Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Disease detection in pigs based on feeding behaviour traits using machine learning

A.T. Kavlak [a,*], M. Pastell [b], P. Uimari [a]

[a] Department of Agricultural Sciences, University of Helsinki, 00014 Helsinki, Finland
[b] Production Systems, Natural Resources Institute Finland (Luke), 00790 Helsinki, Finland

ARTICLE INFO

Article history:
Received 25 December 2021
Received in revised form
15 December 2022
Accepted 5 January 2023
Published online 16 January 2023

Keywords:
Welfare
Disease detection
Pigs
Machine learning
Feeding behaviour

Disease detection is crucial for timely intervention to increase treatment success and reduce negative impacts on pig welfare. The objective of this study was to monitor changes in feeding behaviour patterns to detect pigs that may need medical treatment or extra management. The data included 794,509 observation days related to the feeding behaviour and health information of 10,261 pigs. Feeding behaviour traits were calculated including the number of visits per day (NVD), time spent in feeding per day (TPD), and daily feed intake (DFI). The health status (sick or healthy) of pigs were predicted based on the features including the original feeding behaviour traits and features derived from those using a machine-learning algorithm (Xgboost). The predictions were based either on the features from the same day (one-day window), from the same day and two previous days (three-day window), or from the same day and six previous days (seven-day window). The model based on the seven-day window gave the most robust results and achieved an 80% AUC, 7% F1-score, 67% sensitivity, 73% specificity, and 4% precision. The analyses indicated that the features related to the deviation of a pig's observed TPD and DFI from the expected TPD and DFI were the most informative, as they gained the highest importance score. In conclusion, the feeding behaviour-based features gave good sensitivity and specificity in predicting sickness. However, the precision of the method was very low, possibly due to low prevalence of the monitored sickness symptoms, limiting the application of the approach in real-life.

1. Introduction

Pig welfare has gained more and more attention in recent years and should be improved, according to general consensus (Mellor, 2016). Animals express their wellbeing through feeding, drinking, social behaviour etc. Changes in behaviour can be used as early signs of discomfort and sickness (Matthews et al., 2017). In a commercial farm, only limited time is available to observe the individual behavioural changes in pigs, which only permits detecting considerable behavioural changes. This may lead to the too late treatment of the sick animal or too late improvements in conditions that create discomfort to animals causing losses in production and impaired welfare.

* Corresponding author.
E-mail address: alper.kavlak@helsinki.fi (A.T. Kavlak).
https://doi.org/10.1016/j.biosystemseng.2023.01.004

Nomenclature

AI Artificial intelligence
AUC Area under the ROC curve
CV Cross-validation
delta_DRDFI Difference between the means of the daily rank of daily feed intake from the seven- and three-day windows
delta_DRTPD Difference between the means of daily rank of time spent in feeding per day from the seven- and three-day windows
delta_NVD Difference between the means of number of visits per day from the seven- and three-day windows
delta_p Difference between the means of the proportion of intervals belonging to the first distribution from the seven- and three-day windows
delta_ResDFI Difference between the means of residuals of daily feed intake from the seven- and three-day windows
delta_ResTPD Difference between the means of residuals of time spent in feeding per day from the seven- and three-day windows
delta_μ_1 Difference between the means of the means of intervals belonging to the first distribution from the seven- and three-day windows
delta_μ_2 Difference between the means of the means of intervals belonging to the second distribution from the seven- and three-day windows
delta _σ_1 (s1) Difference between the means of the standard deviations of intervals belonging to the first distribution from the seven- and three-day windows
delta _σ_2 (s2) Difference between the means of the standard deviations of intervals belonging to the second distribution from the seven- and three-day windows
DFI Daily feed intake
DRDFI Daily rank of daily feed intake within the group of pigs
DRTPD Daily rank of time spent in feeding per day within the group of pigs
ML Machine learning
MLP Multilayer Perceptron
m3_DRDFI Daily rank of daily feed intake with three-day window
m3_DRTPD Daily rank of time spent in feeding per day with three-day window
m3_NVD Number of visits per day with three-day window
m3_p Proportion of intervals belonging to the first distribution with three-day window
m3_ResDFI Residuals of daily feed intake with three-day window
m3_ResTPD Residuals of time spent in feeding per day with three-day window
m3_μ_1 The means of intervals belonging to the first distribution with three-day window
m3_μ_2 The means of intervals belonging to the second distribution with three-day window
m3 _σ_1 (s1) The standard deviations of intervals belonging to the first distribution with three-day window
m3_σ_2 (s2) The standard deviations of intervals belonging to the second distribution with three-day window
m7_DRDFI Daily rank of daily feed intake with seven-day window
m7_DRTPD Daily rank of time spent in feeding per day with seven-day window
m7_NVD Number of visits per day with seven-day window
m7_p Proportion of intervals belonging to the first distribution with seven-day window
m7_ResDFI Residuals of daily feed intake with seven-day window
m7_ResTPD Residuals of time spent in feeding per day with seven-day window
m7_μ_1 The means of intervals belonging to the first distribution with seven-day window
m7_μ_2 The means of intervals belonging to the second distribution with seven-day window
m7 _σ_1 (s1) The standard deviations of intervals belonging to the first distribution with seven-day window
m7_σ_2 (s2) The standard deviations of intervals belonging to the second distribution with seven-day window
NVD Number of visits per day
p Proportion of intervals belonging to the first distribution
ResDFI Residuals of daily feed intake
ResTPD Residuals of time spent in feeding per day
ROC Receiver Operator Characteristics
sd7_DRDFI Standard deviation of the daily rank of daily feed intake within the seven-day window
sd7_DRTPD Standard deviation of daily rank of time spent in feeding per day within the seven-day window
sd7_NVD Standard deviation of number of visits per day within the seven-day window
sd7_p Standard deviation of the proportion of intervals belonging to the first distribution within the seven-day window
sd7_ResDFI Standard deviation of residuals of daily feed intake within the seven-day window
sd7_ResTPD Standard deviation of residuals of time spent in feeding per day within the seven-day window
sd7_μ_1 Standard deviation of the means of intervals belonging to the first distribution within the seven-day window
sd7_μ_2 Standard deviation of the means of intervals belonging to the second distribution within the seven-day window
sd7 _σ_1 (s1) Standard deviation of the standard deviations of intervals belonging to the first distribution within the seven-day window

sd7 $_{-\sigma_2}$ (s2)	Standard deviation of the standard deviations of intervals belonging to the second distribution within the seven-day window	μ_2	The means of intervals belonging to the second distribution
TPD	Time spent in feeding per day	σ_1 (s1)	The standard deviations of intervals belonging to the first distribution
μ_1	The means of intervals belonging to the first distribution	σ_2 (s2)	The standard deviations of intervals belonging to the second distribution

Although small changes in daily behaviour are not easy to quantify, data collected automatically from sensors and feeders may include valuable information concerning signs of welfare problems. As an example, increased restlessness among pigs can signal an outbreak of tail biting up to six days prior, which on a commercial scale would be impossible to detect during daily checks (Matthews et al., 2017). In addition, microphones have been used to monitor the sounds of coughing of pigs to build an intelligent alarm system to detect the disease in its early stage (Guarino et al., 2008), 3D-cameras to predict tail biting outbreaks by identifying lowered tail postures (D'Eath et al., 2018), and deviations in typical feeding patterns to monitor overall welfare of pigs (e.g., Brown-Brandl et al., 2013; Bus et al., 2021; Munsterhjelm et al., 2015).

The data collected from the sensors and feeders create challenges to finding the true signals of behavioural changes out of the noise. The complexity of big data with non-linear dependencies and unknown interactions across multiple variables challenges the assumptions of many standard statistical methods (Valletta et al., 2017). Machine learning (ML) methods are highly efficient at determining non-linear relationships between variables in the data (Hastie et al., 2009). As an example, Pandey et al. (2021) collected data on movements, vocal sound, and temperature of pigs using ear sensors and applied ML models to predict the health and welfare status of pigs based on the collected data. Based on their results, the ML approach is a powerful tool for monitoring the health status of pigs leading to reduced medical treatments, cost savings and enhanced animal welfare. Thus, ML methods, such as eXtreme Gradient Boosting, Random Forest, and Support Vector Machine, provide a promising approach for detecting behavioural changes in farm animals that are associated with possible welfare problems (Liakos et al., 2018). Regardless of the method, data quality is important to avoid unwanted outcomes and to gain as robust results as possible. Setting criteria for outliers and applying data filtering prior to applying ML methods to data are therefore important (Alsaaod et al., 2012).

The objective of this study was to use ML methods applied on feeding behaviour data to detect pigs that are potentially sick and may need medical treatment or extra management.

2. Material and methods

2.1. Feeding behaviour data and pig housing

The feeding behaviour data were provided by Figen Oy (Pietarsaari, Finland) from their central test station, spanning from 2011 to 2016. Pigs arrived at the test station either on a Tuesday or a Wednesday, and the tests began on a Saturday. The pigs were grouped into different pens according to their arrival age (89 ± 10 days), weight (34.4 ± 6.4 kg), and sex (only boars or a combination of gilts and castrates). The average daily gain was 946 ± 113 g/day in total testing time (on average, 95 ± 3 days), and the average slaughter weight and age were 121.2 ± 12.9 kg and 186 ± 10 days, respectively. The average number of piglets in a pen was 9.8 (±1.19). Water was available ad libitum. Also, feeding type (dry feeding) was ad libitum, consisting of two commercial feedstuffs, and the proportion of the two feedstuffs was based on the growth rate curve of an average pig from the previous test periods. Antibiotics and other drugs were given only for the sick animals based on veterinary prescriptions. The facility had automated ventilation based on pig age and outdoor temperature, and artificial lighting was on from 7 a.m. to 3 p.m. The size of pen was 16.8 m^2 with a concrete floor (2/3 solid, 1/3 slatted). Feedings were recorded automatically using the Schauer Spotmix with Schauer Multilayer Perceptron (MLP) electronic feeders and MLP manager data management software (Schauer Agrotronic GmbH). For further information see Kavlak et al. (2021).

The raw data consisted of 28,826,029 individual feeding visits from 10,261 pigs (Finnish Yorkshire, Finnish Landrace, and F1-crossbred), and included ear tag transponder id, date, time of entering the feeder, time leaving the feeder, and feed intake per visit. The feed intake was measured as a weight of the feed before and after the pig has been in feeder. The number of visits per day (NVD), time spent in feeding per day (TPD), and daily feed intake (DFI) were calculated from the recorded observations. Observations from the first testing day were not included due to the DFI exhibiting some of the pigs as extreme outliers, may have been caused by the feeding recording system. Similar extreme DFIs were not observed on a large scale on any other testing days.

2.2. Sickness data

The sickness data were recorded daily by the test station staff members during routine checks (twice a day) and included the ID of the pig, the symptom(s), and the date. The symptoms were classified as a cough, a limp, loss of appetite (the pigs who have been eating less than 600 g), skin damage, and a bitten tail. Out of 794,509 daily health observations, 13,018 were related to the recorded symptoms. Within any given day a pig could suffer from several symptoms. In the ML models, pigs with any of the recorded symptoms were classified as "sick" for that given day and pigs with no recorded symptoms were classified as "healthy".

2.3. Feature processing

The absolute values of TPD and DFI may not be optimal features for predicting the sickness status of an animal, as they are strongly related to the animal's age. Therefore, we created new features, including daily ranks and residuals. Daily ranks relate the rank of an animal's observation (DRTPD and DRDFI) compared to other pigs within a pen in a given day, and residuals (ResTPD and ResDFI) describe that animal's difference from the expected value of TPD and DFI for a pig of same age. The residuals of TPD and DFI were calculated by fitting a polynomial (quadratic) regression model to the whole data set:

$$yi = b0 + b1^{*}\ agei + b2^{*}\ age_i^2 + ei \quad (1)$$

where y_i is either the TPD or DFI of pig i, b_0 is overall mean, age_i is the age of pig *i* related to observation y_i, b_1 and b_2 are linear and quadratic regression coefficients, and e_i is the residual used in ML.

Regarding animal welfare based on feeding behaviour, short-term visits have been considerably challenging to interpret in animal behaviour analyses when conventional methods are used (Young & Lawrence, 1994). The frequency of visits without eating and intervals between visits can be informative feeding patterns that can contribute to predict the health status of animals (Garrido-Izard et al., 2020). Tolkamp et al. (1998) proposed log-normal distribution to model within and between feeding events. In this study, the intervals betwThe 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.een feeding visits were calculated as the difference between the time of entering the feeder and the time leaving the feeder during the previous feeding. The intervals were first log transformed and then a mixture of two normal distributions was fitted for the log-transformed intervals of each pig for each day. The first distribution relates to short visits during one meal and second representing intervals between meals. The daily features for each pig from this mixture distributions were the proportion of intervals belonging to the first distribution (p), σ_1 (s1) and σ_2 (s2) the standard deviations, and μ_1 and μ_2 the means of the distributions.

Finally, the features used in ML were NVD, TPD, DFI, DRTPD, DRDFI, ResTPD, ResDFI, μ_1, μ_2, σ_1, σ_2 and *p*. The mean and the distribution of the features in the healthy and sick groups over time (Week) are presented in Fig. 1. Prior to creating the features and ML models, extreme values of the NVD, TPD, DFI (outside quantiles 0.5% and 99.5% corresponding to likely registration errors from the feeders) were removed (less than 0.3% of the sick daily observations and less than 5% of the healthy daily observations) from the data.

The health status of a pig was predicted using three different window lengths for determining features: a one-day window, a three-day window, and a seven-day window (number of observations are given in Table 1). In the one-day window approach, the health status of a pig was predicted based on the features from the same day. In the three-day window approach, the health status of a pig was predicted based on the mean of the features from the same day and the previous two days. Similarly, for the seven-day window approach, the health status of a pig was predicted based on the mean of the features from the same day and the previous six days. Based on the three- and seven-day window features, a new features "delta" and "SD" were calculated; delta as a difference between the means of the same feature from the seven- and three-day windows and SD as a standard deviation of the features within the seven-day window. Windows were overlapping.

The number of daily sick and healthy observations are given in Table 1. The number of observations varies between the models because in three- and seven-day window models if any of the daily features within a tree or seven days, respectively, were missing for a given pig, the pig was not included into analysis. In addition, various combinations of symptoms were used; in Alt-1 -model, "a limp" and "loss of appetite" were treated as "sick", while the other symptoms (cough, bitten tail, skin damage) were omitted and in Alt-2 -model "bitten tail" and "skin damages" were treated as "sick" (Table 1). For any given pig, on average there were 7.1 consecutive sick days (an average length of the sickness period).

2.4. Xgboost algorithm

eXtreme Gradient Boosting (Xgboost) is an ML method similar to Random Forest, decision tree, boosting, gradient boosting, etc. It is an ensemble classifier derived from the gradient boosting decision tree. Xgboost combines weak base classifiers into a strong classifier. At each iteration of the training process, the residual of a base classifier is used in the next classifier to optimise the objective function. In this study, the Xgboost algorithm was applied using the R package Xgboost (Chen et al., 2018) in R 3.6.1 software (R Core Team, 2019).

Hyperparameters are optimization parameters that tune the performance of ML algorithms (Bergstra & Bengio, 2012). In this study, the hyperparameters were chosen using a grid search of the number of boosting iterations (*nrounds*), maximum depth of a tree (*max-depth*), *eta* that controls the learning rate as well as gamma, lambda, subsample. The value grid used for the hyperparameters is given in Table 2 and the final (best) hyperparameters are given in Table 3 based on training data. The objective of the classification model was binary *(binary:logistic)* and the model was fitted by minimizing the binary classification error rate.

2.5. Performance testing and cross-validation

For the performance of the Xgboost algorithm, the data were split into training and testing data sets. In this study, 70% of the observations were used in training the model and 30% in testing it (Fig. 2). A random sampling of observations was stratified according to the symptoms and pig ID to ensure that the proportion of sick and healthy observations was the same in both data sets and that data from different pigs were used for training and testing the model.

To optimise the hyperparameters and the features and to avoid overfitting the models, we applied 10-fold cross-validation (CV) during model training. The training data set was divided into 10 sets (folds) of equal size. In each validation step, nine of the sub-sets were used for training the model and

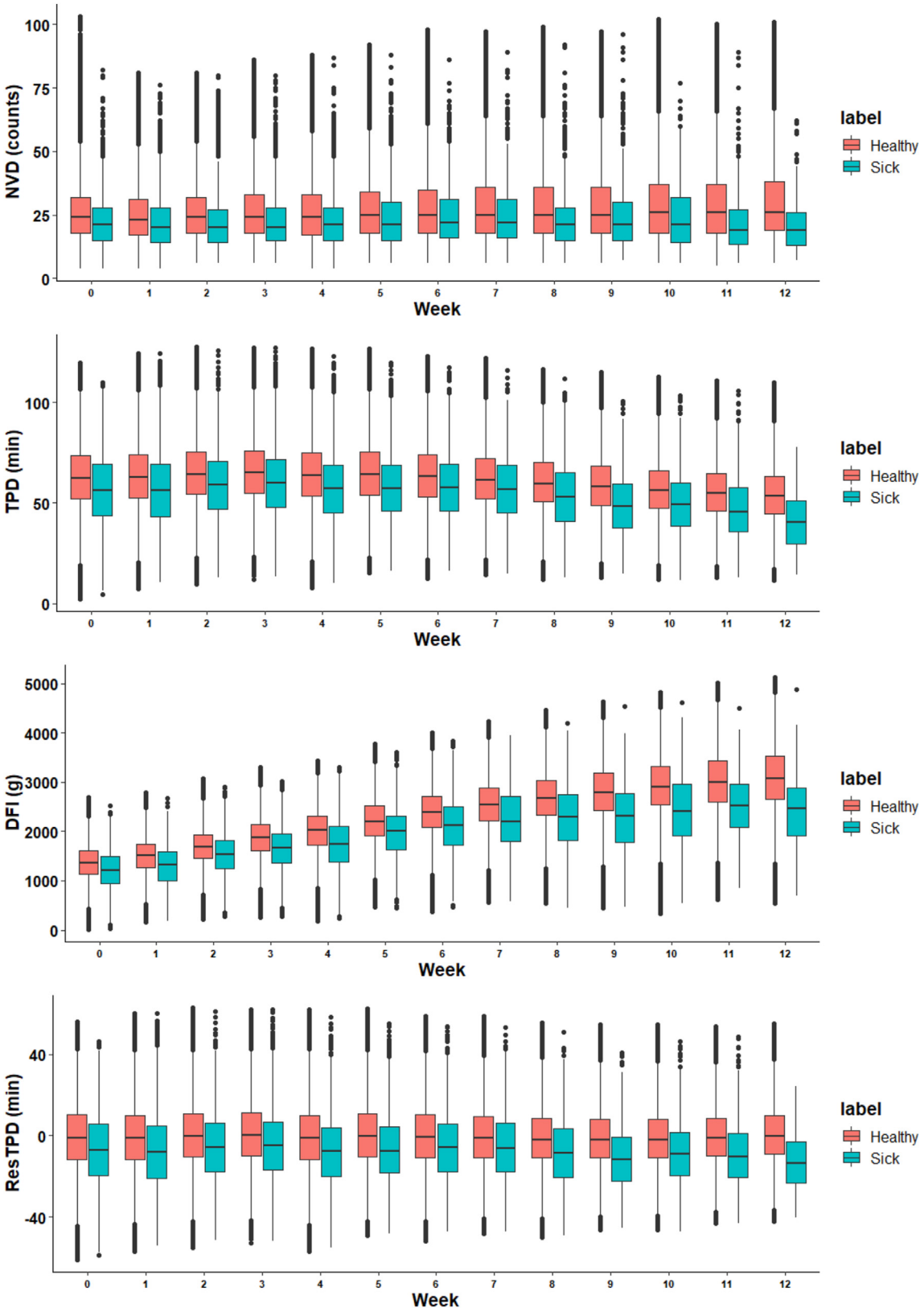

Fig. 1 – Boxplots of the features for the testing period (in weeks) grouped by the disease status of the pigs (13,018 daily sick observations and 781,491 daily non-sick observations). The abbreviations of the features are explained in the text.

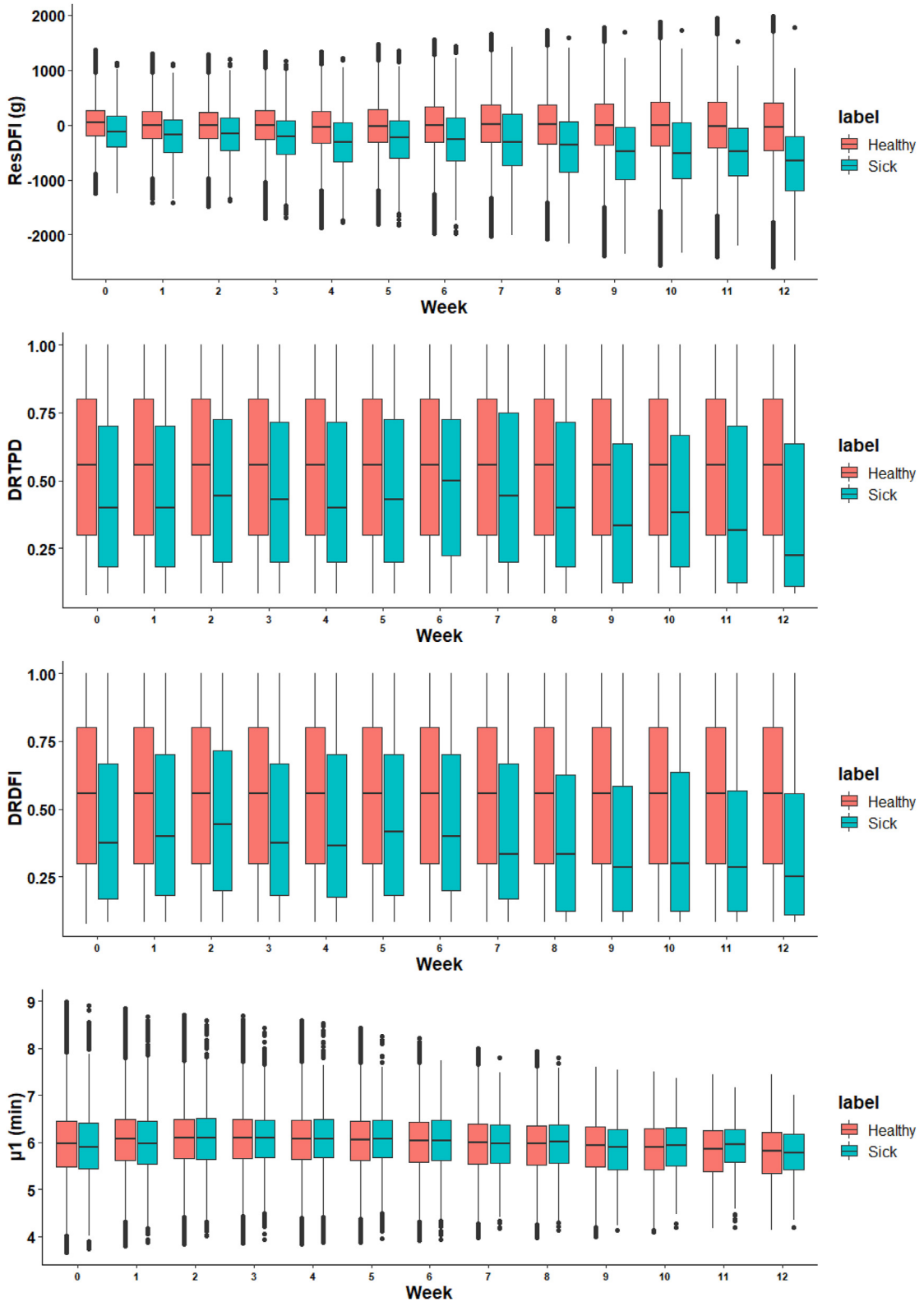

Fig. 1 – Continued

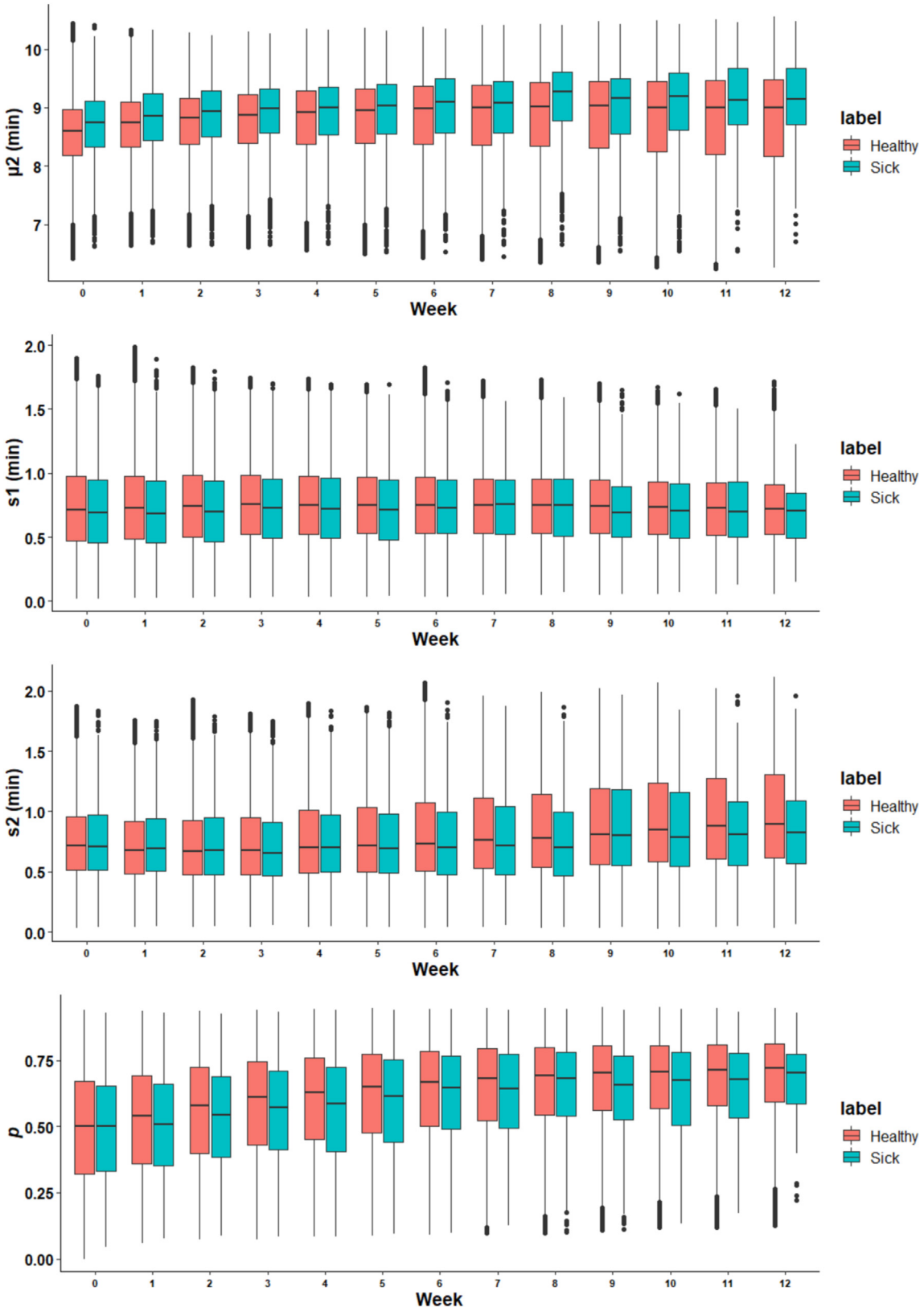

Fig. 1 – continued

Table 1 – Number of symptoms and "sick" and "healthy" observations (pigs x days) for each model.

	Limp	Cough	Bitten tail	Loss of appetite	Skin damage	Sick observations	Healthy observations	Total
Model								
1-day	6603	988	2941	1015	1471	13,018	781,491	794,509
3-day	6377	956	2888	968	1357	12,546	761,477	774,023
7-day	5747	846	2787	824	1060	11,264	722,070	733,334
Alt-1[a]	5747	–	–	824	–	6571	722,070	728,641
Alt-2[b]	–	–	2787	–	1060	347	722,070	725,917
Prevalence	0.008	0.001	0.004	0.001	0.002	0.016		

N = Number of observations; Prevalence = Proportion of total symptoms labelled as "sick" out of the total observations in the data based on the 1-day model; [a] In the Alt-1 model, only "a limp" and "loss of appetite" were labelled as "sick" with the seven-day window model and all other symptoms were omitted; [b] In the Alt-2 model, only "a bitten tail" and "skin damage" were labelled as "sick" with the seven-day window model, and all other symptoms were omitted.

Table 2 – Range of the values of the hyperparameters.

Hyperparameter	Description	Range of values
nrounds	number of boosting iterations	10–20
max_depth	maximum depth of a tree	3–6
eta	controls the learning rate	0.05–0.5
gamma	controls the minimum reduction in the loss function	0–5
lambda	ridge regularization to prevent overfitting	1.0–2.0
subsample	subsample ratio of the training observations	0.5–1.0

Table 3 – The final values (best) of hyperparameters based on training data.

	Window length (day)				
Final Hyperparameters	1	3	7	Alt-1[a]	Alt-2[b]
max_depth	4	4	4	3	4
Eta	0.45	0.45	0.45	0.45	0.45
Gamma	4	5	4	2	3
Lambda	1	2	1.4	2	2
Subsample	1	0.9	0.8	0.9	1

one sub-set was used for testing the model (Fig. 2). In addition, we used an additional parameter (scale_pos_weight: the ratio of number of negative class to the positive class) in the models to control the balance of classes weights due to the imbalanced data set. The parameter was calculated as the proportion of the number of sick observations to number of healthy observations. From each validation step, the area under the ROC curve (AUC) was calculated from the holdout cross-fold (Validation-fold) (Hastie et al., 2009). The set of hyperparameters that gave the best performance metric (AUC) of the model was selected to train the model in the training set and then applied the obtained model for predicting the health status in the testing set (Testing data set in Fig. 2).

Using the test data set, the models were evaluated based on precision (proportion of predicted true positives (an animal predicted as sick) out of all positive predictions; TP/(TP + FP)), sensitivity (proportion of positives (sick) that were identified correctly; TP/(TP + FN)), and specificity (proportion of negatives (healthy) that were identified correctly; TN/(TN + FP)). In addition, the harmonic means of the precision and sensitivity (F1-score = 2 x precision x sensitivity/(precision + sensitivity)), and AUC (receiver operating characteristics) curve were calculated. The model was considered non-informative with an AUC ≤0.50, weak with an AUC of 0.50–0.70, accurate with an AUC of 0.70–0.90, and highly accurate with an AUC ≥0.90 (Swets, 1988; Greiner et al., 2000).

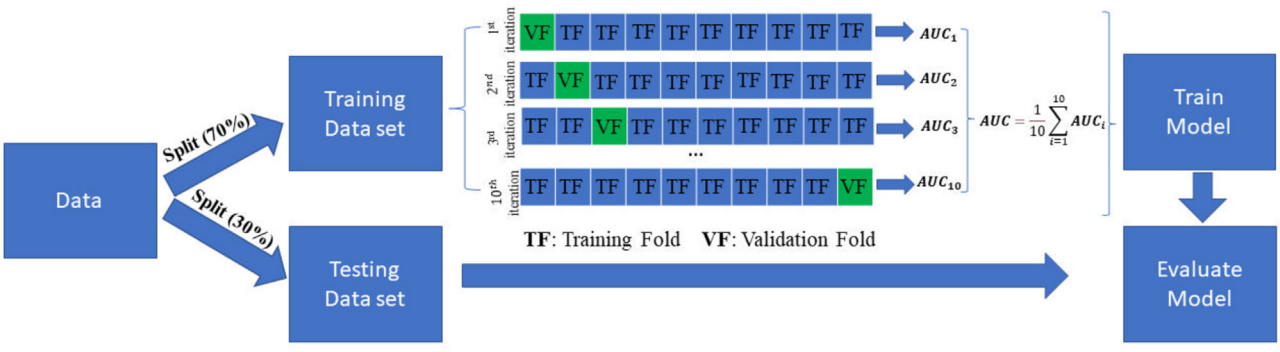

Fig. 2 – Overview of 10-fold cross-validation and model testing.

We also calculated feature importance using the ability of Xgboost to remove the non-informative or redundant predictors from the model (Chen et al., 2018). While fitting the Xgboost models, an importance matrix was produced from each model. The "gain" metric indicates the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. The "cover" metric indicates the relative number of observations related to this feature and the frequency, which is the percentage of the relative number of times a particular feature occurs in the trees of the model. An obtained score of each feature is based on how much more information about the class is gained when using that feature. We quantified the importance of features by "feature gain" (Fig. 3). The steps given above were carried out with the R package caret (Kuhn et al., 2018) in R software (R Core Team, 2019).

3. Results

3.1. *Classification performance of the models*

The models were evaluated based on classification performance metrics, including AUC. The best performance according to AUC was obtained with the model applied in the seven-day window (Table 4). In addition, the difference for accuracy of the performance metrics with the training and testing data sets were small which indicated that over- or under-parametrization of the models was avoided. The best hyperparameters were obtained based on data that provided during training and used in prediction of the models (Table 3).

The sensitivity and specificity of the models were acceptable with all window lengths. However, precision and F1-score were quite low. Again, the best performance (67% sensitivity and 73% specificity) was obtained with the model applied in the seven-day window. Unlike the seven-day window model, other models performed at slightly lower efficiency according to the performance metrics. Overall, the results show that by increasing the window length, the performance of the classification models increases.

Alternative labelling of sick animals was tested with two alternative models. For the first alternative model (Alt-1 model), we only labelled "a limp" and "loss of appetite" as "sick" and omitted all other symptoms (cough, bitten tail and skin damage). This model gave 3–4% better performance based on AUC than the performance of the actual seven-day window model (Table 4). On the other hand, the second alternative model (Alt-2 model), where "a bitten tail" and "skin damage" were categorized as "sick" and omitted all other symptoms (cough, limp and loss of appetite), gave a similar

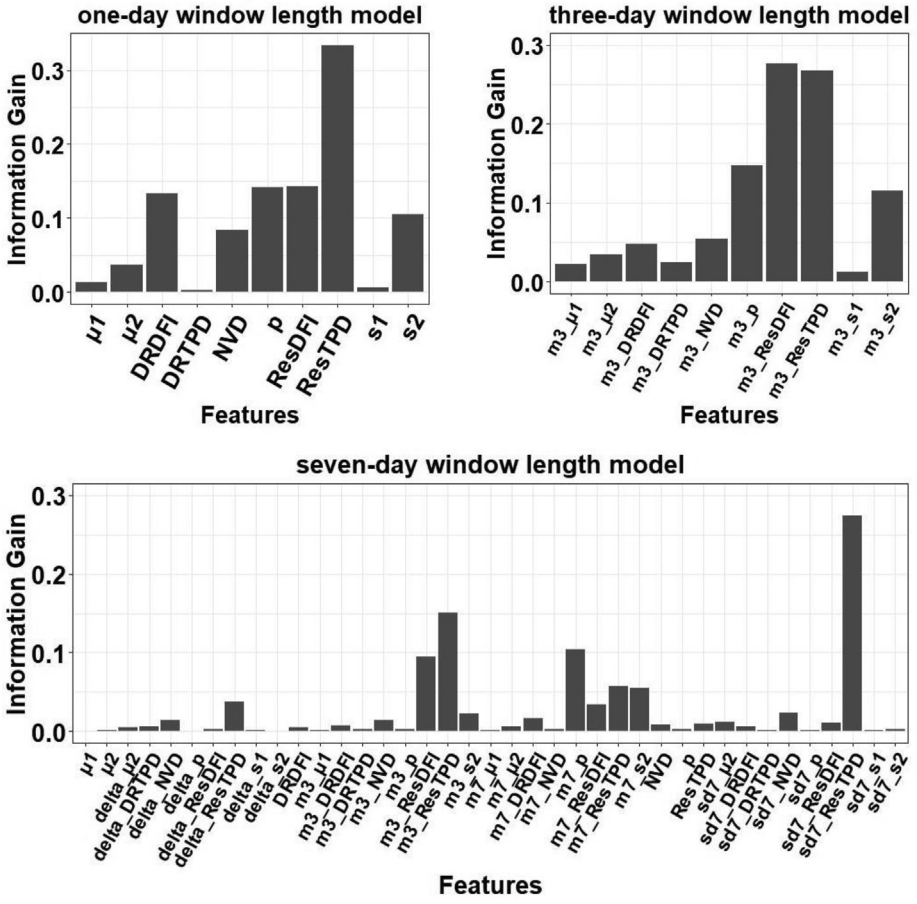

Fig. 3 – Importance of the Xgboost features for the different window length. The "Information Gain" implies the relative contribution of the corresponding feature to the model. The abbreviations of the features are explained in the text.

Table 4 – Results from the models based on testing data and training (average from the 10-fold CV) (given in parentheses).

	Window length (day)				
Metrics	1	3	7	Alt-1	Alt-2
AUC	0.70 (0.71)	0.73 (0.75)	0.80 (0.81)	0.83 (0.85)	0.77 (0.80)
Precision	0.03 (0.03)	0.03 (0.03)	0.04 (0.04)	0.03 (0.03)	0.01 (0.01)
Sensitivity	0.60 (0.61)	0.63 (0.65)	0.67 (0.72)	0.67 (0.71)	0.67 (0.74)
Specificity	0.67 (0.67)	0.69 (0.69)	0.73 (0.73)	0.78 (0.81)	0.70 (0.70)
F1-score	0.06 (0.06)	0.06 (0.06)	0.07 (0.08)	0.05 (0.06)	0.02 (0.02)

In the Alt-1 model, only "a limp" and "loss of appetite" were labelled as "sick" with the seven-day window model and all other symptoms were omitted.
In the Alt-2 model, only "a bitten tail" and "skin damage" were categorized as "sick" with the seven-day window model, and all other symptoms were omitted.

Table 5 – Tabular visualization of observed versus predicted values from the ALT-1 model based on testing data.

		Observed values		
		Sick	Healthy	Total
Predicted Values	Sick	TP (1,321)	FP (46,682)	48,003
	Healthy	FN (650)	TN (169,939)	170,589
	Total	1971	216,621	218,592

The observation number of the observed and predicted values are given in parentheses. TP = true positives, TN = true negatives, FP = false positives, FN = false negatives.

performance as labelling all symptoms as "sick" (Table 4). A detailed distribution of assessments for Alt-1 is presented in Table 5, showing that the proportion of animals predicted as being sick was approximately 22%, despite the prevalence based on observed data being around 1%, resulting in low precision.

3.2. *The most important features*

The most informative features were those related to daily feeding time and daily feed intake: ResTPD and ResDFI in the one-day and three-day window models and SD_ResTPD in the seven-day window model. They alone explained between 20% and 35% of the information gain (Fig. 3). The importance of the other features was less than 10% (Fig. 3). In general, the new features calculated from the NVD, TPD, and DFI were more important in predicting the health status than the absolute values of NVD, TPD, and DFI. As we expected, using the seven-day window model with only the best 10 important features instead of all 40 features increased model performance (based on AUC) slightly (by 1–2%) and reduced the model's run time (results are not shown).

4. Discussion

In this study, the Xgboost algorithm, with features based on feeding station records, was applied to predict the possible sickness of pigs in a test station environment. The considered symptoms were limping, coughing, a bitten tail, loss of appetite, and skin damage, or any combination of these symptoms. In general, the models reached relatively high AUC (0.7–0.83). However, model precision was very low (the models predict more sick animals than are reported in the data). Similar to our study, Thomas et al. (2021) predicted diarrhea based on weight dependent water and feed intake using a machine learning approach with seven different methods. Most of the tested methods failed to detect diarrheic pigs due to substantial individual instability on feeding or water related to weight. Even with the best model, 25% of the sick piglets were not detected. Similar to our study, Maselyne et al. (2018) investigated if unusual behavioural changes in the feeding pattern in pigs can be utilised as an indicator of health, welfare, and productivity problems. Although they had considerably high specificity (98.7%) and accuracy (96.7%), sensitivity (58.0%) and precision (71.1%) were lower causing false alerts of health problems and lack of confidence of the farmers for the system. A higher precision (an average of 80%) has been achieved also in some other studies, e.g., in Alsaaod et al. (2012) and Gertz et al. (2020). Gertz et al. (2020) reported very good classification performance (86% AUC, 81% F1-score, 78% specificity, and 81% sensitivity) using the Xgboost algorithm, where locomotion-related diseases were predicted using locomotion data collected from leg and neck sensors in a commercial farm of 397 dairy cows. The health status of cows was monitored by on-farm staff and veterinarians during their daily routine. Based on their findings, using various models with different features and window segments increased model performance and sickness-related behaviours were accurately identified. Moreover, Alsaaod et al. (2012) reported better classification accuracy (76%) for predicting lameness in dairy cows using features created from the pedometric activity and behaviour data on lying down compared to classification accuracy (65%) achieved with the raw data by using the Support vector machine classification model. Thus, in line with our findings, creative new features calculated from the raw data are more informative than the actual sensor data in predicting the sickness of animals.

In our study, pig health was monitored by station staff during the daily routine check. It is possible that only the most severe cases were detected by the staff and some milder ones were missed, and thus the true prevalence of symptoms may be higher than the observed 2% (depending on what symptoms were classified as "sick") in the data (Table 1). Thus, some of the true negatives (indicated as healthy in the data) could have been sick instead. Higher actual prevalence is supported by Munsterhjelm et al. (2015), where 2672 pigs in the same test station (Längelmäki, Finland) were monitored in

detail for symptoms three to four times daily by farm staff, who were supervised by a herd veterinarian, between November 2007 and December 2008. During that period, the prevalence of tail biting was 13%, 11% for limping, 2% for skin damage, and 6.1% for other symptoms (including diarrhoea, weight loss, vomiting etc.). Another possible explanation for the low precision in our study could be that the classifier did not learn the optimal decision boundary with our highly imbalanced data set despite the weighing we used for the samples from the minority class. Any real dataset may have several imbalanced classes causing biased classification in machine learning. Various techniques have been developed to deal with this problem such as undersampling methods, oversampling methods, ensemble methods etc. that improve the performance of classifiers (He & Garcia, 2009; Japkowicz & Stephen, 2002; Provost, 2000). Although we scaled the class weights according to the prevalence of observations in each class to solve the imbalanced classification problem, we should try out other suggested methods and find the best one in future for our dataset. However, the most effective technique still may vary depending on the dataset.

Unusual behavioural changes in pigs may indicate sickness. These behavioural changes may be rapid and indicate sickness immediately after the behavioural changes have occurred or the changes may begin several days prior to sickness. Therefore, we applied models with different window lengths. We found a clear tendency that considering records from several previous days instead of a single day was beneficial (AUC increased from 0.70 to 80). Gertz et al. (2020) also reported that using various window lengths allows the classifier to select the amount of data leading to the best prediction performance. However, the Xgboost preferably selected shorter window lengths in their study compared to our study. Thus, it is always a good practise to test several window lengths because method performance depends on the features and nature of the data, and the long window may not always be optimal. Also, in other behavioural studies (e.g., Piette et al., 2020; Riaboff et al., 2020; Smith et al., 2016), the sliding window length approach has had a positive impact on algorithm performances.

Selecting the optimal hyperparameters is important for successful model performance, as the ML methods have a high risk of under/overfitting the training data. However, there is no optimal way to tune the hyperparameters. In our study, the hyperparameters were tuned using the grid search method (Bergstra & Bengio, 2012) with 10-fold CV, and the best hyperparameters were selected for further analyses. Thus, even though tuning the hyperparameters requires extra computing time, obtaining good prediction performance is recommended.

Finally, the set of features available for prediction is crucial for improving the performance of the classification. In our study, the features were calculated from the feeding behaviour data with short and long window segmentations. The most important features were ResTPD and ResDFI along with SD_ResTPD in the seven-day window length model (Fig. 3), which indicate that using the residuals of feeding behaviour traits is more beneficial in predicting pig sickness than absolute values. Thus, a deviation from a typical daily feeding time or daily feed intake compared to the feeding time and daily feed intake of an average pig at the same age is a good indication of a possible health problem. Similarly, daily ranks of TPD and DFI were informative and are easier to calculate than the residuals of TPD and DFI. Hoy et al. (2012) also suggested that daily ranks based on feeding must be classified because many pigs have access to one feeding place in a pen. Therefore, we propose using features that indicate a difference of an animal's feeding behaviour from its pen mates (rank) or from pigs of the same age (residual) rather than raw observations (NVD, TPD, DFI). Furthermore, solely using the most important features in the model instead of all available features improved algorithm performance slightly (1–2%).

From a practical standpoint, high sensitivity is more important than high precision because the final assessment of an animal's sickness would be based on a re-check by the management staff if the applied algorithm suggests that the animal may be sick. The cost of re-checking additional animals should be smaller than treating a sick animal that was not detected early enough. Despite this, the precision should be far higher than what was achieved here to gain trust in users of the algorithm on a routine basis. Features derived from other automatic data recording systems, such as locomotion sensors, could improve the predictive performance of the method.

5. Conclusions

Based on the performance metrics (AUC, sensitivity, and specificity), pig sickness is detectable by applying the Xgboost algorithm to the feeding behaviour data. However, very low precisions were obtained, possibly due to imbalanced data. Using the observations from several days (seven days) gave more accurate predictions than predictions based on a single day, even though the results did not differ considerably. When the prediction was based on one- or three-day observations (one- and three-day windows), the most important features were ResTPD and ResDFI. Overall, we examined a vast, but limited set of features, and our results can be improved by calculating new features, considering interactions between features, using different window length(s), different methods etc. This would require more research.

CRediT authorship contribution statement

Alper Tuna Kavlak: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Software, Writing-Original draft. **Matti Pastell**: Methodology, Formal analysis, Software, Data curation, Writing- Reviewing and Editing. **Pekka Uimari**: Supervision, Conceptualization, Methodology, Formal analysis, Data curation, Writing- Reviewing and 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.

Acknowledgement

The study was funded by Raisio Oyj Tutkimussäätiö (Finland) and Suomen Sianjalostuksen Säätiö (Finland). The authors are grateful to Natural Resources Institute Finland (Luke) for collaborating in this project and to Timo Serenius (Figen Oy), Marja-Liisa Sevón-Aimonen (Luke), and Jarmo Valaja (University of Helsinki) for their opinions as thesis advisory committee members.

REFERENCES

Alsaaod, M., Römer, C., Kleinmanns, J., Hendriksen, K., Rose-Meierhöfer, S., Plümer, L., & Büscher, W. (2012). Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. *Applied Animal Behaviour Science, 142*, 134–141.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.

Brown-Brandl, T., Rohrer, G. A., & Eigenberg, R. A. (2013). Analysis of feeding behavior of group housed growing–finishing pigs. *Computers and Electronics in Agriculture, 96*, 246–252.

Bus, J. D., Boumans, I. J. M. M., Webb, L. E., & Bokkers, E. A. M. (2021). The potential of feeding patterns to assess generic welfare in growing-finishing pigs. *Applied Animal Behaviour Science, 241*, Article 105383.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2018). *xgboost: Extreme gradient boosting*.

D'Eath, R. B., Jack, M., Futro, A., Talbot, D., Zhu, Q., Barclay, D., & Baxter, E. M. (2018). Automatic early warning of tail biting in pigs: 3D cameras can detect lowered tail posture before an outbreak. *PLoS One, 13*, 18.

Garrido-Izard, M., Correa, E. C., Requejo, J. M., & Diezma, B. (2020). Continuous monitoring of pigs in fattening using a multi-sensor system: Behavior patterns. *Animals, 10*, 17.

Gertz, M., Große-Butenuth, K., Junge, W., Maassen-Francke, B., Renner, C., Sparenberg, H., & Krieter, J. (2020). Using the XGBoost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. *Computers and Electronics in Agriculture*, 173.

Greiner, M., Pfeiffer, D., & Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine, 45*, 23–41.

Guarino, M., Jans, P., Costa, A., Aerts, J. M., & Berckmans, D. (2008). Field test of algorithm for automatic cough detection in pig houses. *Computers and Electronics in Agriculture, 62*, 22–28.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE, 21*(9), 1263–1284.

Hoy, S., Schamun, S., & Weirich, C. (2012). Investigations on feed intake and social behaviour of fattening pigs fed at an electronic feeding station. *Applied Animal Behaviour Science, 139*, 58–64.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449.

Kavlak, A. T., Strandén, I., Lidauer, M. H., & Uimari, P. (2021). Estimation of social genetic effects on feeding behaviour and production traits in pigs. *Animal, 15*(3), Article 100168.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2018). *caret: Classification and regression training*.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*, 2674.

Maselyne, J., Van Nuffel, A., Briene, P., Vangeyte, J., De Ketelaere, B., Millet, S., Van den Hof, J., Maes, D., & Saeys, W. (2018). Online warning systems for individual fattening pigs based on their feeding pattern. *Biosystems Engineering, 173*, 143–156.

Matthews, S. G., Miller, A. L., Plötz, T., & Kyriazakis, I. (2017). Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. *Scientific Reports, 7*(1), Article 17582.

Mellor, D. J. (2016). Updating animal welfare thinking: Moving beyond the "five freedoms" towards "A life worth living". *Animals, 6*, 21.

Munsterhjelm, C., Heinonen, M., & Valros, A. (2015). Effects of clinical lameness and tail biting lesions on voluntary feed intake in growing pigs. *Livestock Science, 181*, 210–219.

Pandey, S., Kalwa, U., Kong, T., Guo, B., Gauger, P. C., Peters, D. J., & Yoon, K. J. (2021). Behavioral monitoring tool for pig farmers: Ear tag sensors, machine intelligence, and technology adoption roadmap. *Animals, 11*(9), 2665.

Piette, D., Norton, T., Exadaktylos, V., & Berckmans, D. (2020). Individualised automated lameness detection in dairy cows and the impact of historical window length on algorithm performance. *Animal, 14*, 409–417.

Provost, F. (2000). *Machine learning from imbalanced data sets 101. Workshop on learning from imbalanced data sets*. Texas, US: AAAI.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Riaboff, L., Poggi, S., Madouasse, A., Couvreur, S., Aubin, S., Bedere, N., Goumand, E., Chauvin, A., & Plantier, G. (2020). Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data. *Computers and Electronics in Agriculture, 169*, Article 105179.

Smith, D., Rahman, A., Bishop-Hurley, G. J., Hills, J., Shahriar, S., Henry, D., & Rawnsley, R. (2016). Behavior classification of cows fitted with motion collars: Decomposing multi-class classification into a set of binary problems. *Computers and Electronics in Agriculture, 131*, 40–50.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293.

Thomas, J., Rousselière, Y., Marcon, M., & Hémonic, A. (2021). Early detection of diarrhea in weaned piglets from individual feed, water and weighing data. *Frontiers in Animal Science, 2*, 2673–6225.

Tolkamp, B. J., Allcroft, D. J., Austin, E. J., Nielsen, B. L., & Kyriazakis, I. (2016). Satiety splits feeding behaviour into bouts. *Journal of Theoretical Biology, 194*, 235–250.

Valletta, J. J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour, 124*, 203–220.

Young, R. J., & Lawrence, A. B. (1994). Feeding behaviour of pigs in groups monitored by a computerized feeding system. *Animal Science, 58*, 145–152.