INTERPRETATIVE SUMMARY

Applying additive logistic regression to data derived from sensors monitoring behavioral and physiological characteristics of dairy cows to detect lameness

By Kamphuis et al., page XXXX. Lameness negatively impacts cow welfare and decreases farm profitability. Detecting lame cows visually is difficult and will become more challenging as herd sizes are increasing. Automated detection of lame cows may be a viable aid or alternative to visual detection. The hypothesis was that data derived from sensors monitoring behavioral and physiological characteristics of dairy cows are useful to automate the detection of lameness. Univariable and multivariable models were developed using a data-mining technique. The multivariable model outperformed the univariate models but still the detection performance was not sufficiently high to warrant implementation in practice on large, pasture-based, dairy farms.

15 **DETECTING LAME COWS AUTOMATICALLY**

16

17 **Applying additive logistic regression to data derived from sensors monitoring**

18 **behavioral and physiological characteristics of dairy cows to detect lameness**

19

20 **C. Kamphuis,[*,1] E. Frank,[†] J.K. Burke,[*] G.A. Verkerk,[*] and J.G. Jago,[*]**

21

22 [*]DairyNZ Ltd., Private Bag 3221, Hamilton 3240, New Zealand

23 [†]Department of Computer Science, University of Waikato, Private bag 3105, Hamilton 3240,

24 New Zealand

25

26 [1]Corresponding author: C. Kamphuis, Private Bag 3221, Newstead, Hamilton 3240, New

27 Zealand, phone: +64 (0)7 858 3750, fax: +64 (0)7 858 3751, email:

28 claudia.kamphuis@dairynz.co.nz

29

ABSTRACT

The hypothesis was that sensors currently available on-farm that monitor behavioral and physiological characteristics have potential for the detection of lameness in dairy cows. This was tested by applying additive logistic regression to variables derived from sensor data. Data were collected between November 2010 and June 2012 on five, commercial, pasture-based, dairy farms. Sensor data from weigh scales (live-weight), pedometers (activity) and milk meters (milking order, unadjusted and adjusted milk yield in the first two minutes of milking, total milk yield, and milking duration) were collected at every milking from 4,904 cows. Lameness events were recorded by farmers who were trained in detecting lameness before the study commenced. A total of 318 lameness events affecting 292 cows were available for statistical analyses. For each lameness event, the lame cow's sensor data for a time period of 14 days prior to observation date were randomly matched by farm and date to 10 healthy cows, i.e., cows that were not lame and had no other health event recorded for the matched time period. Sensor data relating to the 14-day time periods were used for developing univariable (using one source of sensor data) and multivariable (using multiple sources of sensor data) models. Model development involved the use of additive logistic regression by applying LogitBoost with a regression tree as base learner. The model's output was a probability estimate for lameness given the sensor data collected during the 14-day time period. Models were validated using leave-one-farm-out cross validation and as a result of this validation each cow in the dataset (318 lame and 3,180 non-lame cows) received a probability estimate for lameness. Based on the area under the curve (**AUC**), results indicated that univariable models had low predictive potential with the highest AUC values found for live-weight (AUC = 0.66), activity (AUC = 0.60), and milking order (AUC = 0.65). Combining these three sensors improved AUC to 0.74. Detection performance of this combined model varied between farms but it consistently and significantly outperformed

univariable models across farms at a fixed specificity of 80%. Still, detection performance was not high enough to be implemented in practice on large, pasture-based, dairy farms. Future research may improve performance by developing variables based on sensor data of live-weight, activity, and milking order but that better describe changes in sensor data patterns when cows go lame.

Key words: sensor data, data mining, dairy cow, lameness detection

INTRODUCTION

Lameness has been grouped with mastitis and infertility as the top three dairy cow health issues related to economic losses in the dairy industry (Juarez et al., 2003). Lameness affects welfare negatively as it is associated with pain (Whay et al., 1997; Bicalho et al., 2007), and decreases farm profitability due to poorer reproductive performance, loss of milk production, and increased costs due to treatment and culling (Tranter and Morris, 1991; Sprecher et al., 1997; Green et al., 2002). Lame cows are usually detected by visual observation of gait and back posture (Sprecher et al., 1997); however, in larger herds, along with the number of cows managed per farm labor unit, visual detection of lame cows becomes more challenging.

Previous studies reported that lameness affects the cow's normal behavior and physiology: lame cows are less active (Juarez et al., 2003; Walker et al., 2008), enter the milking parlor later (Walker et al., 2008), produce less milk (Green et al., 2002), and lose body condition (Walker et al., 2008). Sensing technologies are available that can monitor these behavioral and physiological characteristics of cows on a daily basis. For example, with milk meters and weigh scales, a cow's milk production and live-weight can be regularly monitored. Kamphuis et al. (2013) demonstrated that cows becoming clinically lame have sensor data trends that are significantly different for live-weight, activity, milking order, milk

yield (produced in the first two minutes after teat cup attachment and total milk yield), and milking duration compared to cows that do not become clinically lame. Although there was considerable variation in sensor data values between and within lame and non-lame cows, results indicated that sensor data are potentially useful in the detection of lameness.

Datasets containing sensor data are often noisy, due to sensor drift or malfunctioning, and incomplete due to missing values. Additionally, datasets are often imbalanced as the incidence of lameness is low (Tranter and Morris, 1991; Gibbs, 2010). To analyze this noisy, incomplete, and imbalanced data it is essential that the modeling technique used can process data with these anomalies and can model nonlinear relationships. Examples of these more sophisticated models in the field of lameness detection are neural networks applied by Pastell et al. (2007) and principal component analyses used by Miekley et al. (2013). A model that has not been used in automated detection of lameness is a data-mining technique called decision-tree induction, a commonly used technique for classification problems (Quinlan, 1986) in combination with a boosting process that is known to improve accuracy of classification models (Freund and Schapire, 1996). Decision-tree induction with boosting has proven useful to analyze data with similar anomalies in automated clinical mastitis detection (Kamphuis et al., 2010) where it was able to improve detection performance to a level suggested to be of practical relevance being > 80% sensitivity and > 99% specificity (Hogeveen et al., 2010). These values mean that a model should find at least 80% of the cows that do have clinical mastitis and at the same time indicate fewer than 1% of the healthy cows erroneously. It is unknown what performance targets should be set for a lameness detection model.

The hypothesis for the current study was that sensors currently available on-farm to monitor behavioral and physiological characteristics of dairy cows can be used for the detection of lameness. This was tested by applying a boosting technique based on additive

105 logistic regression (Friedman et al., 2000) in combination with a specific type of decision-tree
106 algorithm (regression tree) to variables derived from one sensor (univariable models) and
107 multiple sensors (multivariable models) and assessing their detection performance using
108 leave-one-farm-out cross validation.
109
110 ## MATERIALS AND METHODS
111 Ethics approval was obtained through the Ruakura Animal Ethics Committee (Ruakura,
112 Hamilton, New Zealand; application number 12210) before commencement of the study.
113 Data were collected from five pasture-based dairy farms in the Waikato region of
114 New Zealand between November 2010 and June 2012 (Table 1). All farms except one
115 applied a seasonal spring-calving regime; one farm had cows calving in spring and autumn.
116 On all farms cows were milked on a rotary milking-platform (Waikato Milking Systems,
117 Hamilton, New Zealand) fitted with automatic weigh-scales and electronic milk meters. All
118 cows had a pedometer (Afikim, Kibutz Afikim, Israel) fitted to one hind leg for measuring
119 cow activity. The pedometers also contained an electronic cow identification (**EID**) unit.
120 Individual cow and sensor data from each milking session were automatically recorded on
121 herd management software (Frontier, Afikim, Kibutz Afikim, Israel) with data files generated
122 daily and transferred via the internet to a central database at DairyNZ (Hamilton, New
123 Zealand). Cow data included cow identification number and DIM. Sensor data at the cow
124 level included (1) live-weight, (2) activity as the average number of steps per hour between
125 milking sessions, (3) milking order, (4) milk yield in the first two minutes after teat cup
126 attachment, (5) total milk yield, and (6) milking duration. Participating farmers were trained
127 (Healthy Hoof Programme, DairyNZ Ltd, Hamilton, New Zealand) by accredited
128 veterinarians in detecting and diagnosing lame cows before the study commenced. When a
129 cow was identified as lame, farmers recorded cow identification number, date of observation,

affected limb, and severity of lameness using a 5-point lameness scoring system (adapted from Sprecher et al., 1997) with scoring categories being (1) Normal, (2) Mildly lame, (3) Moderately lame, (4) lame, and (5) Severely lame. Farmers were visited monthly to collect farmer-recorded data on lameness and during these visits lameness scoring forms were discussed to ensure standardized recording throughout the study period. At the end of the study, data on other health events (e.g., clinical mastitis events and data on artificial insemination or natural breeding events) that occurred during the collection period were extracted from the herd management software.

Data Preparation

Cow and sensor data were automatically recorded in two separate data-sets. The first dataset included information on date, cow identification number, DIM, and data on live-weight and activity measured at both morning and afternoon milkings for each cow for each DIM. The second dataset included date, cow identification, and data on milking order, milk yield in the first two minutes after teat cup attachment, total milk yield, and milking duration. These were also measured during morning and afternoon milking for each cow for each DIM. Milking order was made proportional to the number of cows milked during that particular milking session and, therefore, values for milking order range between 0 and 100. The two datasets were merged by date and cow identification number; data on a particular cow and date that were present in just one of the two datasets were excluded (2.9% of the data). Criteria described in Edwards et al. (2013) were used to identify outlier sensor values for milk yield (produced in the first two minutes of teat cup attachment and total milk yield) and milking duration. Sensor values for live-weight and activity were plotted and conservative threshold values were set based on visual judgment of these plots: values for live-weight less than 250 kg or more than 750 kg were set as missing (< 0.5% of all sensor measurements) as were

values for activity less than 25 or more than 1100 steps per hour (< 0.5% of all sensor measurements). A seventh variable was created in which the milk yield in the first two minutes after teat cup attachment was made proportional to the total milk yield production for a particular cow during a particular milking session on a particular date. Finally, sensor data measured at morning and afternoon milkings were averaged to get one sensor value per day for each variable for each cow in milk. If sensor data were available for one milking only, that value was used for that cow for that day.

In total, 466 lameness events were recorded. Lameness events that had no date of observation recorded were excluded (n = 39). Separate lameness events for the same cow for the same affected limb were defined when the time lag between two lameness events was > 31 days; if the period was ≤31 days then the second lameness event recording was excluded (n = 12). Records on other health events were collected including cow identification number, date of health event and type of health event (including clinical mastitis, artificial insemination, and natural breeding).

Sensor data, information on lameness events and other health information were merged based on cow identification number and date. Lameness events (n = 30) and other health events without any sensor data were excluded.

Defining lame and non-lame Cows

Lame cows were defined as cows with at least one lameness event recorded. To ensure that sensor data used for statistical analyses were not affected by health events other than lameness or by calving events, lameness events were excluded from further analyses when the cow was recorded as having another health event occurring within the interval from 14 days prior ($D_{minus14}$) to the date of detection (D_0) till seven days (D_{plus7}) after detection (n = 27) and when $D_{minus14}$ fell within the first 30 DIM of that lactation (n = 34).

The 22 day period (from $D_{minus14}$ to D_{plus7}) was considered a Lameness Episode. Lameness Episodes with fewer than 10 days of sensor data from $D_{minus14}$ through D_0 were excluded from further analyses (n = 6). Each Lameness Episode was randomly matched by farm and date with 10 non-lame cows, creating Lameness Blocks. Non-lame cows were chosen from those cows without a lameness or health event recorded during the matched time period and with at least 10 days of sensor data from $D_{minus14}$ through D_0. Each Lameness Block, therefore, contains sensor data from one lame cow and 10 non-lame cows. The selection procedure allowed lame cows that experienced a lameness event during the study period to contribute sensor data as a non-lame cow during periods they were not lame and non-lame cows to contribute sensor data to more than one Lameness Block.

Model Development and Validation

From each selected lame and non-lame cow, only sensor data from $D_{minus14}$ till D_{minus1} (that is, the 14 days prior to the day that the lame cow was detected by the farmer) were included for further analyses as management intervention might have influenced sensor data from D_0 onwards. Proportional differences in sensor values were calculated between D_{minus1} and the previous thirteen days for each of the six sensor variables and the one derived variable. The proportional differences with previous days (n = 13 for each variable) together with the absolute sensor data value on D_{minus1} (n = 1 for each variable) were considered as independent variables to be used for model development.

All the data were used for model development and validation; however, by applying leave-one-farm-out cross validation the model was validated using data completely independent from data used for model development. With this cross validation approach, five different models were developed excluding data from one farm in each run; the developed model was then validated using the independent data from the left-out farm. As a result of

this cross validation, each record in the dataset (n = 3,498 consisting of 318 records from lame and 3,180 records for non-lame cows) was used four times for model development and once for validation.

Models were developed using a form of additive logistic regression implemented in an algorithm called LogitBoost (Friedman et al., 2000). LogitBoost uses the boosting process for model development in which a sequence of models or a so-called ensemble model is generated. This ensemble model is built by repeatedly invoking a classification algorithm (the base learner) and by doing so the ensemble model will fit the training data more closely as the number of boosting iterations grows. The number of boosting iterations used to form the ensemble model is user-defined. The boosting process applied by LogitBoost is based on maximizing the conditional likelihood of the ensemble model. The final classification of records is obtained by combining the output of all classification models forming the ensemble model. In the current study, LogitBoost was applied in conjunction with WEKA's REPTree regression tree learner (Hall et al., 2009) as the base learner and the number of iterations was set at 100. A regression tree is a form of decision tree, which is a hierarchical structure consisting of a number of nodes connected by directed edges (Figure 1). Each node apart from the root node has exactly one incoming edge and each node apart from the leaf nodes has at least two outgoing edges. These are the internal nodes. At each internal node, an independent variable is tested to decide which edge to follow. In the case of numeric predictors, as in the application used in the current paper, the record's value for the variable that is tested is compared to a numeric threshold stored at the node. If the value is smaller than the threshold, one proceeds along the first edge emanating from the node, otherwise the second one. The REPTree algorithm grows a regression tree from weighted training records by greedily expanding the tree and choosing tests at internal nodes and numeric outputs at leaf nodes such that squared error is minimized. To obtain the regression tree's output for a

particular record, assessment starts at the root node and traverses along directed edges at the internal nodes until a leaf node containing a numeric value is reached. This numeric value is returned as the tree's output for that record and LogitBoost transforms the output to a probability estimate using

$$P(Lame = 1) = \frac{1}{1 + e^{-x}}$$

where *x* is the numeric output from the first regression tree in the ensemble model. The larger this numeric output, the closer the probability estimate gets to 1; the smaller the value, the closer the probability estimate gets to 0. To calculate the final probability estimate for lameness in the current study, the LogitBoost algorithm combined the output of the trees from 100 iterations to obtain a probability estimate for lameness using

$$P(Lame = 1) = \frac{1}{1 + e^{-(x1 + x2 + \ldots + x100)}}$$

where x_1 to x_{100} are the numeric output values for a record received from a leaf node for each of the 100 regression trees forming the ensemble model. To avoid overfitting the ensemble model to the training data, which would negatively affect predictive performance on new data, the complexity of the trees generated by REPTree was limited in the current study. More specifically, trees were grown such that no more than two edges had to be traversed before reaching a leaf node from the root node. As each edge traversal involved a test on one predictor variable, this meant that at most two independent variables were inspected before a numeric output was returned from a leaf node (note that it was possible for the same independent variable to be tested at multiple nodes with different numeric thresholds in each node, Figure 1).

Model development and validation were done using independent variables derived from one sensor only (univariable models) and using independent variables derived from a combination of sensors (multivariable models). Models were developed to produce a binary

252 outcome for lameness as well as outcomes for specific Lameness Scores. In addition, models
253 were developed using all lame and non-lame cows, and using lame cows with a Lameness
254 Score ≥3 and their matched controls only. It should be noted that the 14 independent
255 variables per sensor are clearly correlated. Correlation may imply that relative prominence of
256 these variables in the model built using LogitBoost does not accurately reflect their true
257 relatively importance. However, in this study, measures of variable importance were not
258 considered; detection performance was solely evaluated when applying the model to
259 independent data.
260
261 ***Performance Measures***
262 The area under the receiver operating characteristic (**ROC**) curve was used to evaluate
263 detection performance of different models. This curve is a graphical representation of a
264 model's true positive rate (or sensitivity; in this study, the proportion of lame cows that were
265 correctly classified as lame by the model) against its false positive rate (in this study, the
266 proportion of non-lame cows that were incorrectly classified as lame by the model). When
267 discriminating between lame and non-lame cows the points on the ROC curve are obtained
268 by changing the threshold value that is used by the model to classify records as lame from the
269 largest possible value to the smallest one (Detilleux et al., 1999; Cortes and Mohri, 2005). In
270 the current study, the additive logistic regression model yielded a probability estimate for
271 lameness for each record in the validation set. These probability estimates were used as
272 possible threshold values to create the ROC curve. When discriminating between Lameness
273 Scores, a different ROC curve can be drawn for each Score by treating that Score as the
274 positive category and the union of the other Scores as the negative category.
275 The area under the ROC curve (**AUC**) summarizes the graphical information into a
276 single quantity. When discriminating between lame and non-lame cows, the AUC can be

277 interpreted as the probability that the model generates a higher probability estimate for a
278 randomly selected lame cow than for a randomly selected non-lame cow (Hanley and
279 McNeil, 1982). Values for the AUC range between 0.5 and 1: when a model produces
280 probability estimates at random it will have an AUC of 0.5 whereas a perfect ranking (i.e., all
281 lame cows receive a higher probability estimate for lameness than the non-lame cows) will
282 yield an AUC of 1 (Swets, 1988; Cortes and Mohri, 2005).

283 Two ROC curves may have different shapes within certain threshold value limits but
284 still have the same AUC value (Detilleux et al., 1999). This makes comparison of models'
285 performances using AUC difficult. As a second evaluation measure, the sensitivity of the
286 model was also calculated at two predetermined specificity levels (80% and 90%), where
287 specificity represents the proportion of non-lame cows that were correctly identified as being
288 non-lame by the model. Differences between models in sensitivities at predetermined
289 specificity levels were tested for significance using analysis of variance (ANOVA) including
290 farm and model as fixed effects.

291 Data cleaning, preparation, selection of lame and non-lame cows, and ANOVA were
292 done using SAS (version 9.2, SAS Institute Inc., Cary, NC). Model building, validation and
293 retrieving AUC values were done in WEKA (version 3.7.7, Waikato University, Hamilton,
294 New Zealand; Hall et al., 2009)

295

296 **RESULTS**

297 Sensor data were collected from ~1.5 million cow-milkings from 4,904 dairy cows (Table 1).
298 There were 385 lameness events with both an observation date and sensor data available; 318
299 of these were eligible for inclusion in the statistical analyses (Table 1). Most (43%) of these
300 eligible lameness events came from Farm 1 (Table 2). The biggest proportion (46%) of the
301 318 lameness events involved cows that were Moderately lame (Score 3), 23% of lameness

events involved cows that were Mildly lame (Score 2) and 26% of the lameness events involved cows that were lame or Severely lame (Score 4 or 5; Table 2).

When all lame and non-lame cows were included, AUC values ranged between 0.51 and 0.66 for univariable models (Table 3). The univariable models including the variables related to milk production had slightly lower AUC values than the univariable models including live-weight, activity or milking order (Table 3). Combining sensor data using variables with an AUC > 0.60 had consistently higher AUC values compared with each univariable model and combining all three variables with an AUC > 0.60 (live-weight, activity and milking order and further referred to as the combined model) improved AUC to 0.74 (Table 3). Compared to this combined model the AUC increased by just 0.04 when all variables were included for model development (Table 3). Using the combined model, AUC values for specific Lameness Scores were 0.59 for Score 2, 0.73 for Score 3 and 0.74 for Score ≥ 4. The AUC for the 15 lameness events without a Lameness Score was 0.52. Analyses were repeated including only lame cows with a Lameness Score ≥ 3 (n = 229) and their matched controls (n = 2,290; Table 3). The general trend was similar for this subgroup: univariable models including variables related to milk production had slightly lower AUC values and combining variables yielded higher AUC values than using univariable variables. The combined model increased the AUC to 0.75 (Table 3) and using this model, the Lameness Score specific AUC values were 0.72 for Score 3 and 0.71 for Score ≥ 4. The differences between AUC values using all lame cows or only those with Lameness Score ≥ 3 were minor.

For the three variables with an AUC > 0.60 and for the combined model, sensitivity levels were calculated at two fixed specificity levels using all lame cows and using lame cows with a Lameness Score ≥3 (Table 4). Sensitivity levels at 80% specificity were greater than at 90% specificity, and sensitivity levels for the combined model were greater than for

the univariable models for both specificity levels. When only lame cows with a Lameness Score ≥3 were included sensitivity levels were greater than when all lame cows were included. The combined model detected 50% of all lame cows at 80% specificity and 30% of all lame cows at 90% specificity. When including lame cows with Lameness Score ≥3 only, sensitivity increased to 55% at 80% specificity and to 40% at 90% specificity. Including all variables in the model resulted in slightly higher sensitivities compared to the combined model (Table 4).

Figure 2 demonstrates the sensitivity of each of the five models at two fixed specificity levels (80% and 90%) resulting from the leave-one-farm-out cross validation, i.e., Farm 1 demonstrates the performance of the model built with data from Farm 2 through 5 and validated with data from Farm 1. Important is that Figure 2 shows that at 80% specificity the combined model consistently outperformed the three univariable models including live-weight, activity, and milking order respectively, and that the model using activity only had the poorest performance on four of the five farms. Statistical analyses showed that the difference in performance between the combined model and each univariable model was statistically significant ($P < 0.01$) and that the model only using activity data performed significantly worse than the three other models ($P < 0.05$). There was no significant difference ($P = 0.59$) between the model using live-weight only and the one using milking order only. Results were different at 90% specificity, where the combined model outperformed the others for four of the five farms, and where activity performed worst in four of the five farms. In general, model performance tended to be less consistent at 90% specificity than at 80% specificity (Figure 2). At 90% specificity, the only significant difference ($P < 0.01$) was between the combined model and the model using activity only. When models were developed and validated including lame cows with Lameness Score ≥ 3 only, the combined model significantly ($P < 0.05$) outperformed the models using activity

and milking order only at 80% specificity, with no other significant differences between the models. At 90% specificity and including lame cows with Lameness Score ≥ 3, the combined model significantly ($P < 0.05$) outperformed the model using activity only and the model using live-weight only. There were no other significant differences between models at this specificity level.

DISCUSSION

Results of this study have confirmed the hypothesis that sensor data available on a growing number of farms are potentially useful for the detection of lameness. Sensor data that gave the best prediction of lameness, based on their AUC value, were live-weight, activity, and milking order. These variables have been associated with lameness in previous studies (e.g., Juarez et al., 2003; Walker et al., 2008). In this study on pastured cows, their AUC values varied between 0.60 and 0.66 and on their own are not sufficiently high to be of practical use. Combining data from these three sensors, however, increased detection performance. The AUC of the combined model was 0.74 when all lame cows were included and 0.75 when only lame cows with a Lameness Score of ≥3 were included (Table 3). This result should be interpreted as the combined model having a 74 to 75 percent probability generating a higher probability estimate for lameness for a randomly selected lame cow than for a randomly selected non-lame cow (Hanley and McNeil, 1982). The slightly better detection performance can be explained by two reasons that occur simultaneously when excluding cows with locomotion score 2 and their controls. The first reason is the decreased risk of mislabeling cows with lameness score ≥3 as being healthy and vice versa. The second reason is that differences in sensor data patterns between lame and non-lame cows are more pronounced with increasing lameness score (Kamphuis et al., 2013). Both reasons are linked with the reduction of noise in the dataset and this reduction will make it easier for an algorithm to

model the data. Adding information based on sensors measuring aspects of milk yield did not improve AUC further. This suggests that sensor data based on milk yield have limited potential for lameness detection on the study farms despite previous reports that milk yield is negatively associated with lameness (Green et al., 2002).

The AUC of 0.74 was achieved by combining sensor data from readily available sensors, some of which are already installed on farm but used for different purposes. Sensors that measure activity are usually installed for automated estrus detection (Roelofs et al., 2005; Hockey et al., 2010) and weigh-scales can be used to monitor changes in live-weight as a guide for adjusting feeding programs (Alawneh et al., 2011). In the current study, this data were combined for a different purpose (detecting lameness), which would remove the need for farmers to invest in additional (expensive) sensors to automate parts of the farm management processes. The approach used in the current study allowed a large amount of sensor data and a large number of lameness events to be collected in a relatively short period of time. The dataset is, therefore, more extensive than in previous reports: 18 cases from one farm (Pastell et al., 2007), 58 cases from one farm (Bicalho et al., 2007), and 210 cases from one farm (Miekley et al., 2013). Furthermore, data collection on multiple farms allowed for leave-one-farm-out cross validation, thus using completely independent data for model validation. The sensors in this study were not calibrated as part of the study design and little data cleaning was done to identify erroneous or malfunctioning sensors. This was done deliberately, to study if sensors as used on-farm could be useful for lameness detection. Even with the noisy data, the combined model outperformed the univariable models significantly at a specificity of 80% and did this consistently across farms. These results suggest that live-weight, activity, and milking order are likely to be the most useful to identify lame cows, and that lameness can be detected by utilizing information already available on farm better.

401 The combined model with an AUC of 0.74 used variables based on proportional
402 differences analyzed with a modeling technique called additive logistic regression. These
403 differences, obtained by comparing sensor measurements on different days, are clearly
404 correlated, and we cannot preclude that further improvements in detection performance are
405 possible using other techniques. Nevertheless, the AUC of the combined model was higher
406 than the 0.60 reported by Miekley et al (2013) who applied principal component analysis to
407 develop a lameness detection model that was allowed to assign alerts for up to three days
408 before cows were diagnosed visually as lame. The AUC of the combined model was also
409 higher than the 0.62 reported by Bicalho et al. (2007) who studied a commercially available
410 lameness detection system based on force-measurements. Bicalho et al. (2007), however,
411 used the cow's reaction to gentle pressure applied by hand to lesions identified at hoof
412 trimming as the gold standard for lameness diagnosis, which is unlikely to be the same as
413 lameness as defined in the current study. Pastell et al. (2007) used a probabilistic neural
414 network to detect lame cows using leg-load distribution and reported an AUC of 0.86. Their
415 model was based on data collected at only one farm with limited numbers of cases for model
416 development (n = 9) and validation (n = 9). As a consequence, their model may have
417 overfitted the data used for development. It would be of great interest to test their model
418 using independent data from other farms to confirm the reported performance.
419 The study farmers were trained to identify lameness but it is still likely that the overall
420 incidence of lameness was underestimated. Previous studies have reported that farmers fail to
421 identify ~75% of lame cows (Whay et al., 2002; Fabian, 2012). Previous work has also
422 reported poor agreement between farmers when assigning locomotion scores (Bicalho et al.,
423 2007). These factors both increase the risk that particularly mildly lame cows (locomotion
424 score 2) are mislabelled as non-lame, hence visually undetected but mildly lame cows were
425 likely to have been selected erroneously as non-lame cows. Despite these caveats, the current

study demonstrated that sensor data were useful to predict lameness, and that AUC values increased with increasing Lameness Score (from 0.59 for Score 2 to 0.74 for Score ≥ 4). This result indicates a level of consistency in diagnosing lameness between the enrolled farmers, most likely due to the training that was carried out before the study commenced.

The current study developed prediction models with a binary outcome (lame vs. non-lame) where lame cows were those with a locomotion score ≥2. Whether the applied threshold to define lame and non-lame cows was the most appropriate one is debatable. The current study considered one alternative and that was to exclude mildly lame cows (locomotion score 2) and their controls. This approach may be attractive for performance evaluation purposes; excluding these cows will decrease the noise and with that improve detection performance of a mode. However, from a practical point of view, these mildly lame cows should not be excluded as they are part of real-life. This study labelled the mildly lame cows as lame and together with the model having a binary outcome seemed a fair approach: when applied in practice, the model would alert for only a small proportion of these mildly lame cows (lameness score 2 had the lowest AUC). This, however, will likely not be a significant problem given the mild symptoms and the likelihood of the cow to self-cure without human intervention. When the model would alert for these mildly lame cows, there is a risk that farmers that have difficulties with diagnosing mildly lame cows will not recognize the cow as being lame and thus perceive the alert as false positive. In this situation, the cow can either self-cure or she will deteriorate and the model will pick her up at a later, more severe lameness state. However, for farmers that are capable of diagnosing mildly lame cows, the alerts will be appreciated as early intervention is possible. Should future research focus on developing a detection model that predicts specific locomotion scores, mislabelling of cows should be minimized as much as possible.

The AUC obtained in the current study compared to previous reports is a promising result; however, it does not mean that the performance of the combined model is high for practical use. When the specificity was fixed at 80%, the combined model had an overall sensitivity of 50% which increased to 55% when only lame cows with Lameness Score ≥3 were included (Table 4). Translating these figures into practice and assuming an incidence rate of 25 cases of lameness per 100 cows per lactation (Gibbs, 2010) a farmer could expect approximately 12 lame cows per month in a 500-cow herd and, by the end of each month, six to seven of these lame cows would be detected correctly. At the same time, however, a specificity of 80% corresponds to approximately 200 false alerts per 1,000 measurements (Sherlock et al., 2008), i.e., 100 cows would be falsely identified as lame each day on a 500-cow herd. Therefore, although promising, translating the results into practice demonstrates that, based on the current sensor-baser variables that represent simple proportional differences, detection performance is not high enough to be implemented on larger pasture-based dairy farms. Future research may improve detection performance by developing variables that better describe the changes in sensor data patterns from live-weight, activity, and milking order when cows go lame.

CONCLUSION

Sensor data that gave the best prediction of lameness, based on their AUC value, were live-weight, activity, and milking order. By combining these three sensor data sources, performance increased and this was consistent across farms. Nevertheless, sensor-based variables explored in the current study did not result in a model with a detection performance high enough for practical implementation. To improve detection performance, future research should focus on developing variables based on sensor data of live-weight, activity, and milking order but that better describe changes in sensor data patterns when cows go lame.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Barbara Dow (DairyNZ Ltd, Newstead, Hamilton 3240, New Zealand) for her statistical support. Also acknowledged are the contributions to the acquisition of the data used in this study by participating farmers. This study was funded by New Zealand Government through the Primary Growth Partnership research programme and by New Zealand dairy farmers through DairyNZ Inc.

REFERENCES

Alawneh, J. I., M. A. Stevenson, N. B. Williamson, N. Lopez-Villalobos, and T. Otley. 2011. Automatic recording of daily walkover liveweight of dairy cattle at pasture in the first 100 days in milk. J. Dairy Sci. 94: 4431 - 4440. http://dx.doi.org/10.3168/jds.2010-4002.

Bicalho, R. C., S. H. Cheong, G. Cramer and C. L. Guard. 2007. Association between a visual and an automated locomotion score in lactating Holstein cows. J. Dairy Sci. 90: 3294 - 3300. http://dx.doi.org/10.3168/jds.2007-0076.

Cortes, C., and M. Mohri. 2005. Confidence intervals for the area under the ROC curve. Pages 305 - 312 in Advances in Neural Information Processing Systems 17. LK Saul, Y. Weiss, L. Bottou (eds.), MIT Press, Cambridge, MA, USA.

Detilleux, J., J. Arendt, F. Lomba, and P. Leroy. 1999. Methods of estimating areas under receiver operating characteristic curves: illustration with somatic-cell scores in subclinical intramammary infections. Prev. Vet. Med. 14: 75 - 88. http://dx.doi.org/10.1016/S0167-5877(99)00054-9.

498 Edwards, J. P., J. G. Jago, and N. Lopez-Villalobos. 2013. Large rotary dairies achieve high
499 cow throughput but are not more labour efficient than medium sized rotaries. Anim.
500 Prod. Sci. http://dx.doi.org/10.1071/AN12312.
501 Fabian, J. 2012. The prevalence of lameness on New Zealand dairy farms: a comparison of
502 farmer perception and mobility scoring. MVS thesis. Institute of Veterinary, Animal
503 and Biomedical Sciences, Massey University, Palmerston North, New Zealand.
504 Freund, Y., and R.E. Schapire 1996. Experiments with a new boosting algorithm. Pages 148 -
505 156 in Proceedings of the Thirteenth International Conference on Machine Learning.
506 L. Saitta (ed.). Morgan Kaufmann Publishers, San Fransisco, CA.
507 Friedman, J., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: a statistical
508 view of boosting. The Annals of Statistics. 28: 337 - 407.
509 Gibbs, S. J. 2010. Dairy Lameness in the South Island. Pages 424 – 427 in Proceedings of the
510 4th Australasian Dairy Science Symposium. Caxton Press, Christchurch, New
511 Zealand.
512 Green, L. E., V. J. Hedges, Y. H. Schukken, R. W. Blowey, and A. J. Packington. 2002. The
513 impact of clinical lameness on the milk yield of dairy cows. J. Dairy Sci. 85: 2250 –
514 2256. http://dx.doi.org/10.3168/jds.S0022-0302(02)74304-X.
515 Hall M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The
516 WEKA Data Mining Software: An Update; SIGKDD Explorations. 11: 10 - 18.
517 Hanley, J., and B. J. McNeil. 1982. The meaning and use of the area under the receiver
518 operating characteristic (ROC) curve. Radiology. 143: 29 – 36.
519 Hockey, C. D., J. M. Morton, S. T. Norman, and M. R. McGowan. 2010. Evaluation of a
520 neck mounted 2-hourly activity meter system for detecting cows about to ovulate in
521 two paddock-based Australian dairy herds. Reprod. Dom. Anim. 45: c107 – c117.
522 http://dx.doi.org/10.1111/j.1439-0531.2009.01531.x.

523 Hogeveen, H., C. Kamphuis, W. Steeneveld, and H. Mollenhorst. 2010. Sensors and clinical
524 mastitis – the quest for the perfect alert. Sensors. 10: 7991 - 8009.
525 http://dx.doi.org/10.3390/s100907991.
526 Juarez, S. T., P. H. Robinson, E. J. DePeters, and E.O. Price. 2003. Impact of lameness on
527 behaviour and productivity of lactating Holstein cows. Appl. Anim. Behav. Sci. 83: 1
528 – 14. http://dx.doi.org/10.1016/S0168-1591(03)00107-2.
529 Kamphuis C, J. K. Burke, and J. G. Jago. 2013. Cows becoming clinically lame differ in
530 changes in behaviour and physiology compared to cows that do not become clinically
531 lame. Proceedings of the New Zealand Society of Animal Production. Accepted for
532 publication.
533 Kamphuis, C., H. Mollenhorst, J. A. P. Heesterbeek, and H. Hogeveen. 2010. Detection of
534 clinical mastitis with sensor data from automatic milking systems in improved by
535 using decision-tree induction. J. Dairy. Sci. 93: 3616 – 3627.
536 http://dx.doi.org/10.3168/jds.2010-3228.
537 Miekley, B., I. Traulsen, and J. Krieter. 2013. Principal component analyses for the early
538 detection of mastitis and lameness in dairy cows. J. Dairy. Res. In press.
539 Pastell, M. E., and M. Kujala. 2007. A probabilistic neural network model for lameness
540 detection. J. Dairy Sci. 90: 2283 – 2292. http://dx.doi.org/10.3168/jds.2006-267.
541 Quinlan, J. R. 1986. Induction of decision trees. Machine Learning. 1: 81 – 106.
542 Roelofs, J. B., F. J. C. M. van Eerdenburg, N. M. Soede, and B. Kemp. 2005. Pedometer
543 readings for estrous detection and as predictor for time of ovulation in dairy cattle.
544 Theriogenology. 64: 1690 – 1703.
545 http://dx.doi.org/10.1016/j.theriogenology.2005.04.004.
546 Sherlock, R., H. Hogeveen, G. Mein, and M. Rasmussen. 2008. Performance evaluation of
547 systems for automated monitoring of udder health: analytical issues and guidelines.

548 Pages 275 – 282 in Proceedings of the International Conference of Mastitis Control:
549 from science to practice. T. J. G. M. Lam (ed.), Wageningen Academic Publishers,
550 Wageningen, the Netherlands.
551 Sprecher, D. J., D. E. Hostetler, and J. B. Kaneene. 1997. A lameness scoring system that
552 uses posture and gait to predict dairy cattle reproductive performance. Theriogenoloy.
553 47: 1179 – 1187. http://dx.doi.org/10.1016/S0093-691X(97)00098-8.
554 Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. Science 240: 1285-1293
555 Tranter, W. P., and R. S. Morris. 1991. A case study of lameness in three dairy herds. NZVJ.
556 39: 88 – 96.
557 Walker, S. L., R. F. Smith, J. E. Routly, D. N. Jones, M. J. Morris, and H. Dobson. 2008.
558 Lameness, activity time-budgets, and estrus expression in dairy cattle. J. Dairy. Sci.
559 91: 4552 – 4559. http://dx.doi.org/10.3168/jds.2008-1048.
560 Whay, H. R., A. E. Waterman, and A. J. F. Webster. 1997. Associations between locomotion,
561 claw lesions and nociceptive threshold in dairy heifers during the peri-partum period.
562 The Vet. J. 154: 155 – 161.
563 Whay, H. R., D. C. J. Main, L. E. Green, and A. J. F. Webster. 2002. Pages 355 – 358 in the
564 Proceedings of the 12th International symposium on lameness in ruminants, Orlando,
565 USA.
566

Table 1. Details of data collected from the five pasture-based Waikato dairy farms including the start date, the number of unique cows milked, the number of cow-milkings, the total number of lameness recordings with sensor data and the number of lameness events included in the statistical analyses after excluding the first 30 DIM, lame cows that had other simultaneous health events and lame cows with less than 10 days of sensor data before lameness was observed

Farm	Start Date	Unique cows (n)	Cow-milkings (n)	Lameness recordings (n)	Lameness events in analyses (n)
1	27/11/2010	914	261,311	164	138
2	04/12/2010	539	185,594	24	22
3	04/12/2010	814	290,956	66	54
4	30/11/2010	543	146,587	73	57
5	03/12/2010	2,094	581,142	58	47
Total		4,904	1,465,590	385	318

Table 2. Number of lameness events per farm included in the analysis and distribution of the Lameness Scores assigned to these events by farmers. Lameness Scores represent the severity of the recorded lameness event with Score 2: Mildly lame, Score 3: Moderately lame, Score 4: lame, and Score 5: Severely lame. Lameness events without a Lameness Score assigned by the farmer were considered missing. Figures between brackets represent the number of lameness events as a percentage of the total number of lameness events (n = 318).

Farm	**Lameness Score**					**Total**
	2	**3**	**4**	**5**	**Missing**	
1	38	66	22	4	8	138 (43)
2	8	6	6	1	1	22 (7)
3	11	27	10	1	5	54 (17)
4	11	24	19	3	--	57 (18)
5	6	22	18	--	1	47 (15)
Total	74 (23)	145 (46)	75 (23)	9 (3)	15 (5)	318

584 **Table 3.** Area under the receiver operating characteristic curve (AUC) values for univariable
585 models and multivariable LogitBoost models for detecting lameness. The AUC is calculated
586 using all lame cows (n = 318) and using lame cows with a Lameness Score ≥ 3 (n = 229)
587 only.

LogitBoost model based on	**AUC**	
	All lame cows	**lame cows with Score ≥ 3**
Univariable analyses		
Live-weight (kg)	0.658	0.654
Activity (average steps taken per hour)	0.603	0.617
Milking order	0.646	0.64
Milk yield (kg) produced in the first 2 minutes	0.543	0.514
Adjusted milk yield (kg) produced in the first 2 minutes[a]	0.508	0.535
Total milk yield (kg)	0.597	0.614
Milk duration (min)	0.537	0.568
Multivariable analyses including variables with AUC > 0.6		
Live-weight x Activity	0.699	0.714
Live-weight x Milking order	0.718	0.712
Activity x Milking order	0.664	0.676
Live-weight x Activity x Milking order (Combined model)	0.735	0.746
Multivariable analyses including all variables[b]	0.739	0.749

588 [a] adjusted for the total milk yield for that particular cow-milking

589 [b] including live-weight, activity (average steps taken per hour), milking order, milk yield in
590 the first two minutes after teat cup attachment, adjusted milk yield in the first two minutes
591 after teat cup attachment, total milk yield, and milking duration

Table 4. Sensitivity (SN) at two fixed specificity (SP) levels (80% and 90%) for three univariable models with an AUC > 0.6, for a multivariable model combining these three variables (combined model) and for a multivariable model using all seven variables. The SNs are calculated using all lame cows (n = 318) and using lame cows with Lameness Score of ≥ 3 (n = 229) only.

LogitBoost model based on	**SN at SP of 80%**		**SN at SP of 90%**	
	All lame cows	**lame cows with Score ≥ 3**	**All lame cows**	**lame cows with Score ≥ 3**
Live-weight (kg)	37.2	38.4	23.4	27.5
Activity (average steps taken per hour)	26.3	38.9	13.7	26.6
Milking order	33.2	43.7	21.1	27.5
Combined model	50.1	55.0	28.6	40.2
All seven variables[a]	48.0	56.8	31.3	41.0

[a] including live-weight, activity (average steps taken per hour), milking order, milk yield in the first two minutes after teat cup attachment, adjusted milk yield in the first two minutes after teat cup attachment, total milk yield, and milking duration

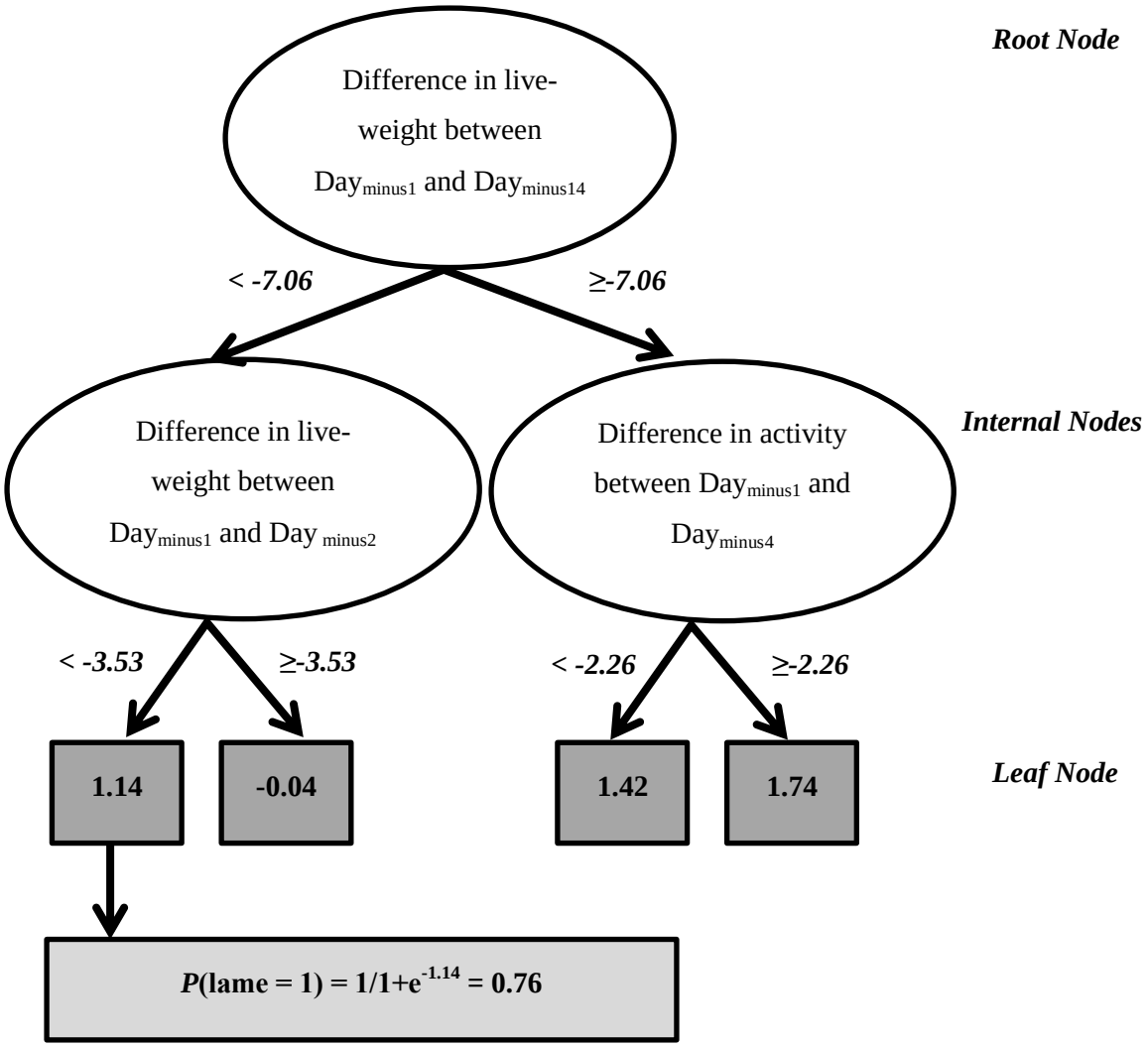
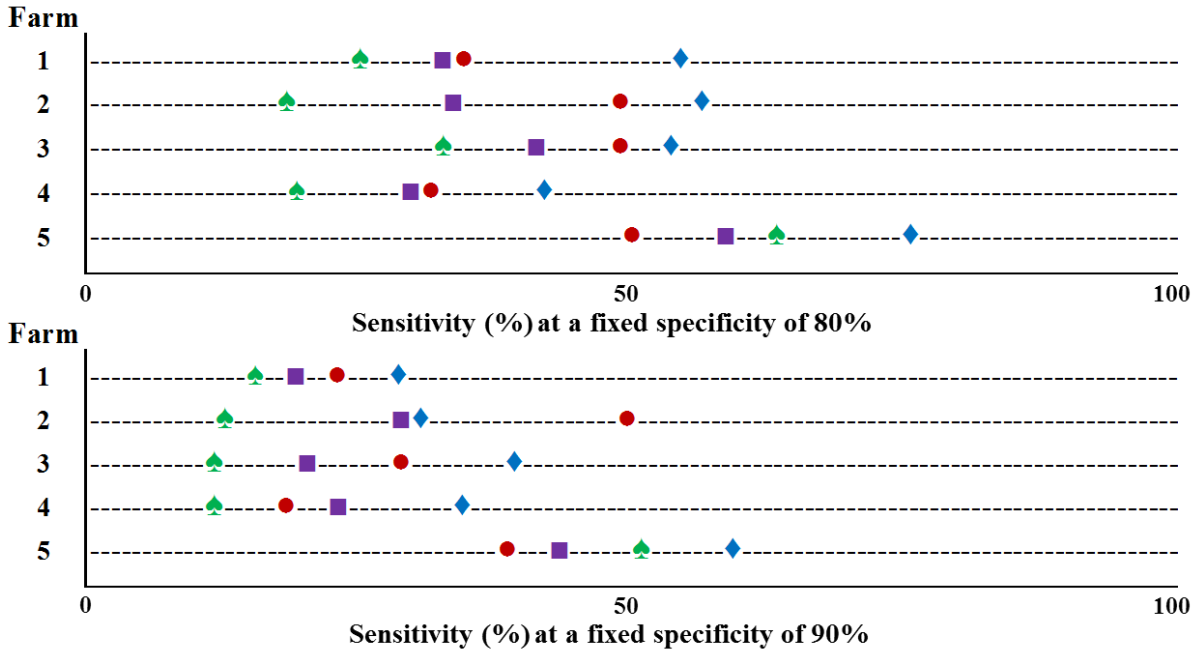


Figure 1. Example of a hypothetical regression tree generated by LogitBoost (Friedman et al., 2000) using a maximum of two independent variables to estimate averaged (weighted and rescaled) residuals at the leaf nodes. These residuals are transformed into a probability estimate for lameness.



627

628 **Figure 2.** Dot-plots of sensitivities at fixed specificities (80% and 90%) for three univariable
629 models (Live-weight, ●; Activity, ♠; Milking order, ■) and a multivariable model combining
630 these three variables (Combined model, ♦) for each farm. Data from all lame cows (n = 318)
631 were used in these calculations