Preventive Veterinary Medicine 163 (2019) 14–23

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Prediction model optimization using full model selection with regression trees demonstrated with FTIR data from bovine milk

M. Tremblay[a,d,*], M. Kammer[b], H. Lange[c,e], S. Plattner[c,e], C. Baumgartner[c], J.A. Stegeman[d], J. Duda[b], R. Mansfeld[e], D. Döpfer[a]

[a] Department of Medical Science, School of Veterinary Medicine, University of Wisconsin, 2015 Linden Dr., Madison,53706, United States
[b] LKV Bayern e.V., Landsberger Straße 282, 80687, München, Germany
[c] Milchprüfring Bayern e.V., Hochstatt 2, 85283, Wolnzach, Germany
[d] Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, PO Box 80151, 3508 TD, Utrecht, the Netherlands
[e] Clinic for Ruminants with Ambulatory and Herd Health Services, Ludwig-Maximilians-Universität München, Sonnenstr. 16, D-85764 Oberschleissheim, Germany

ARTICLE INFO

Keywords:
Full model selection
Regression tree
Preprocessing
Prediction model
Fourier-transform infrared spectra

ABSTRACT

Predictive modeling is the development of a model that is best able to predict an outcome based on given input variables. Model algorithms are different processes that are used to define functions that transform the data within models. Common algorithms include logistic regression (LR), linear discriminant analysis (LDA), classification and regression trees (CART), naïve Bayes (NB), and k-nearest neighbor (KNN). Data preprocessing option, such as feature extraction and reduction, and model algorithms are commonly selected empirically in epidemiological studies even though these decisions can significantly affect model performance. Accordingly, full model selection (**FMS**) methods were developed to provide a systematic approach to select predictive modeling methods; however, current limitations of FMS, such as its dependency on user-selected hyperparameters, have prevented their routine incorporation into analyses for model performance optimization.

Here we present the use of regression trees as an innovative method to apply FMS. Regression tree FMS (rtFMS) requires the development of a model for every combination of predictive modeling method options under consideration. The iterated, cross-validation performances of these models are then passed through a regression tree for selection of a final model. We demonstrate the benefits of rtFMS using a milk Fourier transform infrared spectroscopy dataset, wherein we build prediction models for two blood metabolic health parameters in dairy cows, nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA). The goal for building NEFA and BHBA prediction models is to provide a milk-based screening tool for metabolic health in dairy cattle that can be incorporated automatically in milk analysis routines. These models could be used in conjunction with physical exams, cow side tests, and other indications to initiate medical intervention.

In contrast to previously reported FMS methods, rtFMS is not a black box, is simple to implement and interpret, it does not have hyperparameters, and it illustrates the relative importance of modeling options. Additionally, rtFMS allows for indirect comparisons among models developed using different datasets. Finally, rtFMS eliminates user bias due to personal preference for certain methods and rtFMS removes the dependency on published comparisons of methods. Thus, rtFMS provides clear benefits over the empirical selection of data preprocessing options and model algorithms.

1. Introduction

Predictive modeling is the development of a model that is best able to predict an outcome based on given input variables (Geisser, 1993; Kuhn and Johnson, 2013). Model algorithms are different processes that are used to define functions that transform the data within models (Burger, 2018). Common algorithms include logistic regression (LR), linear discriminant analysis (LDA), classification and regression trees (CART), naïve Bayes (NB), and k-nearest neighbor (KNN). Currently, empirical selection is the standard method to select among predictive modeling method options including different preprocessing techniques, and model algorithms (Harrell et al., 1996; Kuhn and Johnson, 2013);

* Corresponding author at: Department of Medical Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, 2015 Linden Dr., Madison, WI, 53706, United States.
E-mail address: mtremblay@wisc.edu (M. Tremblay).

https://doi.org/10.1016/j.prevetmed.2018.12.012
Received 26 June 2018; Received in revised form 19 October 2018; Accepted 18 December 2018

Table 1
Overview of regression tree full model selection (rtFMS) methods.

Step	Method
1	**Data preparation:** A dataset is prepared by formatting variables, and removing outliers, repeats, and errors.
2	**Outcome selection:** An outcome variable is selected according to the project's goal.
3	**Standard methods:** A set of "standard methods", which are methods that will be applied to all models and will not benefit from comparisons of different options, are selected reflective of the data. For example, if a dataset had missing data, applying imputation would be selected as a standard method, although different imputation functions could be compared in step 4.
4	**Comparison categories:** Categories relating to prediction model methods and multiple options within each category are selected for comparison. Examples of categories include input data subsets, feature extraction and model algorithms (see Table 2).
5	**Modeling:** A model is run for every combination of options per category described in step 4 using standard methods described in step 3.
6	**Performance measure:** A performance measure is selected depending on a dataset's characteristics and how the final prediction model will be used. For example, if a dataset is imbalanced, balanced accuracy or kappa value would be the preferred performance of the final model.
7	**Regression tree:** Then the models' performance measures are run through a regression tree to visualize the best combination of options, which selections made significant differences and in which order these selections were prioritized.
8	**Final model:** A final model is selected based on the regression tree selections. If there were no significant difference among options, then personal preference can be used and justified in making a selection.

however, these options and the order of decisions about predictive modeling methods can significantly influence model performance (Weissenbacher et al., 2009; Han et al., 2011; Horn et al., 2018; Rinnan, 2014; Shi and Yu, 2017). Consequently, full model selection (**FMS**) was developed to provide a systematic approach to eliminate bias in selecting predictive modeling method options for machine learning (Escalante et al., 2009). In short, FMS requires the development of a model for every combination of modeling methods under consideration (i.e., options). Then, the FMS method compares the models' iterated cross-validated performances to select a final optimized model. This system has been implemented in machine learning, but has largely been overlooked in predictive modeling in applied epidemiology.

All datasets, no matter if large or small, could benefit from optimizing the selection of predictive modeling method options for increased performance and reliability, especially given the recent increase availability of machine learning algorithms. One could fit a model for every different combination of options and select the best performing model without FMS, however, one would miss out on important information. Knowing if a decision, such as using all variables or only a small subset, was significant for increasing the performance of a model is crucial when planning future study designs and modeling efforts. In addition, it might not be correct to always select the highest performing model without considering its reliability. Selecting methods that perform well consistently is important when a model will be adjusted and expanded over time.

Current FMS methods in machine learning use multi-agent based and stochastic algorithms (Sun, 2014; Bansal and Sahoo, 2015). The most notable method, particle swarm optimization (PSO), is a FMS search method based on the behavior of individuals in swarms such as fish and birds (Eberhart and Kennedy, 1995; Escalante et al., 2009). PSO is a black box method, thus, the options' influence on making this selection is not visible to the user. In addition, PSO has hyperparameters, those are parameters of a prior distribution (e.g. inertia weight, acceleration coefficients, velocity clamping), that can change the output and be difficult to select. These facts have slowed the incorporation of FMS into applied epidemiology.

In this paper, we describe the use of conditional inference recursive partitioning regression trees as an innovative FMS method (rtFMS) in applied epidemiology predictive modeling during supervised learning. Regression trees were first described by Breiman et al. in, 1984 and are now a common employed technique. Regression trees use recursive partitioning to discover which independent variables are most associated with statistically significant differences in the continuous outcome variable. It then uses that variable to separate the data into two subsets (i.e., nodes) that maximizes their difference. The process is repeated for each node until no more significant associations between independent and outcome variables are found (Hothorn et al., 2006). We propose the use of regression trees for the separation of options associated with significantly different model performance measures to optimize a final combination of options that results in the best final model. We propose the use of regression trees for the separation of options associated with significantly different model performance measures to optimize a final combination of options that results in the best final model. Since only one model was fit per combination of options linear algorithms such as, generalized linear model, would face rank deficiencies and singularities. Therefore we selected a non-linear method for FMS. Finally, unlike PSO and other non-linear methods such as neural network, rtFMS is not a black box method, it does not have hyperparameters, and it is straightforward to interpret.

Our objective was to illustrate that rtFMS is easy to implement and it results in an optimized prediction model and in addition provides the following information and benefits: (i) rtFMS illustrates the relative importance of modeling options. (ii) It allows indirect comparisons among models that were fit to different datasets by examining their terminal node location in the regression tree. (iii) rtFMS allows for the comparison of a much larger number of preprocessing and model algorithm options simultaneously than would be feasible without FMS. (iv) Finally, it also removes user bias due to familiarity or personal preference for certain methods on prediction performance.

Our aim was to demonstrate the benefits of rtFMS using a milk Fourier transform infrared spectroscopy dataset, wherein we build and optimize prediction models for two blood metabolic health parameters in dairy cows, nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA). A FTIR dataset was selected because spectrometry research has many preprocessing options that are commonly chosen empirically (Etzion et al., 2004; Dehareng et al., 2012; Botelho et al., 2015; Belay et al., 2017). The goal for building NEFA and BHBA prediction models is to provide a milk-based screening tool for metabolic health in dairy cattle that can be incorporated automatically in milk analysis routines. These models could be used in conjunction with physical exams, cow side tests, and other indications to initiate medical intervention.

2. Materials and methods

A methods overview is available in Table 1. All data analyses were done in R 3.4.2 (R Core Team, 2018). The following packages were used: DMwR, MLmetric, party, partykit, glmnet, randomForest, gbm, earth, klaR, epiR, caret (Liaw and Wiener, 2002; Weihs et al., 2005; Hothorn et al., 2006; Friedman et al., 2010; Torgo, 2010; Hothorn and Zeileis, 2015; Yachen, 2016; Ridgeway, 2017; Kuhn et al., 2018; Milborrow, 2018; Stevenson et al., 2018).We used the following functions available within the caret package: preprocess, groupKFold, findCorrelation, trainControl, and train. The functions ctree and SMOTE were available within the party and DMwR packages, respectively.

Step 1 Data preparation:

The FMS approach is described using the example of a data set previously reported by Tremblay et al. (2018). The data set includes cow information, milk FTIR data, fatty acid predictions, FOSS predictions for milk BHBA and milk acetone, blood measurements, and milk components. Please see the supplemental material for more information. The final dataset for the BHBA model contained 1035 observations and 1034 observations in the final dataset for the NEFA outcome. The final datasets included data from 26 farms, 346 cows and 115 sampling days.

Step 2

Outcome selection NEFA: Blood NEFA ≥ 0.7 mmol/L served as the case definition for the prediction models (Andrews et al., 2008; Tremblay et al., 2018). This would allow the detection of poor metabolic adaptation syndrome (PMAS) and also conditions such as displaced abomasum (NEFA ≥ 1.0 mmol/L) and ketosis (> 1.5 mmol/L) where cows are off-feed or have decreased feed intake, and increased fat mobilization (LeBlanc et al., 2005; Andrews et al., 2008; Tremblay et al., 2018). In the final dataset, 210 observations had blood NEFA ≥ 0.7 mmol/L, and 824 observations had blood NEFA < 0.7 mmol/L.

Outcome selection BHBA: Blood BHBA ≥ 1.2 mmol/L was used as the case definition for the prediction models (McArt et al., 2012; Suthar et al., 2013; Overton et al., 2017). In the final dataset, 105 observations had blood BHBA ≥ 1.2 mmol/L, and 930 observations had blood BHBA < 1.2 mmol/L.

Step 3 Standard methods: Standard methods are methods that will be applied to all models and will not benefit from comparisons of different options as done in step 4. For the data presented here, standard methods included removing wavenumbers representing water: the O—H bending region 1615-1692 cm^{-1}, the O—H stretching region 3057-3689 cm^{-1} (Afseth et al., 2010). Also, observations were flagged for potential FTIR equipment errors if they did not have a max absorbance within the instrument's working range of 0.1–1.0 absorbance units (Beleites and Sergo, 2012). No error observations were identified in this dataset. Variables with zero or near zero variances needed to be removed from the analysis, but none were present in this dataset (Kuhn and Johnson, 2013).

We performed 10 repeated iterations of 10-fold cross-validation by specifying method = "repeatedcv", number = 10, and repeats = 10 in the trainControl command within the caret package (Bali and Sarkar, 2016; Kuhn et al., 2018). Cross-validation is used to get an estimate of the model performance using data the model has not yet been exposed to (i.e., hold-out set). In k-fold cross-validation the dataset is partitioned into k subsets (i.e., folds). In our case, using 10-fold cross-validation, our data were separated into 10 subsets. The models are then trained using 9 of the 10 subsets of data as the training set and the last subset is used as the hold-out test set. The procedure is repeated 10 times, each time using a different fold as the test set for validation. In the end all of the observations will have been used in the test set once. We repeated this 10-fold cross-validation 10 times, with different splits into the 10 subsets each time to make sure the performance was not influenced by how the data were randomly split into subsets. Therefore, the 10 repeated 10-fold cross-validation resulted in performance data from 100 models, giving a better overall estimate of model performance than a one-time hold-out cross-validation (Chollet, 2017). All 11 model algorithms are run using the same separation of observations into training and test sets per k-fold and repeated iterations. We also wanted to make sure that a model was not biased toward better performance in the test set due to both the training and test sets including observations originating from the same farm. Therefore, the groupKFold function within the caret package in R was used to make sure observations were separated into training and test sets by farm for each cross-validation fold (Kuhn et al., 2018).

Out of 1034 total observations, cross-validation folds for the NEFA model averaged 930.6 (SD 23.6) observations in the training sets. Out of 1035 total observations, cross-validation folds for the BHBA model averaged 931.5 (SD 24.3) observations in the training sets. Since the variables were on different scales, auto-scaling was used to obtain zero mean values and standard deviations equal to one (Gelman and Hill, 2006).

Our datasets were faced with class imbalance due to the low prevalence within the outcome classes, only 20.3% and 10.1% of observations being in the NEFA and BHBA minority class, respectively (He and Ma, 2013). To address the class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was used at each fold of cross-validation to balance the number of observations in each training set previously established (Chawla et al., 2002). The minority classes were over-sampled by 200% as suggested by Chawla et al. (2002), and the majority classes were under-sampled by 150% to obtain a one-to-one ratio between the majority and minority classes' observations. The final balanced BHBA training sets had an average of 567 (SD 41.1) observations and the final balanced NEFA training sets had an average of 1134 (SD 35.6) observations.

Step 4

All of the options per categories of predictive modeling methods are listed in Table 2. Categories of predictive modeling methods were separated into 3 areas: (4.1) input subsets, (4.2) preprocessing methods, and (4.3) algorithms (Table 2).

2.1. Input subset

The milk data subset category (4.1.A) includes 4 options: the component (COMP), FTIR, fatty acid predictions (FA), and FOSS's ketosis screening tool predictions (FOSS) subset. The selection of these 4 options was guided by how milk data are generated and their availability for future model application; Milk testing agencies and automatic milking systems generate the COMP subset, the milk analyzers produce the FTIR data, and Qlip N.V. (Leusden, the Netherlands) and FOSS Analytical A/S (Hillerød, Denmark) calibration models using FTIR data produce the FA and FOSS subsets, respectively. The list of variables included in each of these input subsets are available in the supplemental materials.

Finally, including (+CowInfo) or not including cow information (-CowInfo) were compared within the cow information category (4.1.B). Cow information includes the following: days in milk (DIM), milk production (kg/day), and lactation number.

2.2. Preprocessing

A standardization method is necessary to adjust for instrumental differences since the data in this dataset come from 12 different machines, and the goal is to apply the model to external data that will also be from different machines calibrated at different times. The standardization category (4.2.A) compared the raw absorbance FTIR values (raw-IR) with two baseline corrections: first derivative (FD) and second derivative transformations (SD) (Duckworth, 2004; Beleites and Sergo, 2012; Baker et al., 2014; Smith et al., 2018). The FD is very effective for removing baseline offset and SD is very effective for both the baseline offset and linear trends in spectra (Duckworth, 2004; Rinnan, 2014).

Three categories of dimension reduction methods, also a pre-processing method, were included for comparision using FMS as part of pre-processing: feature extraction (4.2.A), wavenumber subset (4.2.B), and high correlation (4.2.C).

A feature extraction category (4.2.B) compared performing principal component analysis (PCA) (+PCA) or not performing PCA (-PCA). When PCA was applied, the number of components representing 95% of the features' variance were selected (Kuhn and Johnson, 2013). As a feature selection step, a wavenumber subset category (4.2.C) was used to compare performance with (AllWN) or without wavenumber variables (excl.no-infoWN) that are thought to not represent any information ("no information regions"). These are regions from 1800.285 cm^{-1}

Table 2
Options per corresponding category and area (step 4) selected for comparison using regression tree full model selection when applied to a milk Fourier transform infrared spectroscopy dataset to build prediction models for two blood metabolic health parameters in dairy cows: nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA).

Area	Category	Options
1. Input Subset	A Milk Data Subset	Component (COMP)
		Fatty acid predictions (FA)
		Fourier transform infrared spectroscopy (FTIR)
		FOSS's ketosis screening tool predictions (FOSS)[a]
	• Cow Information	Include cow information (+CowInfo)
		Exclusion of cow information (-CowInfo)
2. Pre-processing	A Standardization	Raw absorbance values (Raw-FTIR)
		1 st derivative (FD)
		2nd derivative (SD)
	• Feature Extraction	Performed a PCA (+PCA)
		Did not perform a PCA (-PCA)
	• Wavenumber Subset	Removed "no-information" wavenumbers (excl.WN)
		Included all wavenumbers (AllWN)
	• High Correlation	Removed highly correlated predictors (excl.HighCorr)
		Did not remove highly correlated predictors (incl.HighCorr)
3. Algorithm	Algorithm	Logistic generalized linear models (GLM)
		lasso and elastic-net regularized generalized linear models (GLMNET)
		linear discriminant analysis (LDA)
		linear support vector machines (SVM)
		nearest neighbor methods (KNN)
		naive Bayes (NB)
		classification trees (RPART)
		neural networks (NNET)
		gradient boosting machine (GBM)
		random forests (RF)
		multivariate adaptive regression splines (MARS)

[a] BHBA models only.

to 2798.73 cm^{-1} and 3693.09 cm^{-1} to 5007.645 cm^{-1} (Andersen et al., 2002; Iñón et al., 2004; Dagnachew et al., 2013). A high correlation category (4.2.D) was also included that compared including (incl.HighCorr) or excluding highly correlated variables (excl.HighCorr). A high correlation filter was applied using the findCorrelation function within caret (Kuhn et al., 2018) with a tolerance set to 0.1 (limit at 0.9), which corresponds to a VIF of 10 (Hair, 2007).

2.3. Algorithms

The *algorithm* category included 11 algorithms to compare: logistic generalized linear models (GLM), lasso and elastic-net regularized generalized linear models (GLMNET), linear discriminant analysis (LDA), linear support vector machines (SVM), nearest neighbor methods (KNN), naive Bayes (NB), classification trees (RPART), neural networks (NNET), gradient boosting machine (GBM), random forests (RF), and multivariate adaptive regression splines (MARS). These algorithms were run using the caret model methods "glm", "glmnet", "lda", "svmLinear", "knn", "nb", "rpart", "nnet", "gbm", "rf", and "earth", respectively (Kuhn et al., 2018). Although the random tree method used for model selection does not have hyperparameters, the following 7 of the 11 prediction model algorithms used to model NEFA and BHBA have hyperparameters: GLMNET, SVM, GBM, NB, RF, NNET, KNN. For these 7 model algorithms, the models' hyperparameters, such as alpha and lambda for glmnet, were automatically selected (i.e., fine-tuned) using a grid search. The grid search compares 10 values for each hyperparameter spanning meaningful values (Bergstra and Bengio, 2012; Kuhn et al., 2018). The default convergence criteria for each model algorithm were used (Kuhn et al., 2018).

Step 5 & 6. A total of 660 and 704 models for the NEFA and BHBA outcome, respectively, were run for every combination of options per category described in step 4 (Table 2). Running all of the models took approximately 1 week of computational time for each outcome, BHBA and NEFA. Balanced accuracy was the selected performance parameter because it performs well when the data sets are imbalanced (Japkowicz and Stephen, 2002). See Table 3 for a list of possible performance measures that were available. The average of the 100 cross-validation folds' balanced accuracies were used as the models' point estimate to be used in the regression tree.

Step 7. The models' performance measures were run through a nonparametric regression tree (Eq. (1)), available through the party R package, using Eq. (1) (Hothorn et al., 2006). The factor levels of the variables used in Eq. (1) are described in Table 2. A nonparametric regression tree was selected to be the most inclusive in cases where the outcome variable does not follow a normal distribution (Hothorn et al., 2006).

$$\text{Balanced Accuracy} \sim \text{Milk Data Subset} + \text{Cow Information} + \text{Standardization} + \text{Feature Extraction} + \text{Wavenumber Subset} + \text{High Correlation} + \text{Algorithm} \quad (1)$$

For each node, the regression tree finds the variable most associated with the outcome (with a p-value less than 0.05), splits the data into two branches, and then repeats these steps for each node until no more significant differences are found between independent variable and the outcome variable. A bonferonni correction for multiple comparisons of means was applied. Since this tree is grown using a so called hypothesis test-based stopping rule, pruning of the tree is not needed (Hothorn et al., 2006).

Step 8. The regression tree was inspected to locate the terminal node with the best performance, i.e. highest balance accuracy. The decision nodes leading to the best performing terminal node were described. The number of models per terminal node is the number of different combinations of options that have been left unspecified by the regression tree on the path to the terminal node. If a category did not have an option selected by the regression tree then personal preference was justified in making those decisions since they would not make a statistically significant difference in model performance. The selected final model was applied to the entire original dataset for final performance measures and measures of uncertainty. The 20 most influential

Table 3
Final models' performance measures with 95% confidence intervals.

	Blood nonesterified fatty acids final model		Blood β-hydroxybutyrate acid final model	
Performance measure	estimate	95% CI	estimate	95% CI
Apparent prevalence, %	33.7	(30.8-36.6)	29.2	(26.4-32.1)
True prevalence, %	20.3	(17.9-22.9)	10.1	(8.4-12.1)
Sensitivity, %	77.1	(70.9-82.6)	84.8	(76.4-91)
Specificity, %	77.4	(74.4-80.2)	77.1	(74.3-79.8)
Diagnostic accuracy, %	77.4	(74.7-79.9)	77.9	(75.2-80.4)
Balanced accuracy, %	77.3	(72.6-81.4)	80.9	(75.3-85.4)
Positive predictive value, %	46.6	(41.2-51.9)	29.5	(24.4-35)
Negative predictive value, %	93.0	(90.8-94.8)	97.8	(96.5-98.7)
Likelihood ratio of a positive test	3.42	(2.95-3.96)	3.70	(3.21-4.27)
Likelihood ratio of a negative test	0.295	(0.230-0.380)	0.198	(0.126-0.311)
Kappa	0.438	(0.381-0.496)	0.338	(0.288-0.388)

CI = confidence interval.

predictors were extracted and ranked for each final model using the varImp function in caret (Kuhn et al., 2018). Their importances were scaled to 100 so that the most influential predictor had a value of 100 and the least influential had a value near zero.

3. Results

NEFA

Step 5 & 6. Nine NEFA models did not converge according to each model's convergence criteria (Kuhn et al., 2018). The remaining 651 NEFA models had a mean balanced accuracy of 66.48 (SD 4.84).

Step 7. The NEFA FMS regression tree had 25 decision nodes and 26 terminal nodes (Fig. 1). The average number of models per terminal node was 25.0 (SD 16.4). The 25th terminal node had the highest performance. It contained 8 models with an average balanced accuracy of 74.5 (SD 0.56). The final model was selected after 3 decision nodes. (i) The first decision node selected the FA subset over the FTIR and component milk data subset (p-value < 0.001). Therefore, only the models that only used the 79 variables in the fatty acid subset as predictors, with or without cow information, were used further. (ii) The second decision node selected the following model algorithms (p-value < 0.001): GLMNET, SVM and NNET. (iii) The final decision node selected the GLMNET model algorithm (p-value = 0.017) (Fig. 1).

If the FA subset had not been available, as is often the case in practice, then the models represented in the fifth terminal node would have resulted in the best performance. It contained 16 models with an average balanced accuracy of 73.20 (SD 0.79). This model would have been selected after 4 decision nodes: (i) the following model algorithms were selected (p-value < 0.001): MARS, GBM, GLMNET, LDA, NNET, SVM. (ii) The derivative-transformed (FD, SD) FTIR input subsets were selected (p-value < 0.001). (iii) The next decision node selected the GLMNET model algorithm (p-value < 0.001). (iv) Finally, not performing a PCA (-PCA) was selected (p-value < 0.001) (Fig. 1).

Step 8. Options within the cow information, feature extraction and high correlation categories were not selected by the NEFA FMS regression tree. This leaves room to make these decisions empirically. It was decided to include cow information, to not remove highly correlated variables, and to not perform a PCA. When the selected NEFA model (options: FA, +CowInfo, -PCA, incl.HighCorr, GLMNET) was applied to the entire original dataset it had a final balanced accuracy of 77.3 (95% CI: 72.6–81.4), sensitivity of 77.1 (95% CI: 70.9–82.6), specificity of 77.4 (95% CI: 74.4–80.2) and diagnostic accuracy of 77.4 (95% CI: 74.7–79.9) (Table 3). The final hyperparameter values used in the model were alpha = 0.4 and lambda = 0.0117. The most influential predictors in the final NEFA model were ranked and listed in Table 4.

BHBA

Step 5 & 6. The 704 BHBA models had a mean balanced accuracy of 66.31 (SD 4.58).

Step 7. The BHBA FMS regression tree had 27 decision nodes and 28 terminal nodes (Fig. 2). The average number of models per terminal node was 25.1 (SD 29.7). The eighteenth node had the highest performance. It contained 8 models with an average balanced accuracy of 74.2 (SD 1.03). The final model was selected after 5 decision nodes. (i) The first decision node selected the following model algorithms (p-value < 0.001): MARS, GBM, GLMNET, LDA, NNET, SVM. (ii) The second decision node selected the derivative-transformed (FD, SD) FTIR, COMP and FA subset (p-value < 0.001). (iii) The third decision node selected the GLMNET model algorithm (p-value < 0.001). (iv) Next, the derivative-transformed (FD, SD) FTIR subsets were selected over the COMP and FA subsets (p-value < 0.001). (v) Finally, not performing a PCA (-PCA) was selected (p-value < 0.024).

Step 8. Options within the cow information, wavenumber subset and high correlation criteria were not selected by the BHBA FMS regression tree. Therefore, it was appropriate to make these decisions empirically. It was decided to include cow information, to not subset the no-information wavenumbers, and to not remove highly correlated variables. The BHBA FMS regression tree did not discern between the FD and SD FTIR standardizations. Therefore, we empirically decided to select the FD standardization.

The final BHBA model (options: FTIR, +CowInfo, FD, -PCA, incl.HighCorr, AllWN, GLMNET) had a final balanced accuracy of 80.9 (95% CI: 75.3–85.4), sensitivity of 84.8 (95% CI: 76.4–91.0), specificity of 77.1 (95% CI: 74.3–79.8) and diagnostic accuracy of 77.9 (95% CI: 75.2–80.4) (Table 3). The final hyperparameter values used in the model were alpha = 0.3 and lambda = 0.0735. The most influential predictors in the final BHBA model were ranked and listed in Table 5.

4. Discussion

FMS

Our proposed rtFMS method provides a systematic and unbiased approach to optimizing prediction model performance given many possible options for algorithms and preprocessing methods. Our method demonstrated how different combinations of decisions led to statistically significant differences in model performance. The rtFMS selected different preprocessing options for different model outcomes (NEFA, BHBA) within the same dataset, which illustrates the importance of incorporating this technique into all prediction modeling efforts.

Unlike PSO-FMS, rtFMS does not contain hyperparameters and user-friendliness is further improved by the visual representation of the

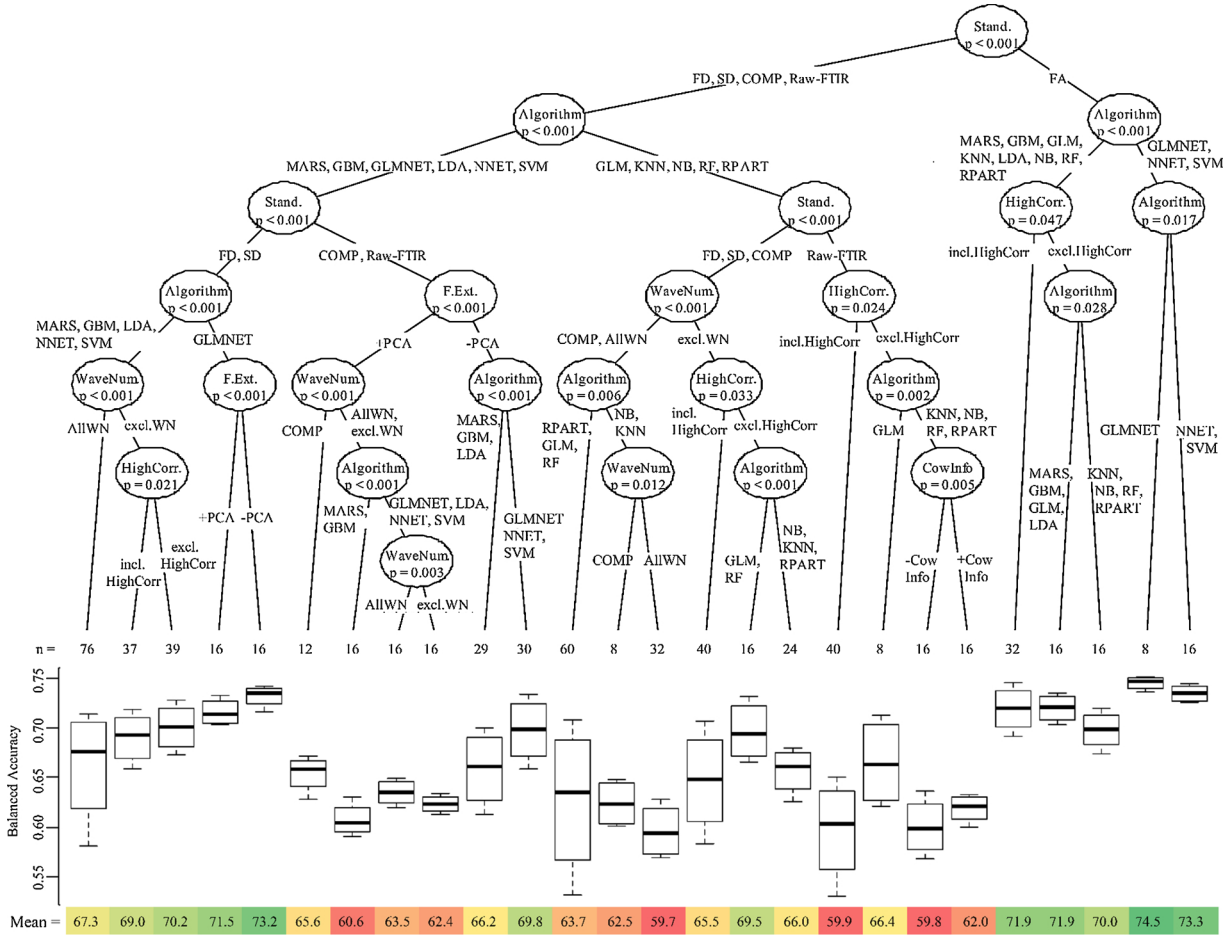

Fig. 1. Blood nonesterified fatty acids (NEFA) rtFMS regression tree results.
The balanced accuracies of the 651 NEFA predictions models were inputted into the regression tree; n = number of models in each terminal node; boxplots visually represent the balanced accuracy of models per terminal model. The bottom and top of the box represent the 25th and 75th percentiles, respectively, and the horizontal line inside the box is the median; Subset = Milk Data Subset category; COMP = component; FA = fatty acid predictions; FTIR = Fourier transform infrared spectroscopy; FOSS = FOSS's ketosis screening tool predictions; Cow Info = Cow information category; +CowInfo = Include cow information; -CowInfo = Exclusion of cow information; Stand. = Standardization category; Raw-FTIR = Raw absorbance values; FD = 1 st derivative; SD = 2nd derivative; F.Ext. = Feature Extraction category; +PCA = Performed a PCA; -PCA = Did not perform a PCA;WaveNum. = Wavenumber Subset category; excl.WN = Removed "no-information" wavenumbers; AllWN = Included all wavenumbers; HighCorr = High Correlation category; excl.HC = Removed highly correlated predictors; incl.HighCorr = Did not remove highly correlated predictors; Algorithm = Algorithm category; GLM = logistic generalized linear models algorithm; GLMNET = lasso and elastic-net regularized generalized linear models algorithm; LDA = linear discriminant analysis algorithm; LDA = linear discriminant analysis algorithm; SVM = linear support vector machines algorithm; KNN = nearest neighbor methods algorithm; NB = naive Bayes algorithm; RPART = classification trees algorithm; NNET = neural networks algorithm; GBM = gradient boosting machine algorithm; RF = random forests algorithm; MARS = multivariate adaptive regression splines algorithm.

results. In addition, the rtFMS method provides information about the relative importance of options when selecting the final model selection. The relative location of nodes in the tree reflects the importance of the decision on the performance of the prediction model. Our method also allows indirect comparisons among models developed using different datasets, by examining the terminal node location of different option combinations in a regression tree. The ability to eliminate bias by performing these indirect comparisons is important when teams with different personal preferences and experiences are collaborating. This information is key when developing future study designs, and when determining future exploration of additional modeling methods. FMS removes user bias due to familiarity and personal preference with regards to certain prediction models methods. In contrast to PSO-FMS, rtFMS allows for empirical decisions when appropriate; however, it removes this source of bias on performance when a significantly superior performing model would be possible. We expect that the benefits and flexibility of rtFMS will accelerate its incorporation into the field of applied epidemiology.

rtFMS only depends on being supplied an outcome variable such as a goodness of fit or performance measure; therefore, any type of model can be optimized using rtFMS. rtFMS can be applied to parameter estimation models, longitudinal models, multinomial models, and even unsupervised learning. When applied to prediction models, the performance measure used as the outcome variable in the regression tree can be selected according to the user's needs regarding performance (e.g. high sensitivity, specificity, accuracy, positive predictive value, or negative predictive value). Categorical performance outcomes could also be accommodated by using classification trees (Therneau et al., 2015). The current rtFMS method optimized a single performance measure but multiple performance measures could be optimized simultaneously using multi-target regression tree (Aho et al., 2012; Osojnik et al., 2015). The resulting performance landscapes of multi-target regression tree would reflect the many facets of model preferences, selection and application.

Table 4
The 20 most important predictors in the blood nonesterified fatty acids (NEFA) final prediction model (options: FA, +CowInfo, -PCA, incl.HighCorr, GLMNET) and their relative importance.

Predictor	Importance[a]
C14:0, μmol/L	100
C30:1, μmol/L	83.35
Milk production, kg	71.94
Lactation number	66.08
C20:4, μmol/L	61.14
BHBA, mmol/L	54.83
C16:0, μmol/L	47.72
C22:5, μmol/L	40.42
Days in milk (DIM)	35.73
C25:1 result 1, μmol/L	24.89
C25:0, μmol/L	22.30
C17:0, μmol/L	18.54
C29:4 result 1, μmol/L	14.81
C24:5, μmol/L	14.71
C25:1 total, umol/L	11.54
C22:6, μmol/L	11.19
C19:1 total, umol/L	10.56
C15:0 result 1, μmol/L	9.94
C25:3, μmol/L	9.30
C23:0, μmol/L	9.25

FA = fatty acid predictions; +CowInfo = Include cow information; -PCA = Did not perform a PCA; incl.HighCorr = Did not remove highly correlated predictors; GLMNET = lasso and elastic-net regularized generalized linear models algorithm.

[a] Importance scaled to 100.

Final models

A general overview of rtFMS findings

Glmnet was consistently one of the best-performing model algorithms. This is most likely because both the FTIR and FA input subsets have many highly correlated variables, which glmnet addresses with the elastic-net penalty (Zou and Hastie, 2005; James et al., 2013). The final selection of a glmnet algorithm is in contrast to Fernández-Delgado et al., 2014 who found the random forest algorithm performed the best when applied to over 100 datasets. However, this study assumed that preprocessing would affect all algorithms similarly and that algorithms would be ranked similarly for all dataset. The differences between this study and ours suggests that findings from published non-FMS comparisons of model algorithms or preprocessing options cannot be applied to other datasets without FMS comparisons.

NEFA model

The NEFA regression tree selected the fatty acid input subset for the final model. This indicated that the additional calibrations for more than 60 different fatty acids by Qlip NV (Leusden, the Netherlands) improved the information gathered by FTIR. We hypothesize that these calibrations are acting as a targeted feature extraction step. Fatty acids that are synthesized *de novo* from ketones in mammary epithelial cells and are distinguished by the presence of fewer than 16 carbon atoms (Bauman and Davis, 2013). Pre-formed fatty acids on the other hand, have more than 16 carbons and originate from NEFA or lipoproteins in the circulation (Barber et al., 1997; Neville and Picciano, 1997). Thirdly, mixed fatty acids have 16 carbons and can be pre-formed or synthesized *de novo*. Blood NEFA has been found to be highly correlated with milk C18:1 cis^{-9}, and also inversely correlated with the proportion of *de novo* fatty acids in milk (Bell, 1995; Jorjong et al., 2014; Friedrichs et al., 2015). The use of ratios between the different fatty acids in milk has been shown to perform better than measurements of single fatty acid in predicting blood NEFA (Mann et al., 2016; Dórea et al., 2017). These findings support our results that the most important predictors in our NEFA model represent all types of fatty acids including *de novo*, preformed, and mixed fatty acids.

BHBA model

Some of the most important predictors in the final BHBA model are located in the acetone region of the FTIR spectra between 1450 and 1200 cm^{-1} (Hansen, 1999; Heuer et al., 2001). This is in line with the previous finding of a high correlation between milk acetone and blood BHBA (Steger et al., 1972). Acetone information was not available in the fatty acid or component datasets, which could explain why the rtFMS selected the FTIR input subset to predict blood BHBA. In addition to acetone, the other highly important predictor of blood BHBA was wavenumber 1542 cm^{-1} that represents milk protein. Other important predictors in the final model were wavenumbers in the "no information regions" of the spectra. Fatty acids have been shown to increase the baseline of the spectra in the "no information" spectral regions, (e.g. wavenumbers greater than 4000 cm^{-1}) (Grabska et al., 2017). We conclude that it is necessary to further investigate FTIR patterns in the so-called 'non-information" regions. In contrast, the FTIR wavenumbers associated with milk fat (i.e., 2927, 2862, 1743, 1454 and 1390 cm^{-1}) were not among the important predictors in our model (Socrates, 1980). This finding suggests that fatty acid information is more important than overall fat composition when predicting blood BHBA. These findings will improve the recommendations for cow health and well-being that can be made based on milk testing data in the near future.

Previous BHBA prediction models based on FTIR data used datasets with different breeds of cattle, geographic regions, sampling structure (DIM), and cross-validation methods, wherein direct comparisons were not possible. However, our rtFMS method allows for indirect comparisons of models using different datasets. Our final BHBA model performance measures overlapped or were significantly better than those of previously published prediction models of blood BHBA that used FTIR data including van Knegsel et al. (2010) and Chandler et al. (2017) that used FOSS milk acetone and milk BHBA predictions in their model. The FOSS input subset did not perform as well in our analysis compared to the FA, COMP, and derivative-transformed IR input subsets. We suspect that this is the case because FOSS calibrations were developed for milk BHBA as the outcome variable rather than blood BHBA used for the current analysis, wherein the correlation between milk and blood BHBA varies widely from 0 to 0.88 (Geishauser et al., 1998). Belay et al., (2017), reported a regression prediction model for blood BHBA that applied feature extraction via partial least squares regression (PLS) regression, akin to PCA. Similarly, our rtFMS results showed that eliminating feature extraction using PCA yielded a better performing BHBA model. Most recently, Pralle et al. (2018) compared 3 model algorithms and 2 data inputs subsets to predict blood hyperketonemia (BHBA ≥ 1.2 mmol/L). Based on our results, improved predictive performance could be achieved for this dataset by the addition of a derivative transformation of the spectral data and the use of the glmnet algorithm without variable reduction.

Discussion of findings for the FTIR data sets and their application in practice

We searched for a standardization method that was rapid, simple, outcome dependent, require no additional samples, applicable to data already collected, and applicable to new observations individually without depending on the remaining dataset. This search resulted in standardization by preprocessing methods, such as a derivative transformation (Feudale et al., 2002). First or 2nd derivative transformations were consistently favored over the raw FTIR data in both models. This suggests that standardization is needed to adjust for changes in calibration over time and differences among instruments. The next step for the prediction modeling of FTIR data sets is to perform a crtFMS omparison among more standardization methods including the piece-wise direct standardization method that maps the response of a 'slave' FTIR instrument onto a 'master' instrument (Wise, 1996; Wise et al., 2007;

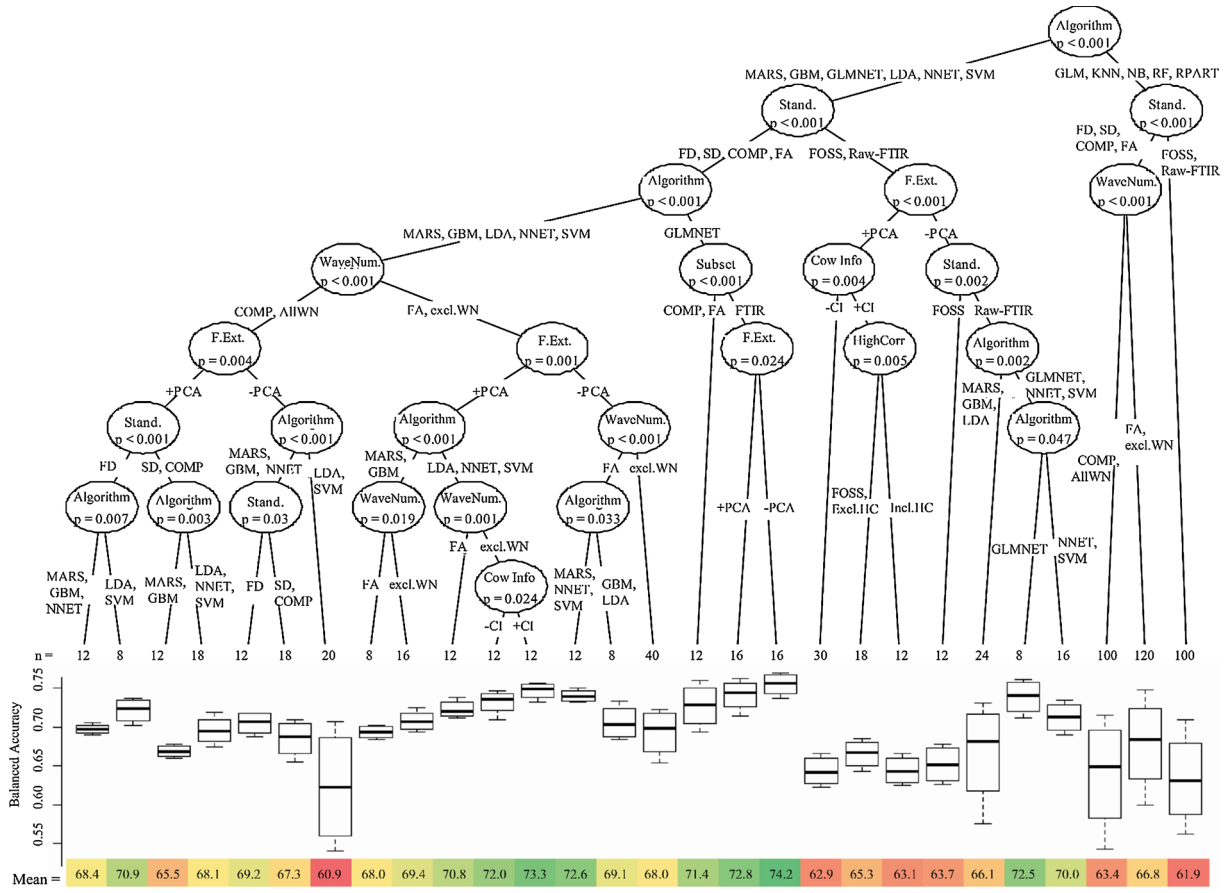

Fig. 2. Blood β-hydroxybutyrate acid (BHBA) rtFMS regression tree results.
The balanced accuracies of the 704 BHBA predictions models were inputted into the regression tree; n = number of models in each terminal node; boxplots visually represent the balanced accuracy of models per terminal model. The bottom and top of the box represent the 25th and 75th percentiles, respectively, and the horizontal line inside the box is the median; Subset = Milk Data Subset category; COMP = component; FA = fatty acid predictions; FTIR = Fourier transform infrared spectroscopy; FOSS = FOSS's ketosis screening tool predictions; Cow Info = Cow information category; +CowInfo = Include cow information; -CowInfo = Exclusion of cow information; Stand. = Standardization category; Raw-FTIR = Raw absorbance values; FD = 1 st derivative; SD = 2nd derivative; F.Ext. = Feature Extraction category; +PCA = Performed a PCA; -PCA = Did not perform a PCA;WaveNum. = Wavenumber Subset category; excl.WN = Removed "no-information" wavenumbers; AllWN = Included all wavenumbers; HighCorr = High Correlation category; excl.HC = Removed highly correlated predictors; incl.HighCorr = Did not remove highly correlated predictors; Algorithm = Algorithm category; GLM = logistic generalized linear models algorithm; GLMNET = lasso and elastic-net regularized generalized linear models algorithm; LDA = linear discriminant analysis algorithm; LDA = linear discriminant analysis algorithm; SVM = linear support vector machines algorithm; KNN = nearest neighbor methods algorithm; NB = naive Bayes algorithm; RPART = classification trees algorithm; NNET = neural networks algorithm; GBM = gradient boosting machine algorithm; RF = random forests algorithm; MARS = multivariate adaptive regression splines algorithm.

Grelet et al., 2017). In addition, rtFMS could be used to compare SMOTE to other methods of balancing the minority and majority classes of observations.

The component dataset did not perform as well as the FTIR and fatty acid subsets in both the NEFA and BHBA models. This is most likely because fatty acid and FTIR data provide more detailed information, such as wavenumbers representing the fatty acid composition of milk fat, compared to component data that only reports total milk fat. This finding supports that it is necessary to invest in incorporating FTIR data and fatty acid calibrations for routine milk analysis. Selection of a model utilizing the FTIR subset would require farms to produce in-line FTIR measurements. In contrast, the fatty acid prediction subset (FA) would require the additional step of sending FTIR data to Qlip N.V. (Leusden, the Netherlands) to produce the fatty acid panel prior to its use in a prediction model.

Although, our final BHBA model performance measures overlapped or were significantly better than those of previously mentioned, both the NEFA and BHBA model's performance need to be further improved to be used as a stand-alone test. We suggest the NEFA model be combined with other alerts such as decreased feed intake to reduce false positive results when detecting conditions causing an increase in fat mobilization such as PMAS. In addition, results of such models should always be confirmed by physical exam or cow side tests before an intervention or treatment is initiated.

Outlook

The benefit and robustness of rtFMS should be evaluated with additional types of data including those with various dimensionalities and different dataset sizes. We foresee additional applications of rtFMS to deep learning networks, which are popular for modeling outcome predictions from large data sets. We recognize that access to the computational capacity necessary to apply FMS can be limited; however, we suspect that this issue will be resolved with the advance in cloud computing. Automation of these methods would also be beneficial for incorporating rtFMS in standard prediction modeling efforts in applied epidemiology.

Table 5
The 20 most important predictors in the blood β-hydroxybutyrate acid (BHBA) final prediction model (options: FTIR, +CowInfo, FD, -PCA, incl.HighCorr, allWN, GLMNET) and their relative importance.

Predictor	Importance[a]
1549.71 cm^{-1}	100
1214.325 cm^{-1}	69.72
Lactation number	63.46
1333.83 cm^{-1}	49.45
2629.11 cm^{-1}	40.93
1210.47 cm^{-1}	39.61
1491.885 cm^{-1}	35.16
Milk production, kg	33.38
2043.15 cm^{-1}	30.93
4891.995 cm^{-1}	29.96
1125.66 cm^{-1}	29.72
4668.405 cm^{-1}	29.59
1372.38 cm^{-1}	26.45
1545.855 cm^{-1}	25.64
1299.135 cm^{-1}	24.26
4336.875 cm^{-1}	22.30
4888.14 cm^{-1}	19.97
2270.595 cm^{-1}	15.85
1164.21 cm^{-1}	15.81
2764.035 cm^{-1}	15.25

FTIR = Fourier transform infrared spectroscopy; +CowInfo = Include cow information; FD = 1 st derivative; -PCA = Did not perform a PCA; incl.HighCorr = Did not remove highly correlated predictors; AllWN = Included all wavenumbers; GLMNET = lasso and elastic-net regularized generalized linear models algorithm.

[a] Importance scaled to 100.

5. Conclusion

In conclusion, rtFMS will allow for the consistent application of FMS to applied epidemiology to improve and optimize prediction model performance and rtFMS will eliminate the bias associated with empirical selection of method options. Other research areas depending on prediction models such as diagnostic imaging, spatial analyses, surveillance, single-nucleotide polymorphism, and microbiome analyses will greatly benefit from applying rtFMS. In the future, rtFMS will continue to provide simplicity and structure to FMS during prediction modeling.

Acknowledgements

The authors acknowledge the Bayerisches Staatsministerium für Ernährung, Landwirtschaft und Forsten (i.e. the Bavarian Ministry for Nutrition, Agriculture and Forests) for supporting the collection of the data. The project was supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme.

The authors thank Dr. Cécile Ané, professor at the Departments of Statistics and of Botany, at the University of Wisconsin–Madison, for her help and advice about the statistical analysis. We gratefully acknowledge the information technology (IT) support provided by Jason Brenner at the University of Wisconsin-Madison, School of Veterinary Medicine.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.prevetmed.2018.12.012.

References

Afseth, N.K., Martens, H., Randby, A., Gidskehaug, L., Narum, B., Jørgensen, K., Lien, S., Kohler, A., 2010. Predicting the fatty acid composition of milk: a comparison of two Fourier transform infrared sampling techniques. Appl. Spectrosc. 64, 700–707.
Aho, T., Ženko, B., Džeroski, S., Elomaa, T., 2012. Multi-target regression with rule ensembles. J. Mach. Learn. Res. 13, 2367–2407.
Andersen, S.K., Hansen, P.W., Andersen, H.V., 2002. Vibrational spectroscopy in the analysis of dairy products and wine. Handbook of vibrational spectroscopy.
Andrews, A.H., Blowey, R.W., Boyd, H., Eddy, R.G., 2008. Bovine Medicine: Diseases and Husbandry of Cattle. John Wiley & Sons.
Baker, M.J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H.J., Dorling, K.M., Fielden, P.R., Fogarty, S.W., Fullwood, N.J., Heys, K.A., Hughes, C., Lasch, P., Martin-Hirsch, P.L., Obinaju, B., Sockalingum, G.D., Sulé-Suso, J., Strong, R.J., Walsh, M.J., Wood, B.R., Gardner, P., Martin, F.L., 2014. Using Fourier transform IR spectroscopy to analyze biological materials. Nat. Protoc. 9, 1771–1791. https://doi.org/10.1038/nprot.2014.110.
Bali, R., Sarkar, D., 2016. R Machine Learning By Example. Packt Publishing Ltd.
Bansal, B., Sahoo, A., 2015. Full model selection using Bat algorithm. Proceedings of the 2015 International Conference on Cognitive Computing and Information Processing (CCIP) 1–4.
Barber, M.C., Clegg, R.A., Travers, M.T., Vernon, R.G., 1997. Lipid metabolism in the lactating mammary gland. Biochimica et Biophysica Acta (BBA) - Lipids Lipid Metab. 1347, 101–126. https://doi.org/10.1016/S0005-2760(97)00079-9.
Bauman, D.E., Davis, C.L., 2013. Biosynthesis of milk fat. Lactation: a comprehensive treatise 2, 31–75.
Belay, T.K., Dagnachew, B.S., Kowalski, Z.M., Adnøy, T., 2017. An attempt at predicting blood β-hydroxybutyrate from Fourier-transform mid-infrared spectra of milk using multivariate mixed models in Polish dairy cattle. J. Dairy Sci. 100, 6312–6326.
Beleites, C., Sergo, V., 2018. hyperSpec: a package to handle hyperspectral data sets in R. R package version 0.99-20180627. http://hyperspec.r-forge.r-project.org.
Bell, A.W., 1995. Regulation of organic nutrient metabolism during transition from late pregnancy to early lactation. J. Anim. Sci. 73, 2804–2819.
Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.
Botelho, B.G., Reis, N., Oliveira, L.S., Sena, M.M., 2015. Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA. Food Chem. 181, 31–37.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Inc.
Burger, S., 2018. Introduction to Machine Learning With R. Rigorous Mathematical Analysis. O'Reilly Media, Inc.
Chandler, T.L., Pralle, R.S., Dórea, J.R.R., Poock, S.E., Oetzel, G.R., Fourdraine, R.H., White, H.M., 2017. Predicting hyperketonemia by logistic and linear regression using test-day milk and performance variables in early-lactation Holstein and Jersey cows. J. Dairy Sci. 101, 2476–2491.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 1 (16), 321–357. https://doi.org/10.1613/jair.953.
Chollet, F., 2017. Deep Learning With Python. Manning Publications Co.
Dagnachew, B.S., Kohler, A., Ådnøy, T., 2013. Genetic and environmental information in goat milk Fourier transform infrared spectra. J. Dairy Sci. 96, 3973–3985. https://doi.org/10.3168/jds.2012-5972.
Dehareng, F., Delfosse, C., Froidmont, E., Soyeurt, H., Martin, C., Gengler, N., Vanlierde, A., Dardenne, P., 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal 6, 1694–1701.
Dórea, J.R.R., French, E.A., Armentano, L.E., 2017. Use of milk fatty acids to estimate plasma nonesterified fatty acid concentrations as an indicator of animal energy balance. J. Dairy Sci. 100, 6164–6176.
Duckworth, J., 2004. Mathematical data preprocessing. Near-Infrared Spectrosc. Agric. 115–132.
Eberhart, R.C., Kennedy, J., 1995. A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micro Machine and Human Science(MHS' 95) 39–43.
Escalante, H.J., Montes, M., Sucar, L.E., 2009. Particle swarm model selection. J. Mach. Learn. Res. 10, 405–440.
Etzion, Y., Linker, R., Cogan, U., Shmulevich, I., 2004. Determination of protein concentration in raw milk by mid-infrared Fourier transform infrared/attenuated total reflectance spectroscopy. J. Dairy Sci. 87, 2779–2788.
Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems. J. Mach. Learn. Res. 15, 3133–3181.
Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D., Ferré, J., 2002. Transfer of multivariate calibration models: a review. Chemom. Intell. Lab. Syst. 64, 181–192. https://doi.org/10.1016/S0169-7439(02)00085-0.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22. http://www.jstatsoft.org/v33/i01/.
Friedrichs, P., Bastin, C., Dehareng, F., Wickham, B., Massart, X., 2015. Final OptiMIR Scientific and Expert meeting: from milk analysis to advisory tools (Palais des Congrès, Namur, Belgium, 16-17 April 2015): optimir-a project aiming the development of novel mid-infrared based management tools for dairy herds. Biotechnologie, Agronomie, Société et Environnement 19, 97.
Geishauser, T., Leslie, K., Kelton, D., Duffield, T., 1998. Evaluation of five cowside tests for use with milk to detect subclinical ketosis in dairy cows. J. Dairy Sci. 81, 438–443.

Geisser, S., 1993. Predictive Inference: An Introduction. Chapman & Hall/CRC Press, New York.
Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
Grabska, J., Beć, K.B., Ishigaki, M., Wójcik, M.J., Ozaki, Y., 2017. Spectra-structure correlations of saturated and unsaturated medium-chain fatty acids. Near-infrared and anharmonic DFT study of hexanoic acid and sorbic acid. Spectrochim. Acta A. Mol. Biomol. Spectrosc. 185, 35–44.
Grelet, C., Pierna, J.F., Dardenne, P., Soyeurt, H., Vanlierde, A., Colinet, F., Bastin, C., Gengler, N., Baeten, V., Dehareng, F., 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. J. Dairy Sci. 100, 7910–7921.
Hair, 2007. Multivariate Data Analysis. Pearson Education.
Han, J., Pei, J., Kamber, M., 2011. Data Mining: Concepts and Techniques. Elsevier.
Hansen, P.W., 1999. Screening of dairy cows for ketosis by use of infrared spectroscopy and multivariate calibration. J. Dairy Sci. 82, 2005–2010.
Harrell, F.E., Lee, K.L., Mark, D.B., 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. 15, 361–387.
He, H., Ma, Y. (Eds.), 2013. ImbAlanced Learning: Foundations, Algorithms, and Applications, 1 edition. Wiley-IEEE Press ed.
Heuer, C., Luinge, H.J., Lutz, E.T.G., Schukken, Y.H., Van Der Maas, J.H., Wilmink, H., Noordhuizen, J., 2001. Determination of acetone in cow milk by Fourier transform infrared spectroscopy for the detection of subclinical ketosis. J. Dairy Sci. 84, 575–582.
Horn, B., Esslinger, S., Pfister, M., Fauhl-Hassek, C., Riedl, J., 2018. Non-targeted detection of paprika adulteration using mid-infrared spectroscopy and one-class classification – Is it data preprocessing that makes the performance? Food Chem. 257, 112–119. https://doi.org/10.1016/j.foodchem.2018.03.007.
Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15 (3), 651–674.
Hothorn, T., Zeileis, A., 2015. Partykit: a modular toolkit for recursive partytioning in R. J. Mach. Learn. Res. 16, 3905–3909. http://jmlr.org/papers/v16/hothorn15a.html.
Iñón, F.A., Garrigues, S., de la Guardia, M., 2004. Nutritional parameters of commercially available milk samples by FTIR and chemometric techniques. Anal. Chim. Acta 513, 401–412. https://doi.org/10.1016/j.aca.2004.03.014.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer Texts in Statistics. Springer, New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7.
Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6, 429–449.
Jorjong, S., Van Knegsel, A.T.M., Verwaeren, J., Lahoz, M.V., Bruckmaier, R.M., De Baets, B., Kemp, B., Fievez, V., 2014. Milk fatty acids as possible biomarkers to early diagnose elevated concentrations of blood plasma nonesterified fatty acids in dairy cows. J. Dairy Sci. 97, 7054–7064.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer Science & Business Media.
Kuhn, M., Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt, T., 2018. caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret.
LeBlanc, S.J., Leslie, K.E., Duffield, T.F., 2005. Metabolic predictors of displaced abomasum in dairy cattle. J. Dairy Sci. 88, 159–170. https://doi.org/10.3168/jds.S0022-0302(05)72674-6.
Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.
Mann, S., Yepes, F.L., Duplessis, M., Wakshlag, J.J., Overton, T.R., Cummings, B.P., Nydam, D.V., 2016. Dry period plane of energy: effects on glucose tolerance in transition dairy cows. J. Dairy Sci. 99, 701–717.
McArt, J.A.A., Nydam, D.V., Oetzel, G.R., 2012. Epidemiology of subclinical ketosis in early lactation dairy cattle. J. Dairy Sci. 95, 5056–5066.
Milborrow, S., 2018. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. earth: Multivariate Adaptive Regression Splines. R package version 4.6.3. https://CRAN.R-project.org/package=earth.
Neville, M.C., Picciano, M.F., 1997. Regulation of milk lipid secretion and composition. Annu. Rev. Nutr. 17, 159–183. https://doi.org/10.1146/annurev.nutr.17.1.159.
Osojnik, A., Panov, P., Džeroski, S., 2015. Comparison of tree-based methods for multi-target regression on data streams. International Workshop on New Frontiers in Mining Complex Patterns. Springer, pp. 17–31.
Overton, T.R., McArt, J.A.A., Nydam, D.V., 2017. A 100-Year Review: metabolic health indicators and management of dairy cattle. J. Dairy Sci. 100, 10398–10417.
Pralle, R.S., Weigel, K.W., White, H.M., 2018. Predicting blood β-hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network. J. Dairy Sci. 101, 4378–4387.
R Core Team, 2018. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ridgeway G., 2017. gbm: Generalized Boosted Regression Models. R package version 2.1.3. https://CRAN.R-project.org/package=gbm.
Rinnan, Å., 2014. Pre-processing in vibrational spectroscopy – when, why and how. Anal. Methods 6, 7124–7129. https://doi.org/10.1039/C3AY42270D.
Shi, H., Yu, P., 2017. Comparison of grating-based near-infrared (NIR) and Fourier transform mid-infrared (ATR-FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for the determination of crude protein and moisture content in wheat. Food Control 82, 57–65. https://doi.org/10.1016/j.foodcont.2017.06.015.
Smith, B.R., Baker, M.J., Palmer, D.S., 2018. PRFFECT: a versatile tool for spectroscopists. Chemom. Intell. Lab. Syst. 172, 33–42. https://doi.org/10.1016/j.chemolab.2017.10.024.
Socrates, G., 1980. Infrared Characteristic Group Frequencies. J. Wilay and Sons, New York 87.
Steger, H., Girschewski, H., Voigt, G., Piatkowski, B., 1972. Die Beurteilung des Ketosisstatus laktierender Rinder aus der Konzentration der Ketokörper im Blut und des Azetons in der Milch. Arch. Anim. Nutr. 22, 157–162.
Stevenson, M., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reiczigel, J., Robison-Cox, J., Sebastiani, P., Solymos, P., Yoshida, K., Jones, G., Pirikahu, S., Firestone, S., Kyle, R., Popp, J., and Jay, M., 2018. epiR: tools for the analysis of epidemiological data. R package version 0.9–79. https://cran.r-project.org/package=epiR.
Sun, Q., 2014. Meta-learning and the Full Model Selection Problem. Doctoral dissertation. University of Waikato.
Suthar, V.S., Canelas-Raposo, J., Deniz, A., Heuwieser, W., 2013. Prevalence of subclinical ketosis and relationships with postpartum diseases in European dairy cows. J. Dairy Sci. 96, 2925–2938.
Therneau, T., Atkinson, B., Ripley, B., 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1–10.
Torgo, L., 2010. Data Mining With R, Learning With Case Studies. Chapman and Hall/CRC. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.
Tremblay, M., Kammer, M., Lange, H., Plattner, S., Baumgartner, C., Stegeman, J.A., Duda, J., Mansfeld, R., Döpfer, D., 2018. Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis. J. Dairy Sci. 101 (8), 7311–7321. https://doi.org/10.3168/jds.2017-13582.
Van Knegsel, A.T.M., Van Der Drift, S.G.A., Horneman, M., De Roos, A.P.W., Kemp, B., Graat, E.A.M., 2010. Ketone body concentration in milk determined by Fourier transform infrared spectroscopy: value for the detection of hyperketonemia in dairy cows. J. Dairy Sci. 93, 3065–3069.
Weihs, C., Ligges, U., Luebke, K., Raabe, N., 2005. klaR analyzing German business cycles. In: Baier, D., Decker, R., Schmidt-Thieme, L. (Eds.), Data Analysis and Decision Support. Springer-Verlag, Berlin, pp. 335–343.
Weissenbacher, A., Kasess, C., Gerstl, F., Lanzenberger, R., Moser, E., Windischberger, C., 2009. Correlations and anticorrelations in resting-state functional connectivity MRI: A quantitative comparison of preprocessing strategies. NeuroImage 47, 1408–1416. https://doi.org/10.1016/j.neuroimage.2009.05.005.
Wise, B.M., 1996. Introduction to Instrument Standardization and Calibration Transfer. Shedding New Light on Disease: Optical Diaganostics for a New Millenium, Winnipeg, CA June 2000.
Wise, B.M., Gallagher, N.B., Bro, R., Shaver, J., Windig, W., Koch, R.S., 2007. PLS Toolbox 4.0. Eigenvector Research Incorporated. Wenatchee, WA, USA. .
Yachen Y., 2016. MLmetrics: Machine Learning Evaluation Metrics. R package version 1.1.1. https://CRAN.R-project.org/package=MLmetrics.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320.