Predicting dairy cattle heat stress indicators using machine learning and mid infrared spectral data

L.M. Dale[1], E.J.P. Strang[1], P. Lemal[2], M. Jattiot[3], A. Werner[1], HoliCow Consortium[4], ResKuh Consortium[5], H. Amamou[6], N. Gengler[2] and H. Soyeurt[2]

[1]Regional association for performance testing in livestock breeding of Baden-Wuerttemberg Heinrich-Baumann-Str. 1-3, 70190 Stuttgart, Germany
[2]Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés 2, 5030 Gembloux, Belgium
[3]INNOVAL, 35538, Noyal-sur-Vilaine, France
[4]https://holicow.nweurope.eu/, 4 Rue des Champs Elysées, 5590, Ciney, Belgium
[5]https://agroecologie-rhin.eu/de/klimaco-2/, 2 rue de Rome, 67013, Schiltigheim, France
[6]High School of Agriculture of Kef, University of Jendouba, 7119, Le Kef, Tunisia
Corresponding Author: ldale@lkvbw.de

Abstract

Heat stress negatively affects the well-being, productivity, and profitability of dairy cows and farms. Previous studies have primarily focused on indicators such as respiration rate, skin temperature, rectal temperature, and milk yield, with limited research on the impact of heat stress on milk composition. Initial findings from various locations like Belgium, Tunisia, and Germany have explored the mid-infrared (MIR) responses to heat stress in dairy cows. This study aimed to investigate the effects of heat stress on LKV Baden-Württemberg (LKVBW) farms within a 20 km radius of weather stations using all available datasets related to milk production traits and predicted milk biomarkers derived from MIR spectra. Additionally, the study sought to determine if there are differences in heat stress indicators between data collected from barn weather stations versus public weather stations. Meteorological data from public weather stations in Baden-Württemberg (BW) and barn weather stations from MobiMets and Pessl Instruments devices, including temperature and humidity, were combined with data collected by the milk recording organization. THI values were calculated using mathematical calculations for daily averages, and a three-day average was linked to farm data based on the day of milk recording collection. These datasets were then linked to each animal using monthly spectral data for each cow from 500 selected LKVBW farms. The model was developed as part of the HappyMoo project using MIR spectral data from Bentley Instruments devices collected in the LKVBW database from 2012 to 2019, with external validation conducted on a dataset containing MIR spectral data from 2020 to 2022. Barn weather data was collected in Projekt KlimaCO, with MobiMets data from 2020 to 2022 and Pessl Instruments data from 2021 to 2023. A machine learning algorithm was implemented in R using the “glmnet” package. The spectral data were standardized using the EMR method and preprocessed with the first derivative algorithm using the Savitzky-Golay filter. Differences were observed in MIR spectra recorded under THI and thermoneutral conditions, with certain wavenumbers of the MIR spectrum showing varying responses. The THI index was established based on the relationship between the THI value of individual cows and the mean THI value of the farm. Pearson correlations were calculated using the THI index and milk parameters in the R environment with the “corrplot” library. The THI index showed negative correlations with milk yield (0.15), lactose (0.12), acetate (0.33), blood NEFA (0.2) and positive correlations with fat content (0.59), protein content (0.40), blood BHB (0.25), blood glucose (0.30), blood calcium (0.21) and fatty acids (0.35). No differences were found between public weather stations and barn weather stations. Additional analysis is required within the scope of upcoming projects like HoliCow and ResKuh to identify possible MIR heat stress phenotypes derived from milk. These phenotypes could be utilized for herd management and breeding purposes to pinpoint animals that are resilient to heat stress.

Keywords: MIR, spectral data, dairy cows, heat stress.
Presented at the ICAR Anual Conference 2024 in Bled at the Session 2: New tools to extend the horizon of milk mid-infrared spectrometry

Introduction

The decline in small and medium-sized dairy farms has led to the HoliCow1 project's emphasis on assisting these farmers. The primary objective is to transform intricate Big Data into accessible, cost-effective, and remote decision-making tools for farmers. This project aims to combat farm abandonment, improve animal welfare, and tackle climate-related challenges. By offering a novel solution, HoliCow strives to equip farmers with the tools needed to make informed choices, fostering sustainability in agriculture, enhancing livestock welfare, and safeguarding the environment.

To achieve this goal, a cross-border database has been set up to encompass various dairy breeds on farms across North-West Europe. The objective is to gather sufficient data to establish a reliable alert system. It is crucial to acknowledge the significant role that farms play in fostering robust and eco-friendly communities. Farmers are key advocates for animal welfare and climate resilience. HoliCow aims to consolidate all incoming data sources and streamline them effectively to provide a comprehensive overview of the welfare status of cows. In addition, HoliCow will incorporate animal-related data, milk spectral predictions refined through enhanced equations from previous projects like Robust Milk, OptiMIR, GplusE, HappyMoo, or D4Dairy, as well as external data such as climate information sourced from public weather stations. Moreover, the project will leverage interactive machine learning models to promote transparent farming practices. Therefore, the efficient processing of vast amounts of Big Data, including climate and weather data, is essential to assist farmers in making informed decisions and implementing sustainable practices that enhance their operations and benefit the broader community. This integration will empower farmers to remain engaged with their local communities and embrace modern agricultural techniques using cost-effective tools.

HoliCow will enable farmers to remotely monitor individual cows and their surroundings. Recognizing the importance of each animal compared to larger farms, receiving "early warnings" about an individual's status is economically vital. HoliCow will drive the collaborative adoption of integrated and innovative methods to facilitate this transformation for farmers, as well as the integration of these methods into regional agricultural innovation strategies and advisory services. HoliCow will help establish NWE as a pioneering force in innovative solutions for rural areas with a diverse cultural landscape. The approach will also address climate change mitigation and resilience by incorporating climate data, nitrogen efficiency, and methane predictions. Heat stress is a crucial factor affecting dairy cow performance and productivity, leading to decreased milk yields and metabolic disorders. As global warming trends continue, the impact of heat stress is receiving more attention even in temperate regions like central Europe. Traditionally quantified by the temperature-humidity index (THI), heat stress is also known as the discomfort index. Different countries have varying THI thresholds for negative effects on milk production; for example, the US uses a threshold of 72, while Luxembourg and Germany use thresholds of 62-60 that are often exceeded during the summer months. Studies have shown that for each THI unit increase, there can be a loss of 0.08 to 0.26 kg of milk per cow. In the HappyMoo project, a THI of 68 resulted in a 21% decrease in milk production and a 9.6% reduction in dry matter intake on average.

Material and methods

Experimental data

Through the HappyMoo Project, Amammou *et al*. (2021) have created a milk MIR spectra based THI model to forecast the cow's THI as a 3-day average before the testing day, aiming to determine the cow's resilience to heat stress. The primary goal of the HappyMoo research was to analyse the changes in milk MIR spectra linked to THI within a specific subset of the southwest German dairy herd under milk monitoring and explore the potential for identifying clear indicators of heat stress in individual milk samples.

The data utilized for the modelling was sourced from 120 farms across Baden-Württemberg, representing a subset of data from approximately 4500 dairy farms in the region. THI values were determined using data from 67 local German weather stations (DWD) for each farm. In Figure 1, the map on the left displays coloured dots representing all 4,500 farms participating in milk recording, while the map on the right highlights selected farms and weather stations with red dots. The geographical coordinates and measurements from these sensor points are openly available on the DWD server. Milk analysis and milk MIR spectral data were accessible for the farms between 2012 and 2019, encompassing the primary breeds and production systems in the area. To link to temperature and humidity data, the closest active sensor point at the time of milk recording was selected.

Statistical approach

The statistical analyses and machine learning utilized spectral data from Bentley FTIR analysers that were standardized using the EMR/CRA-W procedure. Absorbance values from the spectra were used to calculate the first derivative, and 212 relevant wave-numbers were selected. Additional input parameters included breed, parity, milking time, days in milk (DMI) categories, and age at calving. As a reference point, a three-day average of THI values prior to the day of milk recording was determined. Various linear regression methods such as PCR, PLS, CPPLS (pls package in R), and GLMNET (glmnet package in R) were employed. Three different calibration and validation subsets were created for validation purposes, based on spectra, animals, and cross-validation selection.

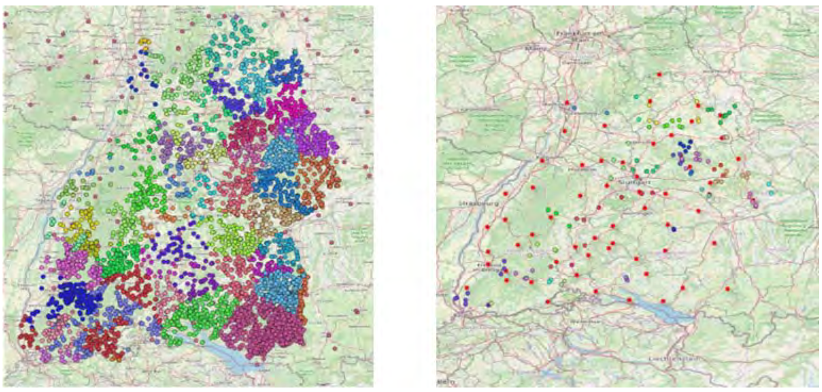

Figure 1. Map showing the positions of farms in relation to weather stations.

Spectra selection

Soyeurt (2024) has introduced an innovative approach to creating a globally representative spectral database (WRSD) to tackle the challenges associated with spectral extrapolation in predicting milk yield traits for new samples. This method is efficient and space-efficient, employing a two-stage selection process. Initially, a decomposition matrix is generated through principal component analysis using a dataset of approximately 2,137,394 records. Subsequently, an iterative spectral selection is conducted based on a sample location index derived from the principal components (PC). The frequency of spectra occurrence for each location index is calculated to influence the subsequent barycentre calculations. Despite 10 PCs explaining 95% of spectral variability, the barycentre pattern of selected spectra accurately represents the entire dataset, showcasing the effectiveness of a location index based on only 3 PCs. Finally, a WRSD was created using the HoliCow data samples with the same algorithm, selecting around 103,477 spectral data points from a total of 2,137,394 million spectral data collected from 503 farms.

Clustering approach

Hierarchical clustering was employed to identify patterns among futures rather than independent samples based on the predicted spectral dataset. The Ward's agglomerative method with Euclidean distance (Ward, 1963) was selected for its ability to effectively differentiate groups in a multivariate Euclidean space. However, due to the extensive amount of data, only records from April to October and from the years 2018-2023 were included, around 55,759 records used for calculating distances for each observation and cluster analysis. To address the clustering issue, a divide-and-conquer strategy was implemented by dividing the data into 100 subsets, clustering them separately, and then merging the centroids obtained from each subset (Wang *et al*., 2016). This process was repeated five times to ensure the reliability of the final clustering outcomes. The inertia of the dendrogram was utilized to determine the optimal number of groups by identifying a point where there is a significant decrease in inertia gain. This iterative approach enabled efficient analysis of the data despite its magnitude, showcasing a practical solution for managing challenges associated with big data.

Cluster interpretation

Traditional statistical methods such as ANOVA were not appropriate for this analysis due to the high volume of records, which could result in all effects being considered significant. Instead, the emphasis was on interpreting the clusters to understand their implications. To compare the clusters, their distinctions were highlighted by calculating the least squares means (LSM) per group and displaying them in bar graphs on a standardized scale (Franceschini *et al*., 2022).

Results and discussion

The calibration methods were chosen based on the information provided in Table 1, which includes the statistical parameters for the Calibration data of the Weather Station THI3mean equations. A total of 140,618 records were utilized for each model, with a THI value ranging from 25 to 75, a mean of 51, and a standard deviation of 8.03.

The root squared mean ranged from 0.76 to 0.89, with RPDs varying between 2.05 and 3.08. Ultimately, the GLMNET model emerged as the most robust option, primarily due to its ability to eliminate irrelevant or noisy input parameters, reducing the number of variables to the essential minimum required. The model developed by LKV-BW was the first model to incorporate spectral data analysis and weather data from public stations;

Table 1. Statistical parameters for the calibration data of the THI3mean equation

Model	Number of samples	Min	Mean	Max	SD	SEC	R2	RPD
PCR	140,618	25	51	75	8.03	2.51	0.76	2.05
PLS	140,618	25	51	75	8.03	2.47	0.77	2.09
MVR	140,618	25	51	75	8.03	2.47	0.77	2.09
CPPLS	140,618	25	51	75	8.03	2.44	0.77	2.11
GLMNET	140,618	25	51	75	8.03	2.58	0.89	3.08

Table 2. Identification results of the final model – THI3mean based on the spectral model (1st Calibration), animal model (2nd Calibration) and Cross Validation model.

Model	Number of samples	Min	Mean	Max	SD	SEC	R2	RPD
1st Calibration	98.434	25	51	75	7.92	2.58	0.89	3.08
1st Validation	42.184	25	50	75	7.93	2.58	0.89	3.08
2nd Calibration	98.435	25	51	75	7.93	2.58	0.89	3.07
2nd Validation	42.183	25	50	75	7.93	2.57	0.90	3.09
Cross Validation	140.618	25	51	75	8.03	2.58	0.89	3.08

although it does not predict heat stress, it forecasts the temperature and humidity index. The prediction results are promising, with an R^2 value close to 0.89 based on THI3mean and the predicted CowTHI3mean value. *Table 2* displays the calibration and validation datasets for the GLMNET model. To randomly select different datasets, the Mahalanobis distance was calculated for the first calibration and validation model among all spectral data points. This distance is determined by measuring the test point's distance from the centre of mass divided by the width of the ellipsoid in the direction of the test point (Mahalanobis, 1936). Based on this distance calculation, 70% of spectral data points were chosen for the calibration model and 30% for the validation model. For the second calibration and validation dataset, random selection was performed among all animals. Based on the total number of animals in the dataset, 70% were included in the calibration model and 30% in the validation model. The cross-validation model is based on the K-fold cross-validation process, which can be implemented using the cv.glmnet function.

In addition to the standard glmnet parameters, cv.glmnet introduces its own unique parameters such as nfolds (indicating the number of folds), fold id (allowing for user-supplied folds), and type.measure (specifying the loss metric used for cross-validation): "deviance" or "mse" for squared loss, and "mae" for mean absolute error (Friedman *et al.*, 2010).

Additionally, predictions of MIR spectral data for fatty acids and minerals (RobustMilk and OptiMIR), ketone bodies (OptiMIR and OptiKuh), as well as health and immunity markers (HappyMoo, GplusE, and D4Dairy) have been calculated using various MIR models from the afore mentioned projects. These predictions, along with milk components, were used to calculate Pearson correlations based on CowTHI3mean and WSmean3THI (see Figure 2).

Research conducted by Dale *et al.*, 2023 and Lemal. P., *et al.*, 2023 identified a positive correlation between the CowTHI3mean indicator and WSmean3THI in Germany and Belgium. This suggests that utilizing this indicator may be more effective than traditional traits such as milk yield, protein, or fatty acids.

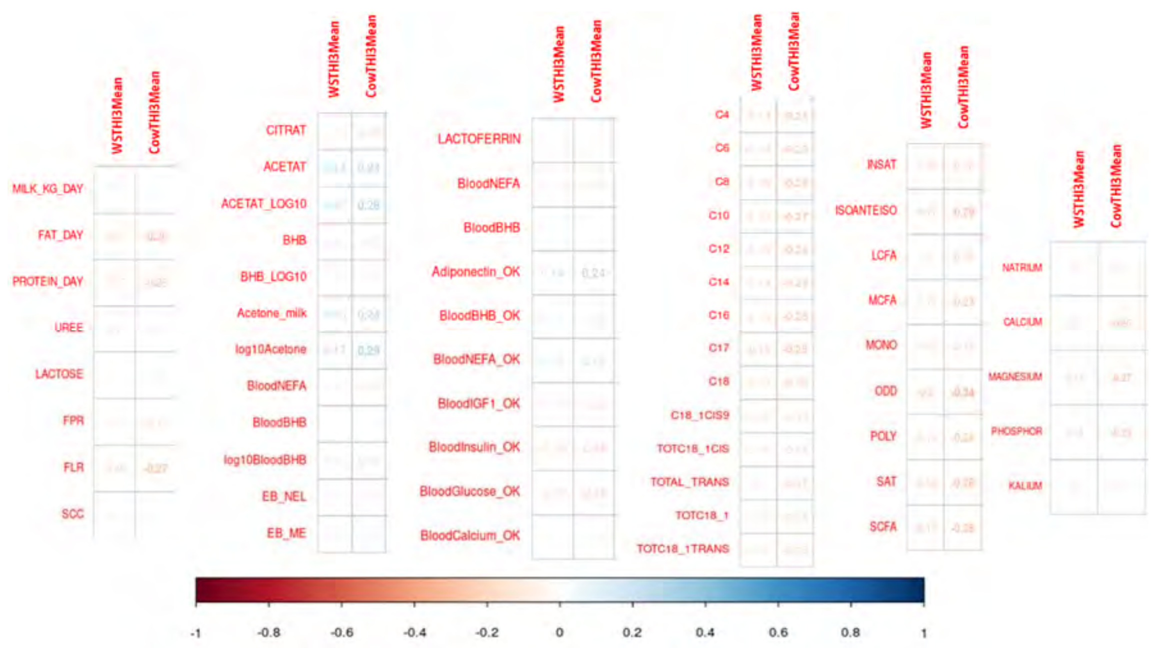

Figure 2. Pearson correlations between milk parameters and WSTHI3Mean and CowTHI3mean values.

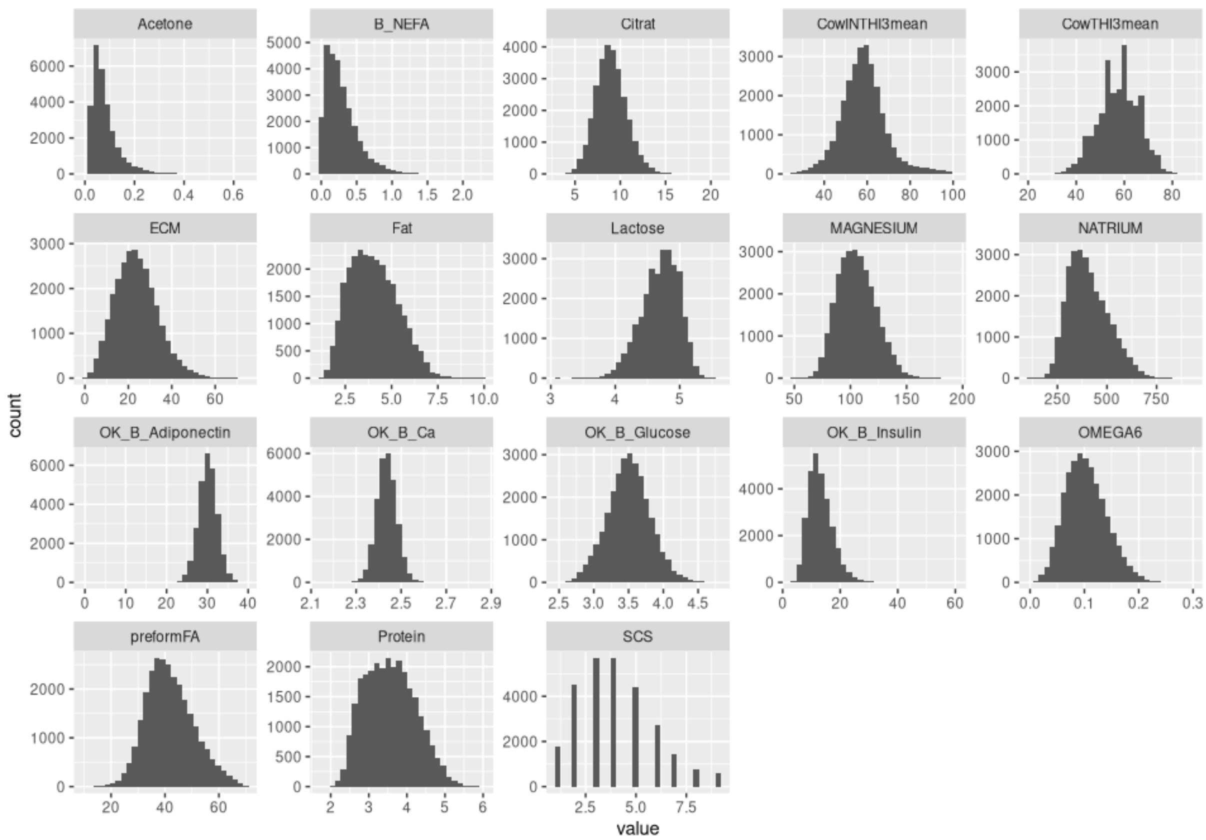

Figure 3. Data used in unsupervised clustering.

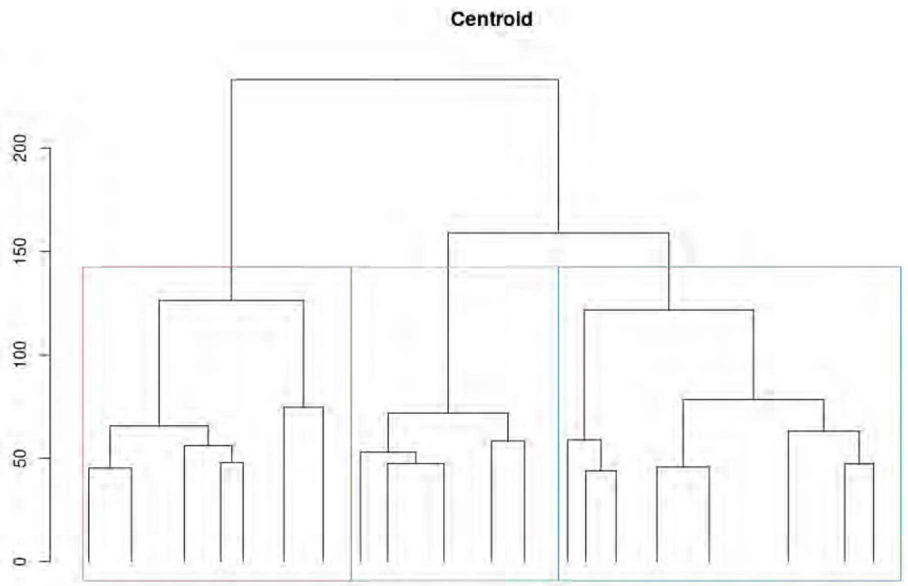

Figure 4. Cluster Analysis: Hierarchical clustering – LKVBW data set

For the unsupervised cluster analysis, the following mainly milk MIR based prediction variables were utilized: acetone in milk, blood NEFA (B_NEFE), citrate, CowTHI3mean, model for inside barns THI prediction based on Pessl and MobiMets devices (CowINTHI3mean), energy-corrected milk (ECM), fat, lactose, magnesium, sodium (natrium), blood adiponectin (O_B_Adiponenctin), blood calcium (OK_B_Ca), blood glucose (OK_B_Glucose), blood insulin (OK_B_Insulin),as well as omega-6 fatty acid, preformed fatty acid, protein, and somatic cell score (SCS). The distribution of the dataset used can be observed in Figure 3.

As accurately identifying heat stress based solely on CowTHI3mean values is challenging, cluster analysis was conducted in the HoliCow predictions dataset using these predictions as variables specifically between April and October, as well as from 2018 to 2023. The results revealed three distinct groups (refer to Figure 4). To further comprehend the cluster analysis, Classification and Regression Trees (CART) were utilized. It was observed that the trees' classification depended on whether the dependent variable was a numeric value.

In the CART analysis, the variables utilized for classification into the three groups were acetone in milk, blood NEFA, CowTHI3mean, lactose, blood glucose, omega-6 fatty acid, preformed fatty acid, and protein. These specific variables were chosen based on their significance in determining heat stress levels within the dataset. The inclusion of these factors allowed for a more comprehensive and low accurate classification of the data into distinct groups based on their respective values. As shown in Figure 5, class 1 was only predicted well at 54%, with 30% of the entire dataset classified as class 1. Class 2 was classified between 53% and 77% with a balanced accuracy of 63%, while class 3 was classified between 48% and 76% with a balanced accuracy of 71%. The cluster with the highest accuracy from the CART model was class 3, which also exhibited higher mean standard prediction (msp) for all components considered in the modelling process (Figure 6). Protein, magnesium, blood insulin, blood calcium, fat, and blood adiponectin had a positive msp higher than 0.75, while acetone, CowTHI3mean, ECM, preformed FA, and lactose had a negative msp lower than -0.35.

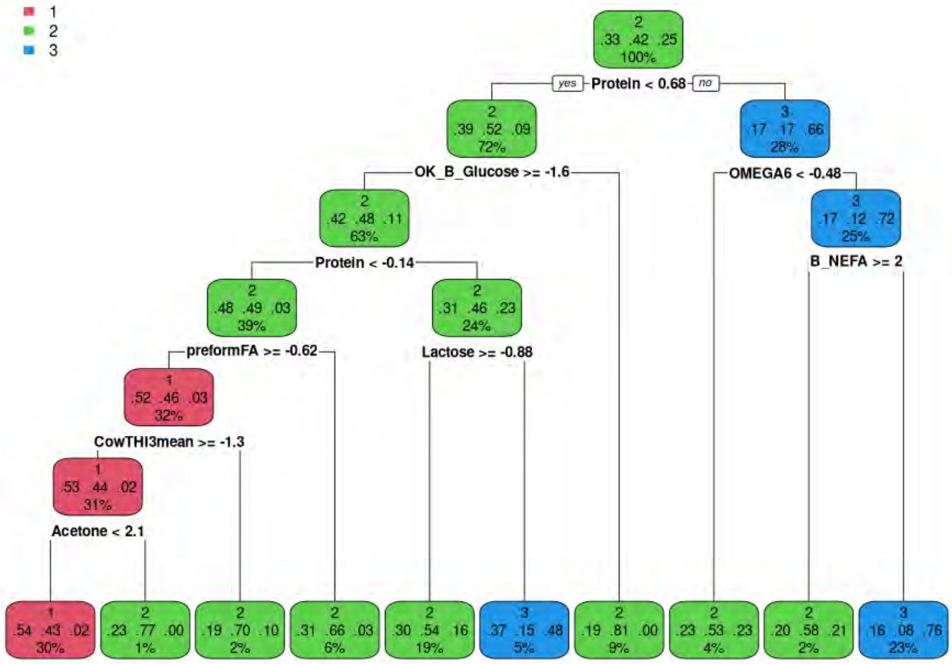

Figure 5. Classification and regression trees based on the cluster analysis.

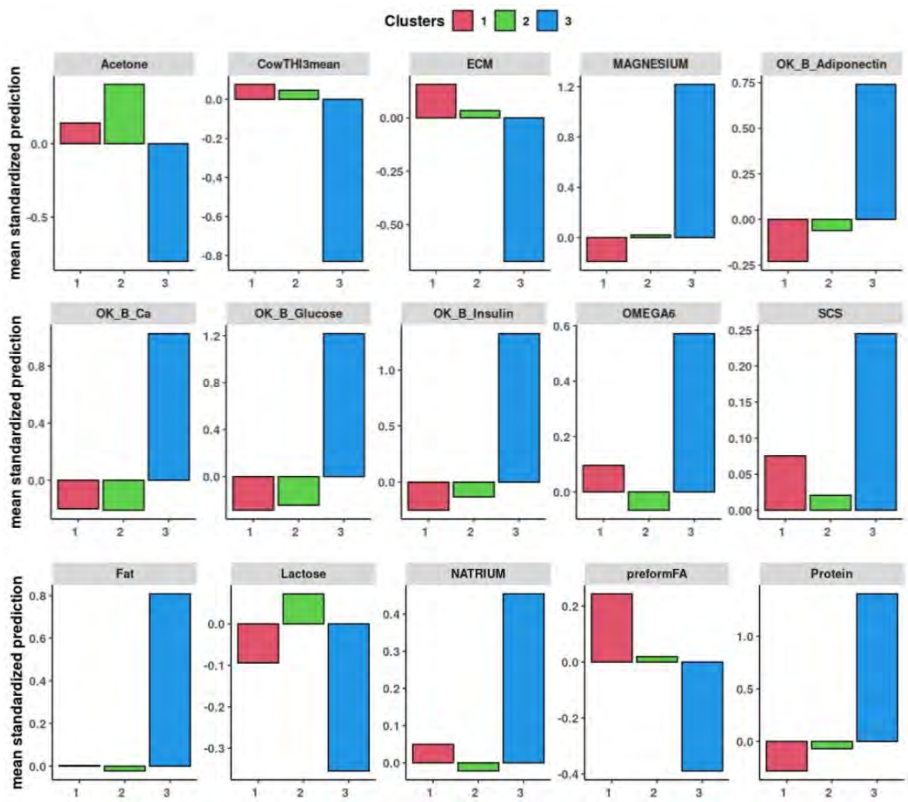

Figure 6. Correlation between cluster response and milk recording data and MIR spectra predictions used for clustering.

Conclusions

In the initial phase of the HoliCow project, significant progress and promising results have been noted. The project has shown great potential in achieving its objectives and goals, indicating a positive trajectory for future developments. The early findings suggest that the project is on track to deliver valuable insights and advancements in the field, setting a strong foundation for further research and innovation. The next steps involve validating the predictions against heat stress or health problems situations by utilizing the same animals in the time series analysis and closely monitoring the development of the clusters. The challenge of selecting the most suitable clustering algorithm arose due to the array of options available, including K-means, hierarchical clustering, and DBSCAN. The decision-making process considered our data and research objectives, with valuable input from the HoliCow research team and insights from previous studies (Franceschini *et al*., 2022). Cluster analysis is an iterative procedure, and plans are underway to interpret the results effectively to gain a deeper understanding of each cluster's characteristics in the upcoming phase of the project. It is crucial to refine the outcomes and improve the quality of clustering for practical implementation in pilot farming scenarios.

List of references

Amammou, H. Mahouachi, M., Werner, A., Gengler, N., Bieger, J., Drössler, K., Hammami, H., HappyMoo Consortium, Dale, L.M. 2021. In Book of Abstracts of the 72th Annual Meeting of the European Federation of Animal Science, Davos, Switzerland, 582.

Dale, L.M., Jattiot, M., Werner, A., Strang, E.J.P., Lemal, P., HappyMoo Consortium, KlimaCO Consortium, Gengler, N. 2023. MIR spectral prediction based on heat stress in dairy cattle. In Book of Abstracts of the 74th Annual Meeting of the Euro-pean Federation of Animal Science, Lyon, France, 1031.

D4Dairy, Digitalisation, Data integration, Detection and Decision support in Dairying, Project 322 number: 872039, COMET-Project, Oktober 2018 to September 2022.

Franceschini, S., Grelet, C., Leblois, J., Gengler, N., and Soyeurt, H. 2022. Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health?. Journal of Dairy Science, 105(8), 6760-6772.

Friedman J., Tibshirani R., Hastie T. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33(1) 1-22.

GplusE, Genotype and Environment contributing to the sustainability of dairy cow pro-duction systems through the optimal integration of genomic selection and novel management protocols based on the development. FP7 project funded by the Eu-ropean Union. January 2014 to December 2018.

HappyMoo, Delivering NWE dairy farmers, vets and extension workers with innovative and holistic monitoring tool to screen animals with regards to animal welfare and more especially freedom from disease, hunger and stress, through new smart data opportunities and across a network of European farmers' organizations. INTER-REG North West Europe. October 2018 to December 2022.

Lemal, P., Dale, L. M., Jattiot, M., Leblois, J., Schroyen, M., and Gengler, N.. Use of MIR spectra-based indicator for genetic evaluation of heat stress in dairy cattle [Paper presentation]. EAAP 2023, Lyon, France, 31 August 2023

Mahalanobis, P.C. 1936. On the Generalized Distance in Statistics. Proceedings of the National Institute of Science of India, 2, 49-55.

OPTIMIR - New tools for a more sustainable dairy sector. INTERREG North West Europe. September 2009 to December 2015

ROBUSTMILK: Innovative and Practical Breeding Tools for Improved Dairy Products from More Robust Dairy Cattle. European Commission, Directorate-General for Agriculture and Rural Development, under Grant Agreement 211708. April 2008 to September 2012.

Soyeurt, H. 2024. Rapid selection of milk mid-infrared spectra for creating a world representative spectral database of dairy cow population. WRSD - ICAR webinar.