Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for Bovine Respiratory Disease

David E. Amrine, Brad J. White, Robert L. Larson

Published Version Information

Citation: Amrine, D. E., White, B. J., & Larson, R. L. (2014). Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. Computers and Electronics in Agriculture, 105, 9-19.

Digital Object Identifier (DOI): doi:10.1016/j.compag.2014.04.009

Publisher's Link:
http://www.sciencedirect.com/science/article/pii/S016816991400091X

Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for Bovine Respiratory Disease

David E. Amrine[a,b,] Brad J. White[b,c,] Robert L. Larson[b]

[a] Department of Diagnostic Medicine and Pathobiology, College of Veterinary Medicine, Kansas State University, Manhattan, KS 66506

[b] Department of Clinical Sciences, College of Veterinary Medicine, Kansas State University, Manhattan, KS 66506

[c] Address correspondence to Dr. White (bwhite@vet.k-state.edu), 785-532-4243

This manuscript represents a portion of a dissertation submitted by the senior author to the Kansas State University Department of Diagnostic Medicine and Pathobiology as partial fulfillment of the requirements for a Doctor of Philosophy degree.

This project was funded by the National Research Initiative of the Cooperative State Research, Education, and Extension Service, United States Department of Agriculture (Grant #2007-35204-18320).

Abstract

Bovine respiratory disease (BRD) continues to be the primary cause of morbidity and mortality in feedyard cattle. Accurate identification of those animals that will not finish the production cycle normally following initial treatment for BRD would provide feedyard managers with opportunities to more effectively manage those animals. Our objectives were to assess the ability of different classification algorithms to accurately predict an individual calf's outcome based on data available at first identification of and treatment for BRD and also to identify characteristics of calves where predictive models performed well as gauged by accuracy.

Data from 23 feedyards in multiple geographic locations within the U.S. from 2000 to 2009 representing over one million animals were analyzed to identify animals clinically diagnosed with BRD and treated with an antimicrobial. These data were analyzed both as a single dataset and as multiple datasets based on individual feedyards and partitioned into training, testing, and validation datasets. Classifiers were trained and optimized to identify calves that did not finish the production cycle with their cohort. Following classifier training, accuracy was evaluated using validation data. Analysis was also done to identify sub-groups of calves within populations where classifiers performed better compared to other sub-groups.

Accuracy of individual classifiers varied by dataset. The accuracy of the best performing classifier by dataset ranged from a low of 63% in one dataset up to 95% in a different dataset. Sub-groups of calves were identified within some datasets where accuracy of a classifiers were greater than 98%; however these accuracies must be interpreted in relation to the prevalence of the class of interest within those populations. We found that by pairing the correct classifier with the data available, accurate predictions could be made that would provide feedlot managers with valuable information.

Keywords

Bovine respiratory disease

Machine learning

1. Introduction

Bovine respiratory disease continues to be the most important syndrome affecting post-weaned cattle and is associated with approximately 75% of the morbidity and 50% of the mortality in feedyards(Smith, 1998). The overall incidence of BRD has was reported as 14.4% in 1999 and 16.2% in 2011 and the estimated cost of treating a single case of BRD has nearly doubled from $12.59 to $23.60 (USDA, 1999, 2013). Feedlots collect large amounts of individual and cohort level data; however, most of these data are used retrospectively in analysis of trends and to provide guidance for future management practices. Data is frequently collected on individual animals at the time of health events such as treatment for BRD. Previous authors have advocated the use of daily feedlot data in the prediction of overall outcomes (Babcock et al., 2013a). However, there is no literature that uses both individual and cohort level data to make predictions regarding an individual animal's response to treatment. The ability to use real-time information to predict an individual animal's response at the time of respiratory disease treatment would provide tremendous advantages and offer the ability to tailor treatment programs for individual animals based on the estimated accuracies of the individual classifiers.

Using 'smart' sensors coupled with complex mathematical models to aid livestock production is not new (Berckmans, D., 2004) and the application of these technologies to the livestock industry has the potential to aid in the detection of ill animals (Wathes, C.M., 2008); however, most feedyards are not equipped with such monitoring capabilities. Many feedyards do however, collect real-time information on animal treatments and this information could be combined with historical cohort and feedyard data, processed through predictive classification algorithms, and then used these prediction to provide real-time guidance to managers and treatment personnel regarding probabilistic outcomes for individual animals. Our primary

objective was to assess the ability of different classification algorithms to accurately predict an individual calf's post-treatment outcome based on data available at first identification of and treatment for BRD. As many classification algorithms have complex mathematical models, no attempt was made to determine which variables were important to each classifier, only to determine if an algorithm could use the data provided to accurately predict our outcome of interest. Our secondary objective was identification of calf characteristics or situations where predictive models performed well as gauged by accuracy.

2. Materials and Methods

The research strategy involved evaluating of predictive ability of several classification algorithms and data management methodologies both within and across multiple feedyards; therefore, project goals were achieved through an iterative process using multiple datasets. Creation, revision, and evaluation of predictive algorithms based on existing and generated data from multiple sources were accomplished in a stepwise fashion (Fig 1). Multiple datasets were used to create an independent series of classification algorithms (based on training and test data) and to allow comparison of predictive accuracy (validation data).

2.1 Data source

Individual and cohort-level data from 23 feedlots in multiple geographic locations throughout the US were collected on cattle that arrived from 2000 to 2009. A population of cattle (or lot) that were purchased, managed and marketed in a similar fashion was defined as a cohort, although the entire cohort may or may not have been housed in the same pen throughout the production phase. Cohort-level data included demographic characteristics known about the population at feedyard arrival (e.g. arrival date, arrival weight). Individual animal data were collected at the

time a calf was treated for any disease and included characteristics of the individual at that time point (e.g. treatment date, diagnosis, rectal temperature).

2.2 Data preparation

Individual and cohort data were combined into an original dataset containing 1,400,437 event records from 804,631 individual calves within 35,737 cohorts. This dataset contained 27 unique variables and several combination, derived, and redundant variables. The 27 variables consisted of cohort level data and individual information recorded when an animal was pulled for any event such as suspected illness. Cohort level variables available within our dataset for most animals were date they arrived at the feedyard, total number of head within that cohort and average arrival weight (weight of all animals within the cohort divided by total head in that cohort). Some feedyards also recorded the gender of the lot (male, female, mix, or designated the lot as Holstein if they were dairy breeds) and the risk code assigned to that lot (high, medium, low) which represents the feedyard's perceived risk of those animals developing respiratory disease. Some variables such as individual animal weight, and rectal temperature at the time of treatment were not consistently recorded by all feedyards. The goal of this project was to develop and compare the accuracy of models for predicting our outcome of interest, animals that had been treated for BRD with an antimicrobial and did not finish (DNF) the production cycle with their cohort. Our case definition for DNF was similar to previous work describing feedyard mortality (Babcock et al., 2013b) and included any animal that died following BRD treatment or any animal that was removed from the feeding phase prior to cohort harvest following initial treatment for BRD. A binary variable (DNF) was created and populated with values of 0 and 1 (finished the production cycle normally and did not finish normally, respectively). Our study population dataset (SPD) was a subset of the original data and included all calves identified as

being treated for BRD with an antimicrobial. This subset included 468,734 animals of the total 1,400,437 events. Calves could have been diagnosed and/or treated for other conditions prior to or after the initial diagnosis and treatment for BRD. In the SPD, 8.5% (39,699 / 468,734) of the calves did not finish the production cycle normally.

2.3 Variable creation

New variables were derived in an effort to capture predictive characteristics relative to an individual animal's outcome of DNF and thereby enhance predictive models accuracy. Cohort level variables were created that identified trends in the incidence of calves identified as diseased and treated within a cohort over the course of time on feed. For each day a cohort was in the feedyard, variables were created that calculated the daily incidence proportion of calves diagnosed and treated for BRD. The daily incidence proportion of calves identified as ill for any reason within a lot was also calculated. As changes in BRD morbidity over time can be important to understanding an outbreak, cumulative proportions of BRD incidence were calculated for the previous 2, 5, 15, 20, and 30 days for each day a cohort was in the feedyard. Cumulative proportion variables using the same structure were created representing the incidence of all disease within the cohort, not just BRD incidence. Temporal patterns of BRD within cohorts have been shown to be associated with cattle health and performance (Babcock et al., 2009); therefore, we were aiming to capture this information and make it available for the various predictive models.

Variables were created using the previously created cohort incidence proportion variables, querying only those animals within those variables having rectal temperatures at the time of treatment greater than or equal to common industry breakpoints of 39.4, 39.7 and 40 degrees Celsius. Cohort level variables involving the proportion of animals within a lot meeting

certain criteria such as dying or having been identified as diseased for any reason were also created. Previous studies have highlighted the poor diagnostic performance of using clinical signs of illness in combination with rectal temperature; therefore the goal of using incidence proportion variables at different temperature cutoffs was to capture any additional information that might increase the predictive ability of the various classifiers. While the accuracy of rectal thermometers used at the various feedyards is not known, and these variations in accuracy could impact treatment decisions for individual animals, we these measurement inaccuracies to be equally spread across all feedyards and would impact all results equally. Development of all algorithms was based on temperatures reported in Fahrenheit, final values reported in this manuscript were converted to Celsius. Table 1 displays cohort-level variables and specific definitions used to create each variable.

To capture temporal information associated with changes in cohort-level incidence of BRD, variables were created calculating differences in incidence proportions from one time point to another. Changes in the incidence of BRD over time have been associated with cattle health (Babcock et al., 2009). The change from the previous calendar day's incidence proportion of BRD was calculated as well as changes in cumulative incidence proportions from the previous 3, 5, 10, 15, 20, and 30 days. Variables representing changes for incidence rates associated with diseases of all causes within a cohort, were also created.

Derived variables were also created using individual animal information. The type(s) of antimicrobial calves received were classified into one of seven categories: cephalosporins, tetracyclines, fluoroquinolones, macrolides, mixed (animal was treated with more than one class of antimicrobial), ampicillins, or older drugs (e.g. penicillins). A binary variable (0 = no, 1 = yes) was created for all animals indicating if they had received a non-steroidal anti-inflammatory

(NSAIDYN). Variables capturing information relating to the time of year, month, and week when an animal was treated were also created. Table 2 lists all individual animal level variables.

The variable building process resulted in a dataset combining the original and the newly created variables. Each record in the final SPD dataset contained the 27 original variables, 126 newly created cohort-level variables, and 13 newly created individual animal or event variables (Tables 1and 2). Data were not consistent across feedyards and if data were not available to calculate one of the new variables, the resulting fields were treated as null or 0 depending on the variable structure.

Prior to predictive model building, a pair-wise correlation analysis was performed on all variables within the dataset using the linear correlation node within KNIME (Berthold, 2008). If the value of the correlation statistic between any two variables was |0.9| or higher, only one of the variables was selected and included in any subsequent predictive classifiers. Variables identifying a specific feedyard or containing information pertaining to specific dates (i.e. year of feedlot entry, or treatment date) were not used when training classifiers. The goal was to train classifiers that could be used on new data and not be tied to only the original datasets.

2.4 Data partitioning

The SPD dataset was randomly partitioned into training, testing and validation datasets representing 40%, 30%, and 30% of the full dataset, respectively. Within each partitioned dataset, 23 subset datasets each representing only data from an individual feedyard were created. All training, testing and validation steps were performed 24 times, once using the dataset with all feedyards combined (COMBO) and once for each of the individual feedyard datasets (1 thru 23). The validation datasets were saved and evaluated with the final trained algorithms only once to evaluate classifier accuracy.

2.5 Classification algorithms

All classification algorithms were implemented using The Waikato Environment for Knowledge Analysis (WEKA)(Hall, 2009) nodes available as extensions within KNIME.

2.5.1 Decision Trees

Decision tree classification is a learning process that recursively partitions a training dataset and is then used to determine the appropriate class for each example within a test dataset (Zhang, 2012). Each node or branch within a tree splits the data into two or more categories usually based on a single attribute. Each leaf of a node is then assigned to a class that represents the most appropriate target value and calculates a probability that an individual belongs to that node (Rokach, 2005). We evaluated two variations of decision trees classification algorithms, Random forests (RF) and Decision stump (DS). Random forests build several individual classification trees using random samples of the data (i.e. bagging) and then vote for the most popular class (Breiman, 2001). Decision stump finds a single attribute that provides the best discrimination between the classes and then bases future predictions on this attribute (Iba, 1992).

2.5.2 Bayesian networks

In general, Bayesian classifiers estimate the conditional probability distributions of each attribute within the training dataset and then assign cases within the test datasets to the class with the highest posterior probability using Bayes' Theorem (Sebastiani, 2005). We used bayesnet (BN) with a K2 search algorithm and Naïve bayes (NB) with all the default settings in our study. Bayesian network classifiers use directed acrylic graphs where each node in the graph represents a random variable and the edges represent probabilistic dependencies among those random variables. Naïve bayes classifiers analyze the relationship between each variable and the class of interest to then determine a conditional probability for the relationship (Williams, 2006).

2.5.3 Meta-classifiers

Boosting is a type of meta-learning that classifies subsets of the initial training dataset. As each subset is used to train the classifier, the algorithm attempts to use information from the previous subset cases that were incorrectly classified (Vilalta, 2005). Multiboost (MB) and logitboost (LB) are algorithms that are constructed by multiplying the individual conditional probabilities from each feature to get the total probability of a class (Webb, 2000). The class with the highest probability is then selected as the winner. In our application, the base learner for MB was the random tree algorithm and DS was used as the base learner for LB. The filtered classifier (FC) algorithm used a J48 tree algorithm after the data was passed through a discretization filter.

2.5.4 Functions/Neural Networks

The VotedPerceptron (VP) classifier is based on the perceptron algorithm as described by Feund and Schapire in 1999. Neural networks predict outcomes based on relationships between variables that may be complex and multidimensional and are well suited to our data structure, as they do not require *a priori* assumptions about the underlying data structure (Zhang, 2005). The VP takes advantage of data that are linearly separable with large margins (Freund and Schapire, 1999). The VP classifier we used contained all of the default settings.

2.5.6 Statistical methods

Logistic regression models were developed and prediction equations were used to evaluate the test and validation datasets. Only variables significantly associated ($P < 0.05$) with our outcome of interest (DNF) in a univariable screening were included when training our logistic classifiers. No further attempt was made to create a more parsimonious model as the goal was to evaluate logistic regression classification in a similar manner to the other classifiers we had trained.

2.5.7 Classifier selection

Approximately 25 to 30 potential classifiers were trained using their default settings on the COMBO training dataset and initial accuracies were evaluated after classifying the test COMBO data. Classifiers with the lowest accuracies and or lowest sensitivity (Se) or specificity (Sp) values were eliminated. To further eliminate classifiers providing similar information, pairwise correlation coefficients between predicted probabilities from each classifier were calculated using the linear correlation node in KNIME. Classifiers with correlation statistics of > |0.5| were removed leaving twelve classification algorithms. From these 12 algorithms, nine were selected that were representative algorithms belonging to one of five general groups of classifiers evaluated in this study (Decision trees, Bayesian Methods, Meta-classifiers, Functions/Neural Networks, and Statistical). Where possible, individual parameters for each of the nine classifiers were modified, retrained and evaluated using the COMBO test data. The accuracy of the classifier was compared with those from the same classifier using different settings. This procedure was repeated for each classifier until optimal accuracy for each classifier had been achieved using the test dataset.

2.6 Sampling of rare events

Determining the optimal training dataset distribution for the class of interest in respect to classifier performance can be challenging. Previous research has determined that classifier performance varies based on the data structure and classifiers used; however, general conclusions favored balanced (equal number of events and non-events) training datasets to optimize the performance of classification algorithms (Japkowicz, 2000; Weiss, 2003). Overall, calves meeting our case definition (DNF = 1) represented a low proportion (8.5%) of the total number of animals. To evaluate the impact of an un-balanced training dataset on overall classifier accuracy, two different types of balanced training datasets were created; over-sampled

and under-sampled. The over-sampled training dataset was created by selecting all of the calves in our dataset belonging to the minority class (DNF = 1) and creating exact duplicates of them until the distribution of DNF = 1 in the oversampled training data was approximately 50%, or equal to the distribution of our majority class (DNF = 0). The under-sampled training dataset preserves all of the minority class rows and was created by randomly removing rows belonging to the majority class until there are an equal number of rows belonging to both the majority and minority classes within the final dataset (Japkowicz, 2000b). Each classification algorithm was created using the three different training datasets, and accuracy of each classification algorithm was determined with the same test data. Analysis of the variance among dataset sampling techniques was performed on Area Under the receiver operating Curves (AUC) using the Kruskal-Wallis test allowing for multiple comparisons using Steel-Dewass methods in JMP (JMP, SAS Inc.) The sampling technique resulting in the highest AUCs was selected as the method to train all classifiers.

2.7 Classifier accuracy

Classifier accuracy was determined by allowing each algorithm to classify the validation datasets. Classifier predicted probabilities of DNF = 0 and 1 were created for each calf for each classifier. Using these probabilities, receiver-operating characteristic curves (ROC) were created using the LOGISTIC procedure in SAS (SAS 9.3, SAS Institute. Inc.). The ROC curve allows for evaluation of the trade-off between correctly identifying animals that meet the case-definition for DNF (DNF = 1) and falsely identifying DNF = 0 calves as DNF = 1 (false positives) for each classification algorithm (Gardner and Greiner, 2006). As the optimal cutoff varies based on the application of the diagnostic test (classifiers), we elected to identify the point on the ROC curve where Se and Sp are maximized by calculating Youden's index (Youden, 1950) and using the

corresponding classifier generated probability as the cutoff between DNF = 1 and DNF = 0. Youden's index (J) ranges between 0 and 1, with a value of 1 indicating a test with perfect sensitivity and specificity. The J_{max} is the point on the ROC curve that has the greatest vertical distance from the diagonal or chance line (Schisterman et al., 2005).

Logisitic regression models for each classifier were fit in SAS using the LOGISTIC procedure with the animal's true status (DNF) as the dependent variable and the predicted probabilities out of KNIME as the independent variable of interest. The OUTROC statement was included in the MODEL statement to output a dataset for each classifier with all distinct predicted probabilities and their corresponding sensitivity and specificity values. Youden's index (J) = Se + (Sp - 1) was calculated for each possible Se, Sp combination and the J_{max} was identified along with its corresponding probability (P). The probability P represents the cutoff that maximizes Se and Sp for an individual classifier. Final predicted classification for each calf (predicted DNF = 1 or 0) was based on their predicted probability from that specific classifier in relation to P. Calves with predicted probabilities greater than or equal to P were classified as DNF = 1 and all others were assigned DNF = 0. Classifier diagnostic performance was then assessed using the final predicted DNF status to calculate true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), Se, Sp, and accuracy = (TP+TN)/(TP+TN+FP+FN) for each classifier.

2.8 Accuracy among sub-groups within populations

Logistic regression models were employed to evaluate potential changes in classifier accuracy based on sub-groups within each dataset population. For each dataset the classifier that provided the highest overall accuracy after identifying the cutoff yielding the highest combined Se and Sp was selected. A binary variable (CORR) for each calf within a classified dataset was

created and populated with a value of 1 if the classifier predicted DNF status agreed with the true value of DNF for that calf, otherwise the value was 0. For example: if the Naïve Bayes classifier provided the highest overall accuracy for the COMBO dataset, then a CORR variable for each calf was created and populated with a value of 1 where that calf's true status for DNF agreed with the Naïve Bayes prediction, otherwise the variable was populated with 0. Independent variables of interest were gender, arrival weight (WTIN), rectal temperature at treatment (TEMP), days on feed at treatment (TDOF), and the interaction of all variables with TDOF. Gender was categorized into four categories representing all genders of calves in our population with gender information provided; males (MAL), females (FEM), mixed (MIX), and Holsteins (HOL). Arrival weight was categorized into five categories; less than 181kg (400lbs), 181 to 226 kg (400 to 500 lbs), 227 to 272 kg (501 to 600 lbs), 273 to 318 (601 to 700 lbs) and greater than 318 kg (700lbs). Rectal temperature was also categorized into 4 categories; less than 39.1, 39.1 to 39.4, 39.41 to 40 and greater than 40 degrees Celsius. Days on feed at treatment was categorized into animals less than 15 days on feed, 15 to 30 days on feed, 30 to 45 days on feed and greater than 45 days on feed. The true status of each calf, DNF was also offered to each model as an independent variable of interest. The dependent variable of interest was predictive model accuracy or agreement between the model prediction for DNF and true calf status (CORR).

A multi-variable logistic regression model was fit using the GENMOD procedure in SAS with a binary distribution and logit link function. Manual backwards elimination was performed keeping only those variables associated to the outcome at a 5% significance level ($P< 0.05$). Main effects remained in the model regardless of significance if their corresponding interaction terms were significantly associated with the outcome. Least squares means for each predictor

remaining in the model were calculated and then transformed back to probabilities using the formula: P = exp(logit)/(1+exp(logit). Probabilities for a given predictor represented the agreement/accuracy of calves within that group of the population represented by that variable. The true status of the animals DNF if significant (P< 0.05) in the model represented the model adjusted positive predictive value (PPV) for each DNF category.

3. Results

3.1 Descriptive statistics

A total of 468,734 individual calves from 23 different feedyards representing multiple geographic locations in the United States were included in our study population. The mean number of individual animals per feedyard meeting our case definition of having been treated with an antimicrobial for BRD was 20,379 (SE = 3168) with a median of 17,823. The prevalence of DNF = 1 within the COMBO dataset was 8.5 % with a range among feedyards of 0.5% to 14.5% and averaging 9.1% (SE = 0.7%) and a median of 9.9%.

3.2 Comparison of dataset balancing techniques

To determine the optimal training dataset balancing technique for our data, the AUC from each of the nine classifiers were analyzed using the native, under-sampled, and over-sampled test datasets. Variance of the nine classifier AUCs was compared among datasets. There were no differences (P > 0.05) between the native dataset and the over-sampled datasets as well as between the over-sampled and under-sampled datasets. The AUC's using under-sampled data were higher (P < 0.05) when compare to those using the native dataset (Fig. 2). The under-sampled COMBO training dataset contained 15,821 animals in each DNF category and was used to perform all evaluations of classifier accuracy.

3.3 Classification accuracy

The accuracy of the nine classification algorithms was evaluated using the validation data sets. Predictions were generated using the COMBO dataset and individual data sets for each feedyard. Accuracies were based on predictions from each classifier following the use of Youden's index to determine the cutoff that maximizes Se and Sp. Variation in accuracies of each classifier and each dataset are displayed in Fig 3. Accuracies of the nine classifiers using the COMBO dataset ranged from 52% to 77% for BN and NB, respectively. Classifier accuracy using dataset 23 ranged from 6% (DS) to 79% (LB). While classification of the dataset 23 using the DS algorithm resulted in a sensitivity of 99% (not displayed) the algorithm predicted almost every calf as DNF = 1 (4097/4156). The prevalence of DNF = 1 in dataset 23 was 4.7% (197/4156). The FC algorithm achieved the highest accuracy of 95% when applied to dataset five; however the prevalence of DNF = 1 in this dataset was less than 1% (21/3957) (Table 3). Of the 24 datasets analyzed, six achieved accuracies greater than 80% using one of the nine classification algorithms with 50% (3/6) of those algorithms using classifiers with Bayesian network architecture. Logistic regression was the highest performing classifier in only one dataset and overall accuracy was 73% in a population where the prevalence of DNF = 1 was high (14.5%).

3.4 Sub-group analysis

Logistic regression was employed to evaluate potential sub-groups within each dataset where classification accuracy was better than overall accuracy. The classifiers with the highest overall accuracy by dataset were analyzed to identify these potential sub-groups. For the COMBO dataset, the main effects of gender, WTIN, TEMP, TDOF and DNF were significantly associated ($P < 0.05$) with classifier accuracy as well as all interactions with TDOF (Table 4). However, for dataset 15, DNF was the only effect associated ($P < 0.05$) with agreement. Overall, TDOF or the interaction of WTIN and TDOF were associated ($P < 0.05$) with agreement/accuracy in 22 of

the 24 models. Rectal temperature recorded at treatment was associated ($P < 0.05$) with agreement/accuracy in 16 of the 24 models.

Prevalence of calves DNF = 1 in our study was low, 8.5% overall and varied by dataset. Sub-groups within dataset populations with model adjusted significant ($P < 0.05$) accuracies greater than 1-prevalence represent calves within that population where classifiers performed better than guessing all animals would finish the production cycle normally. Using the BN classifier on calves in dataset 1, the accuracy of predicting DNF for lightweight calves on arrival (less than 181 kg) that were 15 to 30 days on feed at their initial treatment for BRD was 95 ± 4 % while the overall prevalence in this population was 12%. The prevalence of DNF = 1 in dataset five was less than 1%; however, accuracy using the FC algorithm was near 100% in all categories of gender, TEMP, WTIN, and TDOF (Table 5). For calves in dataset 19 that were greater than 45 TDOF and had rectal temperatures greater than 40 °C the MB classifier was over 97 ± 2 % accurate in identifying calves DNF = 1 (Table 5).

4. Discussion

Bovine respiratory disease continues to adversely impact cattle health with an estimated 16.2% of all cattle placed in feedlots showing signs of respiratory disease at some point during the feeding period (USDA, 2013). Accurately predicting health outcomes is an important component in increasing performance within feedyards (Babcock et al., 2013a; Corbin and Griffin, 2006). Characteristics of cohorts and individual animals upon arrival and individual animal treatment records are frequently recorded. These data have previously been analyzed for risk factors associated with developing BRD;(Babcock et al., 2009; Step et al., 2008) however, to our knowledge, using this information to make prognostic predictions for an individual animal at the time of first treatment for BRD has not been reported. Others have also advocated the use of

mathematical modeling to aid in animal disease detection and predicting animal responses (Wathes, C.M., 2008). In this study we evaluated the ability of several classification algorithms to accurately predict calves within cohorts that would not finish the production cycle normally. Accuracy of classification algorithms was relatively low using combined data from all feedyards; however, when applied to datasets representing individual feedyards, accuracy of some classifiers improved. This is not surprising given the inconsistency in data recorded among feedyards. There were sub-groups of calves within individual datasets where classifier accuracy was quite good considering the prevalence of animals meeting our case definition was relative low in most datasets.

When learning from imbalanced datasets, some classifiers can learn to provide adequate distinction between FPs and FNs while others simply learn to predict the majority class (Maloof, 2003). Sampling techniques have been developed to minimize the impact of learning with imbalanced data by changing the distributions within the training sets (Maalouf and Trafalis, 2011). Several methods have been proposed to handle imbalanced datasets and results differ based on the classification algorithm chosen;(Japkowicz, 2000b) however, with large datasets it is generally accepted that a balanced class distribution performs better than un-balanced (Weiss G.M., 2003). Two frequently used techniques involve under-sampling and over-sampling on the class of interest. While sampling techniques may provide benefits in accuracy they can also introduce bias due to choice based sampling. However, if this choice based sampling introduces bias that impacts a classifiers ability to accurately predict minority events, then overall accuracy should be negatively impacted when validation data (naturally imbalanced) is classified. There were minor differences in classification accuracy from test to validation data (data not shown) indicating the choice based sampling method used to train classifiers was not introducing an

important source of bias. In our study, following validation, only those classifiers with the highest within-dataset accuracies were used for subsequent sub-group analyses.

Previous research has discovered that training classifiers on imbalanced data frequently produces classifiers that favor the majority class (Weiss G.M., 2003). In our study, the prevalence of calves that received treatment for BRD and then DNF was relatively low (8.5%) resulting in an imbalanced dataset. We evaluated the AUC for each of the classifiers using 3 versions of the COMBO dataset (native, under-sampled, over-sampled). Area under the curve, unlike accuracy, provides a measurement of a classifiers abilities using all possible cutoffs in Se and Sp and has previously been used to distinguish among sampling methods (Maloof, 2003). While no difference was found in AUCs between the native and over-sampled datasets, we found AUC's were significantly higher ($P < 0.05$) using the under-sampled data in relation to our native distribution. Under-sampling and over-sampling techniques are both appropriate methods to balance the class distribution; however, some have noted that over-sampling can lead to over-fitting due to making exact copies of the minority class records (Weiss, 2003). This potential of over-fitting was avoided by using under-sampled data to train all classifiers in our study.

Area under the ROC curve does summarize the ROC curve and provides an overall method to discriminate among potential classifiers; however, it does not directly supply a classifiers predictive ability (accuracy) given a specific trade-off in Se and Sp (Greiner et al., 2000). Given the imbalanced nature of datasets and the complex nature of BRD within feedyards, the cost of a FP is likely not the same as that of a FN. The impact of predicting a calf would not finish the production cycle normally could involve changes in management procedures for that animal that minimize further expenses. False positive calves represent those where significant economic loss could be realized due to lost potential while FN animals

managed normally could result in increased expenses of feed and treatments that will provide negative returns on investment. We attempted to minimize FPs and FNs by altering the decision threshold for each classifier by using the point on the ROC curve furthest from the line of chance (Fluss et al., 2005). While this decision threshold may not be the most appropriate for every situation (i.e. the cost difference in a FN in relation to FP), by selecting the same threshold across all classifiers we were able to compare predictive ability among classifiers using overall accuracy.

We evaluated the accuracies of multiple classifiers to predict our outcome of interest (DNF = 1). Accuracies varied by classifier within a dataset and among datasets. Within datasets, the variation in classifier accuracies ranged from 15% up to 75%; however, in some datasets where variation in accuracy was relatively low (15%), the accuracy of the best classifier was only 63%, indicating in this dataset, classifiers lacked the appropriate training to provide useful predictions. Accuracy of individual classifiers varied by greater than 49% considering all the datasets. Filtered classifier for example, achieved the highest overall accuracy of classifiers evaluated when using dataset five and was only 5% accurate using dataset 23. Bayesian network and meta-classifiers each achieved the highest within dataset accuracies in 33% (8/24) of datasets but no one type of classifier outperformed all the others. These large variations in accuracies within and among datasets as well as differences in individual classifier accuracies across datasets are likely due to differences in variables recorded at each feedyard (represented by different data available among datasets) as well as differences in management practices at each feedyard that were not represented within these data. The complex epidemiology of BRD within feedlot production systems makes it highly plausible that management of BRD varies by

feedyard (Taylor et al., 2010). Differences in accuracies discovered here highlight the importance of pairing the classifier that works bests given the data available.

Evaluating accuracy alone can lead to misleading results when the class of interest within the dataset is imbalanced (Chawla, 2005). In our study, the overall prevalence of DNF =1 in the COMBO dataset was 8.5% and ranged from < 1 % to 14.5 % in individual datasets. As mentioned previously, the FC algorithm achieved an accuracy of 95% within one dataset (5) and appeared to be useful in identifying calves meeting our case definition. However, further analysis reveals the prevalence of calves within dataset five of DNF = 1, was less than 1%. A default strategy of guessing every calf in this population will finish the production cycle normally would have resulted in a predictive accuracy of greater than 99%. Therefore, in this population, to achieve performance better than guessing a classifier would need to be greater than 99% accurate.

Overall accuracies of individual classifiers and for individual datasets within our study were relatively low; however, we identified sub-groups of calves within some datasets where classification accuracy was considered good (greater than 1- prevalence of DNF = 1). The characteristics that we used to discriminate among sub-groups within dataset (gender, WTIN, TEMP, TDOF) would be known at time of first pull and could be used to guide selection of the appropriate classifier given characteristics of that specific animal. Using a Bayesnet algorithm on dataset 1, accuracies were greater than 1-prevalence for all categories of WTIN although they were modified by the TDOF. This makes sense, as we would expect animals visually identified as suffering from BRD would express different clinical signs based on the amount of time they have been in the feedyard. We also found that in some datasets using an animal's TEMP and TDOF provided accuracies that were better than (1-prev of DNF =1) for that yard. While these

four sub-groups provided insight into instances where classifiers performed well, there are likely other sub-groups that could be analyzed and included in sub-group analysis because several risk factors have previously been associated with the risk of developing BRD (Babcock et al., 2010_ENREF_3; Cernicchiaro et al., 2012). By understanding sub-groups of cattle where classification algorithms are known to perform well, one could tailor classifier selection at the time of treatment based upon the characteristics of the population . Further research is needed to more clearly define these populations and specific classifiers that would optimize prediction performance.

5. Conclusion

The objective of this study was to evaluate the ability of several different classification algorithms to identify individual calves that would not finish the production cycle normally. As with many real-world classification problems, these data with respect to our class of interest were highly imbalanced. Under-sampling dataset balancing was performed prior to classifier training to give algorithms the best opportunity to learn the class of interest. We compared the ability of these classifiers by using accuracy after adjusting the decision threshold for each classifier by maximizing Se and Sp in relation to each other and found it varied not only by classifier but also by the data analyzed. We identified sub-groups within each population where specific classifiers performed well considering the prevalence of our class of interest. These sub-groups of calves were based on demographic characteristics available at the time an animal was pulled and treated for BRD indicating there are specific characteristics that could be used to tailor classification accuracy.

Information recorded among feedyards was not always consistent as was partially evident in the ranges of accuracy for individual classifiers among datasets. The predictive accuracy of a

classifier is directly related to the data provided when training. If information provided during training does not help distinguish the class of interest then results of classification using validation data will not be useful. Methodology used here has provided insight into the capability of predictive models to be used in a production setting and the importance of pairing the data available with the correct classifier can lead to accurate predictions of calves of interest. Many feedyards do collect vast amounts of cohort and individual animal data and we have outlined one potential method of using that data in a real-time setting to make treatment decisions. Classification methods described could be used with currently available data allowing feedyards to more accurately tailor their treatment protocols based on the probability and animal would finish the production cycle normally or not.

Literature Cited

Babcock, A., White, B., Renter, D., Dubnicka, S., Scott, H.M., 2013a. Predicting cumulative risk of bovine respiratory disease complex (BRDC) using feedlot arrival data and daily morbidity and mortality counts. Can J Vet Res 77, 33-44.

Babcock, A.H., Cernicchiaro, N., White, B.J., Dubnicka, S.R., Thomson, D.U., Ives, S.E., Scott, H.M., Milliken, G.A., Renter, D.G., 2013b. A multivariable assessment quantifying effects of cohort-level factors associated with combined mortality and culling risk in cohorts of U.S. commercial feedlot cattle. Prev Vet Med 108, 38-46.

Babcock, A.H., Renter, D.G., White, B.J., Dubnicka, S.R., Scott, H.M., 2010. Temporal distributions of respiratory disease events within cohorts of feedlot cattle and associations with cattle health and performance indices. Prev Vet Med 97, 198-219.

Babcock, A.H., White, B.J., Dritz, S.S., Thomson, D.U., Renter, D.G., 2009. Feedlot health and performance effects associated with the timing of respiratory disease treatment. J Anim Sci 87, 314-327.

Berckmans, D., 2004. Automatic on-line monitoring of animals by precision livestock farming. In: Proceedings of the ISAH Conference on Animal Production in Europe: The Way Forward in a Changing World, vol. 1, Saint-Malo, France, October 11-13, pp. 31-37.

Berthold, M. R., N.C., Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, Bernd Wiswedel, 2008. KNIME: The Konstanz

Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization 2008. Springer Berlin Heidelberg, pp. 319-326.
Breiman, L., 2001. Random Forests. Mach Learn 45, 5-32.
Cernicchiaro, N., White, B.J., Renter, D.G., Babcock, A.H., Kelly, L., Slattery, R., 2012. Associations between the distance traveled from sale barns to commercial feedlots in the United States and overall performance, risk of respiratory disease, and cumulative mortality in feeder cattle during 1997 to 2009. J Anim Sci 90, 1929-1939.
Chawla, N.V., 2005. Data mining for Imbalanced Datasets: An overview, In: Maimon, O., Rokach, L. (Ed.), Data Mining and Knowledge Discovery Handbook. Springer, New York.
Corbin, M.J., Griffin, D., 2006. Assessing performance of feedlot operations using epidemiology. Vet Clin North Am Food Anim Pract 22, 35-51.
Fluss, R., Faraggi, D., Reiser, B., 2005. Estimation of the Youden Index and its associated cutoff point. Biometrical journal. Biometrische Zeitschrift 47, 458-472.
Freund, Y., Schapire, R.E., 1999. Large margin classification using the perceptron algorithm. Mach Learn 37, 277-296.
Gardner, I.A., Greiner, M., 2006. Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. Vet Clin Pathol 35, 8-17.
Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med 45, 23-41.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. SIBKDD Explorations 11.
Iba, W.L., P., 1992. Induction of One-LevelDecision Trees, Ninth International Conference on Machine Learning, Aberdeen.
Japkowicz, N., 2000. Learning from Imbalanced Data sets: A comparison of Various Strategies., AAAI' 2000 Workshop on Learning from Imbalanced Data Sets, , Austin, TX.
Maalouf, M., Trafalis, T.B., 2011. Robust weighted kernel logistic regression in imbalanced and rare events data. Comput Stat Data An 55, 168-183.
Maloof, M.A., 2003. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown, Workshop on Learning from Imbalance Data Sets II, Washington DC.
Rokach, L., Oded, Maimon, 2005. Decision Trees, In: Maimon, O., Rokach, L. (Ed.), Data Mining and Knowledge Discovery Handbook. Springer, New York.
Schisterman, E.F., Perkins, N.J., Liu, A., Bondell, H., 2005. Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples. Epidemiology 16, 73-81.
Sebastiani, P., Abad, M.M., Ramoni, M.F., 2005. Bayesian Networks, In: Maimon, O., Rokach, L. (Ed.), Data Mining and Knowledge Discovery Handbook. Springer, New York.
Smith, R.A., 1998. Impact of disease on feedlot performance: a review. J Anim Sci 76, 272-274.
Step, D.L., Krehbiel, C.R., DePra, H.A., Cranston, J.J., Fulton, R.W., Kirkpatrick, J.G., Gill, D.R., Payton, M.E., Montelongo, M.A., Confer, A.W., 2008. Effects of commingling beef calves from different sources and weaning protocols during a forty-two-day receiving period on performance and bovine respiratory disease. J Anim Sci 86, 3146-3158.

Taylor, J.D., Fulton, R.W., Lehenbauer, T.W., Step, D.L., Confer, A.W., 2010. The epidemiology of bovine respiratory disease: What is the evidence for predisposing factors? Can Vet J 51, 1095-1102.
USDA, 1999. Part III: Health Management and Biosecurity in U.S. Feedlots, p. 25.
USDA, 2013. Types and Costs of Respiratory Disease Treatments in the U.S. Feedlots, In: USDA:APHIS:VS (Ed.). National Animal Health Monitoring System, Fort Collins, CO.
Vilalta, R., Giraud-Carrier, C., Brazdil, P., 2005. Meta-Learning Concepts and Techniques, In: Maimon, O., Rokach, L. (Ed.), Data Mining and Knowledge Discovery Handbook. Springer, New York.
Wathes, C.M., Kristensen, H.H., Aertjs, J.-M., Berckmans, D. 2008. Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall? Comput Electron Agric 64, 2-10.
Webb, G.I., 2000. MultiBoosting: A technique for combining boosting and wagging. Mach Learn 40, 159-196.
Weiss G.M., P.F., 2003. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. Journal of Artificial Intelligence Research, 315-354.
Williams, N., 2006. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. Computer communication review 36, 5-16.
Youden, W.J., 1950. Index for rating diagnostic tests. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 3, 32-35.
Zhang, G.P., 2005. Neural Networks for Data Mining, In: Maimon, O., Rokach, L. (Ed.), Data Mining and Knowledge Discovery Handbook. Springer, New York.
Zhang, S., 2012. Decision tree classifiers sensitive to heterogeneous costs. J Syst Software 85, 771-779.

List of Figures

Fig 1. Schematic flow of data refinement, partitioning and classification algorithm evaluation

[a] Study Population dataset

Fig 2. Box and whisker plots of classifier[a] Area under the Receiver-operating characteristic Curve from three different versions of the full dataset each with different distributions of the classification variable of interest (percent of calves that did not finish feeding period within cohorts). Native (raw, observed data; 8.5% rate of calves not finishing the production cycle normally; total n= 187,493), Over-sample (duplicates of calves within the minority class until distributions are balanced; 50% calves did not finish; total n= 374,986, Under-sample (removal

of records from the majority class until distributions are balanced; 50% calves did not finish; total n= 31,642).

[a] BN = Bayesnet, DS = Decisionstump, FC = Filteredclassifier, LB = Lobitboost, LR = Logistic Regression, MB = Multiboost, NB = Naïve Bayes, RF = Random forest, VP = VotedPerceptron Datasets with different letter superscripts represent differences ($P < 0.05$) determined using Kruskal-Wallis analysis of variance accounting for multiple comparisons using the Steel-Dewass method.

Fig 3. Box and whiskers plots of accuracies for nine classification algorithms[a] by dataset[b]. Boxes represent the 25th and 75th quartiles and whiskers span from minimum to maximum accuracy values.

[a] BN = Bayesnet, DS = Decisionstump, FC = Filteredclassifier, LB = Lobitboost, LR = Logistic Regression, MB = Multiboost, NB = Naïve Bayes, RF = Random forest, VP = VotedPerceptron

[b] COMBO represents combined dataset with all feedyards. Individual numbers represent datasets containing one feedyard

Table 1. Cohort animal level variables

Variable	Description
arrivalmonth	Month of lot arrival (1,2,3,4,5,6,7,8,9,10,11,12)
arrivalquarter	Quarter of the year of lot arrival (1,2,3,4)
arrivalyear	Year of lot arrival
brdcasestothispoint	Sum of animals diagnosed with BRD and administered an antimicrobial as of the previous treatdate
distbrdcasestothispoint	sum of distinct animals diagnosed with BRD and administered an antimicrobial as of the previous treatdate
treatment failure #1 (txfailure#1)	1 = any animal requiring re-treatment for BRD or didnotfinish after receiving antimicrobial treatment for BRD
treatment failure #2 (txfailure#2)	1 = any animal requiring treatment with an antimicrobial for any reason after an initial treatment for BRD or didnotfinish after receiving antimicrobial treatment for BRD
treatment failure #3 (txfailure#3)	1 = any animal being pulled for any event after their initial antimicrobial treatment for BRD
treatment failure #4 (txfailure#4)	1 = same as txfailure#3, but includes animals on the same event day
1st treatment success rate #1 (1ssttxsuccessrate_p1)	((distbrdcasestothispoint -sum of txfailure#1) / distbrdcasestothispoint) * 100
1st treatment success rate #2 (1ssttxsuccessrate_p2)	((distbrdcasestothispoint -sum of txfailure#2) / distbrdcasestothispoint) * 100
1st treatment success rate #3 (1ssttxsuccessrate_p3)	((distbrdcasestothispoint -sum of txfailure#3) / distbrdcasestothispoint) * 100
1st treatment success rate #4 (1ssttxsuccessrate_p4)	((distbrdcasestothispoint -sum of txfailure#4) / distbrdcasestothispoint) * 100
propbrdcasestothispoint	Proportion of BRD cases to this point = (brdcasestothispoint/headin)*100
propdistcaestothispoint	Proportion of distinct BRD cases to this point = (distbrdcasestothispoint/headin)*100
propbrdcasestothispoint (time and temperature cutoffs)	(brdcasestothispoint/headin)* 100; New variable created for each combination of day (2,3,5,10,15,20,30) and temperature cutoffs (≥ 103, ≥ 103.5, ≥ 104)
propdistcaestothispoint	(distbrdcasestothispoint/headin)* 100; New variable created for each combination of day

(time and temperature cutoffs)	(2,3,5,10,15,20,30) and temperature cutoffs (≥ 103, ≥ 103.5, ≥ 104)
deathstothispoint	Sum of deaths to this point
propdeathstothispoint	Proportion of deaths to this point = (deathstothispoint / headin) * 100
propdeathstothispoint (time cutoffs)	(deathstothispoint / headin) * 100 New variable created for each combination of the previous days (2,3,5,10,15,20,30)
propdailybrdpulls	Proportion of lot pulled for BRD on this event day = (Dailybrdpulls/headin) *100
propdailyallpulls	Proportion of lot pulled for any reason on this event day = (Dailyallpulls/headin) * 100
deltapropdailybrdpulls	Change in propdailybrdpulls from previous calendar day
deltapropdailyallpulls	Change in propdailyallpulls from previous calenday day
deltapropdailybrdpulls (time cutoffs)	Change in propdailybrdpulls for the previous (2,3,4,10,15,20,30 days)
deltapropdailyallpulls (time cutoffs)	Change in propdailyallpulls for the previous (2,3,4,10,15,20,30 days)
dailyyardpopulation	Total number of head on feed for that calendar day
propryardbrdpulls	Proportion of yard pulled for BRD on this event day (totalbrdpullsforyard/dailyyardpopulation)*100
deltapropyardbrdpulls	Change in propyardburdpulls from previous calendar day
deltapropyardbrdpulls (time cutoffs)	Change in propyardburdpulls for the previous (2,3,4,10,15,20,30 days)
Exponential moving averages (EMA)	EMAs were calculated for the average daily BRD pulls for the lot for 3, 5,10,15,20, and 30 days
Moving average convergence divergence (MACD)	MACDs were calculated for differences in all EMA combinations

Table 2. Individual animal level variables and their origin

Variable	Description
tagno	Individual calf's tag number
treatdate	Date calf was treated
eventno	Running count of events per calf
diag	Diagnosis assigned to calf for this event
dxcode	Diagnosis code assigned to calf for this event
dieddxcode	Diagnosis code for calf's death
antimicrobial	1 = antimicrobial was administered at this event, 0 = no antimicrobial administered
deaddate	Date animal died if known and recorded
didnotfinish	1= calf did not finish production cycle, 0 = finished production cycle
treatdof	Days on feed at treatment for this animal
wt	Individual animal weight
temp	Rectal temperature
overallclass	Class of antibiotic administered (CEPH, TET, FLU, MAC, MIX, AMP, OLD)
nsaidyn	1 = NSAID administered during this event, 0 = no NSAID administered
brdabxyn	1 = calf diagnosed with BRD and administered an antimicrobial (denominator of CFR)
treatnobrdabxyn	Count of the times this calf has been diagnosed with BRD and treated with an antimicrobial
abxtreatno	Running total of the number of times calf treated with antimicrobial for any reason
eventspriortobrdyn	1 = calf had events recorded prior to brdabxyn = 1, 0 = no events recorded prior to brdabxyn = 1
nbreventspriortobrd	Sum of eventspriortobrdyn
pulldayofweek	The day of the week for this event (M,Tu, We, Thr, F, Sa, Su)
pullweekdayYN	1 = animal pulled on a weekday, 0 = pulled on a weekend
pullonmondayYN	1 = animal pulled on a Monday, 0 = pulled any other day
pullmonth	Month of the year for this event (1,2,3,4,5,6,7,8,9,10,11,12)
pullquarter	Quarter of the year for this event (1,2,3,4)
pullyear	Year the calf was pulled

Table 3. Diagnostic performance of classifiers[a] achieving the highest accuracy by dataset

Dataset	Classifier	TP[b]	FP[b]	TN[b]	FN[b]	Sensitivity	Specificity	Accuracy	Prev DNF=1[c]
0	NB	3240	24345	104322	8714	27.1%	81.1%	76.5%	8.5%
1	BN	315	1512	6234	770	29.0%	80.5%	74.2%	12.3%
10	NB	155	1039	1768	170	47.7%	63.0%	61.4%	10.4%
11	BN	7	33	994	102	6.4%	96.8%	88.1%	9.6%
12	BN	73	275	2466	206	26.2%	90.0%	84.1%	9.2%
16	NB	235	1786	4875	575	29.0%	73.2%	68.4%	10.8%
18	NB	55	211	3192	388	12.4%	93.8%	84.4%	11.5%
21	NB	297	1780	7474	655	31.2%	80.8%	76.1%	9.3%
7	VP	103	745	4034	459	18.3%	84.4%	77.5%	10.5%
17	VP	312	2033	6637	664	32.0%	76.6%	72.0%	10.1%
4	LB	221	1149	1568	135	62.1%	57.7%	58.2%	11.6%
5	FC	18	191	3745	3	85.7%	95.1%	95.1%	0.5%
13	FC	329	2604	6900	407	44.7%	72.6%	70.6%	7.2%
14	LB	68	578	2152	46	59.6%	78.8%	78.1%	4.0%
15	LB	6	19	154	23	20.7%	89.0%	79.2%	0.6%
19	MB	400	4576	15508	640	38.5%	77.2%	75.3%	4.9%
22	LB	379	2772	4776	271	58.3%	63.3%	62.9%	7.9%
23	LB	62	724	3235	135	31.5%	81.7%	79.3%	4.7%
8	LR	109	394	1521	216	33.5%	79.4%	72.8%	14.5%
2	DS	38	141	4997	603	5.9%	97.3%	87.1%	11.1%
3	RF	337	2104	3824	329	50.6%	64.5%	63.1%	10.1%
6	DS	90	291	1330	106	45.9%	82.0%	78.2%	10.8%
9	DS	85	297	6364	725	10.5%	95.5%	86.3%	10.8%
20	DS	257	1532	8453	707	26.7%	84.7%	79.6%	8.8%

[a] BN = Bayesnet, DS = Decisionstump, FC = Filteredclassifier, LB = Lobitboost, LR = Logistic Regression, MB = Multiboost, NB = Naïve Bayes, RF = Random forest, VP = VotedPerceptron

[b] TP = true positives, FP = false positive, TN = true negatives, FN = false negatives

[c] Prevalence of calves within each dataset that did not finish (DNF) the production cycle normally

Table 4. Type 3 fixed effects results of logistic regression models evaluating the associations between algorithm accuracy and gender[a], temperature category (temp_cat)[b], arrival weight category (wtin_cat)[c], days on feed at first treatment category (tdof_cat)[d] and selected interactions of gender, temp_cat, wtin_cat all with tdof_cat.

Dataset	FYYDNO	Fixed effects offered to each model							
		DNF	gender	temp_cat	wtin_cat	tdof_cat	gender*tdof	temp_cat*tdof_cat	wtin_cat*tdof_cat
COMBO	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	0.02	< 0.01
1	NA	-	-	< 0.01	0.03	0.04	-	-	0.02
2	NA	< 0.01	-	-	-	< 0.01	-	-	-
3	NA	< 0.01	0.04	-	< 0.01	< 0.01	-	-	< 0.01
4	NA	< 0.01	-	< 0.01	< 0.01	< 0.01	-	-	-
5	NA	-	0.01	0.01	< 0.01	< 0.01	-	-	-
6	NA	< 0.01	0.01	< 0.01	0.80	0.94	-	-	< 0.01
7	NA	< 0.01	< 0.01	-	< 0.01	0.27	-	-	< 0.01
8	NA	< 0.01	-	< 0.01	-	< 0.01	-	-	-
9	NA	< 0.01	< 0.01	-	< 0.01	0.08	-	-	< 0.01
10	NA	< 0.01	-	0.02	< 0.01	< 0.01	-	-	< 0.01
11	NA	-	-	< 0.01	-	-	-	-	-
12	NA	< 0.01	< 0.01	0.04	-	< 0.01	-	-	-
13	NA	< 0.01	0.06	< 0.01	0.75	0.05	0.02	0.01	0.02
14	NA	< 0.01	-	< 0.01	< 0.01	< 0.01	-	-	0.01
15	NA	< 0.01	-	-	-	-	-	-	-
16	NA	< 0.01	-	-	< 0.01	< 0.01	-	-	0.02
17	NA	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	-	-	-
18	NA	< 0.01	-	-		< 0.01	-	-	-
19	NA	< 0.01	-	0.270	< 0.01	< 0.01	-	< 0.01	-
20	NA	< 0.01	-	< 0.01		0.012	-	-	-
21	NA	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	-	-	-

22	NA	0.03	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	-	< 0.01
23	NA	< 0.01	-	< 0.01	< 0.01	< 0.01	-	-	-

(-) Indicates P > 0.05

[a] Gender was categorized into: MAL = males, FEM = females, MIX = mix of males and females, HOL = Holstein

[b] Rectal temperature (temp_cat) was categorized into: < 39.1, 39.1 to 39.4, 39.41 to 40, and > 40 degrees Celsius.

[c] Arrival weight was categorized into 5 categories: < 181kg (400lbs), 181 to 226 kg (400 to 500 lbs), 227 to 272 kg (501 to 600 lbs), 273 to 318 (601 to 700 lbs), and > 318 kg (700lbs).

[d] Days on feed at treatment was categorized into animals less than 15 days on feed, 15 to 30 days on feed, 30 to 45 days on feed and greater than 45 days on feed.

Table 5. Results of logistic regression models of associations between classifier accuracy and gender[a], temperature category (temp_cat)[b], arrival weight category (wtin_cat)[c], days on feed at first treatment category (tdof_cat)[d] and known status of animal finishing the production cycle (DNF).

dataset	Variable	DNF	gender	temp_cat (°C)	wtin_cat (lbs)	tdof_cat	n	Agreement	SE
1	temp_cat			39.1 to 39.4			456	0.922	0.014
1	temp_cat			39.41 to 40.0			2221	0.897	0.010
1	wtin_cat				227 to 272		1843	0.886	0.009
1	wtin_cat				273 to 318		2956	0.901	0.007
1	wtin_cat				>700		2992	0.909	0.007
1	tdof_cat					< 15 d	3535	0.902	0.012
1	tdof_cat					15 to 30 d	2397	0.916	0.017
1	wtin_cat*tdof_cat				<181	< 15 d	38	0.926	0.041
1	wtin_cat*tdof_cat				<181	15 to 30 d	21	0.955	0.044
1	wtin_cat*tdof_cat				181 to 226	< 15 d	328	0.886	0.018
1	wtin_cat*tdof_cat				181 to 226	> 45 d	160	0.890	0.023
1	wtin_cat*tdof_cat				227 to 272	15 to 30 d	564	0.891	0.013
1	wtin_cat*tdof_cat				227 to 272	30 to 45 d	225	0.904	0.019
1	wtin_cat*tdof_cat				227 to 272	> 45 d	314	0.891	0.017
1	wtin_cat*tdof_cat				273 to 318	< 15 d	1278	0.900	0.009
1	wtin_cat*tdof_cat				273 to 318	15 to 30 d	743	0.912	0.011
1	wtin_cat*tdof_cat				273 to 318	30 to 45 d	315	0.902	0.017
1	wtin_cat*tdof_cat				273 to 318	> 45 d	620	0.891	0.013
1	wtin_cat*tdof_cat				>318	< 15 d	1140	0.926	0.008
1	wtin_cat*tdof_cat				>318	15 to 30 d	741	0.925	0.010
1	wtin_cat*tdof_cat				>318	30 to 45 d	270	0.896	0.018

1	wtin_cat*tdof_cat				>318	> 45 d	841	0.883	0.012
2	didnotfinish	0					5138	0.983	0.002
5	gender		HOL				122	0.970	0.018
5	gender		MIX				55	1.000	0.000
5	temp_cat			< 39.1			188	1.000	0.000
5	temp_cat			39.1 to 39.4			175	1.000	0.000
5	temp_cat			39.41 to 40.0			1381	1.000	0.000
5	temp_cat			> 40.0			2087	1.000	0.000
5	wtin_cat				<181		218	1.000	0.000
5	wtin_cat				181 to 226		1276	1.000	0.000
5	wtin_cat				227 to 272		1200	1.000	0.000
5	wtin_cat				273 to 318		718	1.000	0.000
5	wtin_cat				>318		544	0.999	0.000
5	tdof_cat					< 15 days	1704	1.000	0.000
5	tdof_cat					15 to 30 days	898	1.000	0.000
5	tdof_cat					30 to 45 days	374	1.000	0.000
5	tdof_cat					> 45 days	981	1.000	0.000
6	didnotfinish	0					1621	0.930	0.012
6	temp_cat			39.41 to 40.0			377	0.924	0.023
6	wtin_cat*tdof_cat				<181	< 15 d	13	0.905	0.079
6	wtin_cat*tdof_cat				227 to 272	15 to 30 d	53	0.945	0.027
7	didnotfinish	0					4779	0.927	0.026
7	gender		MIX				20	0.923	0.081
7	wtin_cat*tdof_cat				<181	30 to 45 d	15	0.902	0.080
8	didnofinish	0						0.861	0.012
9	didnotfinish	0					6661	0.970	0.006
9	wtin_cat*tdof_cat				<181	15 to 30 days	81	0.911	0.056
11	temp_cat			< 39.1			14	0.929	0.069
11	temp_cat			39.1 to 39.4			22	0.955	0.044

11	temp_cat		39.41 to 40.0		48	0.958	0.029
12	didnotfinish	0			2741	0.936	0.007
13	temp_cat*tdof_cat		<102.5	15 to 30 days	93	0.928	0.035
13	temp_cat*tdof_cat		39.1 to 39.4	15 to 30 days	86	0.941	0.033
13	temp_cat*tdof_cat		39.1 to 39.4	> 45 days	167	0.937	0.025
13	temp_cat*tdof_cat		39.41 to 40.0	15 to 30 days	560	0.930	0.015
13	temp_cat*tdof_cat		39.41 to 40.0	30 to 45 days	348	0.935	0.019
14	didnotfinish	0			2730	0.983	35.754
14	temp_cat		< 39.1		74	0.966	68.032
14	temp_cat		39.1 to 39.4		77	0.979	43.554
14	temp_cat		39.41 to 40.0		356	0.970	61.386
14	wtin_cat		227 to 272		189	0.999	5.021
14	tdof_cat			30 to 45 days	388	1.000	1.543
14	wtin_cat*tdof_cat		227 to 272	30 to 45 days	22	1.000	0.000
18	didnotfinish	0			3403	0.950	0.006
19	temp_cat*tdof_cat		≥ 104	> 45 d	227	0.973	0.021
20	didnotfinish	0			9985	0.952	0.004
22	wtin_cat*tdof_cat		>318	15 to 30 d	1522	0.924	0.008
22	wtin_cat*tdof_cat		>318	30 to 45 d	749	0.936	0.009

[a]Gender was categorized into: MAL = males, FEM = females, MIX = mix of males and females, HOL = Holstein

[b] Rectal temperature (temp_cat) was categorized into: < 39.1, 39.1 to 39.4, 39.41 to 40, and > 40 degrees Celsius.

[c] Arrival weight was categorized into 5 categories: < 181kg (400lbs), 181 to 226 kg (400 to 500 lbs), 227 to 272 kg (501 to 600 lbs), 273 to 318 (601 to 700 lbs), and > 318 kg (700lbs).

[d] Days on feed at treatment was categorized into animals less than 15 days on feed, 15 to 30 days on feed, 30 to 45 days on feed and greater than 45 days on feed.