Predicting dairy cattle heat stress using machine learning techniques

C. A. Becker,¹ A. Aghalari,² M. Marufuzzaman,² and A. E. Stone^{1*}

¹Department of Animal and Dairy Sciences, Mississippi State University, Mississippi State 39762

²Department of Industrial and Systems Engineering, Mississippi State University, Mississippi State 39762

ABSTRACT

The objectives of the study were to use a heat stress scoring system to evaluate the severity of heat stress on dairy cows using different heat abatement techniques. The scoring system ranged from 1 to 4, where 1 = no heat stress; 2 = mild heat stress; 3 = severe heat stress; and 4 = moribund. The accuracy of the scoring system was then predicted using 3 machine learning techniques: logistic regression, Gaussian naïve Bayes, and random forest. To predict the accuracy of the scoring system, these techniques used factors including temperature-humidity index, respiration rate, lying time, lying bouts, total steps, drooling, open-mouth breathing, panting, location in shade or sprinklers, somatic cell score, reticulorumen temperature, hygiene body condition score, milk yield, and milk fat and protein percent. Three different treatments, namely, portable shade structure, portable polyvinyl chloride pipe sprinkler system, or control with no heat abatement, were considered, where each treatment was replicated 3 times with 3 second-trimester lactating cows. Results indicate that random forest outperformed the other 2 methods, with respect to both accuracy and precision, in predicting the sprinkler group's score. Both logistic regression and random forest were consistent in predicting scores for control, shade, and combined groups. The mean probability of predicting non-heat-stressed cows was highest for cows in the sprinkler group. Finally, the logistic regression method worked best for predicting heat-stressed cows in control, shade, and combined. The insights gained from these results could aid dairy producers to detect heat stress before it becomes severe, which could decrease the negative effects of heat stress, such as milk loss.

Key words: heat stress, machine learning, shade, sprinklers

INTRODUCTION

Heat stress can be experienced in all climate zones, depending on the time of year (Beede and Collier, 1986). However, dairy cows in tropical, subtropical, and Mediterranean climates experience extended periods of heat stress and are therefore unable to recover from the negative effects of heat stress as quickly compared with cows in temperate climates. The southeastern United States is classified as humid and subtropical with high ambient temperatures, humidity, and rainfall (Johnson, 1987). In regions with a subtropical climate, the high relative humidity limits the rate of heat loss via evaporative cooling (Silva et al., 2007). When natural heat dissipation methods become insufficient, the increased heat load can increase body temperature (Wheelock et al., 2010), decrease milk yield (Wheelock et al., 2010), decrease fertility (De Rensis and Scaramuzzi, 2003), and in extreme cases, can result in mortality (Stull et al., 2008; Vitali et al., 2009).

Temperature-humidity index (THI) is typically used to summate the intensity of heat stress on dairy cows. This calculation includes values for both ambient temperature and relative humidity (Mader et al., 2006; Bohmanova et al., 2007; Morton et al., 2007). When $THI \geq 68$, lactating cows will begin experiencing negative effects of heat stress (Collier et al., 2012). Temperature-humidity index does not incorporate cow-specific factors, which also play a role in heat stress. Individual cow factors that can contribute to the severity of heat stress experienced include but are not limited to the level of production, disease, age, body condition, and hair coat characteristics (Collier et al., 2009).

Several phenotypic indicators have been identified as heat stress indices, including respiration rate (RR; Gaughan et al., 2000), drooling, open-mouth breathing, and panting (Tresoldi et al., 2018), and body temperature (Kendall et al., 2006). Observing phenotypic indicators of heat stress could be advantageous when assessing the severity of heat stress experienced in individuals or groups of dairy cows compared with using THI alone. A scoring system could aid in heat stress detection by categorizing specific heat stress indicators into different levels. The scoring system could be used a

Received April 4, 2020.

Accepted August 14, 2020.

*Corresponding author: amanda.stone@msstate.edu

tool to provide early intervention with cooling measures to prevent cows from experiencing an increased heat load.

An increase in RR is one of the most sensitive phenotypic indicators of heat stress. An RR >60 breaths/min is an indicator of heat stress in lactating dairy cows. A non-stressed dairy cow will have a RR range of 26 to 59 breaths/min. When cows are experiencing heat stress, they may begin drooling, panting, or open-mouth breathing with a protruding tongue (Shultz, 1984; Berman et al., 1985). The observation of RR and panting characteristics can be accessible tools for identifying cows that are experiencing high heat load (Tresoldi et al., 2018).

Animal health professionals and dairy producers use changes in behavioral patterns of cows as a tool for identifying poor health and welfare (Mattachini et al., 2013). In hot conditions, cows will spend more time standing and decrease activity to increase surface area for heat abatement, sensible water loss, radiating surface area, and air movement via convection (Cook et al., 2007; Allen et al., 2015; Polsky and von Keyserlingk, 2017). Increased standing time in a 24-h period caused by heat stress may contribute to a decrease in milk production, an increase in disease prevalence (Cook et al., 2007; Polsky and von Keyserlingk, 2017), and an increase in body temperature (Anderson et al., 2013). Blood flow to the udder is limited in a standing position compared with a lying position, which limits the amount of nutrient uptake by the mammary gland and contributes to a decrease in milk production (Rulquin and Caudal, 1992; Delamaire and Guinard-Flament, 2006).

When provided shade, heat-stressed cows increase time spent ruminating (Blackshaw and Blackshaw, 1994), have a higher milk yield (West, 2003a), and have decreased body temperatures compared with unshaded heat-stressed cows (Kendall et al., 2006). To thermoregulate, dairy cows will spend the majority of their time seeking and standing in shaded areas in the pasture (Schütz et al., 2008; Vizzotto et al., 2015). Cows will continue seeking shade and standing in an attempt to reduce internal body temperature, even when deprived of lying for 12 h (Schütz et al., 2008). However, heat-stressed cattle tend to lie down and reduce activity during daylight hours to reduce energy expenditure. When experiencing heat stress, cattle will actively seek shade during the heat of the day, and if no shade is available, an animal will change posture or behavior in an attempt to dissipate the accumulating heat (Bianca, 1968; Finch, 1986).

Cooling with sprinklers can decrease some of the negative effects associated with heat stress (Valtorta and Gallardo, 2004; Kendall et al., 2007; Chen et al.,

2013). Researchers reported that cows with access to sprinklers had a reduced RR compared with cows with access to shade (Tarazón-Herrera et al., 1999; Kendall et al., 2007). Various cooling strategies have been evaluated in confinement settings of dairy cattle in response to heat stress, but minimal research has been conducted to assess how different heat abatement methods can affect the level of heat stress of grazing dairy cows. Therefore, the objectives of this study are to (1) design and utilize a heat stress scoring system to evaluate heat stress severity on grazing dairy cows with access to different heat abatement strategies; (2) analyze and predict the accuracy of the scoring system with machine learning methods; and (3) validate the machine learning methods by using a real-life case study.

Machine learning has become a valuable tool for prediction in many fields due to its versatility and ability to derive a model from available data without previous knowledge of the relationship between the variables (McQueen et al., 1995; Kotsiantis et al., 2007). Large data sets that may have non-normally distributed data are able to be interpreted using machine learning with fewer assumptions made about the data (Gahegan, 2003; Gianola et al., 2014) compared with traditional linear methods. Machine learning has been used in the field of dairy science to predict traits such as milk production, mastitis (Kamphuis et al., 2010; Ebrahimie et al., 2018), and methane production (Zheng, 2018). Traditional linear methods, like logistic regression, are still utilized for predictions because machine learning may not always be the best fit (Van Hertem et al., 2014; Hempstalk et al., 2015; Ghafouri-Kesbi et al., 2017). The different methods can be compared in some circumstances, but it can be difficult to determine which method will result in the highest accuracy beforehand (White et al., 2018). Many different machine learning techniques exist and may be suitable to predict the variable of interest (van der Heide et al., 2019). Therefore, a trial-and-error approach can be used to find the best method for each prediction (Amrine et al., 2014; Libbrecht and Noble, 2015).

The naïve Bayes and random forest machine learning methods are commonly used in animal science research. These 2 methods use very different approaches, where the naïve Bayes is a family of classifiers that uses Bayesian techniques to form a simple network based on previous probabilities (Jensen, 1996) and the random forest method makes use of decision trees, or a sequence of rules that splits the data in a way that most optimally reduces variation (Breiman, 2001). To the authors' knowledge, research on heat stress in dairy cows using machine learning methods does not yet exist. In this study, traditional linear regression method will be com-

pared with machine learning methods (Gaussian naïve Bayes and random forest) to find the most accurate method to predict heat stress score using a novel heat stress scoring system.

MATERIALS AND METHODS

This experiment was conducted at the Mississippi State University Bearden Dairy Research Center in Starkville, Mississippi, from July 1, 2019 to August 8, 2019. All animals were treated according to Institutional Animal Care and Use Committee standards and protocols for Mississippi State University (IACUC-19-227). A randomized experimental design was used with 3 treatments. The treatments consisted of control (**CON**, no heat abatement), shade (**SH**, portable shade structure with 80% protective shade cloth; Shade Cloth Store, Mundelein, IL), and sprinklers (**SP**, polyvinyl chloride sprinkler system, 3 m high; Figure 1). The study consisted of 9 pens (0.4 ha each), with each treatment replicated 3 times. Each shade pen had one shade structure that provided 4.6 m² of shade space per cow (Higgins et al., 2011). The shade structures were moved around each of the pens 4 times a week to prevent accumulation of mud. Each sprinkler pen had 2 sprinkler systems and provided each cow with 6 m of sprinkler space. The sprinklers remained on for the duration of the study as THI was anticipated to be ≥ 68 based on 11 yr of climate data (US Climate Data, 2020), and were rotated 3 times per week. All treatment pens included 2 pregnant lactating Holstein cows and 1 pregnant lactating Jersey cow, with 27 cows total enrolled in the study. All cows were in their second trimester. Table 1 depicts the mean parity, DIM, days in gestation, and SCC for each treatment pen at the start of the study. All groups walked 630.94 m to the parlor twice each day at 0330 and 1530 h. All cows had access to shade in holding pens before milking, access to water ad libitum, and were fed the same concentrate ration at 15.11 kg/animal per d after each milking (Table 2; Ware Milling, Houston, MS).

The study took place on a 7.39-ha pasture. The east side of the pasture was divided into five 0.40-ha pens and the west side of the pasture was divided into four 0.40-ha pens (Figure 2). Pens were divided using electric fence wire (0.16 cm Poly Wire, Gallagher USA Electric Fencing, Riverside, MO). All pens were planted with Hybrid Pearl Millet (Tifleaf III; 9.1 kg/ha; Wax Company Inc., Amory, MS). Prior to the start of the study, cows were housed in a freestall barn on the same farm, and pastures were not grazed. Sixteen days before the study began, cows were gradually acclimated to pasture with shade access and acclimated to the concentrate ration. The acclimation pasture was

in a separate location from where the treatment pens were located.

Two days before cows began grazing, forage samples were hand-collected by 2 individuals. Walking in a zig-zag pattern, 9 forage samples were collected from each pen using hand shears. Available forage in a 0.25-m² area was cut to a ground level (Hughes et al., 2010). Samples were placed in plastic bags and hand mixed to create a sample representative of the whole pen. All samples were stored in a -80°C freezer until after the study was completed. Forage samples were removed from the freezer and dried for 48 h in a 55°C forced-air oven to determine DM concentration and calculate forage mass. The dried samples were then sent to Dairy

Figure 1. The polyvinyl chloride pipe sprinkler system (3 m tall) with a shrub sprinkler nozzle on each end (4.57 m water spray distance each) created for this project. The study evaluated the performance of 3 methods (Gaussian naïve Bayes, random forest, and logistic regression) in predicting heat-stressed and non-heat-stressed grazing dairy cattle subjected to different cooling methods.

Table 1. Mean $\pm$ SD at the start of the study for parity, DIM, days in gestation, and SCC for each treatment group (control, shade, and sprinklers)^{1,2}

Variable	Treatment			P-value		
	Control	Shade	Sprinkler	Control vs. shade	Control vs. sprinkler	Shade vs. sprinkler
Parity	2.22 $\pm$ 0.16	2.66 $\pm$ 0.47	2.22 $\pm$ 0.16	0.47	1.0	0.47
DIM	208.22 $\pm$ 29.83	202.33 $\pm$ 3.78	183.99 $\pm$ 33.71	0.98	0.66	0.79
Days in gestation	104.55 $\pm$ 9.70	95.11 $\pm$ 17.89	81.77 $\pm$ 11.68	0.83	0.35	0.69
SCC (100 cells/mL)	111.0 $\pm$ 196.14	163.83 $\pm$ 349.13	47.56 $\pm$ 83.28	0.79	0.71	0.33

¹The study evaluated the performance of 3 methods (Gaussian naïve Bayes, random forest, and logistic regression) in predicting heat-stressed and non-heat-stressed grazing dairy cattle subjected to different cooling methods. Treatment pen: control = no heat abatement; shade = portable shade structure with 80% protective shade cloth; sprinkler = portable, 3 m tall, polyvinyl chloride pipe sprinkler system.

²Significance was considered at $P \leq 0.05$.

One Forage Lab (Dairy One Cooperative Inc., Ithaca, NY) to be analyzed for nutritive value.

A grazing stick was used to measure forage height in each treatment pen 1 d before study start and once per week during the study. Walking in a zigzag pattern, forage was measured from 9 different locations throughout the pen and averaged to calculate mean forage height for each pen. If forage height was grazed to an average of 18 cm or less, the specified treatment pen was expanded by 12 m in one direction to provide optimal grazing space. All pens that were expanded 12 m in one direction were shortened by 12 m on the opposite side to keep all pastures the same size. Sprinkler pen 1 and pen 2 were expanded according to the above description on July 13, 2019, and July 27, 2019, respectively. Cows in CON pen 6a were all moved to CON pen 6b

on July 7, 2019, due to flooding that occurred from the run-off created by the 2 SP pens located up a slight slope to the north of pen 6a (Figure 2).

The temperature and humidity was measured in a central location between the pens at 10-min intervals using a weather station with a data logger (Vantage Vue, WeatherLink USB data logger, Davis Instruments, Hayward, CA). Temperature-humidity index was calculated to measure thermal comfort of cows using the following equation from Mader et al. (2006): $0.8 \times \text{temp}^{\circ}\text{C} + [(\text{HUM}/100) \times (\text{temp}^{\circ}\text{C} - 14.4)] + 46.4$, where $\text{temp}^{\circ}\text{C}$ = ambient temperature in $^{\circ}\text{C}$ and HUM = relative humidity.

Physiological Variables

Reticulorumen boluses (SmaXtec, Graz, Austria) recorded reticulorumen temperature (**RT**) for all cows every 10 min. Respiration rates were visually observed by recording the time in seconds for 10 full breaths by watching flank movements, then multiplying by 60 to calculate breaths per minute (Tresoldi et al., 2018). The same 3 d each week, all cows were observed 3 times/day for 2-h periods at 0630 to 0830, 1100 to 1300, 1600 to 1800 h. Two trained researchers observed RR throughout the study. Before the study started, both researchers practiced observing RR together for 2 wk to ensure similarity between each observation. One researcher observed RR for 2 d each week and the other researcher observed RR for 1 d each week. Body condition score as described by Ferguson et al. (1994) and hygiene score (**HS**) as described by Schreiner and Ruegg (2002) were collected once weekly. For hygiene scoring, a score of 1 represented udder and legs that were free from dirt, 2 represented slightly dirty udder and legs and 2% to 10% of the areas were covered with dirt, 3 represented moderately dirty udder and legs with 10 to 30% of the area covered with dirt, and 4 represented >30% of the udder and legs were covered with dirt. Automated milk yields (**MY**; kg/d) were recorded for every cow at each

Table 2. List of ingredients in the pelleted concentrate ration that was fed to all cows throughout the 39-d study¹

Ingredient ²	Percent
Ground corn	39.00
20% high-fiber dairy pellet	28.97
Cottonseed hulls	10.17
Dried distillers grain	8.13
Soybean meal	6.75
Molasses	3.13
Feed-grade limestone	1.63
Sodium bicarbonate	1.25
Clear mixing salt	0.63
Magnesium oxide	0.16
Clarifly 0.67%	0.07
Beef and dairy vitamin mix	0.06
Vitamin A, D, E premix	0.03
Zinpro Avalia 4	0.03
Vitamin E 50%	0.005

¹The study evaluated the performance of 3 methods (Gaussian naïve Bayes, random forest, and logistic regression) in predicting heat-stressed and non-heat-stressed grazing dairy cattle subjected to different cooling methods. Control = no heat abatement; shade = portable shade structure with 80% protective shade cloth; sprinkler = portable 3 m tall, polyvinyl chloride pipe sprinkler system.

²Clarify 0.67%: Central Life Sciences, Schaumburg, IL; Beef and dairy vitamin mix: Ware Milling, Houston, MS; Vitamin A, D, E premix: Ware Milling; Zinpro Avalia 4: Zinpro Corporation, Eden Prairie, MN.

milking (Dairyplan DP5, Westfalia Surge, GEA Group, Düsseldorf, Germany). Milk samples were collected once per week and shipped to Mid-South Dairy Records (Springfield, MO) for analysis of milk components (fat and protein, %) and SCC. Somatic cell count was log-transformed to SCS using the following equation: $SCS = 1,000 + 100 \times [\log_2(SCC/1,000)]$ to fit the Gauss normal distribution (Ali and Shook, 1980).

Behavioral Variables

Cows were fitted with AfiAct II (Afimilk, Kibbutz Afikim, Israel) leg sensors to continuously monitor activity (e.g., lying time, number of lying bouts, steps) throughout the study. One day before the study start, leg sensors were attached to the left rear leg of each cow using the AfiAct II leg strap. Panting characteristics were recorded for a 2-min interval 3 times/day for every day of the study by the same observer. Panting characteristics included drooling, determined by saliva hanging from the cow's mouth when she was not ruminating; open-mouth panting, determined by an open mouth when the cow was not ruminating; and open-mouth panting with tongue protruding, determined if at least the tip of the tongue crossed the edge of the

bottom lip (Martin and Bateson, 2007; Tresoldi et al., 2016).

Heat Stress Scoring

The study was planned for 60 d unless the cows were judged to experience repeated episodes of severe heat stress. As the identification of severe heat stress is to some degree subjective, the researchers created a heat stress scoring system based on aspects of other scoring systems (Tresoldi et al., 2016; US Meat Animal Research Center, 2017; Woolums et al., 2018). This scoring system was used to score all cows 3 times each day. The heat stress scoring system used RR and various behavioral measures to assess the cow's level of heat stress. The scoring system ranged from 0 to 4, where 0 = no signs of heat stress, ≤ 60 breaths/min up to 79 breaths/min; 1 = mild heat stress, elevated RR (80 to 99 breaths/min), increased time spent standing, slight drooling (1 string of saliva hanging from mouth), and restlessness; 2 = moderate heat stress, elevated RR (100 to 119 breaths/min), excessive drooling (2 to 3 strings of saliva hanging from mouth) or foaming, lethargic, but looks more alert when approached, head carriage low, and most animals in pen are standing;

Figure 2. Drawing of the treatment pen setup and the path used to walk to the milking parlor. The study evaluated the performance of 3 methods (Gaussian naïve Bayes, random forest, and logistic regression) in predicting heat-stressed and non-heat-stressed grazing dairy cattle subjected to different cooling methods. All treatment pens were 0.4 ha with all cows using the same path (630.94 m) to walk to the milking parlor. Sprinkler pen 1 and sprinkler pen 2 were both expanded by 12 m on one side to increase forage availability (indicated by red arrows), and space was decreased on the opposite side to keep pen at 0.4 ha. Cows in control pen 3a were all moved to control pen 3b on July 7, 2019, due to flooding that occurred from the run-off of mister pen 2 and mister pen 3 that were located up a slight slope to the north of control pen 3a.

3 = severe stress, increased breathing effort, elevated RR (≥ 120 breaths/min), open-mouth breathing or panting with tongue protruding, possible drooling (>3 strings of saliva hanging from mouth), head carriage low, lethargic, and may not look more alert when approached; and 4 = moribund, open-mouth breathing with tongue protruding, breathing is labored and RR may decrease, pushing from flank when breathing, not necessarily drooling, individual animals may be isolated from others.

Heat stress score was used as a means to assess the level of heat stress and as a means of assessing when an animal might warrant removal from the study to prevent excessive animal suffering. The same researcher scored all cows at all scoring periods to maintain consistency. If a cow had a score of 3 on any day, the researcher would rescore the cow 1 h later. If the cow remained a 3 or was scored a 4 in the second scoring, a veterinarian was called to examine the cow. A heat stress score of 4, determined by the scorer or veterinarian, warranted immediate removal from the study. This was considered an emergency and any animal experiencing this was to be moved to the freestall barn and treated in the manner deemed necessary by the veterinarian. If the veterinarian scored the cow a 3 after being called, the cow would remain on the study. However, these cows were taken to the holding pen and hosed off with cold water to cool them to prevent them from advancing to a score of 4. If any cow received a score of 3 for 3 consecutive scoring times, the cow was removed from the study. To balance the importance of both sample size and animal welfare, the removal of 3 cows from any one treatment (CON, SH, or SP) was set as a predetermined point before the study began to warrant ending the study before the 60 d.

Statistical Analysis

Organization of Raw Data. The MEANS procedure of SAS version 9.4 (SAS Institute Inc., Cary, NC) was used to calculate the average daily THI and wind speed and RR, RT, and panting characteristics (drooling, mouth open, tongue protruding) for each cow on all 3 treatments during the study period. Lying time, number of lying bouts, and steps were summed per cow per day. Morning and afternoon MY were summed per cow per day. Percent of milk fat, percent of milk protein, SCS, BCS, and HS data were only collected once per week and therefore were used for that entire study week in the analysis.

Data Preprocessing for Machine Learning. Three methods were evaluated: logistic regression, Gaussian naïve Bayes, and random forest. Logistic regression is a linear method, whereas the other 2 are

machine learning methods. Researchers continue to use traditional linear methods over machine learning because, despite the great potential for machine learning, these methods have not always proven superior to traditional linear modeling (Cortez et al., 2006; Van Hertem et al., 2014). The 3 methods included in this study were chosen through a trial-and-error approach to evaluate different machine learning methods. This may provide different accuracy on the prediction of the variable of interest for the same data set. Both random forest and Gaussian naïve Bayes have been successfully implemented in animal science (Jensen, 2012; Shahinfar et al., 2014). Random forest and Gaussian naïve Bayes represent 2 families of machine learning classifiers, one using decision trees and the other using Bayesian-based methods, respectively. By comparing a traditional linear method to the machine learning classifiers, these results could help dairy researchers understand the advantages and disadvantages of each classifier in the prediction of heat stress in dairy cows and help gain a better understanding of the wide variety of factors that affect heat stress.

Prior to fitting models according to these methods, feature scaling was taken into account. The 2 common feature scaling approaches are standardization and normalization. Standardization maintains information about outliers and makes the algorithm less sensitive in contrast to normalization. Optimization algorithms used inside the logistic regression method perform better if features are on the same scale. Therefore, with logistic regression, the features are scaled through standardization.

The 18 features that were included into the analysis were RR, RT, daily milk yield, THI, wind speed, panting characteristics (drooling, mouth open, tongue protruding), milk fat percent, milk protein percent, HS, BCS, lying time, number of lying bouts, number of steps, breed, parity (primiparous or multiparous), and DIM. The original data set consisted of 21 features. In this data set, in addition to the 18 features that are mentioned above, the data included heat stress score, heat index, and dew point. The visualization of the correlation of the features is plotted in Figure 3. After finding the correlations for those features, the correlation is >0.9 ; therefore, only one of the variables was kept in the data set. It can be observed that the response variable is highly correlated with heat stress score, drooling, and RR. However, among these 3 features, the correlation between the heat stress score and the response variable is 0.92; therefore, this feature was removed from the data set. In addition, a strong positive correlation existed between the THI, dew point, and heat index. The correlation between the THI and the dew point was 0.96, and the correlation between the

THI and heat index was 0.92. Given this result, the dew point and heat index were removed from the data set. All other correlations were <0.9 and were kept in the data set. The visualization of the correlation between features after removing some of the features depicted in Figure 4.

Description of Methods. The analyses were performed in Python version 2.7, using scikit-learn libraries. To apply logistic regression on the data set, LogisticRegression classifier in the scikit-learn package was used. In this method, the regularization or penalizing of coefficients was applied by default. Two regularization methods could be applied: (1) L1 regularization (LASSO) and (2) L2 regularization (RIDGE). Using L1 regularization, some of the coefficients associated with features in the data set can become zero and will be eliminated from the model. This type of regularization

is suitable for data sets that possibly have collinearity among features. On the other hand, applying L2 regularization does not lead to the elimination of the features. To avoid the possibility of collinearity among features in the data set and to maintain the interpretability of models, L1 regularization was used. In the LogisticRegression classifier, using hyperparameter C, it can handle the inverse of regularization strength. This parameter is set to 0.1, 1, 10, 100, and 1,000, with 1,000 representing the strength of regularization as very low, and therefore the method can be considered as the logistic regression without regularization. The GridSearchCV library was used to assess the performance of the logistic regression model based on different values of the C hyperparameter.

This library has a parameter cv, which determines the cross-validation strategy of the optimization pro-

Figure 3. Visualization of the correlation of 21 features from the original data set for evaluating 3 machine learning methods (logistic regression, random forest, and Gaussian naïve Bayes) for predicting heat stress in cows subjected to 3 heat abatement treatments (no heat abatement, shade, or sprinklers). THI = temperature-humidity index.

Figure 4. Visualization of the correlation of the 18 features that were included in the analysis when evaluating 3 machine learning methods (logistic regression, random forest, and Gaussian naïve Bayes) to predict heat stress in cows subjected to 3 heat abatement treatments (no heat abatement, shade, or sprinklers).

cedure. The default value of this hyperparameter was used, which means that 5-fold cross-validation was used. To report each one of the metrics using regularized logistic regression, 2 nested loops were used. The outer loop was in charge of creating 20 trials or 20 random segmentation of the data set, and the inner loop was in charge of performing the GridSearchCV with 5-fold cross-validation. For example, in trial 1 (outer loop), the first random segmentation of the data set was used. In this trial, the data set was split into a test set and a rest set: the test set was set aside, and the rest set was used in the inner loop. In the inner loop, the rest set was again split into a validation set and a training set. Then, the training set was used to develop a series of regularized logistic regression mod-

els with different values of C , and these regularized logistic regression models were used to predict samples in the validation set, which is further used in the selection of an optimal value of C hyperparameter for the current trial. In the inner loop, which was responsible for optimizing the value of C hyperparameter for the current trial, we set the number of folds of the cross-validation of GridSearchCV to 5. In other words, the inner loop was repeated until all samples from the rest set were included in the validation set once. Having obtained optimal C , for the rest set, then a separate regularized logistic regression model with optimal C was calculated, and this model is further used in the outer loop to predict the test set samples. This procedure was repeated for the remaining trials, and for each

trial, a list of metrics based on the optimized value of C was gathered. Finally, by averaging the performance metrics that were obtained in 20 trials, the corresponding metrics of the regularized logistic regression model were reported. This procedure was designed in a way that the assessment of model quality and the model optimization was independent, and samples that were used in the final model assessment were not used in the model optimization procedure.

The Gaussian naïve Bayes is the Bayesian method that was used in this analysis where no features needed to be scaled before fitting the model. For the random forest method, the number of trees is fixed to 500, and the number of features to consider when looking for the best split was set to the square root of the total number of available variables. These values were used because they resulted in the highest area under the curve (AUC) values for the experiments. To obtain the best value for the number of trees before performing all the experiments, GridSearchCV (this library by default performs 5 fold cross-validation) was used by calculating a model for each one of the 20 trials on the combined (COMB) data set. To do so, the metric based on GridSearchCV attempted to find the optimized number of trees set to the area under the receiver operating characteristic (ROC) curve. Further, the number of trees that were evaluated through GridSearchCV was set to 500, 600, 700, 800, 900, and 1,000. According to this experiment, the frequency of selection of 500 trees by naïve GridSearchCV for the 20 trials was greatest. With this, 500 was set as the number of trees in experiments with the random forest classifier. To run the experiments and get validation of the results, the data set was randomly split into 2 sets: training and testing. The training set consisted of 70% of the data set, and the testing set consisted of the remaining 30% of the data set. This random segmentation was performed 20 times on the data set and with different groups of the treatments. The CON group consisted of 119 non-heat-stressed samples and 153 heat-stressed samples; the SH group had 118 non-heat-stressed samples and 155 heat-stressed samples; and the SP group had 261 non-heat-stressed samples and 50 heat-stressed samples. The number of total samples for the aggregation of the 3 treatment groups (COMB) was 856. For each class, the number of samples was 498 samples for non-heat stressed and 358 samples for heat stressed. To further maintain the information in the data set, stratified sampling was used. Applying such splitting generates training and test subsets that have the same proportions of class labels as the input data set.

Validation of Methods. For this data set, the variable in focus was HS. Based on this variable, the whole data set was divided into 2 major classes: heat-stressed

and non-heat-stressed animals. If the value of HS for an animal was zero, they were entered in the non-heat-stressed class (class label = 0). If the value of HS for an animal was 1, 2, 3, or 4, they were classified as a heat-stressed animal (class label = 1, positive class).

To assess the performance of each method, 4 metrics were used: accuracy, sensitivity, specificity, and F_1 score. The positive class in the data set represents the animals that displayed signs of heat stress, where accuracy indicated the rate of correctly predicted animals in the data set, sensitivity indicated the proportion of heat-stressed animals that were correctly predicted as heat-stressed animals, and specificity indicated the proportion of non-heat-stressed animals that were correctly predicted as non-heat-stressed animals. Precision is the sensitivity rate and recall is the rate of correctly identified heat-stressed animals divided by the number of relevant animals (the heat-stressed animals that are truly predicted as heat-stressed plus the non-heat-stressed animals that are falsely predicted as heat-stressed animals). Using the following equation, F_1 score was calculated:

$$F_1 \text{ score} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}.$$

To further evaluate the performance of these methods, the area under the ROC curve was determined using the roc_auc_score metric in scikit-learn (Pedregosa et al., 2011). A ROC curve is created by plotting the true positive rate versus the false positive rate at various thresholds of the classifier. True positive rate and false positive rate are illustrated as follows:

true positive rate = sensitivity,

false positive rate = 1 – specificity.

Consistency of the methods or the ability of these methods to provide similar probability scores for samples in the test set were of interest. After applying each one of the machine learning methods, the probability of the samples for each class in the model was obtained. The output is a 2-dimensional matrix where the columns represent the 2 classes, and the rows represent samples in the testing set. For each one of the 20 trials (k), the first column of the probability matrix, denoted by V , represents the probability of a positive class (heat-stressed class). The following vectors of probabilities, V_{RF}^k , V_{LR}^k , and V_{GNB}^k , were obtained, where RF = random forest, LR = logistic regression, and GNB =

Gaussian naïve Bayes. For each pair of these 3 methods, the respective correlations are obtained as follows, where P and S are the functions using Pearson's r and Spearman's ρ correlation coefficients between 2 vectors:

$$\begin{aligned} P(V_{RF}^k, V_{LR}^k) &= r_{RF-LR}^k, \quad P(V_{RF}^k, V_{GNB}^k) = r_{RF-GNB}^k, \\ P(V_{LR}^k, V_{GNB}^k) &= r_{LR-GNB}^k, \\ S(V_{RF}^k, V_{LR}^k) &= \rho_{RF-LR}^k, \quad S(V_{RF}^k, V_{GNB}^k) = \rho_{RF-GNB}^k, \\ S(V_{LR}^k, V_{GNB}^k) &= \rho_{LR-GNB}^k. \end{aligned}$$

After finding the Pearson and Spearman correlations for each one of the trials and then taking the average of the 20 trials, the reported correlation values were calculated as follows:

$$\begin{aligned} r_{RF-LR} &= \frac{\sum_{i=1}^{20} r_{RF-LR}^k}{20}, \quad r_{RF-GNB} = \frac{\sum_{i=1}^{20} r_{RF-GNB}^k}{20}, \\ r_{LR-GNB} &= \frac{\sum_{i=1}^{20} r_{LR-GNB}^k}{20}, \\ \rho_{RF-LR} &= \frac{\sum_{i=1}^{20} \rho_{RF-LR}^k}{20}, \quad \rho_{RF-GNB} = \frac{\sum_{i=1}^{20} \rho_{RF-GNB}^k}{20}, \\ \rho_{LR-GNB} &= \frac{\sum_{i=1}^{20} \rho_{LR-GNB}^k}{20}. \end{aligned}$$

This procedure was repeated for all the data sets. The correlations indicate how similar the probability of the predicted samples is across the 3 methods, and based on these probabilities, cows can be ranked and be provided the appropriate heat abatement.

RESULTS

Cows Removed from the Study

Treatments were applied, and data collection started on July 1, 2019, and ended August 8, 2019 (39 d of data collection). Due to receiving a heat stress score of 3 on 3 consecutive measurements, as described in the Materials and Methods, 2 cows were removed from the study on d 31 (one CON pen 3 cow and 1 SH pen 2 cow). Two more cows were removed on d 38 of the study (one CON pen 1 cow and 1 SH pen 2 cow), again for receiving a heat stress score of 3 on 3 consecutive measurements. On d 39 of the study, 2 more cows met the standards of a severe heat stress concern and were removed (2 CON pen 2 cows). Due to 3 cows being removed from the CON group, the study concluded on the afternoon of August 8, 2019. No cows received a heat stress score of 4 throughout the duration of the experiment. Two cows (1 cow in SH and 1 cow in SP) aborted on July 14 and on approximately July 17, 2019, respectively. Both cows were examined by a veterinarian and were approved to stay in the study. The cause of abortion could not be determined.

Body condition score at the start of the study was 2.89 ± 0.19 for CON, 3.15 ± 0.23 for SH, and 2.75 ± 0.07 for SP. No significant difference in BCS existed between treatment groups at the study start. Nutritive values for each treatment are depicted in Table 3. No significant difference existed between any of the treatments for any of the nutritive values measured.

Climatic Conditions

The mean THI during the study was 76.81 ± 4.99 , with a maximum THI of 88.11 on August 7, 2019, and a minimum of 61.81 on July 16, 2019. On July 13,

Table 3. Nutritive value analysis for Hybrid Pearl Millet (Tifleaf III, Wax Company Inc., Amory, MS) for the 3 treatments (control, shade, and sprinkler)¹

Variable	Treatment			P-value		
	Control	Shade	Sprinkler	Control vs. shade	Control vs. sprinkler	Shade vs. sprinkler
DM, %	23.9 ± 4.9	22.0 ± 2.5	23.0 ± 2.5	0.84	0.96	0.96
CP, %	13.5 ± 2.1	16.8 ± 2.6	17.4 ± 0.7	0.23	0.15	0.96
ADF, %	26.4 ± 3.0	26.5 ± 4.0	26.5 ± 4.0	0.84	1.0	1.0
aNDF, ² %	52.7 ± 3.3	50.9 ± 0.8	50.0 ± 4.3	0.82	0.64	0.95
NFC, %	21.6 ± 2.4	20.0 ± 2.8	20.4 ± 3.4	0.83	0.89	0.99
TDN, %	57.8 ± 1.1	58.3 ± 0.5	58.6 ± 0.9	0.77	0.53	0.92
NE _L , Mcal/kg	0.25 ± 0.01	0.26 ± 0.00	0.26 ± 0.01	0.80	0.62	0.95

¹The study evaluated the performance of 3 methods (Gaussian naïve Bayes, random forest, and logistic regression) in predicting heat-stressed and non-heat-stressed grazing dairy cattle subjected to different cooling methods. Mean ± SD for each treatment (control = no heat abatement; shade = portable shade structure with 80% protective shade cloth; sprinkler = portable 3 m tall, polyvinyl chloride pipe sprinkler system).

²aNDF = ash-free neutral detergent fiber.

Table 4. The average of each performance metric over 20 trials for each of the treatment groups¹

Method	Group	Accuracy	Sensitivity	Specificity	F ₁ score	AUC
Random forest	COMB	0.888	0.928	0.863	0.87	0.945
	CON	0.853	0.902	0.809	0.864	0.934
	SP	0.893	0.952	0.891	0.53	0.917
	SH	0.853	0.899	0.803	0.871	0.894
Logistic regression	COMB	0.875	0.913	0.851	0.855	0.944
	CON	0.85	0.889	0.816	0.865	0.933
	SP	0.86	0.649	0.894	0.501	0.9
	SH	0.837	0.884	0.786	0.858	0.896
Gaussian naïve Bayes	COMB	0.822	0.836	0.812	0.796	0.931
	CON	0.825	0.879	0.771	0.839	0.928
	SP	0.817	0.453	0.883	0.42	0.891
	SH	0.811	0.849	0.768	0.838	0.896

¹The study evaluated the performance of 3 methods (Gaussian naïve Bayes, random forest, and logistic regression) in predicting heat-stressed and non-heat-stressed grazing dairy cattle subjected to different cooling methods. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth). AUC = area under the curve.

2019, hurricane Barry made landfall 632.31 km south of Starkville, Mississippi, on the Gulf Coast of Louisiana. Throughout the study, Starkville received 28.7 cm of rainfall, with 27.13 cm occurring in July. Starkville received 17.9 cm more rainfall for the month of July compared with the average July rainfall for the past 5 yr (US Climate Data, 2020).

Significance of Machine Learning Results

The value of reported metrics could be the result of a lucky random choice of the samples in the 20 submodels. To alleviate this problem and give a measure of statistical significance of the utilized metrics, a permutation test was performed. In the permutation test, these 20 random segmentations of each data set were used as the original submodels. The average performance metrics over 20 submodels for the different data sets are presented in Table 4. In the permutation test, the label of samples is permuted, and a new model is fitted on the permuted data set. The value of performance metrics based on the permuted data set is expected to be lower than the average value of the original submodels. By repeating this process many times, a hypothesis test for each one of the performance metrics is obtained. In this study, accuracy, sensitivity, specificity, F₁ score, and AUC were used as the performance metrics of the model where the upper bound (P) of the *P*-value for accuracy is calculated with the following equation:

$$P = \frac{1 + \#(\text{accuracy}_p \geq \text{accuracy})}{N},$$

where accuracy represents the average value of accuracy from the 20 submodels, accuracy_p is the accuracy of each one of the permuted data sets, and N represents

the number of permutation tests. Further, # (accuracy_p ≥ accuracy) represents the number of the permuted data set in which the accuracy of the model is greater or equal to the average value of the accuracy based on the submodels. For example, to attain a *P*-value less than 0.01, at least 100 permutations are necessary but cannot be sufficient (Churchill and Doerge, 1994). In this study, to obtain an accurate *P*-value, 3,000 permutation tests were performed. The *P*-value of each of the 5 metrics for each of the data sets is represented in Table 5. The value of the significance level was set to *P* ≤ 0.05. The null hypothesis (no difference between the 2 classes) would be rejected if the *P*-value of a test was <0.05. For the rest of the metrics, a similar approach was used to calculate the threshold of the *P*-value.

For COMB, CON, and SH data sets, the *P*-value of the models calculated for the 3 methods and for all 5 metrics were less than 0.05. Therefore, the null hypothesis was rejected, and the observed difference between the 2 classes was statistically significant at α = 0.05. In reference to sensitivity for the SP group, the *P*-value of models calculated for all 3 methods was >0.05. Therefore, when using sensitivity as the performance metric, the null hypothesis cannot be rejected, and no proof exists for the statistical significance of observed differences at α = 0.05. However, for the SP group, when using other metrics, *P* < 0.05. With this, it appears that using the other 4 performance metrics, the observed difference between the 2 classes is statistically significant at α = 0.05.

Machine Learning Results

The random forest and logistic regression methods outperformed the Gaussian naïve Bayes method with respect to accuracy (Figure 5). However, for cows in the CON group, the difference was negligible. The per-

Table 5. The P -value of the permutation tests according to all performance metrics for each treatment group¹

Method	Group	Accuracy	Sensitivity	Specificity	F ₁ score	AUC
Random forest	COMB	0.0003	0.0003	0.0003	0.0003	0.0003
	CON	0.0003	0.0003	0.0006	0.0003	0.0003
	SP	0.0003	0.07	0.0003	0.0003	0.0006
	SH	0.0003	0.0003	0.001	0.0003	0.0003
Logistic regression	COMB	0.0003	0.0003	0.0003	0.0003	0.0003
	CON	0.0003	0.0003	0.0003	0.0003	0.0003
	SP	0.01	0.067	0.0006	0.001	0.003
	SH	0.0003	0.0003	0.0006	0.0003	0.0003
Gaussian naïve Bayes	COMB	0.0003	0.0003	0.0003	0.0003	0.0003
	CON	0.0003	0.0003	0.001	0.0003	0.0003
	SP	0.0425	0.112	0.003	0.005	0.002
	SH	0.0003	0.0003	0.002	0.0003	0.0003

¹The study evaluated the performance of 3 methods (Gaussian naïve Bayes, random forest, and logistic regression) in predicting heat-stressed and non-heat-stressed grazing dairy cattle subjected to different cooling methods. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth); AUC = area under the curve. Significance was considered at $P \leq 0.05$.

formance of random forest and logistic regression with respect to the accuracy is similar for all the treatment groups and for COMB. However, for the SP group, random forest outperformed the logistic regression method. Random forest provided the most consistent and accurate results for all groups.

Of the 3 methods, random forest had the highest sensitivity rate for all the groups (Figure 5). In terms of sensitivity rate (precision) for the COMB, CON, and SH groups, the performance of the logistic regression and Gaussian naïve Bayes method was similar to the performance of the random forest method. However, the logistic regression and Gaussian naïve Bayes methods indicated a poor performance concerning the sensitivity rate in the SP group of animals. In terms of specificity rate, random forest and logistic regression outperformed the results obtained by the Gaussian naïve Bayes. The trend of specificity rate was similar for all 3 methods. In the CON group compared with COMB group, all the classification methods scored less with respect to the specificity rate. Likewise, the specificity rate of the CON group was less than the specificity rate of the SP group. However, all these methods scored high in the SH group compared with the SP group.

The best value of the F₁ score is 1 (perfect precision and recall), and with respect to this metric, the logistic regression and the random forest methods represented a good performance and outperformed the Gaussian naïve Bayes in all the groups (Figure 5). In the SP group, none of the methods indicated promising results with respect to the F₁ score metric. Two metrics to compare various methods are precision and recall. These 2 metrics can show different behaviors for different methods. For example, if one algorithm has higher precision but lower recall than the others, which algorithm is better? One possible way to tackle this question is F₁ score. This metric can be used to determine the superiority

of different classification methods considering the value of both precision and recall metrics. In this data set in terms of precision, random forest shows better performance for the SP group. However, only based on this metric, it would not be reliable to state that using the random forest for the cows in the SP group is superior to other methods. To better understand this issue, the results associated with F₁ score need to be considered as well. Although the performance of the 3 methods in terms of F₁ score is far from the best value (1) in the SP group, the F₁ score associated with random forest outperformed the other 2 methods. This result indicated that the harmonic mean of precision and recall for this method is better than the 2 other methods. As previously mentioned, in terms of precision, the random forest indicated better results. Subsequently, results obtained for F₁ score approved the better performance of the random forest as well. Thus, we can determine that the random forest classifier is superior to the 2 other methods in the SP group of animals. Similarly, for COMB, CON, and SH groups, the random forest and logistic regression outperformed the Gaussian naïve Bayes. However, to confidently determine the best method to be used in these groups the F₁ score has to be evaluated as well. The performance of the random forest and the logistic regression method regarding the F₁ score was better than the Gaussian naïve Bayes in the COMB, CON, and SH groups of animals. Therefore, these 2 methods were preferred over the Gaussian naïve Bayes method for these groups.

In terms of AUC (Figure 6), all 3 methods represented overlapping behavior for animals in CON and SH groups, meaning that no method outperformed the other method significantly. In the COMB group, random forest and logistic regression had overlapping behavior and outperformed the Gaussian naïve Bayes. In the SP group, random forest performed better than

Figure 5. Accuracy, sensitivity, specificity, and F_1 score of logistic regression, Gaussian naïve Bayes, and random forest methods for cows subjected to 3 heat abatement treatments. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth). Error bars represent SD of the performance metrics over 20 trials.

Gaussian naïve Bayes, and Gaussian naïve Bayes outperformed logistic regression. The AUC obtained for all 3 methods was greater than 0.5, which indicates that these methods performed significantly better than the random guessing in classifying the heat-stressed animals.

The mean probability of prediction of non-heat-stressed animals for the 4 groups of animals using the 3 methods is high (Figure 7). Concerning the mean probability of predicting heat-stressed animals, logistic regression works slightly better than other methods. Further, all methods could predict the non-heat-stressed animals appropriately. However, these methods did not make identical predictions for each group and the results obtained by logistic regression outperformed the other methods. In all groups, the logistic regression compared with the other methods had a higher standard deviation. For all of the methods, the mean probability of predicting non-heat-stressed animals decreased from the COMB group to the CON group, increased from the CON group to the SP group, and decreased from the SP group to the SH group. From

these results, the mean probability of predicting non-heat-stressed animals is highest for the SP group.

Further, the mean probability of predicting heat-stressed animals (Figure 7) is considerably high for all the treatment groups except for the SP group. The results obtained for different methods represents the same trend across different groups. This means that when using 3 methods, the mean probability of predicting heat-stressed animals slightly increases from the COMB group to the CON group, decreases from the CON group to the SP group, and again increases from the SP group to the SH group. Although the difference among various methods is not considerable, and the mean probability of prediction of heat-stressed animals can be considered similar for the COMB, CON, and SH groups. In the SP group, logistic regression methods provide the highest probability of prediction of the heat-stressed animals.

To assess the consistency of the 3 different methods, the correlation was obtained using the Pearson and Spearman methods (Table 6). As expected, Spearman's ρ was generally higher than Pearson's r . All the

Figure 6. Area under the curve (AUC) of logistic regression, Gaussian naïve Bayes, and random forest methods for cows subjected to 3 heat abatement treatments. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth). The box of the boxplot indicates the first quartile, mean, and third quartile borders, and the whiskers show the highest and lowest values found.

Figure 7. Mean model output of the logistic regression, random forest, and Gaussian naïve Bayes method for cows subjected to 3 heat abatement treatments. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth). Error bars represent SD of the performance metrics over 20 trials.

correlations between different methods were positive, indicating that, in general, if the probability of being in a specific class was higher in one method, then it is higher in other methods as well.

For animals in the COMB group, the correlations were close to 1 and ranged from 0.85 (between logistic regression and Gaussian naïve Bayes) to 0.97 (between random forest and logistic regression). The correlations between different methods for the other groups, CON and SH, were slightly less than the COMB group, but followed the same trend. The correlations decreased considerably and ranged between medium and high for the SP group.

The lowest correlation for the SP group (0.53) was between the logistic regression and Gaussian naïve Bayes. The highest correlation for the SP group is 0.81 and it is between random forest and logistic regression. Overall, the correlation between the different methods for different groups ranged from moderate to high (0.53 to 0.97), and the lowest correlation can be consistently observed between logistic regression and the Gaussian naïve Bayes.

Figure 8 indicates an example of correlations in one of the 20 validation runs. In general, when the score of a method for a sample approaches to 0 or 1, the respective score of that same sample in other methods

Table 6. Pearson and Spearman correlation coefficients between the model output of the logistic regression, Gaussian naïve Bayes, and random forest models averaged over 20 runs^{1,2}

Treatment	Pearson			Spearman		
	RF-LR	RF-GNB	LR-GNB	RF-LR	RF-GNB	LR-GNB
COMB	0.97	0.87	0.85	0.93	0.89	0.90
CON	0.91	0.90	0.82	0.88	0.89	0.83
SP	0.81	0.65	0.53	0.67	0.76	0.64
SH	0.93	0.88	0.84	0.90	0.87	0.87

¹COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth).

²RF-LR represents the correlation of model output between random forest and logistic regression; RF-GNB represents the correlation of model output between random forest and Gaussian naïve Bayes; LR-GNB represents the correlation of model output between logistic regression and Gaussian naïve Bayes.

Figure 8. Visualization of the correlations between the 3 methods in 1 of the 20 validation runs. Plotted are the model output values (between 0 and 1) for logistic regression, Gaussian naïve Bayes, and random forest methods. The first row depicts correlations in the combined group, the second row shows correlations in the control group, the third row shows correlations in the sprinkler group, and the last row shows the correlation in the shade group. The first column shows the logistic regression method versus the random forest method, the second column shows the Gaussian naïve Bayes method versus the random forest method, and the third column shows the logistic regression method versus the naïve Bayes method.

approaches to 0 or 1 as well. This means that when animals in one group score high with one method, they typically score high with other methods as well. This detail is an indicator of consistency across methods or the similarity of the predicted probabilities between the methods. However, because the correlation of methods for the SP group is moderate, there would be some exceptions with these results. In the third row of Figure 8 illustrates the results for the SP group, the proportion of samples that score extremely different based on different methods was large compared with the other groups of animals.

Investigating the Imbalance of the Data Set

In this study, we used the following data sets to train classification methods that could evaluate heat stress severity on grazing dairy cows: (1) the CON data set, which consisted of 119 non-heat-stressed samples and 153 heat-stressed samples; (2) the SH data set, which consisted of 118 non-heat-stressed samples and 155 heat-stressed samples; (3) the SP data set, which consisted of 261 non-heat-stressed samples and 50 heat-stressed samples; and (4) the COMB data set, which consisted of 498 non-heat-stressed samples and 358 heat-stressed samples. All data sets were imbalanced to some extent. As our sample size for the different groups was limited, the performance metrics such as accuracy, sensitivity, specificity, and F_1 score used for evaluating the models trained by different machine learning methods may reflect the class distribution and not the goodness of the classifier. To alleviate this problem and improve the reliability and quality of the results, a set of new performance metrics were introduced to evaluate the performance of the proposed models in the prediction of the heat-stressed animals. Luque et al. (2019) introduced a method of calculating the imbalance in a data set and proposed a framework to compare the behavior of several performance metrics based on the binary confusion matrix. According to the study by Luque et al. (2019), once the classification error is not of great importance in the application study, 2 performance metrics, geometric mean (**GM**) and bookmaker informedness (**BI**), are the best choices to evaluate the performance of the classifier. However, if the classification errors are of great importance, the Matthews correlation coefficient (**MCC**) may be the best performance metric. The main target of classification methods is to improve the sensitivity rate of the classification method without negatively affecting the specificity rate. However, the aim of these 2 rates is conflicting, and as a result they may not work well when an imbalance in the data set exists. To overcome

this issue, GM and BI can aggregate the sensitivity and specificity rates, and hence are suggested for imbalanced data sets. These metrics were computed as follows (Sokolova et al., 2006):

$$GM = \sqrt{\text{sensitivity} \times \text{specificity}},$$

$$BI = \text{sensitivity} + \text{specificity} - 1.$$

Matthews correlation coefficient was introduced by Matthews (1975) and indicates the correlation between the observed and predicted classifications. This metric value is between -1 and 1 , where -1 represents total disagreement between prediction and true values and 1 means that the prediction provided by the method is ideal. In the following equation used to evaluate MCC, TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Two approaches are proposed in the machine learning community to alleviate the issue of class imbalance in the data set. The first approach assigns a distinguished cost to different classes in the data set (Pazzani et al., 1994). The second approach resamples the original data set, where a set of studies oversamples the minority class, whereas another set of studies undersamples the majority class (Kubat and Matwin, 1997; Ling and Li, 1998). In this study, we adopted the synthetic minority oversampling technique (**SMOTE**), which blends the concept of oversampling the minority class and undersampling the majority class (Chawla et al., 2002). To duplicate instances from minority class, this method first draws random instance from the minority class. Then, it seeks to find the number of nearest neighbors (k) of the selected instance. Finally, one of the selected neighbors is randomly chosen, and the new instance of the minority class is generated as a convex combination of the chosen minority instance and chosen neighbor instance. According to Chawla et al. (2002), fixing the number of selected neighbors (k) to 5 would result in better performance of SMOTE. Additionally, the other factor that could improve the efficiency of SMOTE is undersampling the majority class before generating the new instance for the minority class. In the different data sets used in this study, the minority class had 71%, 77%, 19%, and 77% of the number of examples of the majority class in COMB, CON, SP, and SH data sets, respectively. We generated 5 balanced data sets

for each of the original data sets. For example, for the COMB data set, the 5 balanced data sets had 75%, 80%, 85%, 90%, and 95% of the number of examples of the majority class, for both minority and majority classes. Balanced data sets were generated for other data sets as well. The new data sets associated with CON and SH had 80%, 84%, 88%, 92%, and 95% of the number of examples of the majority class, for both minority and majority classes. The new data sets associated with SP had 35%, 50%, 65%, 80%, and 95% of the number of examples of the majority class, for both minority and majority classes. In Figures 9, 10, and 11, the value of metrics reported for the balanced data sets represents the average value of the metrics over 5 new balanced data sets.

As can be observed from Figure 9, the random forest algorithm is able to provide similar performance with respect to 3 performance metrics, namely GM, BI, and MCC, for all imbalanced and balanced data sets (the difference between balanced and imbalanced data set is less than 5%). Likewise, the logistic regression method indicated similar performance with respect to GM, BI, and MCC for both imbalanced and balanced data sets (Figure 10). The average value of GM, BI, and MCC obtained by logistic regression, however, was slightly

less than of the random forest for all the data sets, which indicates the superiority of random forest over logistic regression in predicting the heat-stressed cows. The Gaussian naïve Bayes method represented similar performance with respect to GM, BI, and MCC for balanced and imbalanced data sets associated with COMB, CON, and SH groups of cows (the difference between balanced and imbalanced data set is less than 5%). For the SP data set, the difference between the imbalanced and balanced data set based on the different metrics varies between 15% and 22% (Figure 11). Given that these sets of performance metrics are suggested to be used with an imbalanced data set, their value based on imbalanced and balanced data sets could be used as a criterion to explore if the imbalance could affect the performance of the machine learning methods. The results from this study indicated that the presence of an imbalance in all the data sets did not affect the results obtained by random forest or logistic regression. However, the effect of imbalance on the performance of the Gaussian naïve Bayes method regarding the SP data set was greater than for the other classifiers, indicating that its capability in classifying heat-stressed cows for COMB, CON, and SH data sets in the presence of imbalance is similar to the other classifiers.

Figure 9. Average value of geometric mean (GM), bookmaker informedness (BI), and Matthews correlation coefficient (MCC) metrics based on the random forest for balanced and imbalanced data sets of cows subjected to 3 heat abatement treatments. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth). Error bars represent the SD of the performance metrics over 20 trials for imbalanced data sets, and SD of the performance over 5 balanced data sets generated by synthetic minority oversampling technique.

Figure 10. Average value of geometric mean (GM), bookmaker informedness (BI), and Matthews correlation coefficient (MCC) metrics based on logistic regression for balanced and imbalanced data sets of cows subjected to 3 heat abatement treatments. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth). Error bars represent SD of the performance metrics over 20 trials for imbalanced data sets, and SD of the performance over 5 balanced data sets generated by synthetic minority oversampling technique.

DISCUSSION

Good health of dairy cows is essential to the well-being of the animal and is indicative of physiological functioning (Fraser et al., 1997). When a cow becomes heat stressed, DMI decreases, and therefore the availability of nutrients for milk synthesis is decreased (West, 2003a; Rhoads et al., 2009). Heat stress at any level can increase maintenance requirements by 7% to 25%, further exacerbating the decrease in milk production (NRC, 2001). The severity of stress experienced depends on various factors, including age, breed, production, body condition, stage of lactation, pregnancy, and housing (Yousef, 1985). When only monitoring environmental stressors during periods of heat stress, intervention for cows that are more sensitive to increased THI may be delayed.

A delay between environmental events and the effect on cow performance has been exhibited (West, 2003a). In a Florida study, black globe temperature (a measure of ambient temperature and radiant energy) had no significant effect on milk production when measured on the same day (Collier et al., 1981). The effects of a given black globe temperature on milk production are most noticeable between 24 and 48 h following the heat stress period (Collier et al., 1981; Spiers et al., 2004).

West et al. (2003b) reported that mean THI had the greatest correlation to decreases in milk production 2 d before milk yield measurement. For each unit increase of THI, Holstein cows decreased milk production by 0.88 kg for the 2 d lag of mean THI. When milk yield was measured on the same day as THI, the decrease in production was substantially less compared with 2 d earlier. The ability to accurately assess heat stress using individual cow indices and a scoring system could be a useful tool for producers to know when to intervene with appropriate heat abatement or veterinary assistance to decrease the production loss experienced in periods of heat stress.

To aid in the process of heat stress detection, many heat stress indices (e.g., body temperature, RR, drooling, panting) have been developed by researchers to evaluate the response of dairy cattle to environmental stressors (e.g., temperature, humidity, wind speed). The relationship between these indices or variables and the environmental stressors can be complex and nonlinear (Hastie et al., 2009). Models constructed with the assumption of linearity may not correctly represent how the environmental stressors are affecting the response of dairy cattle. In this study, a heat stress scoring system was developed and utilized to assess different levels of heat stress a dairy cow was experiencing and

Figure 11. Average value of geometric mean (GM), bookmaker informedness (BI), and Matthews correlation coefficient (MCC) metrics based on Gaussian naïve Bayes for balanced and imbalanced data sets of cows subjected to 3 heat abatement treatments. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth). Error bars represent SD of the performance metrics over 20 trials for imbalanced data sets, and SD of the performance over 5 balanced data sets generated by synthetic minority oversampling technique.

machine learning methods (Gaussian naïve Bayes and random forest) were compared with logistic regression to see which method could most accurately predict a heat-stressed or non-heat-stressed cows in a pen with sprinkler, shade, or no form of heat abatement (CON). The ability to predict if a cow is heat-stressed by using cow-based variables (e.g., RR and RT) could allow producers to provide early intervention before the negative effects of heat stress arise.

The experiment consisted of 3 different treatments or groups of dairy cows that were provided with different heat abatement strategies (CON, SH, SP). The models used to analyze the data set were able to classify the animals into heat-stressed or non-heat-stressed classes. Random forest classifier outperformed other classifiers for all the groups with respect to the mean value of AUC. The AUC value of Gaussian naïve Bayes for the COMB group did not overlap the AUC value of the other 2 methods. The AUC value of all 3 methods overlapped in CON, SH, and SP groups of animals. This result could be due to the subjectivity of some of the variables observed.

The mean value AUC for all the methods in the CON group was greater than the mean values in the SP group. In addition, the average value of AUC computed for all methods increased from the SP group to the SH group.

The average AUC value of classifiers was significantly different than 0.5, and other than the SP group, the average AUC value is above 0.9. Using AUC to evaluate the accuracy of the classifiers can have some drawbacks (Lobo et al., 2008); however, if a classifier scores an AUC value above 0.9 it has good accuracy, if it scores between 0.7 and 0.9 it has moderate accuracy, and if its AUC value is between 0.5 and 0.7 its accuracy is low (Akobeng, 2007). Therefore, other than the SP group where the performance of the classifiers was moderate, for all other groups, the proposed models were able to accurately predict the individual animal heat stress. This result could indicate that the heat-stressed cows in SP were more difficult to detect because they were less heat-stressed compared with the cows in SH and CON.

By using feature importance in the random forest classifier, the average importance of each feature was obtained for the different data sets over the 20 trials to better understand which variables were leading the observed results. The visualization of the importance of each of the features is plotted in Figure 12. Two features, RR and drooling, outperformed the rest of the features for all the data sets. However, it is observed that in the SP group, the average importance of these features is lower than their importance in the rest of

Figure 12. Visualization of the average importance of the features over 20 runs and based on the random forest classifier for the different treatment group data sets. COMB = combined; CON = control (no heat abatement); SP = sprinkler (polyvinyl chloride sprinkler system); SH = shade (shade structure with 80% protective shade cloth).

the groups. Further, the importance of the days in the milk feature in the SP group is greater than its value in the rest of the groups. Therefore, DIM is the third most important feature in SP, which outperforms the rest of the features. These 3 features, drooling, RR, and DIM, could be thought of as the features that drive heat stress in the different groups.

In this work, 3 machine learning methods, namely logistic regression, random forest, and Gaussian naïve Bayes, were used to predict heat-stressed and non-heat-stressed cows. The reason behind the selection of these methods is that they can be considered as a representative of a large set of classifiers. To further the quality of results and to achieve promising results for even the SP group of cows, investigating other advanced machine learning techniques such as neural networks could be a viable option. This method was attempted on the current data set, but did not get promising results. The reason is that this method needs a large amount of data to be trained and worked properly, and that amount of data was not available for this study. In one study by Fenlon et al. (2017), the authors used 4 predictive mod-

els of calving difficulty in dairy heifers and cows. The results from that study indicated that the neural network model outperformed the regression and random forest methods. Although neural networks are powerful predictive tools, they often need a large amount of data to be trained to provide quality predictions. Hence, using a neural network to improve the classification accuracy can be considered as future research when a large data set is available.

Apart from using powerful tools to predict heat-stressed animals, other ways to further improve the prediction performance could be considered. Increasing the amount of data collected could improve the prediction performance. The research conducted by Shahinfar et al. (2014) indicated that with a larger number of samples, pregnancy results could be improved. In the same study, involving more data in the experiments, the researchers reported an AUC score of 0.76, whereas the accuracy was between 72 to 74%. This result could mean that using the same classification in practice could cause a quarter of animals to be classified incorrectly. In addition to collecting more data, another pos-

sible way to improve the accuracy of even the SP group is including more variables. While taking more samples and measuring more variables could improve accuracy, it cannot solve all difficulties. A predictive model is able to recognize a heat-stressed animal only in the case where the reason for the heat stress is the pattern in the data set. This requirement is not certain, and there may be some factors that are not explained by the available information measured. When measuring more variables, limitations exist. For example, to find the appropriate amount of subjects for an experiment, a statistical power test is conducted for each variable to determine at what level an effect could be displayed. In animal science studies, if too many variables are measured then the sample size necessary may exceed the amount of animals that one has available for research purposes. Considering more information can increase accuracy, but because not every variable associated with heat stress can be evaluated, a certain degree of uncertainty will remain.

CONCLUSIONS

Three machine learning methods, logistic regression, Gaussian naïve Bayes, and random forest, were compared to predict if a cow was heat-stressed or non-heat-stressed. In terms of accuracy, random forest outperformed the other 2 methods for the SP group, and both logistic regression and random forest were consistent in prediction for CON, SH, and COMB. The precision of prediction for the methods was best for the random forest, especially for the SP group. The mean probability of predicting non-heat-stressed cows was high using all 3 methods and was highest for cows in the SP group. The logistic regression method worked best for predicting heat-stressed cows, with the mean probability of predicting heat-stressed animals being high for CON, SH, and COMB, but low for SP. The correlations of the consistency of prediction for the different methods ranged from 0.53 to 0.97, with the lowest correlation consistently being between the logistic regression and Gaussian naïve Bayes methods. Machine learning models can predict heat-stressed and non-heat-stressed cows provided with different heat abatement, but had an increased accuracy and precision when using the random forest method. If this study could be conducted over multiple summer seasons, more data would be available to analyze with other advanced machine learning methods (e.g., neural networks) to compare with the methods used in this study. For future studies, utilizing a combination of precision technologies to monitor specific individual cow heat stress indices and machine learning to create an algorithm to alert producers when cows begin to experience heat stress could be beneficial.

ACKNOWLEDGMENTS

The authors extend their sincere thanks to the staff and student workers at the Mississippi State University Bearden Dairy Research Center for their help in project preparation and implementation. The authors are grateful to Amelia Woolums (Mississippi State University College of Veterinary Medicine) for being the veterinarian for the project. We send thanks to Scott Hardin (nutritionist at Ware Milling, Houston, MS) who formulated the concentrate ration that the cows were supplemented with throughout the study. We also thank Mauricio Xavier and the undergraduate student interns Carolina Canestrano, Briana Johnson, Amberly Dennis, and Alexandria Ross (Mississippi State University, Starkville, MS), who helped conduct the project. Finally, thank you to Tami Brown-Brandl (University of Nebraska-Lincoln, Lincoln, NE) for her advice on heat stress scoring and Cassandra Tucker (University of California Davis, Davis, CA) for her insight on assessing panting parameters. The authors have not stated any conflicts of interest.

REFERENCES

- Akobeng, A. K. 2007. Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatr.* 96:644–647. <https://doi.org/10.1111/j.1651-2227.2006.00178.x>.
- Ali, A. K. A., and G. E. Shook. 1980. An optimum transformation for somatic cell concentration in milk. *J. Dairy Sci.* 63:487–490. [https://doi.org/10.3168/jds.S0022-0302\(80\)82959-6](https://doi.org/10.3168/jds.S0022-0302(80)82959-6).
- Allen, J. D., L. Hall, R. Collier, and J. Smith. 2015. Effect of core body temperature, time of day, and climate conditions on behavioral patterns of lactating dairy cows experiencing mild to moderate heat stress. *J. Dairy Sci.* 98:118–127. <https://doi.org/10.3168/jds.2013-7704>.
- Amrine, D. E., B. J. White, and R. L. Larson. 2014. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. *Comput. Electron. Agric.* 105:9–19. <https://doi.org/10.1016/j.compag.2014.04.009>.
- Anderson, S. D., B. J. Bradford, J. P. Harner, C. B. Tucker, C. Y. Choi, J. D. Allen, L. W. Hall, S. Rungruang, R. J. Collier, and J. F. Smith. 2013. Effects of adjustable and stationary fans with misters on core body temperature and lying behavior of lactating dairy cows in a semiarid climate. *J. Dairy Sci.* 96:4738–4750. <https://doi.org/10.3168/jds.2012-6401>.
- Beede, D. K., and R. J. Collier. 1986. Potential nutritional strategies for intensively managed cattle during thermal stress. *J. Dairy Sci.* 69:543–554. <https://doi.org/10.2527/jas1986.622543x>.
- Berman, A., Y. Folman, M. Kaim, M. Mamen, Z. Herz, D. Wolfenson, A. Arieli, and Y. Graber. 1985. Upper critical temperatures and forced ventilation effects for high-yielding dairy cows in a subtropical climate. *J. Dairy Sci.* 68:1488–1495. [https://doi.org/10.3168/jds.S0022-0302\(85\)80987-5](https://doi.org/10.3168/jds.S0022-0302(85)80987-5).
- Bianca, W. 1968. Thermoregulation. Pages 97–118, Chapter 7 in *Adaptation of Domestic Animals*. E. S. E. Hafez, ed. Lea and Febiger, Philadelphia, PA.
- Blackshaw, J. K., and A. W. Blackshaw. 1994. Heat stress in cattle and the effect of shade on production and behaviour: A review. *Aust. J. Exp. Agric.* 34:285–295. <https://doi.org/10.1071/EA9940285>.
- Bohmanova, J., I. Misztal, and J. B. Cole. 2007. Temperature-humidity indices as indicators of milk production losses due to heat stress. *J. Dairy Sci.* 90:1947–1956. <https://doi.org/10.3168/jds.2006-513>.

- Breiman, L. 2001. Random forests. *Mach. Learn.* 45:5–32. <https://doi.org/10.1023/A:1010933404324>.
- Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16:321–357. <https://doi.org/10.1613/jair.953>.
- Chen, J. M., K. E. Schütz, and C. B. Tucker. 2013. Dairy cows use and prefer feed bunks fitted with sprinklers. *J. Dairy Sci.* 96:5035–5045. <https://doi.org/10.3168/jds.2012-6282>.
- Churchill, G. A., and R. W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* 138:963–971.
- Collier, R. J., L. H. Baumgard, R. B. Zimbelman, and Y. Xiao. 2018. Heat stress: Physiology of acclimation and adaptation. *Anim. Front.* 9:12–19. <https://doi.org/10.1093/af/vfy031>.
- Collier, R. J., R. M. Eley, A. K. Sharma, R. J. Pereira, and D. E. Buffington. 1981. Shade management in subtropical environment for milk yield and composition in Holstein and Jersey cows. *J. Dairy Sci.* 64:844–849. [https://doi.org/10.3168/jds.S0022-0302\(81\)82656-2](https://doi.org/10.3168/jds.S0022-0302(81)82656-2).
- Collier, R. J., R. B. Zimbelman, R. P. Rhoads, M. L. Rhoads, and L. H. Baumgard. 2009. A re-evaluation of the impact of temperature humidity index (THI) and black globe humidity index (BGHI) on milk production in high producing dairy cows. Pages 158–168 in *Proc. 24th Annual Southwest Nutrition and Management Conference*.
- Cook, N. B., R. L. Mentink, T. B. Bennett, and K. Burgi. 2007. The effect of heat stress and lameness on time budgets of lactating dairy cows. *J. Dairy Sci.* 90:1674–1682. <https://doi.org/10.3168/jds.2006-634>.
- Cortez, P., M. Portelinha, S. Rodrigues, V. Cadavez, and A. Teixeira. 2006. Lamb meat quality assessment by support vector machines. *Neural Process. Lett.* 24:41–51. <https://doi.org/10.1007/s11063-006-9009-6>.
- De Rensis, F., and R. J. Scaramuzzi. 2003. Heat stress and seasonal effects on reproduction in the dairy cow—A review. *Theriogenology* 60:1139–1151. [https://doi.org/10.1016/S0093-691X\(03\)00126-2](https://doi.org/10.1016/S0093-691X(03)00126-2).
- Delamare, E., and J. Guinard-Flament. 2006. Increasing milking intervals decreases the mammary blood flow and mammary uptake of nutrients in dairy cows. *J. Dairy Sci.* 89:3439–3446. [https://doi.org/10.3168/jds.S0022-0302\(06\)72381-5](https://doi.org/10.3168/jds.S0022-0302(06)72381-5).
- Ebrahimi, E., F. Ebrahimi, M. Ebrahimi, S. Tomlinson, and K. R. Petrovski. 2018. Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. *Comput. Electron. Agric.* 147:6–11. <https://doi.org/10.1016/j.compag.2018.02.003>.
- Fenlon, C., L. O'Grady, J. F. Mee, S. T. Butler, M. L. Doherty, and J. Dunning. 2017. A comparison of 4 predictive models of calving assistance and difficulty in dairy heifers and cows. *J. Dairy Sci.* 100:9746–9758. <https://doi.org/10.3168/jds.2017-12931>.
- Ferguson, J. D., D. R. Galligan, and N. Thomsen. 1994. Principle descriptors of body condition score in Holstein cows. *J. Dairy Sci.* 77:2695–2703. [https://doi.org/10.3168/jds.S0022-0302\(94\)77212-X](https://doi.org/10.3168/jds.S0022-0302(94)77212-X).
- Finch, V. A. 1986. Body temperature in beef cattle: Its control and relevance to production in the tropics. *J. Anim. Sci.* 62:531–542.
- Fraser, D., D. M. Weary, E. A. Pajor, and B. N. Milligan. 1997. A scientific conception of animal welfare that reflects ethical concerns. *Anim. Welf.* 6:187–205.
- Gahegan, M. 2003. Is inductive machine learning just another wild goose (or might it lay the golden egg?). *Int. J. Geogr. Inf. Sci.* 17:69–92. <https://doi.org/10.1080/713811742>.
- Gaughan, J. B., S. M. Holt, G. L. Hahn, T. L. Mader, and R. Eigenberg. 2000. Respiration rate—Is it a good measure of heat stress in cattle. *Asian-Aus. J. Anim. Sci.* 13(Suppl.):329–332.
- Ghafouri-Kesbi, F., G. Rahimi-Mianji, A. Honarvar, and A. Nejati-Javaremi. 2017. Predictive ability of random forests, boosting, support vector machines and genomic best linear unbiased prediction in different scenarios of genomic evaluation. *Anim. Prod. Sci.* 57:229–236. <https://doi.org/10.1071/AN15538>.
- Gianola, D., K. A. Weigel, N. Krämer, A. Stella, and C. C. Schön. 2014. Enhancing genome-enabled prediction by bagging genomic BLUP. *PLoS One* 9:e91693. <https://doi.org/10.1371/journal.pone.0091693>.
- Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, Springer, New York, NY.
- Hempstalk, K., S. McParland, and D. Berry. 2015. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *J. Dairy Sci.* 98:5262–5273. <https://doi.org/10.3168/jds.2014-8984>.
- Higgins, S. F., C. T. Agouridis, and S. J. Wightman. 2011. *Shade options for grazing cattle*. Extension publication. Department of Biosystems and Agricultural Engineering, University of Kentucky, Lexington, KY. AEN-99.
- Hughes, A. L., M. J. Hersom, J. M. B. Vendramini, T. A. Thrift, and J. V. Yelich. 2010. Comparison of forage sampling method to determine nutritive value of Bahiagrass pastures. *Prof. Anim. Sci.* 26:504–510. [https://doi.org/10.15232/S1080-7446\(15\)30638-0](https://doi.org/10.15232/S1080-7446(15)30638-0).
- Jensen, F. V. 1996. *An Introduction to Bayesian Networks*. UCL Press, London, UK.
- Jensen, M. B. 2012. Behaviour around the time of calving in dairy cows. *Appl. Anim. Behav. Sci.* 139:195–202.
- Johnson, H. D. 1987. Bioclimate and livestock. Pages 3–16 in *Bioclimatology and the Adaptation of Livestock*. H. D. Johnson, ed. Elsevier Science Publisher, Amsterdam, the Netherlands.
- Kamphuis, C., H. Mollenhorst, A. J. Feeders, D. Pietersma, and H. Hogeveen. 2010. Decision-tree induction to detect clinical mastitis with automatic milking. *Comput. Electron. Agric.* 70:60–68. <https://doi.org/10.1016/j.compag.2009.08.012>.
- Kendall, P. E., P. P. Nielsen, J. R. Webster, G. A. Verkerk, R. P. Littlejohn, and L. R. Matthews. 2006. The effects of providing shade to lactating dairy cows in a temperate climate. *Livest. Sci.* 103:148–157. <https://doi.org/10.1016/j.livsci.2006.02.004>.
- Kendall, P. E., G. Verkerk, J. Webster, and C. Tucker. 2007. Sprinklers and shade cool cows and reduce insect-avoidance behavior in pasture-based dairy systems. *J. Dairy Sci.* 90:3671–3680. <https://doi.org/10.3168/jds.2006-766>.
- Kotsiantis, S. B., I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. *Informatica (Vilnius)* 31:249–268.
- Kubat, M., and S. Matwin. 1997. Addressing the Curse of Imbalanced Training Sets: One Sided Selection. Pages 179–186 in *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, Nashville, TN.
- Libbrecht, M. W., and W. S. Noble. 2015. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16:321–332. <https://doi.org/10.1038/nrg3920>.
- Ling, C., and C. Li. 1998. Data Mining for Direct Marketing Problems and Solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*. AAAI Press, New York, NY.
- Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* 17:145–151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>.
- Luque, A., A. Carrasco, A. Martín, and A. de las Heras. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* 91:216–231. <https://doi.org/10.1016/j.patcog.2019.02.023>.
- Mader, T. L., M. S. Davis, and T. Brown-Brandl. 2006. Environmental factors influencing heat stress in feedlot cattle. *J. Anim. Sci.* 84:712–719. <https://doi.org/10.2527/2006.843712x>.
- Martin, P., and P. Bateson. 2007. *Measuring Behaviour: An Introductory Guide*. 3rd ed. Cambridge University Press, Cambridge, UK.
- Mattachini, G., E. Riva, C. Bisaglia, J. C. A. M. Pompe, and G. Provolto. 2013. Methodology for quantifying the behavioral activity of dairy cows in freestall barns. *J. Anim. Sci.* 91:4899–4907. <https://doi.org/10.2527/jas.2012-5554>.
- Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405:442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9).
- McQueen, R. J., S. R. Garner, C. G. Nevill-Manning, and I. H. Witten. 1995. Applying machine learning to agricultural data. *Com-*

- put. *Electron. Agric.* 12:275–293. [https://doi.org/10.1016/0168-1699\(95\)98601-9](https://doi.org/10.1016/0168-1699(95)98601-9).
- Morton, J. M., W. P. Tranter, D. G. Mayer, and N. N. Jonsson. 2007. Effects of environmental heat on conception rates in lactating dairy cows: Critical periods of exposure. *J. Dairy Sci.* 90:2271–2278. <https://doi.org/10.3168/jds.2006-574>.
- National Research Council (NRC). 2001. *Nutrient Requirements of Dairy Cattle*. 7th rev. ed. National Academy Press, Washington, DC.
- Pazzani, M., C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. 1994. Reducing misclassification costs. Pages 217–225 in *Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettonhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Machine Learning Res.* 12:2825–2830.
- Polsky, L., and M. A. von Keyserlingk. 2017. Invited review: Effects of heat stress on dairy cattle welfare. *J. Dairy Sci.* 100:8645–8657. <https://doi.org/10.3168/jds.2017-12651>.
- Rhoads, M. L., R. P. Rhoads, M. J. VanBaale, R. J. Collier, S. R. Sanders, W. J. Weber, B. A. Crooker, and L. H. Baumgard. 2009. Effects of heat stress and plane of nutrition on lactating Holstein cows: I. Production, metabolism and aspects of circulating somatotropin. *J. Dairy Sci.* 92:1986–1997. <https://doi.org/10.3168/jds.2008-1641>.
- Rulquin, H., and J. P. Caudal. 1992. Effects of lying or standing on mammary blood flow and heart rate of dairy cows. *Annales de Zootechnie, INRA/EDP Sciences* 41:101–101.
- Schreiner, D. A., and P. L. Ruegg. 2002. Effects of tail docking on milk quality and cow cleanliness. *J. Dairy Sci.* 85:2503–2511. [https://doi.org/10.3168/jds.S0022-0302\(02\)74333-6](https://doi.org/10.3168/jds.S0022-0302(02)74333-6).
- Schütz, K. E., N. R. Cox, and L. R. Matthews. 2008. How important is shade to dairy cattle? Choice between shade or lying following different levels of lying deprivation. *Appl. Anim. Behav. Sci.* 114:307–318. <https://doi.org/10.1016/j.applanim.2008.04.001>.
- Shahinfar, S., D. Page, J. Guenther, V. Cabrera, P. Fricke, and K. Weigel. 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *J. Dairy Sci.* 97:731–742. <https://doi.org/10.3168/jds.2013-6693>.
- Shultz, T. A. 1984. Weather and shade effects on cow corral activities. *J. Dairy Sci.* 67:868–873. [https://doi.org/10.3168/jds.S0022-0302\(84\)81379-X](https://doi.org/10.3168/jds.S0022-0302(84)81379-X).
- Silva, R. G., D. A. E. F. Morais, and M. M. Guilhermino. 2007. Evaluation of thermal stress indexes for dairy cows in tropical regions. *Rev. Bras. Zootec.* 36(4 suppl):1192–1198. <https://doi.org/10.1590/S1516-35982007000500028>.
- Sokolova, M., N. Japkowicz, and S. Szipakowicz. 2006. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. Pages 1015–1021 in *Australasian Joint Conference on Artificial Intelligence*. Springer, Berlin, Germany.
- Spiers, D. E., J. N. Spain, J. D. Sampson, and R. P. Rhoads. 2004. Use of physiological parameters to predict milk yield and feed intake in heat-stressed dairy cows. *J. Therm. Biol.* 29:759–764. <https://doi.org/10.1016/j.jtherbio.2004.08.051>.
- Stull, C. L., L. L. McV. Messam, C. A. Collar, N. G. Peterson, A. R. Castillo, B. A. Reed, K. L. Andersen, and W. R. VerBoort. 2008. Precipitation and temperature effects on mortality and lactation parameters of dairy cattle in California. *J. Dairy Sci.* 91:4579–4591. <https://doi.org/10.3168/jds.2008-1215>.
- Tarazón-Herrera, M., J. Huber, J. Santos, H. Mena, L. Nusso, and C. Nussio. 1999. Effects of bovine somatotropin and evaporative cooling plus shade on lactation performance of cows during summer heat stress. *J. Dairy Sci.* 82:2352–2357. [https://doi.org/10.3168/jds.S0022-0302\(99\)75485-8](https://doi.org/10.3168/jds.S0022-0302(99)75485-8).
- Tresoldi, G., K. E. Schütz, and C. B. Tucker. 2016. Assessing heat load in drylot dairy cattle: Refining on-farm sampling methodology. *J. Dairy Sci.* 99:8970–8980.
- Tresoldi, G., K. E. Schütz, and C. B. Tucker. 2018. Cooling cows with sprinklers: Timing strategy affects physiological responses to heat load. *J. Dairy Sci.* 101:4412–4423. <https://doi.org/10.3168/jds.2018-14917>.
- US Climate Data. 2020. Monthly climate Starkville-Mississippi. <https://www.usclimatedata.com/climate/starkville/mississippi/united-states/usms0761>.
- US Meat Animal Research Center. 2017. Recognizing heat stress. US-DA-ARS. <https://www.ars.usda.gov/plains-area/clay-center-ne/marc/docs/heat-stress/recognizing-heat-stress/>.
- Valtorta, S. E., and M. R. Gallardo. 2004. Evaporative cooling for Holstein dairy cows under grazing conditions. *Int. J. Biometeorol.* 48:213–217. <https://doi.org/10.1007/s00484-003-0196-9>.
- van der Heide, E. M. M., R. F. Veerkamp, M. L. van Pelt, C. Kamphuis, I. Athanasiadis, and B. J. Ducro. 2019. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. *J. Dairy Sci.* 102:9409–9421. <https://doi.org/10.3168/jds.2019-16295>.
- Van Hertem, T., S. Viazzi, M. Steensels, E. Maltz, A. Antler, V. Alchanatis, A. A. Schlageter-Tello, K. Lokhorst, E. C. B. Romanini, C. Bahr, D. Berckmans, and I. Halachmi. 2014. Automatic lameness detection based on consecutive 3D-video recordings. *Biosyst. Eng.* 119:108–116. <https://doi.org/10.1016/j.biosystemseng.2014.01.009>.
- Vitali, A., M. Segnalini, L. Bertocchi, U. Bernabucci, A. Nardone, and N. Lacetera. 2009. Seasonal pattern of mortality and relationships between mortality and temperature-humidity index in dairy cows. *J. Dairy Sci.* 92:3781–3790. <https://doi.org/10.3168/jds.2009-2127>.
- Vizzotto, E. F., V. Fischer, A. Thaler Neto, A. S. Abreu, M. T. Stumpf, D. Werncke, F. A. Schmidt, and C. M. McManus. 2015. Access to shade changes behavioral and physiological attributes of dairy cows during the hot season in the subtropics. *Animal* 9:1559–1566. <https://doi.org/10.1017/S1751731115000877>.
- West, J. W. 2003a. Effects of heat-stress on production in dairy cattle. *J. Dairy Sci.* 86:2131–2144. [https://doi.org/10.3168/jds.S0022-0302\(03\)73803-X](https://doi.org/10.3168/jds.S0022-0302(03)73803-X).
- West, J. W., B. G. Mullinix, and J. K. Bernard. 2003b. Effects of hot, humid weather on milk temperature, dry matter intake, and milk yield of lactating dairy cows. *J. Dairy Sci.* 86:232–242. [https://doi.org/10.3168/jds.S0022-0302\(03\)73602-9](https://doi.org/10.3168/jds.S0022-0302(03)73602-9).
- Wheelock, J. B., R. Rhoads, M. VanBaale, S. Sanders, and L. Baumgard. 2010. Effects of heat stress on energetic metabolism in lactating Holstein cows. *J. Dairy Sci.* 93:644–655. <https://doi.org/10.3168/jds.2009-2295>.
- White, B. J., D. E. Amrine, and R. L. Larson. 2018. Big data analytics and precision animal agriculture symposium: Data to decisions. *J. Anim. Sci.* 96:1531–1539. <https://doi.org/10.1093/jas/skx065>.
- Woolums, A. R., R. D. Berghaus, D. R. Smith, R. F. Daly, G. L. Stokka, B. J. White, T. Avra, A. T. Daniel, and M. Jenerette. 2018. Case-control study to determine herd-level risk factors for bovine respiratory disease in nursing beef calves on cow-calf operations. *JAVMA* 252:989–994. <https://doi.org/10.2460/javma.252.8.989>.
- Yousef, M. K. 1985. *Stress Physiology in Livestock*. Vol. 1. CRC Press. Boca Raton, FL.
- Zheng, H. 2018. Analysis of global warming using machine learning. *Comp. Water Energy Env. Eng.* 7:127–141. <https://doi.org/10.4236/cweee.2018.73009>.

ORCIDS

- C. A. Becker <https://orcid.org/0000-0002-4713-4298>
 A. Aghalari <https://orcid.org/0000-0003-1122-5558>
 A. E. Stone <https://orcid.org/0000-0001-5687-3394>