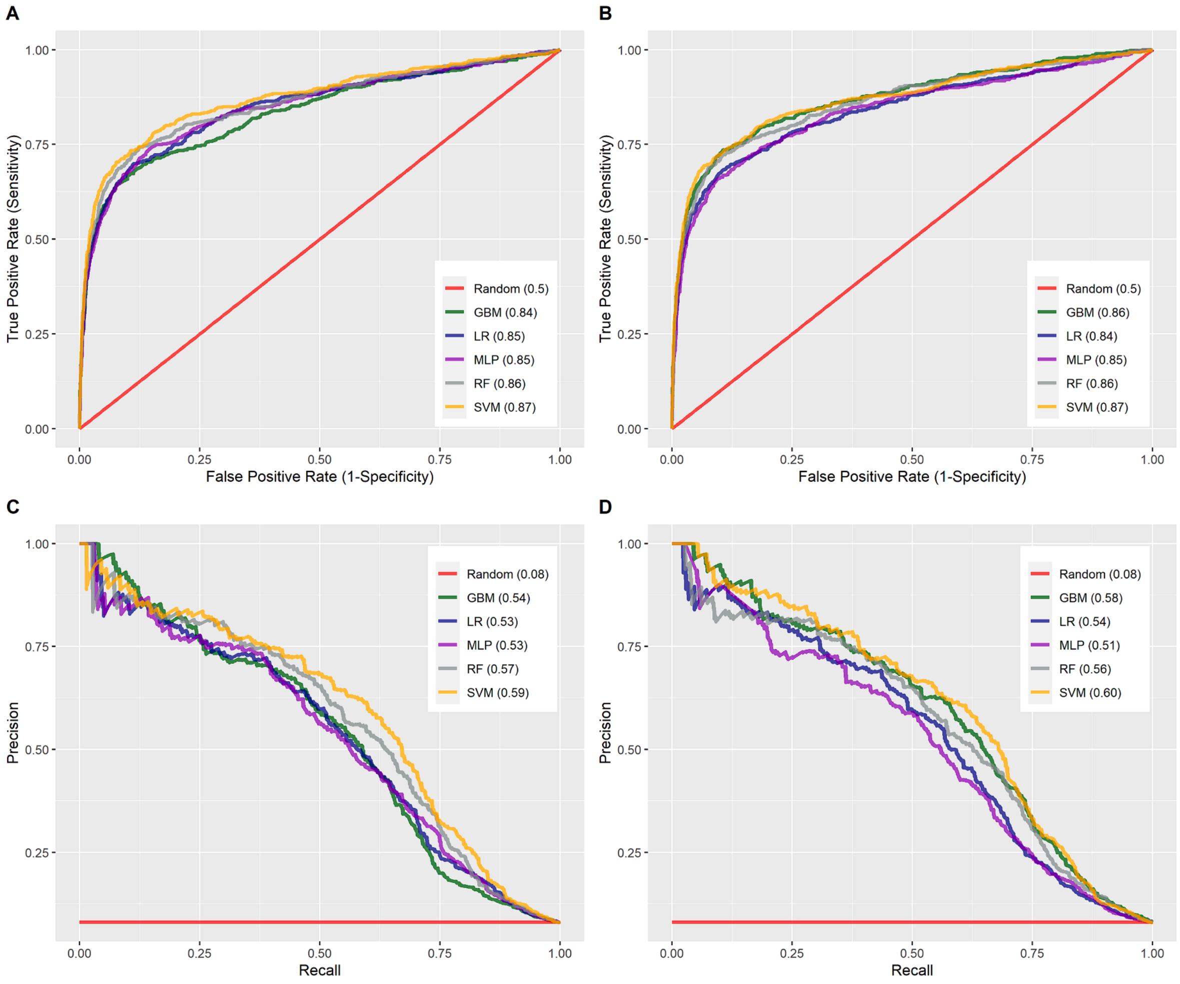

Fig. 6. Performance evaluation of machine learning algorithms for one-day advance prediction of mortality events (animal found dead or welfare culling) in the testing set. The Area under the ROC (Receiver Operating Characteristic) curves (**A** and **B**) and Precision-Recall curves (**C** and **D**) are presented. These algorithms were trained with features extracted from feeding behavior trends measured in broilers. Subplots in the left column depict the performance of models trained with a random under-sampling (RUS) technique. Right column subplots reflect the results obtained for models trained using a combination of under-sampling (RUS) and oversampling (synthetic minority oversampling technique – SMOTE) strategies. Red lines represent the performance obtained for a random classification.

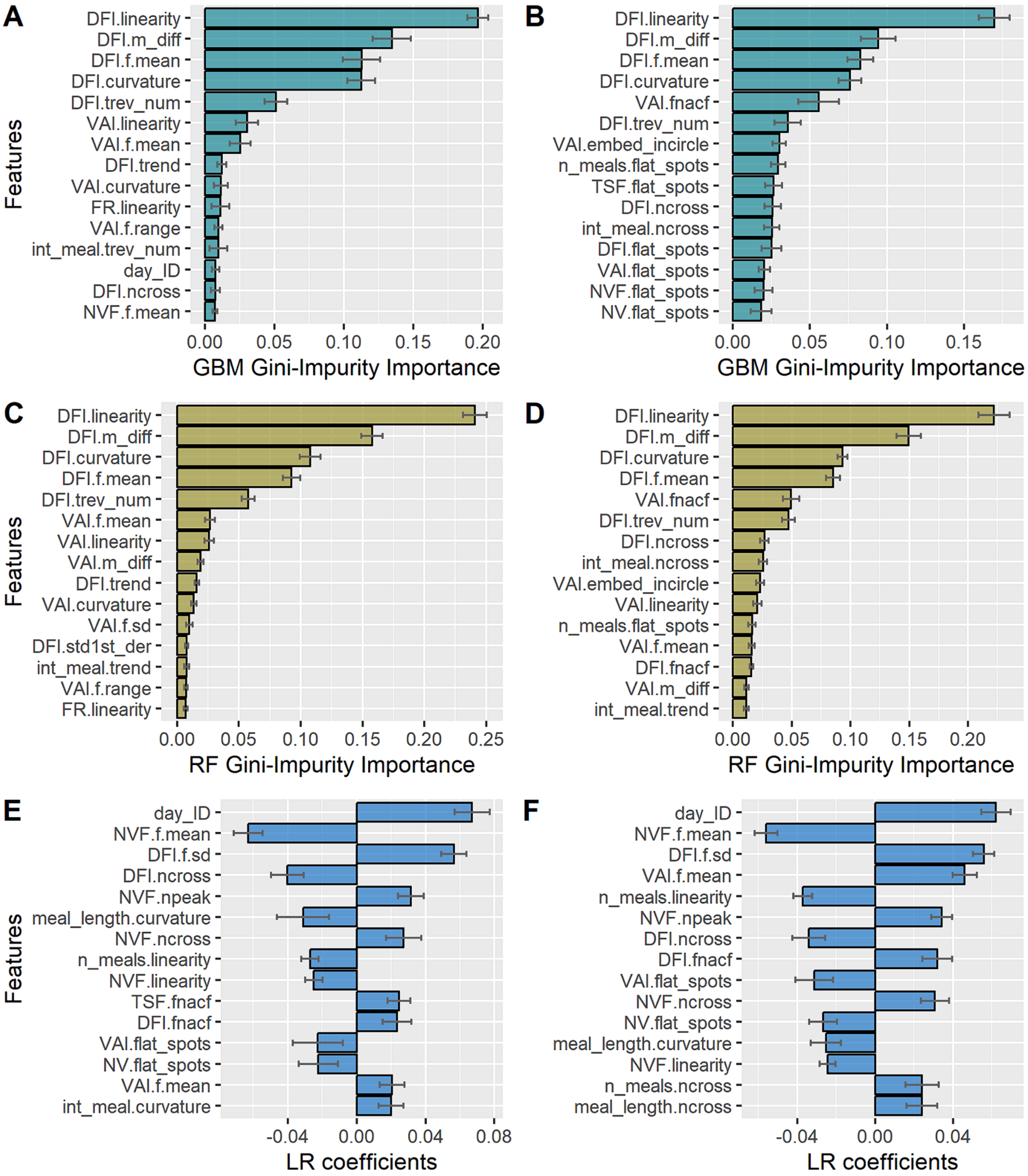

Fig. 7. Top-ranked features used to classify illness-related mortality events in broiler chickens according to the trained model (Gradient Boosting Machine – **A** and **B**; Random Forest – **C** and **D**, and Logistic Regression – **E** and **F**) and sampling techniques (left and right columns). Subplots in the left column depict the variable importance (VI) scores and LR coefficients computed with models trained using a random under-sampling (RUS) technique. Right column subplots reflect the results obtained for models trained using a combination of under-sampling (RUS) and oversampling (synthetic minority oversampling technique – SMOTE) strategies. Error bars represent the standard deviation obtained for the impurity-based scores and LR coefficients across a 20-fold cross-validation.

the LR assigned the heaviest weights to *day_ID*, NVF's *f.mean*, and DFI's *f.sd*, following this order (Fig. 7E and 7F). The LR also assigned high importance to *linearity* (for NMEAL, NVF, and DFI), *flat_spots* (for VAI and NV), *npeak* (for NVF, MLEN, INTMEAL, and DFI), *ncross* (for NVF, MLEN, DFI), among others.

The average values for the impurity-based scores (GBM and RF) and regression coefficients (LR) were used to select the features with the top 20 % highest predictive importance. The percentage of variables in common between feature sets selected with the GBM and RF methods was 84.61 % and 76.92 % for the RUS and RUS + SMOTE strategies, respectively. These percentages were very low when comparing the features selected with impurity-based scores (GBM or RF) and the logistic regression coefficients, ranging between 30.77 % and 33.33 % depending on the sampling technique.

Fig. 8 shows the classification performance in terms of AUPRC achieved by all algorithms in the testing data according to the different feature sets (All variables, Top20%_GBM, Top20%_RF, and Top20%_LR) selected for the two sampling strategies in the two prediction intervals investigated (1 or 3 days before the mortality events). In general, the MLP and RF methods benefited the most from the feature selection performed with impurity-based scores (20 %_GBM and 20 %_RF), especially considering the predictions performed 1 day before the mortality events (Fig. 8A). Conversely, the feature selection based on the logistic regression coefficients considerably worsened the performance of all models compared with the full variable set (Fig. 8A and 8B).

For the RUS strategy, the highest AUPRC values were achieved with the MLP × 20 %_RF (0.60) and GBM × All (0.44) combinations for predictions performed 1 and 3 days before the mortality events,

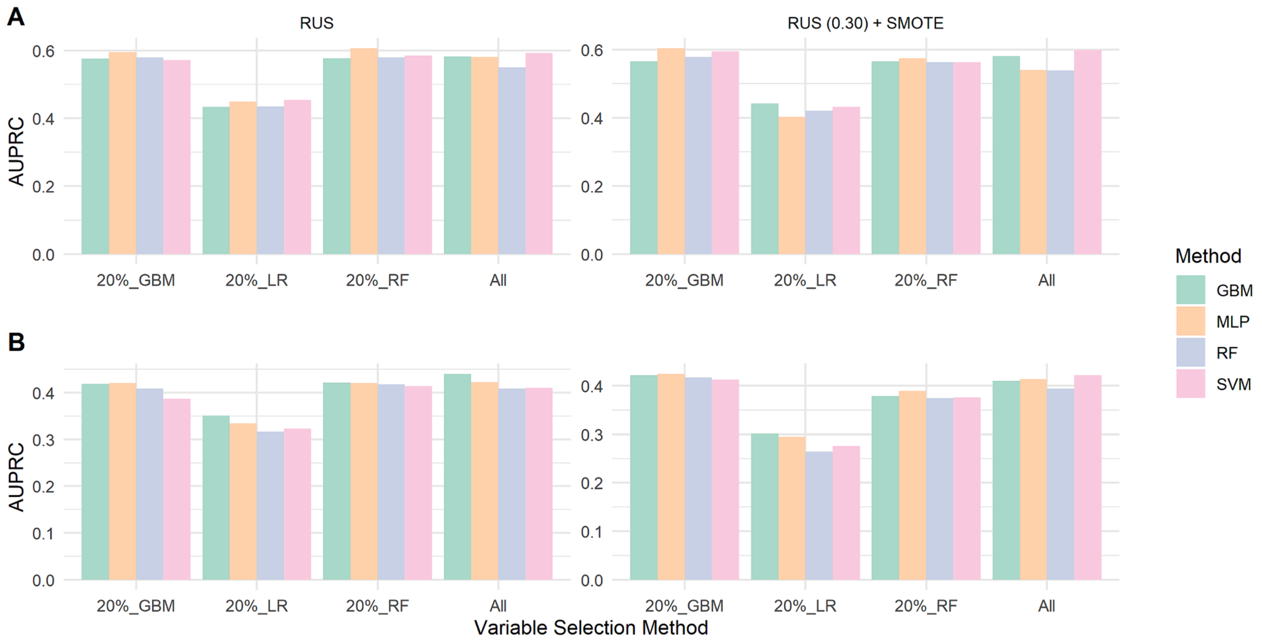

Fig. 8. Barplots for the area under the precision-recall curve (AUPRC) achieved by machine learning algorithms according to the feature selection method (All, Top20%_GBM, Top20%_RF, and Top20%_LR) and sampling strategies (RUS and RUS + SMOTE) adopted to predict mortality in broiler chickens based on time-series features extracted from feeding behavior traits. Predictions were performed considering intervals of 1 day (**A**) or 3 days (**B**) before the mortality events. The performance of models trained with the full variable set (All) is compared with models using feature subsets containing the top 20 % variables with the highest predictive importance, ranked according to different methods (GBM, RF, and LR). The sampling techniques considered were random under-sampling (RUS) and a combination of RUS and synthetic minority oversampling technique (RUS + SMOTE) strategies.

respectively (Fig. 8). In turn, for the RUS + SMOTE strategy, the highest AUPRC values were obtained with both MLP × 20 %_GBM (0.60) and SVM × All (0.60) for the 1-day prediction interval (Fig. 8A), while the GBM × All (0.44) performed the best for predictions 3 days before the mortality events (Fig. 8B). Considering both AUC and AUPRC values, the SVM trained with all data and using the RUS + SMOTE sampling strategy yielded the best overall performance (AUC = 0.87, AUPRC = 0.60) for the one-day prediction interval. For three-day predictions, the GBM using all data combined with the RUS strategy performed best (AUC = 0.80, AUPRC = 0.44). Notably, similar performance could be achieved using only a subset of features, as indicated by the performances achieved with the SVM × 20 %_GBM × RUS + SMOTE combination (AUC = 0.87 and AUPRC = 0.59).

4. Discussion

This study explored the use of electronic feeders equipped with radio-frequency identification systems as a means of non-invasively monitoring illness-related mortality events in poultry production systems. We hypothesized that subtle changes in FB patterns measured in floor-raised broilers could be associated with the individual health status of these birds. For most of the FB traits studied, the statistical tests suggested a consistent divergence between the averages of the two groups (control and withdrawn birds) at least 7 days before the event of interest. Nonetheless, we noted a wide and overlapping variation among the target classes for the FB daily observations, which highlights the challenges of monitoring mortality risk at the individual level based on the feeding behavior observed for specific days.

To overcome these challenges, we extracted different time series features from the FB trends aiming to capture patterns occurring throughout time intervals of the feeding trials. These TS features were used as input information in different ML methods to predict within one day in advance the risk of an animal being removed from the feeding test due to illness-related issues. Due to the imbalanced nature of this classification problem, we also investigate the impact of two sampling strategies used for training the classification algorithms, namely RUS and RUS + SMOTE. The tested classifiers achieved averages (standard deviation) of up to 0.87 (0.02) and 0.55 (0.03) for AUC and AUPRC in the 20-fold cross-validation scheme, which indicates a substantial increase in the classification performance compared to a no-skill classifier. Furthermore, similar values were achieved in the independent set, highlighting the good generalization capability of our approach. Results suggested that the combination of under and oversampling strategies (RUS + SMOTE) delivered slightly better performance than the random under-sampling (RUS) strategy across the classification methods. Overall, the GBM and SVM algorithms achieved the best performance.

Results presented in this study provide important insights into the feasibility of implementing automated data-driven systems for monitoring in near real-time the individual health status of floor-raised broilers. Our findings suggest that the high-throughput measurement of FB through electronic feeders could be a valuable tool in building such systems. We have shown that this information can be combined with efficient classification algorithms to monitor individual health status non-invasively in poultry systems. A significant increase in the number of individuals predicted to be at high risk of illness-related mortality events could signal management failures or the onset of contagious diseases in specific pens. This information could be used by farmers and veterinarians to implement targeted intervention strategies to prevent disease outbreaks across different pens. The information on the expected health status of birds the next day could also guide the suitability of performing stressful management interventions in that specific pen. Additionally, the automated and non-invasive surveillance of animal health could help to improve the general welfare at the individual level by reducing the number of potentially stressful situations generated by unnecessary interventions in the pens (Sassi et al., 2016; Winckler, 2019).

We have shown that there is a trade-off between precision and sensitivity when using FB to predict illness-related withdrawal events as methods that improved precision typically presented lower sensitivity and vice versa. These findings have important practical implications at the farm level. Precision refers to the proportion of TP among all positive predictions (i.e., TP + FP), while recall (also called sensitivity) refers to the proportion of TP among all actual positive cases (i.e., TP + FN) (Saito and Rehmsmeier, 2015). In the context of broiler production, the most important consideration may depend on the specific goals of the monitoring system. If the goal is to identify as many sick birds as possible, then a higher recall rate may be more important than precision. On the other hand, if the goal is to reduce false positives and minimize unnecessary interventions, then higher precision may be preferred. Therefore, the threshold for a classifier to assign animals as positive cases must be adjusted accordingly. For instance, Fig. 6D suggests that it would be possible for the SVM to achieve a sensitivity of 0.67 with a precision of 0.5 in the independent set by increasing the model threshold over 0.5. In other words, this means that 67 % of all birds removed from the pens due to illness-related issues would be identified by the model, with one out of every two animals predicted as positive outcomes being indeed withdrawn from the feeding trial the next day due to illness-related issues. These results are encouraging given the low incidence of the target class, which challenges the development of prediction algorithms with both high precision and sensitivity.

Despite the promising results, this study also has some gaps that need to be addressed further in future research. While achieving good classification performance within one day in advance is important for monitoring purposes, increasing this interval window would be certainly advantageous for preventing disease outbreaks more effectively. Our results revealed that the classification performance dropped considerably when the prediction window was increased to 3 days, although the models still performed better than a random classifier (Fig. 5). This trend is expected to replicate for even bigger prediction windows (*e.g.*, for 5 or 7 days in advance) relative to the day of observed events. Additionally, because we only had available the information on which animals were withdrawn from the feeding test, our target class reflected a general outcome resulting from different illness-related issues. It might be beneficial to develop models trained with more detailed labels to improve both the classification performance and management decisions to prevent specific diseases (Cuan et al., 2022).

Preventing the spread of diseases and improving animal welfare in commercial poultry systems have been a relevant concern for the industry and consumers for years (Hofacre, 2002; Capua and Marangon, 2006; Erian and Phillips, 2017; Hafez and Attia, 2020). Recent advancements in sensor technology and data processing techniques have made possible an unprecedented advance in large-scale identification and prevention of diseases as well as monitoring animal welfare (Brito et al., 2020). Therefore, implementing these high-throughput phenotyping technologies in commercial production systems holds immense potential to benefit the poultry industry (Li et al., 2020). Unsurprisingly, there has been an increasing effort from different research groups to integrate sensor data for disease surveillance and welfare monitoring in poultry species. Some examples include the use of sound data for the identification of respiratory diseases (Carpentier et al., 2019; Cuan et al., 2022), tracking systems based on wearable sensors (Banerjee et al., 2014; Shahbazi et al., 2023), and image processing for health and behavior monitoring (Zhuang et al., 2018; Zhuang and Zhang, 2019; Liu et al., 2021). Ideally, gathering as much information as possible from different data sources could benefit the development of IoT-based intelligent systems for early disease detection in poultry species (Singh et al., 2020; Ahmed et al., 2021). Here, we provided evidence that feeding behavior assessed through RFID systems comprises a useful piece of data for integrating such systems. To the best of our knowledge, this study comprises one of the first efforts toward this direction.

Our results are in line with what has been found previously in cattle by different studies, which provided evidence that abnormal changes in feeding behavior are potentially associated with the onset of several diseases in this species (Gonzalez et al., 2008; Wolfger et al., 2015; Sutherland et al., 2017; Duthie et al., 2021). For instance, it has been shown that a decrease in the mean meal intake, mealtime, and frequency of meals was associated with increased hazard for bovine respiratory disease in mixed-breed steers up to 7 days before clinical symptoms were noticed by the feedlot staff (Wolfger et al., 2015). Similarly, calves presenting sub-clinical or clinical symptoms for respiratory diseases decreased their feeding time and had fewer feeder visits compared to the healthy group (Duthie et. al., 2021). The results of this study suggest that TS features derived from traits such as DFI, NVF, VAI, and NMEAL present high predictive importance for health monitoring in broilers. Furthermore, our results indicate that a similar performance could be achieved by considering only a subset of features based on these traits in comparison with models trained with the full feature set, this result may have important implications for the scalability of this monitoring system. It is broadly documented that sickness behavior is characterized by lethargy, anorexia, and depression in animals and humans (Hart, 1988; Tizard, 2008). Hence, this evidence supports our findings that abnormal changes in traits related to feeding motivation, frequency of feeding bouts, and overall activity of the birds are associated with good health deprival.

Despite the evidence of an association between disease onset and FB, only a few studies have assessed the predictive usefulness of this information with formal validation schemes. Belaid et al. (2019) reported specificity and sensitivity values of 42 % and 92 % in the testing data for the early classification of sick bulls using a monitoring system that combined activity and FB information generated with accelerometer sensors and feed bunks equipped with antenna systems, respectively. The specificity value found by these authors is lower than the values found in our study. Nonetheless, Belaid et al. (2019) used a prediction window of 9 days before the clinical signals, which may have contributed to the high false positive rate (inversely proportional to the specificity) reported in their study.

In summary, the findings of this study provide evidence that FB traits measured through RFID technology are useful for predicting mortality risk in floor-raised broilers. This information could be combined with state-of-the-art technologies (*e.g.*, high-throughput genotyping) in the poultry industry to build automated data-driven systems for monitoring the individual health status of floor-raised broilers. While the presented results are encouraging, there is room for further improvement. Future research addressing the different challenges and limitations discussed in this paper is encouraged.

5. Conclusion

According to our findings, large-scale feeding behavior data measured with electronic feeders comprise valuable information to predict illness-related mortality events in floor-raised broilers using machine learning methods. Our results suggest that GBM and SVM algorithms achieved the best overall performance for this task. Further research is needed to investigate the generalizability of the findings to other populations (*e.g.* other genetic lines) and to test the feasibility and cost-effectiveness of implementing such monitoring systems in commercial settings.

CRediT authorship contribution statement

Anderson A.C. Alves: Conceptualization, Data curation, Formal analysis, Software, Validation, Writing – original draft, Writing – review & editing, Methodology. **Arthur F.A. Fernandes:** Data curation, Methodology, Supervision, Writing – review & editing. **Vivian Breen:** Data curation, Project administration, Supervision, Writing – review & editing. **Rachel Hawken:** Data curation, Project administration, Supervision, Writing – review & editing. **Guilherme J.M. Rosa:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank Cobb Vantress for providing the dataset used in this study.

References

Ahmed, G., Malick, R.A.S., Akhunzada, A., Zahid, S., Sagri, M.R., Gani, A., 2021. An approach towards IoT-based predictive service for early detection of diseases in poultry chickens. Sustainability 13, 13396. https://doi.org/10.3390/su132313396.

Ali, H., Salleh, M., Hussain, K., Ullah, A., Ahmad, A., Naseem, R., 2019. A review on data preprocessing methods for class imbalance problem. IJET 8 (3), 390–397. https://doi.org/10.14419/ijet.v8i3.29508.

Alves, A.A.C., Fernandes, A.F.A., Lopes, F.B., Breen, V., Hawken, R., Rosa, G.J.M., 2024. Genetic analysis of feed efficiency and novel feeding behavior traits measured in group-housed broilers using electronic feeders. Poult. Sci. 103 (7), 103737 https://doi.org/10.1016/j.psj.2024.103737.

Astill, J., Dara, R.A., Fraser, E.D.G., Sharif, S., 2018. Detecting and predicting emerging disease in poultry with the implementation of new technologies and big data: a focus on avian influenza virus. Front. Vet. Sci. 5, 263. https://doi.org/10.3389/fvets.2018.00263.

Aydin, A., Cangar, O., Ozcan, E.S., Bahr, C., Berckmans, D., 2010. Application of a fully automatic analysis tool to assess the activity of broiler chickens with different gait scores. Comput. Electron. Agric. 73, 194–199. https://doi.org/10.1016/j.compag.2010.05.004.

Banerjee, D., Daigle, C.L., Dong, B., Wurtz, K., Newberry, R.C., Siegford, J.M., Biswas, S., 2014. Detection of jumping and landing force in laying hens using wireless wearable sensors. Poult. Sci. 93, 2724–2733. https://doi.org/10.3382/ps.2014-04006.

Belaid, M.A., Rodriguez-Prado, M., Chevaux, E., Calsamiglia, S., 2019. The use of an activity monitoring system for the early detection of health disorders in young bulls. Animals 9, 924. https://doi.org/10.3390/ani9110924.

Bley, T.A.G., Bessei, W., 2008. Recording of individual feed intake and feeding behavior of pekin ducks kept in groups. Poult. Sci. 87, 215–221. https://doi.org/10.3382/ps.2006-00446.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brito, L.F., Oliveira, H.R., McConn, B.R., Schinckel, A.P., Arrazola, A., Marchant-Forde, J. N., Johnson, J.S., 2020. Large-scale phenotyping of livestock welfare in commercial production systems: a new frontier in animal breeding. Front. Genet. 11, 793. https://doi.org/10.3389/fgene.2020.00793.

Capua, I., Marangon, S., 2006. Control of avian influenza in poultry. Emerg. Infect. Dis. 12, 1319–1324. https://doi.org/10.3201/eid1209.060430.

Carpentier, L., Vranken, E., Berckmans, D., Paeshuyse, J., Norton, T., 2019. Development of sound-based poultry health monitoring tool for automated sneeze detection. Comput Electron Agric. 162, 573–581. https://doi.org/10.1016/j.compag.2019.05.013.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2022. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Cobb-Vantress Inc., 2021. Cobb Broiler Management Guide. Accessed Feb. 2024. https://www.cobb-vantress.com/assets/Cobb-Files/4d0dd628b7/Broiler-Guide_English-2021-min.pdf.

Cuan, K., Zhang, T., Li, Z., Huang, J., Ding, Y., Fang, C., 2022. Automatic Newcastle disease detection using sound technology and deep learning method. Comput Electron Agric. 194, 106740 https://doi.org/10.1016/j.compag.2022.106740.

Duthie, C.-A., Bowen, J.M., Bell, D.J., Miller, G.A., Masonc, C., Haskell, M.J., 2021. Feeding behavior and activity as early indicators of disease in pre-weaned dairy calves. Animal 15, 100150. https://doi.org/10.1016/j.animal.2020.100150.

Erian, I., Phillips, C.J.C., 2017. Public understanding and attitudes towards meat chicken production and relations to consumption. Animals (Basel) 7, 20. https://doi.org/10.3390/ani7030020.

Fossum, O., Jansson, D.S., Etterlin, P.E., Vågsholm, I., 2009. Causes of mortality in laying hens in different housing systems from 2001 to 2004. Acta Vet. Scand. 51, 3. https://doi.org/10.1186/1751-0147-51-3.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Fulcher, B.O., 2017. Feature-based time-series analyses. arXiv:1709.08055v. https://doi.org/10.48550/arXiv.1709.08055.

Gonzalez, L.A., Tolkamp, B.J., Coffey, M.P., Ferret, A., Kyriazakis, I., 2008. Changes in feeding behavior as possible indicators for the automatic monitoring of health disorders in dairy cows. J. Dairy Sci. 91, 1017–1028. https://doi.org/10.3168/jds.2007-0530.

Hafez, H.M., Attia, Y.A., 2020. Challenges to the poultry industry: current perspectives and strategic future after the COVID-19 outbreak. Front. Vet. Sci. 7, 516. https://doi.org/10.3389/fvets.2020.00516.

Hart, B.L., 1988. Biological basis of the behavior of sick animals. Neurosci. Biobehav. Rev. 12, 123–137. https://doi.org/10.1016/S0149-7634(88)80004-6.

Haykin, S., 1998. Neural Networks: A Comprehensive Foundation, second ed. Prentice Hall PTR.

Hofacre, C.L., 2002. The health and management of poultry production. Int. J. Infect. Dis. 6, S3–S7. https://doi.org/10.1016/S1201-9712(02)90177-3.

Howie, J.A., Tolkamp, B.J., Avendano, S., Kyriazakis, I., 2009. A novel flexible method to split feeding behavior into bouts. Appl. Anim. Behav. Sci. 116, 101–109. https://doi.org/10.1016/j.applanim.2008.09.005.

Howie, J.A., Avendano, S., Tolkamp, B.J., Kyriazakis, I., 2011. Genetic parameters of feeding behavior traits and their relationship with live performance traits in modern broiler lines. Poult. Sci. 90, 1197–1205. https://doi.org/10.3382/ps.2010-01313.

Hyndman, R., Kang, Y., Montero-Manso, P., Talagala, T., Wang, E., Yang, Y., O'Hara-Wild, M. 2022. tsfeatures: Time Series Feature Extraction. R package version 1.1.0.9000. https://pkg.robjhyndman.com/tsfeatures/.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning: with applications in R. Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7.

Kang, Y., Hyndman, R.J., Smith-Miles, K., 2017. Visualising forecasting algorithm performance using time series instance spaces. Int. J. Forecast. 33 (2), 345–358. https://doi.org/10.1016/j.ijforecast.2016.09.004.

Li, N., Ren, Z., Li, D., Zeng, L., 2020. Review: Automated techniques for monitoring the behaviour and welfare of broilers and laying hens: towards the goal of precision livestock farming. Animal 14, 617–625. https://doi.org/10.1017/S1751731119002155.

Liu, H.-W., Chen, C.-H., Tsai, Y.-C., Hsieh, K.-W., Lin, H.-T., 2021. Identifying images of dead chickens with a chicken removal system integrated with a deep learning algorithm. Sensors 21, 3579. https://doi.org/10.3390/s21113579.

Lu, D., Jiao, S., Tiezzi, F., Knauer, M., Huang, Y., Gray, K.A., Maltecca, C., 2017. The relationship between different measures of feed efficiency and feeding behavior traits in Duroc pigs. J. Anim. Sci. 95, 3370–3380. https://doi.org/10.2527/jas.2017.1509.

Mendes, E.D.M., Carstens, G.E., Tedeschi, L.O., Pinchak, W.E., Friend, T.H., 2011. Validation of a system for monitoring feeding behavior in beef cattle. J. Anim. Sci. 89, 2904–2910. https://doi.org/10.2527/jas.2010-3489.

Millman, S.T., 2007. Sickness behaviour and its relevance to animal welfare assessment at the group level. Anim Welf. 16, 123–125. https://doi.org/10.1017/S0962728600031146.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pérez-Enciso, M., Steibel, J.P., 2021. Phenomes: the current frontier in animal breeding. Genet. Sel. Evol. 53, 22. https://doi.org/10.1186/s12711-021-00618-1.

Python Software Foundation, 2022. Python (Version 3.10.4) [Software]. Available at: https://www.python.org/.

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Rosa, G.J.M., 2021. Grand challenge in precision livestock farming. Front. Animal Sci. 2, 650324 https://doi.org/10.3389/fanim.2021.650324.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROCPlot when evaluating binary classifiers on imbalanced datasets. PLoS One 10 (3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

Sassi, N.B., Averós, X., Estevez, I., 2016. Technology and poultry welfare. Animals 6, 62. https://doi.org/10.3390/ani6100062.

Schwean-Lardner, K., Fancher, B.I., Gomis, S., Van Kessel, A., Dalal, S., Classen, H.L., 2013. Effect of day length on cause of mortality, leg health, and ocular health in broilers. Poult. Sci. 92, 1–11. https://doi.org/10.3382/ps.2011-01967.

Shahbazi, M., Mohammadi, K., Derakhshani, S., Groot, Koerkamp, P.W.G., 2023. Deep learning for laying hen activity recognition using wearable sensors. Agriculture 13, 738. https://doi.org/10.3390/agriculture13030738.

Scikit-learn Developers, 2007. User Guide. Available at: https://scikit-learn.org/stable/user_guide.html.

Singh, M., Kumar, R., Tandon, D., Sood, P., Sharma, M., 2020. Artificial Intelligence and IoT based Monitoring of Poultry Health: A Review. IEEE International Conference on Communication, Networks and Satellite (Comnetsat), Batam, Indonesia, 2020, pp. 50-54, https://doi.org/10.1109/Comnetsat50391.2020.9328930.

Spackman, E., Pantin-Jackwood, M.J., Kapczynski, D.R., Swayne, D.E., Suarez, D.L., 2016. H5N2 highly pathogenic avian influenza viruses from the US 2014–2015 outbreak have an unusually long pre-clinical period in turkeys. BMC Vet. Res. 12, 260. https://doi.org/10.1186/s12917-016-0890-6.

Sullivan, T.W., 1994. Skeletal problems in poultry: estimated annual cost and descriptions. Poult. Sci. 73, 879–882. https://doi.org/10.3382/ps.0730879.

Sutherland, M.A., Lowe, G.L., Huddart, F.J., Waas, J.R., Stewart, M., 2017. Measurement of dairy calf behavior prior to onset of clinical disease and in response to disbudding using automated calf feeders and accelerometers. J. Dairy Sci. 101, 8208–8216. https://doi.org/10.3168/jds.2017-14207.

Tizard, I., 2008. Sickness behavior, its mechanisms and significance. Anim. Health. Res. Rev. 9 (1), 87–99. https://doi.org/10.1017/S1466252308001448.

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CA CreateSpace, Scotts Valley.

Vapnik, V., 1995. The Nature of Statistical Learning Theory, second ed. Springer, New York, NY.

Ventura, R.V., Silva, F.F., Yáñez, J.M., Brito, L.F., 2020. Opportunities and challenges of phenomics applied to livestock and aquaculture breeding in South America. Anim Front. 10 (2), 45–52. https://doi.org/10.1093/af/vfaa008.

Walker, S.H., Duncan, D.B., 1967. Estimation of the probability of an event as a function of several independent variables. Biometrika 54 (1/2), 167–178. https://doi.org/10.2307/2333860.

Wang, X., Smith, K.A., Hyndman, R.J., 2006. Characteristic-based clustering for time series data. Data Min. Knowl. Disc. 13 (3), 335–364. https://doi.org/10.1007/s10618-005-0039-x.

Winckler, C., 2019. Assessing animal welfare at the farm level: Do we care sufficiently about the individual? Anim. Welf. 28, 77–82. https://doi.org/10.7120/09627286.28.1.077.

Wolfger, B., Schwartzkopf-Genswein, K.S., Barkema, H.W., Pajor, E.A., Levy, M., Orsel, K., 2015. Feeding behavior as an early predictor of bovine respiratory disease in North American feedlot systems. J. Anim. Sci. 93, 377–385. https://doi.org/10.2527/jas.2013-8030.

Yan, W., Sun, C., Wen, C., Ji, C., Zhang, D., Yang, N., 2019. Relationships between feeding behaviors and performance traits in slow-growing yellow broilers. Poult. Sci. 98, 548–555. https://doi.org/10.3382/ps/pey424.

Zhang, C., Ma, Y., 2012. Ensemble Machine Learning: Methods and Applications. Springer Publishing Company, Incorporated.

Zhang, X., Tsuruta, S., Andonov, S., Lourenco, D.A.L., Sapp, R.L., Wang, C., Misztal, I., 2018. Relationships among mortality, performance, and disorder traits in broiler chickens: a genetic and genomic approach. Poult. Sci. 97, 1511–1518. https://doi.org/10.3382/ps/pex431.

Zhuang, X., Bi, M., Guo, J., Wu, S., Zhang, T., 2018. Development of an early warning algorithm to detect sick broilers. Comput. Electron. Agric. 144, 102–113. https://doi.org/10.1016/j.compag.2017.11.032.

Zhuang, X., Zhang, T., 2019. Detection of sick broilers by digital image processing and deep learning. Biosyst. Eng. 179, 106–116. https://doi.org/10.1016/j.biosystemseng.2019.01.003.