Assessing traditional and machine learning methods to smooth and impute device-based body condition score throughout the lactation in dairy cows

J. Chelotti[1,2], H. Atashi[1,3], M. Ferrero[2], C. Grelet[4], H. Soyeurt[1], L. Giovanini[2], H. L. Rufiner[2,5], N. Gengler[1,*]

[1] TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium

[2] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, 3000 Santa Fe, Argentina

[3] Department of Animal Science, Shiraz University, 71441-13131 Shiraz, Iran.

[4] Walloon Agricultural Research Center (CRA-W), 24 Chaussée de Namur, 5030 Gembloux, Belgium

[5] Facultad de Ingeniería, Universidad Nacional de Entre Ríos, 3100 Entre Ríos, Argentina

[*]Corresponding Author:

Nicolas Gengler

Passage des Déportés, 2

B-5030 Gembloux

Belgium

+32/81/62 22 06 (office)

+32/81/62 21 15 (fax)

nicolas.gengler@uliege.be

ABSTRACT

Regular monitoring of body condition score (**BCS**) changes during lactation is an essential management tool in dairy cattle; however, the current BCS measurements are often discontinuous and unevenly spaced in time. The imputation of BCS values is useful for two main reasons: i) achieving completeness of data is necessary to be able to relate BCS to other traits (e.g. milk yield and milk composition) that have been routinely recorded at different times and with a different frequency, and ii) having expected BCS values provides the possibility to trigger early warnings for animals with certain unexpected conditions. The contribution of this study was to propose and evaluate potential methods useful to smooth and impute device-based BCS values recorded during lactation in dairy cattle. In total, 26,207 BCS records were collected from 3,038 cows (9,199 and 14,462 BCS records on 1,546 Holstein and 1,211 Montbéliarde cows respectively, and the rest corresponded to other minority cattle breeds). Six methods were evaluated to predict BCS values: the traditional methods of test interval method (**TIM**), and multiple-trait procedure (**MTP**), and the machine learning (**ML**) methods of multi-layer perceptron (**MLP**), Elman network (**Elman**), long-short term memories (**LSTM)** and bi-directional LSTM (**BiLSTM**). The performance of each method was evaluated by a hold-out validation approach using statistics of the root mean squared error (**RMSE**) and Pearson correlation (r). TIM, MTP, MLP, and BiLSTM were assessed for the imputation of intermediate missing values, while MTP, Elman, and LSTM were evaluated for the forecasting of future BCS values. Regarding the machine learning methods, BiLSTM demonstrated the best performance for the intermediate value imputation task (RMSE=0.295, r=0.845), while LSTM demonstrated the best performance for the future value forecasting task (RMSE=0.356, r=0.751). Among the methods evaluated, MTP showed the best performance for imputation of intermediate missing values in terms of RMSE (0.288) and r (0.856). MTP also achieved the best performance for forecasting of future BCS values

in terms of RMSE (0.348) and r (0.760). This study demonstrates the ability of MTP and machine learning methods to impute missing BCS data and provides a cost-effective solution for the application area.

Key words: body condition score, data imputation, machine learning, dairy cows

 3

1. INTRODUCTION

Smoothing and imputing records throughout the lactation is an issue that is often required in dairy cattle to allow optimal use of non-continuously recorded traits. The fat reserves and changes in fat reserves over time are indicators of the cow's energy balance (**EB**) (Edmonson et al., 1989, Beam and Butler, 1999, Collard et al., 2000, Bernabucci et al., 2005). Therefore, it is important to know the energy reserve status (in the form of body fat) and its changes during lactation (Schröder and Staufenbiel, 2006, Roche et al., 2009). Although a negative energy balance (**NEB**) is common in the early lactation of dairy cows, abrupt changes are associated with health and welfare problems in the mid- and late-lactation (Beam and Butler, 1999, Collard et al., 2000, Bernabucci et al., 2005). Recording of body condition through body condition score (**BCS**) is a useful management tool to assess body fat stores of dairy cows (Pryce et al., 2001, Roche et al., 2009) compared to expected status. Regardless of the scale used for the BCS, low BCS values reflect emaciation and high BCS values indicate obesity (Edmonson et al., 1989, Bastin et al., 2007).

The usual procedure to measure BCS value in dairy cows is based on the visualization and touching of the animal by expert technicians visiting the farm and following a scoring protocol (Edmonson et al., 1989, Ferguson et al., 1994). There are various non-continuous scales to assign BCS in dairy cows (Roche et al., 2004, Roche et al., 2009). Two commonly used scales are a five-point scale, with 0.50 or 0.25-point intervals (Wildman et al., 1982) and a nine-point scale system with unit increments, used in the Walloon Region of Belgium (Bastin et al., 2007, Bastin and Gengler, 2013), which is based on and promoted by the ICAR guidelines for the linear type traits (ICAR, 2022). Traditional BCS measurements have been considered subjective and have shown considerable intra- and inter-technician variability (Kristensen et al., 2006). Therefore, new automatic and potentially more objective methods have been proposed to measure BCS. Methods and devices using 3D cameras for body measurements have gained great popularity due to improvements in image quality and processing in recent years (Kuzuhara et al., 2015, Spoliansky et al., 2016, Du et al., 2022, Luo et al., 2023, Zhang et al., 2023). Several studies have used machine learning (**ML**) techniques to assess the BCS from 3D images, achieving high performance rates (Alvarez et al., 2019; Song et al., 2019)Furthermore, there are few commercial devices available to measure BCS. These devices can help experts perform their appraisal, such as the BodyMat system (Ingenera SA, Cureglia, Switzerland) or be installed on the farm to do a continuous automatic recording, such as the DeLaval system (DeLaval International, Tumba, Sweden). The first type of device facilitates recording, but generates records that stay relatively sparse, and still needs a large human investment for BCS scoring. The second type of device provides nearly continuous measurements, but some measurements may fail (i.e., cows may not present themselves correctly to the device).

Device-based scoring data behaves like most real-world data generating datasets containing missing values. A basic strategy to use incomplete datasets is to discard entire rows or samples containing missing values (Rubin, 1976, Meng and Shi, 2012). However, this comes at the price of losing data which, although incomplete, may be valuable (Lobato et al., 2015, Van Buuren, 2018). A better strategy is to impute the missing values, i.e., to infer them from the known part of the data (Graham, 2009, Lobato et al., 2015), using appropriate methods, e.g., based on multiple trait models. Another issue that affects human scores, but also partially device-based scores, is that they are inherently uncertain and potentially erroneous. An important reason for increased random errors was identified in the variation in the presentation of the animal to the device (Coffey et al., 2002). For this reason, strategies of smoothing this type of data may be useful (Coffey et al., 2002).

Smoothed and continuously available BCS measurements would be of major priority for dairy herd management, but also for studies requiring BCS data aligned with other longitudinal traits recorded during the lactation by dairy herd improvement (**DHI**) organizations which are running programs to collect and analyze data related to milk production, cow health, and reproductive performances. Different procedures may be used to smooth and impute BCS records throughout the lactation. In this study, we only used for this purpose endogenous information based on observed BCS data on an individual and lactation level. As a primary objective, successful data imputation would allow missing information to be completed and thus improve conditions for the development of new models to add exogenous information that can also be obtained in routine by DHI. In this context, imputed BCS data can be used, directly or indirectly, for the development of models that predict BCS also from milk yield, milk composition and especially milk mid-infrared (**MIR**)-based fine milk composition (Gengler et al., 2016). A few studies have addressed the regression of BCS values from the MIR spectra using techniques such as partial least squares, random forests and gradient boosting machines (McParland et al., 2011; Mota et al., 2021). However, accurate alignment of smoothed and imputed BCS data and MIR spectra are needed for any MIR prediction equation calibration process which underlines the interest of this research. An important second objective is the forecasting of future BCS values as knowing these expected values can help trigger alerts at critical moments during the whole lactation. The contribution of this study is therefore the evaluation and proposal of traditional and ML methods to smooth and impute device-based BCS throughout the lactation in dairy cows allowing its use through the comparison of observed and expected BCS values.

2. MATERIALS AND METHODS

2.1 Data Sources

Two databases (**DB**) were provided by French DHI organizations. The first DB was created in the Alsace region (**DBA**) and provided by the DHI organization Chambre Conseil Contrôle Elevage (**3CE**) active in this region. The other DB was created in the Bourgogne-Franche-Comté region (**DBB**) and provided by the regional DHI organization Conseil Elevage 25-90. For both databases, automatic BCS measurements were recorded by trained technicians using the same BodyMat system (Ingenera SA, Cureglia, Switzerland) and following the same experimental protocols. The BodyMat is an automated body condition scoring system using a 3D sensor to estimate BCS (Mullins et al., 2019; Leary et al., 2020). The system is based on a stick with a tactile control box in the base and a sensor with an infrared camera, infrared generator and a laser in the extreme. At the time of measurement, the laser pointer must be positioned at the level of the 2nd or 3rd transverse apophysis of the spine of the cow. The device senses and processes a 3D model of the back of the cow, reporting a BCS value in the range of 0 to 5. Details on the collected datasets recorded using this device are given in Table 1. Figure 1 shows the distribution of the data, with BCS data showing a near Gaussian distribution within databases.

2.2 Data Preparation and Distribution

To use homogeneous data on a breed x database level, only data recorded on Holstein cows for DBA, and on Montbéliarde and Holstein cows for DBB were used. Records from given days in milk (**DIM**) greater than 365 d were eliminated. In order to check for and to detect atypical BCS curves, the variance of the residuals between the observed curve for a given cow-lactation and expected curves for each specific population were computed and used as an indicator of the deviation from the expected curves. The threshold of one BCS unit SD in variation of the average residuals was considered to distinguish typical from atypical BCS curves. This was done in order to assess to what extent the available BCS curves showed atypical behavior but not to filter them out as in a real-life situation, except for obvious outliers, no BCS records would be a priori deleted.

2.3 Data Imputation Methods

There are different strategies to impute missing data from known data (Sainani, 2015; Van Buuren, 2018). In this study, six strategies were evaluated to impute missing BCS values.

The first two methods were based on the traditional strategies used in DHI to deal with non-continuous milk yield and component test-day records. These two methods, which are still currently used, were the test interval method (**TIM**) and the multiple-trait procedure (**MTP**). TIM, as a simple linear interpolation, was used as one of the simplest approaches in the area for interpolation purposes, while MTP was included as an enhancement incorporating population information. Additionally, four **ML** based methods were evaluated starting from simple approaches using Multilayer Perceptron (**MLP**) and continuing with recurrent neural networks that incorporate information from the temporal evolution of the data, which is useful in the case of BCS. MLPs can capture complex relationships between input and output features and they can learn a mapping from features derived from the existing data to the target BCS values. They are suitable networks with well-defined features but do not consider time-sequential patterns. The dynamic networks evaluated ranged from basic structures using Elman networks (Elman) to more complex structures using long-short term memories (LSTM) and bi-directional LSTM (BiLSTM). Elman networks handle sequence data better than MLP. However it may struggle with longer-term dependencies. LSTM is included as an advanced recurrent neural network and excellent for capturing long-term dependencies and temporal patterns in sequential data. Finally, we tried Bi-LSTM, which is suitable for capturing both past and future context, providing a more global view for imputation tasks.

Implicitly all strategies, except for TIM which needs by definition adjacent observations (i.e., 45 days maximum), had a more or less direct smoothing effect finding a compromise across observed records to estimate missing ones. Moreover, a common validation strategy was developed to test all these methods in this precise context. As a part of the training stage, selected hyperparameters such as the number of hidden layers, the number of neurons and the learning rate were optimized for MLP and BiLSTM methods prior to their validations.

To the best of our knowledge, there is no other work or study that evaluates, compares and proposes traditional and machine learning methods for BCS imputation using only existing time sequences of BCS. The methods evaluated are described in detail below.

2.3.1 Traditional methods

The approach called test interval method and abbreviated TIM in this document is still used in many countries and DHI systems and has been for many years (Everett and Carter, 1968, Sargent et al., 1968). ICAR (2020) considers TIM as one of the reference methods to calculate accumulated lactation yield, especially in the context of regular approximately 30-d interval testing schemes through the lactation. With special adjustments for the first and last test day records, TIM can be considered as an unbiased measure of actual 305-d milk yield (Schaeffer and Jamrozik, 1996). This method also can estimate missing data points in the process to compute lactation records, a feature that was used in this study. This consists of simple linear interpolation, where boundary points are necessary to predict a point in-between. The TIM approach needs limited distance between records. Therefore, in this work, a separation between two existing points of maximum 45 days was required. Data out of this range were excluded from this research.

The approach called multiple-trait procedure by ICAR (2020), and hereafter abbreviated as MTP, was originally proposed for predicting jointly lactation yields for milk, fat, and protein (Schaeffer and Jamrozik, 1996). This procedure uses a Bayesian estimation for lactation curve parameters of each cow and lactation based on their conditional distribution. The MTP method has the advantage over the use of full random regression models (Mayeres et al., 2004) that it can be used lactation by lactation and that the modeling of the whole population is not necessary. Missing values at a given DIM are then obtained using these lactation curve parameters. Therefore, values between samples can be predicted with long intervals apart or even if there is just one sample during the complete lactation (Schaeffer and Jamrozik, 1996).

Moreover, this method is based on standard lactation curve models (Wilmink, 1987), and covariances between parameters. Here, MTP was adapted to work with BCS values throughout the lactation. MTP can be seen as a combination of the observed BCS values at a given DIM during lactation (**y**) for a given cow in a given lactation, the characteristics of the population to which an animal belongs ($\mathbf{c}_0$) and other parameters (**p**) i.e., related to the covariances among elements of $\mathbf{c}_0$ and among residuals (Figure 2). A priori knowledge of the height and the shape of the BCS curves over the course of the lactation will be used when defining $\mathbf{c}_0$. Thus, the estimated lactation curve parameters $\hat{\mathbf{c}}$ of a given cow and lactation can be expressed as:

$$\hat{c} = f\left(y\left(DIM\right), c_0, p\right) \quad (1)$$

More specifically, this equation as formulated by Henderson (1984) was solved to predict $\hat{\mathbf{c}}$:

$$\left(X' R^{-1} X + G^{-1}\right)\hat{c} = X' R^{-1} y + G^{-1} c_0 \quad (2)$$

where **X** is the incidence matrix linking BCS records for a given cow in a given lactation, **R** represents the residual covariance matrix among BCS records for a given cow in a lactation, **G** is the covariance matrix among $\hat{\mathbf{c}}$ parameters, **y** is, as already explained, the BCS value at a given DIM, and $\mathbf{c}_0$ represents the parameters computed from all cows with similar characteristics such as breed and region. Figure 2, using a real case, illustrates how MTP works using the slightly modified Wilmink function (Wilmink, 1987) as explained above. As illustrated in Figure 2, MTP has a second feature that smooths directly observed records towards population values. The relative importance of population values decreases with the increasing number of direct BCS records which would decrease the importance of $\mathbf{G}^{-1}\mathbf{c}_0$ relative to $\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}$.

We computed the main parameters with complete data according to the strategy outlined in the original study (Schaeffer and Jamrozik, 1996). First, based on the exploratory computations we decided to use a slightly modified Wilmink function (Wilmink, 1987) to predict a given element of **y** here defined as a scalar as:

$$y=\alpha+\delta x+\beta e^{-(\gamma DIM)} \tag{3}$$

where, x = 2(DIM-1)/(365-1) – 1, which varies in the range [-1, 1], and α, β, and δ are the adjustable parameters elements of the vector **c**. The parameter γ which was also estimated in this process, was however kept fixed throughout the rest of the study as the Bayesian linear model used in (2) was not able to update its value for each lactation. These different parameters are related with the evolution of the lactation curve (Macciotta et al., 2005). Thus, α can be seen as an intermediate value, giving an offset to the complete evolution; and β and γ are factors explaining the drop in the early lactation stage; and δ is the general slope after the nadir stage, strongly related with the recovery of the BCS in late lactation. We used the NLIN procedure in SAS (SAS Institute Inc., Cary, NC, USA) to estimate $\mathbf{c}_o$ for each population based on the average BCS per DIM defined as **y** in (3) using the Gauss-Newton method by default. A minimum number of BCS records by DIM was necessary to meet the convergence criteria. Therefore, the stratification of the population could not be very detailed. The parameter γ was obtained a priori and considered fixed throughout the rest of the study. In the next step curve parameters were estimated for each cow by solving a simplified version of equation (2) for $\hat{\mathbf{c}}$:

$$\left(X'R^{-1}X\right)\hat{c}=X'R^{-1}y \tag{4}$$

where (4) produced the ordinary least-square estimator and not the Bayesian linear regression estimator obtained by solving (2). For this purpose, only a group of cows with good records describing their BCS lactation curves was used (i.e., with a minimum of three test day records through the lactation, at least one record before 50 DIM and at least one record after 250 DIM). We estimated **R**, which was considered a diagonal residual matrix expressing the variances of the differences between the expected and the observed values. Expected BCS were predicted by fitting the BCS curve through the lactation using $\hat{\mathbf{c}}$. The residual variance was kept constant because no significant variations were observed throughout the lactation. Simple variances and covariances of elements across cows were computed to obtain the covariance matrix **G** among the model parameters.

2.3.2 Machine Learning Methods. As a type of longitudinal data, this study consists of repeated BCS observations at different DIM in the lactation period for each cow. Thus, given a BCS observation that could be considered as the present, it is straightforward to refer to the past (previous) and the future (following) observations in that specific lactation period. In this study, the performance of ML techniques including MLP, Elman, LSTM and BiLSTM to impute BCS values were evaluated (Figure 3). We addressed two imputation tasks: (I) imputation of intermediate BCS values (i.e. an unknown BCS value that lies between two known BCS values in time) and, also (II) forecasting of BCS values. As input features for the first task, we used DIM (past, present and future) and BCS values (past and future) in order to estimate the BCS at a given DIM in the lactation. For the forecasting task, we only used DIM (past and future) and past BCS values as input to forecast BCS values in the future.

The MLP approach was assessed as one of the simplest ML techniques used for classification and regression problems (Bishop and Nasrabadi, 2006). MLPs consist of several layers of neurons. Each neuron in one layer is connected with all nodes from the previous layer (Figure 3-a). There are three types of layers including the input, hidden and output. Whereas neurons in the input layer represent the features provided to the network, each neuron in the hidden and output layers is a processing element which combines the output of incoming connected neurons using a nonlinear activation function. The strength of these connections is controlled using weights, which are optimized during the training process (Bishop and Nasrabadi, 2006).

 12

Elman, LSTM and bidirectional LSTM networks (BiLSTM) are types of recurrent neural networks (**RNN**) (Rumelhart et al., 1985). A key factor in a RNN is that connections between neurons can create a cycle, making it possible that the outputs of some neurons can affect the subsequent inputs of the same neurons. This recurrence gives RNN certain memory capabilities and makes them more efficient where the data follow temporal sequences as in the case of longitudinal data. RNNs have the ability to learn the evolution of a trait when they are trained with individual evolutions for that trait, even corresponding to several subjects. Moreover, Elman networks are one of the simplest RNN structures. They include hidden neurons and incorporate context (or memory) neurons, which are connected to allow past inputs to influence future computations during the training stage. In these networks the dynamics of the data is learned from the context layer (Figure 3-b) (Elman, 1990).

In practice, classical RNNs such as Elman networks have some limitations in learning complex sequences. To overcome this restriction, LSTM networks use 3 gates in each neuron in order to control how much information should be used from inputs to update the internal state (input gate), how much information should be forgotten from the previous state (forget gate), and how much information should be used directly from inputs to generate the output (output gate) (Figure 3-c). Like classical RNNs, LSTMs are made up of multiple neurons (Hochreiter and Schmidhuber, 1997). Although Elman and LSTM are suitable for forecasting tasks, in some scenarios the goal is to predict an intermediate point of the sequence. In these cases, an alternative method called bidirectional LSTM (BiLSTM) allows combining past and future information to generate a prediction in-between (Graves and Schmidhuber, 2005). This network introduces two identical LSTM, one trained with time sequences forwards and the other with the same sequences backwards (Figure 3-d).

In this work, the hyper-parameters of each method were optimized using a grid search strategy. These hyper-parameters varied with the method but, in general, the common search was considering the number of layers and the number of neurons per layer. We used a standard validation split for each epoch (80/20). The convergence criterion was an early stop based on the RMSE, thus avoiding overfitting during the training phase. An optimized MLP model with 3 hidden layers, with 16, 8 and 16 neurons from shallow to deep layers, and a rectified linear unit (**ReLU**) as the activation function was used. The use of ReLU has shown to improve the network performance significantly because it avoids gradient vanishing problems (Bishop and Nasrabadi, 2006). A linear function was used in the output layer to generate the final prediction. Features were normalized to be included into the model. In the case of Elman, the optimal number of neurons in the hidden layer was 32. In the case of LSTM, the number of hidden layers and the number of neurons per layer were optimized, resulting in 3 hidden layers of 16, 16 and 8 neurons from shallow to deep layers, and using the default parameters as defined in Keras v2.10.0 (Chollet, 2015) and in particular the default activation function (hyperbolic tangent). Finally, a BiLSTM network with a single recurrent layer of 5 neurons and hyperbolic tangent as the activation function was used. The outputs of the BiLSTM were fed and combined into a fully connected dense layer of 10 neurons and a hyperbolic tangent activation function. The output layer was composed of a single neuron with a linear activation function.

2.4 Validation Strategy

To evaluate the performance of each method, the combined dataset (Holstein data of DBA + Holstein data of DBB + Montbéliarde data of DBB) was split into calibration and validation sets, often called training and test sets in the field of machine learning, respectively. The same calibration and validation datasets were kept for the different methods. As we tested in this context essentially the capacity to fill in gaps, the validation data was a subset of the original data based on test-days within a given cow. Then, we compared predicted values against the real observed values in the validation set. According to the objectives of this work, the methods were divided and evaluated for two tasks: (I) imputation of intermediate BCS values and (II) forecasting the future BCS values.Using the configuration proposed for each method, only MTP is suitable for both types of tasks (Schaeffer and Jamrozik, 1996). TIM is straightforward, easy to implement and computationally efficient method for imputation of values in-between. However, it does not capture complex patterns or dependencies beyond a simple linear trend, making it unsuitable for predicting future values where such complexity is often present. MLPs can capture non-linear relationships between inputs and missing values. When combined with other features or lagged values, MLPs can effectively impute missing values by learning patterns in the data. However, they may not model sequential dependencies as well as recurrent networks, which are more suited for time-series forecasting. Bi-directional LSTMs are capable of utilizing context from both past and future states, making them effective for imputation in temporal sequences where knowing future context (within the sequence) can help better estimate missing values. While powerful, bi-directional LSTMs are typically not used for forecasting because they consider data in both directions, which is not available in a forecasting context. LSTMs are specifically designed to handle long-term dependencies in sequential data. They are highly effective and primarily designed to predict future values in a time series based on learned patterns. Elman networks are suitable for forecasting because they can model sequential dependencies over time. They are not robust for imputation tasks where bidirectional context or more advanced memory handling is required.

Thus, two different settings were proposed in terms of the selection of records for the calibration and validation sets (Figure 4). For both tasks, we kept only one point per each cow-lactation curve for the validation set, which implied 8-10% of the total points. Points were reserved for the validation set only when there were at least three points for that cow-lactation. For the data imputation task, the selection of points for the validation set was random (orange points in Figure 4-a) in each execution, while the rest of the points were included in the calibration set (green points in Figure 4-a). Due to the random process involved, we decided to train and validate each method during 10 executions to finally obtain stable average values. Thus, in each execution each method was calibrated and validated with the same set of points, allowing a direct comparison among the methods. For this task, extremes in time (i.e. first or last record) were never selected because could represent a drawback for some of the techniques. For example, TIM cannot perform linear interpolation without extreme values. Following these rules, we kept around 20,000 records for the calibration set and around 2,000 records for the validation set. The number of records in each set varied slightly across each random execution. Finally, we reported the macro-average across executions of the root mean squared error (**RMSE**) and the Pearson correlation (r) for each method using the observed BodyMat values present in the validation set as the reference. RMSE is defined as:

$$RMSE=\sqrt{\sum_{i=1}^{n}\frac{\left(\hat{y}_i-y_i\right)^2}{n}}, \tag{5}$$

Where $\hat{y}_i$ are predicted values, y_i are observed values and n is the number of observations. Pearson correlation is defined as:

$$r=\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2}}, \tag{6}$$

Where x_i are samples of the x variable, $\bar{x}$ is the mean of the x variable, y_i are samples of the y variable, $\bar{y}$ is the mean of the y variable.

On the other hand, to forecast future BCS values we only kept the last values in the lactation to build the validation set, while the rest of the points were kept for the calibration set (Figure 4-b). This setup allowed methods to be trained on past values (green points in Figure 4-b) to predict future values (orange points in Figure 4-b).

16

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

After data preparation, 83.2% of the total original BodyMat records were kept showing a mean of 2.50 and a **SD** of 0.59 BCS units. This included 4,286 records on 755 Holstein cows for DBA and 17,518 records on 1,951 animals for DBB (4,053 records on 753 Holstein and 13,465 records on 1,198 Montbéliarde cows). Means for BCS found for both databases were very similar with values around 2.50 BCS units. However, the SD of BCS found for DBB was considerably lower than that found for DBA (0.56 vs 0.70 BCS units for DBB and DBA, respectively). A potential explanation for this difference is the high number of Montbéliarde cows for the DBB, which is a breed with different characteristics from Holstein. Figure 5 shows the average BCS by DIM and corresponding modelled mean curves using the modified Wilmink function. We found similar evolutions of lactation curves between both databases (Figure 5a). It can be seen that the DBA was noisier, which could be due to a lower number of points by DIM contributing to averages for this database.

As explained above, we only kept the majority breeds for each database which results in three groups: I) DBA-Holstein, II) DBB-Holstein, and III) DBB-Montbéliarde (Figure 5b). The inclining slopes after the nadir (the lowest value of BCS throughout the lactation) were similar for Holsteins from DBB and DBA (0.0022 and 0.0025 BCS units / DIM, respectively), but different from that found for DBB-Montbéliarde (0.0013 BCS unit / DIM). Each population showed a particular global distribution regarding BCS (Figure 6). A lower variance (i.e., lower density at the ends of the distribution) was observed for the Montbéliarde population compared to that found for Holstein populations in both datasets. The SD of BCS records was 0.50 for DBB-Montbéliarde, 0.68 for DBB-Holstein and 0.70 for DBA-Holstein.

 17

In this sense, Montbéliarde seems to be a more stable breed throughout the lactation. These findings support the general accepted hypothesis that, Montbéliarde cows keep their body condition better than Holstein cows, indicating a higher resilience in terms of body condition through the lactation for this breed (Walsh et al., 2008, Berghof et al., 2019, Poppe et al., 2020, Poppe et al., 2021). On the other hand, the behavior of DBB-Holstein and DBA-Holstein populations was similar (i.e., similar shapes), with a minimal difference between median values (2.3 and 2.5 BCS units respectively).

Training the methods with the combined data allowed us to build a more general model and this is an advantage when, for example, there are crossbreeds or a large variety of parities in the population. Based on the raw data summaries of both datasets (DBA, and DBB), we concluded that they are mostly compatible. Also, BCS data were acquired with the BodyMat system and following the same experimental protocols. In the following, the datasets were combined to a single dataset with which methods were calibrated and evaluated. Due to the similar behavior found for each breed, we decided to analyze the data by breed, without a division by region. Figure 5c shows the behavior of each breed through the lactation and considering two parity classes: primiparous and multiparous. Statistical description of the used datasets considering parity classes and breed is shown in Table 2. It was observed that primiparous animals presented a higher mean of BCS throughout lactation (2.68 and 2.66 BCS unit for Montbéliarde and Holstein breeds, respectively) compared with multiparous animals (2.51 and 2.32 BCS unit for Montbéliarde and Holstein breeds, respectively). In addition, the nadir values of BCS were higher and expressed earlier for primiparous cows compared to multiparous cows. However, it was observed that the recovery BCS rates (delta in equation 4) found for multiparous cows were almost double those found for primiparous cows in both breeds (Table 2).

The analysis by breed and parity classes showed that the primiparous cows tend to be more resilient than multiparous cows (Poppe et al., 2021). It could be due, at least in part, to the fact that primiparous cows mobilize less body energy than multiparous cows during their lactations and they produce less milk (Friggens et al., 2007, Wathes et al., 2007). On the other hand, we observed that multiparous cows generally express the nadir stage later than primiparous cows (Truman et al., 2022). Primiparous cows presented higher BCS at the nadir time than multiparous cows, which is consistent with previous works. (Mao et al., 2004, Sakaguchi, 2009). For both breeds, the recovery BCS rate during mid- and late-lactation for multiparous cows was higher than the corresponding to primiparous cows.

In this work, no formal analyses were performed to look for statistically significant differences due to breed, dataset and parity. Comparisons between breeds and parities were not the main aim of this study, rather just comparing data collected across these categories for analyzing the suitability of models for these categories.

3.2 Identification of Atypical Curves

The variance of the residual between observed and expected curves for each specific population was computed and used as an indicator of the deviation from the expected curves. Higher variance of the residual indicated that beyond a translation (i.e., constant shift) of the curve, which will not show up in the variance, its shape was not as expected. During the data analysis, we found typical curves but also a considerable number of atypical curves (Figure 7). We sorted the curves according to the variance of the residuals and the curves with the lowest and highest variances were plotted. In the left side of Figure 7 we can see a typical evolution, even considering that the observed cow is thinner than expected for her population indicating a translation. In contrast, in the right side of Figure 7 the observed points follow a very messy curve with a behavior far from that expected for that population, even considering potential health issues. We found 11% of observations that were over one BCS unit SD in variation of the residual. Moreover, our data did not allow us to determine if this could be related to device problems or measurement problems or reflected real variability of underlying BCS status. For this reason, and in order to keep the study close to a real-life situation, we did not filter based on this aspect but used all the previously selected (i.e., pre-filtered) data.

Finally, if an atypical BCS value is obtained in practice, the first thing that should be done is to identify that value and then analyze it. This value could be compared with the expected for that cow (e.g. using an imputation method). An atypical value could be due to measurement error or a pathological condition of the animal, which is an objective of the use of BCS. In the first case, it could be directly discarded. On the other hand, if this BCS value is due to an atypical condition of the animal, it should be saved for detection of relevant animals in bad condition. These BCS values will also be useful for future adjustments of the methods or models used for BCS prediction.

3.3 Performance Evaluation

The performance of methods including TIM, and MTP, and the ML methods of MLP, Elman, LSTM and BiLSTM to predict BCS values were evaluated. The proposed methods were divided into those suitable for the imputation of intermediate values such as TIM, MTP, MLP and BiLSTM and those suitable for forecasting tasks such as MTP, Elman and LSTM. Each method was calibrated using the calibration data and then evaluated using the validation data. Performance measures were computed between the reference values and the values predicted by each method. The average RMSE and the average correlation for each method suitable for the imputation task are presented in Table 3.

Figure 8 shows the distribution of the RMSE and r for each imputation method. Among the evaluated methods, MTP achieved the best performance (Table 3; Figure 8). The results showed that MTP achieved the lowest RMSE (median of 0.288) followed by LSTM (median of 0.295) and MLP (median of 0.297). Regarding the Pearson correlation, MTP achieved the highest value (median of 0.849) followed by BiLSTM (median of 0.845) and MLP (median of 0.843). These results proved to be significantly different from each other ($p < 0.05$) under the Wilcoxon test (Woolson, 2007), except between MLP and biLSTM. The Wilcoxon test is a non-parametric test that compares paired samples or two related groups, offering the advantage of not requiring normal distribution, making it suitable for small or non-normally distributed data. The poorest results for this task were observed for TIM (medians of 0.302 and 0.837 for RMSE and r, respectively).

In addition to its advantage to be able to extrapolate values, a task that TIM cannot do; MTP can impute missing values even when the distance between existing points is large (Schaeffer and Jamrozik, 1996). In some cases, MTP allows a smoothing effect on the messy curves, resulting from atypical measurements, by incorporating information from the population. This could imply an advantage to process data from noisy automatic systems, but it could be a disadvantage when there are real abrupt changes in the body condition. Regarding the ML methods, MLP and BiLSTM showed comparable results to MTP and provided better performance than those provided by TIM. MLP can be considered as a non-linear interpolation for data imputation (Bishop and Nasrabadi, 2006). In this sense, this superiority over a linear method like TIM is not surprising. MLP is a simple ML method that was not designed to directly handle longitudinal data. However, MLP can be used for that, and its use is common and accepted (Anglart et al., 2020). On the other hand, a recurrent approach like BiLSTM allows past and the future sequences of measurements to be received as inputs, which makes BiLSTM ideal for longitudinal data, and useful as a tool to impute missing values in between known values. A practical advantage of this method is that it can receive input sequences of variable length as past or future measurements, which would be common for data collected in the field. (Graves and Schmidhuber, 2005). Table 4 shows the results of the forecasting methods. For this task, again MTP achieved the lowest RMSE (0.348) followed by LSTM (0.356) and Elman network (0.373). This difference was also observed when r was evaluated, where MTP achieved the highest correlation (0.760), followed by LSTM (0.751) and Elman networks (0.728). Due to the validation methods used for this task, already mentioned in the validation strategy section, no random executions were obtained over all the methods. Due to the deterministic nature of MTP, a single execution is reported for this method, while for the ML methods the average of 5 executions is reported. The reason is the random process involved in the initialization of the weights for a neural network. Although MTP showed that best performance for imputation and forecasting BCS values, its performance for forecasting was generally lower than the corresponding to the imputation of intermediate values. This is logical due to the greater difficulty of predicting future values only from past data, which becomes even more challenging when the temporal distance between measurements increases. Although Elman and LSTM showed lower performance than MTP, these results are of great interest considering that unlike the other methods, which use past and future information to predict intermediate missing values, Elman and LSTM only use past information to predict future information. This is important because one application of interest is to predict the future information using the historical data for purposes of evaluation and as a tool to provide early warning indicators of the body condition of an animal. RNNs like Elman or LSTM learn the temporal relationships in the evolution of the BCS through lactation. Unlike MTP, these networks do not assume a previous evolution, but instead they learn from the data sequence during the training stage. Finally, this information persists in the weights of the network.

A limitation of the validation strategy used in this work for the forecasting task is that by keeping only the last points of the sequence for the validation set, these were found mostly

 22

in mid- and late- lactation. Although it could be interesting to evaluate the forecasting of points in early lactation, in the present study this was not possible because to train only with previous points, many later points would have had to be discarded to keep the natural sequence of recordings. The latter was not possible due to the limited amount of data to train some of the methods.

3.4 Comparing Methods and Perspectives

The studied deep learning methods like LSTM or BiLSTM did not outperform MTP, which may be due to the limited amount of data available for the training phase. Also, MTP is directly using information available across (sub-)populations inside a Bayesian framework. However, this key feature of MTP may also generate an issue as it is potentially oversmoothing the observed BCS records towards the expected BCS curve which might not reflect the correct expectations. This was already reported as a major issue in yield traits and this fact explains the changed lactation curve model used in the practical application as reported by ICAR (2020). The improvement of the parametrization of MTP which controls the weight of prior curves and observed BCS, or the use of finer expected curves for different subpopulations could be available strategies. As shown in this study, the definition of such subpopulations needs enough data, or innovative strategies as clustering of lactations by features which could include not only breed, as done in this study, but also genetic differences between animals.

In the context of machine learning, particularly when faced with limited data, the relationship between the number of parameters in a model and the amount of available data is crucial. Small-structured networks, characterized by fewer parameters, are often employed to mitigate the risk of overfitting when data is scarce. However, this trade-off necessitates a careful balance; too few parameters may hinder the model's ability to capture complex patterns in the data. To address this, data augmentation techniques can be invaluable, as they artificially expand the training dataset by introducing variations through different types of transformations. This not only increases the effective size of the dataset but also enhances the model's robustness and generalization capabilities. One possibility in the future is also to combine the strengths of the different methods shown. Methods such as TIM, MTP or others, could be used as data augmentation tools to obtain extended datasets. Moreover, domain transfer strategies can provide significant advantages by allowing the model to leverage knowledge from related domains or tasks. By pre-training on larger, relevant datasets, we can improve performance even in scenarios with limited data. Future work should focus on optimizing the interplay between model complexity and data augmentation while also exploring effective domain transfer methods to further enhance predictive performance. By combining these strategies, models capable of achieving better outcomes in data-constrained environments can be developed.

While few device-based methods for routine body condition scoring (Martins et al., 2020) are available, they entail significant initial capital and ongoing maintenance costs. Consequently, animal scientists and producers seek a cost-effective method for regularly predicting accurate body condition scores (BCS). One proposed solution is to utilize mid-infrared (MIR) milk spectra to estimate BCS in dairy cows. However, this approach requires precise alignment between BCS data and MIR spectra for effective calibration. Successful data imputation allows missing information to be completed and thus improve conditions for the development of new models to add exogenous information that can also be obtained in routine by DHI. Therefore next steps will be to use these imputed BCS data, directly or indirectly, in the context of the development of models that predict BCS using exogenous information from milk yield, milk composition and especially milk mid-infrared (**MIR**)-based fine milk composition (McParland et al., 2011; Gengler et al., 2016; Mota et al., 2021). This requires further developments and needs additional research even if the present work provided

24

insight into strategies to align smoothed and imputed reference BCS data with DHI data containing relevant potential predictors. Even if the setting of this study did not favor their use, random regression models (e.g., Mayeres et al., 2004) and alternative approaches such as generalized additive models (e.g., Ankinakatte et al., 2013) have specific advantages to become alternatives to the methods proposed in this study.

This work is not conclusive since more experimentation might be needed. However, we can conclude that ML can avoid some initial assumptions that limit conventional interpolation methods and possess great potential in advanced intelligent applications over traditional techniques. Particularly, it is the case of the predictive capability of RNNs for longitudinal data without requiring any or much domain knowledge about the phenomenon of study. ML methods and especially deep learning methods are promising for the future development and use in the field of study. However traditional methods such as TIM or MTP, which are defined by known equations, facilitate the interpretation of the obtained model. This is often not straightforward for ML methods and particularly deep learning, in which model explainability is a known weak point (Arrieta et al., 2020).

4. CONCLUSIONS

The application of data imputation is of particular interest in the precision livestock farming area. In this study six different methods were evaluated to impute BCS values throughout the lactation in Holstein and Montbéliarde dairy cows. These methods were classified into traditional methods (TIM and MTP), and ML methods (MLP, Elman, LSTM and BiLSTM). Two tasks were addressed: the imputation of intermediate BCS values and the forecasting of future BCS values. For both tasks, MTP provided the best performance in terms of RMSE and Pearson correlation. The studied deep learning methods like LSTM or BiLSTM

25

did not outperform MTP, but this may also be due to non-optimal context (i.e., amount of available data) of their use.

This study analyzes methods for successful BCS imputation, allowing missing information to be completed and thus improving conditions for the development of new models to add exogenous information that is also obtained in routine by DHI. The proposed methods also provide expected BCS values, which are useful for triggering early warnings in the event of atypical or unexpected conditions.

ACKNOWLEDGMENT

The authors acknowledge the INTERREG NWE HappyMoo project, grant agreement NWE 730, co-financed by the Walloon Government (Service Public de Wallonie, Namur, Belgium). The providing of BCS data by the DHI organizations 3CE (Strasbourg, France) and Conseil Élevage 25-90 (Roulans, France) is recognized. Nicolas Gengler, as a former senior research associate, acknowledges the support of the National Fund for Scientific Research (Brussels, Belgium) also through the grant no. T.0095.19 (PDR “DEEPSELECT”). The authors have not stated any conflicts of interest.

 26

REFERENCES

Ankinakatte, S., E. Norberg, P. Løvendahl, D. Edwards, and S. Højsgaard (2013). Predicting mastitis in dairy cows using neural networks and generalized additive models: A comparison. Comput. Electron. Agric. 99:1-6. https://doi.org/10.1016/j.compag.2013.08.024.

Alvarez, J.R., M. Arroqui, P. Mangudo, J. Toloza, D. Jatip, J.M. Rodriguez, A. Teyseyre, C. Sanz, A. Zunino, C. Machado, and C. Mateos. (2019). Estimating body condition score in dairy cows from depth images using convolutional neural networks, transfer learning and model ensembling techniques. Agronomy 9:90. https://doi.org/10.3390/agronomy9020090

Anglart, D., C. Hallén-Sandgren, U. Emanuelson, and L. Rönnegård. (2020). Comparison of methods for predicting cow composite somatic cell counts. J. Dairy Sci. 103:8433-8442.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58:82-115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bastin, C. and N. Gengler. 2013. Genetics of body condition score as an indicator of dairy cattle fertility. A review. Biotechnol. Agron. Soc. Environ. 17:64-75.

Bastin, C., L. Laloux, A. Gillon, C. Bertozzi, S. Vanderick, and N. Gengler. 2007. First results of body condition score modeling for Walloon Holstein cows. Interbull Bull. 37:170-174. https://journal.interbull.org/index.php/ib/article/view/1371/1371

Beam, S. and W. Butler. 1999. Effects of energy balance on follicular development and first ovulation in postpartum. J. Reprod. Fertil. Suppl. 54:411-424.

27

Berghof, T. V., M. Poppe, and H. A. Mulder. 2019. Opportunities to improve resilience in animal breeding programs. Front. Genet. 9:692. https://doi.org/10.3389/fgene.2018.00692.

Bernabucci, U., B. Ronchi, N. Lacetera, and A. Nardone. 2005. Influence of body condition score on relationships between metabolic status and oxidative stress in periparturient dairy cows. J. Dairy Sci. 88:2017-2026. https://doi.org/10.3168/jds.S0022-0302(05)72878-2.

Bishop, C. M. and N. M. Nasrabadi. 2006. Pattern recognition and machine learning. Vol. 4. No. 4. Springer, New York, NY.

Chollet, F. 2015. Keras. https://github. com/fchollet/keras.

Coffey, M. P., G. Simm, and S. Brotherstone. 2002. Energy balance profiles for the first three lactations of dairy cows estimated using random regression. J. Dairy Sci. 85:2669-2678. https://doi.org/10.3168/jds.S0022-0302(02)74352-X.

Collard, B. L., P. J. Boettcher, J. C. M. Dekkers, D. Petitclerc, and L. R. Schaeffer. 2000. Relationships between energy balance and health traits of dairy cattle in early lactation. J. Dairy Sci. 83:2683-2690. https://doi.org/10.3168/jds.s0022-0302(00)75162-9.

Du, A., Guo, H., Lu, J., Su, Y., Ma, Q., Ruchay, A., et al. (2022). Automatic livestock body measurement based on keypoint detection with multiple depth cameras. Computers and Electronics in Agriculture, 198, Article 107059. https://doi.org/10.1016/j.compag.2022.107059

Edmonson, A. J., I. J. Lean, L. D. Weaver, T. Farver, and G. Webster. 1989. A body condition scoring chart for Holstein dairy cows. J. Dairy Sci. 72:68-78. https://doi.org/10.3168/jds.S0022-0302(89)79081-0.

Elman, J. L. 1990. Finding structure in time. Cogn. Sci. 14:179-211. https://doi.org/10.1016/0364-0213(90)90002-E.

 28

Everett, R. W. and H. W. Carter. 1968. Accuracy of test interval method of calculating dairy herd improvement association records. J. Dairy Sci. 51:1936-1941. https://doi.org/10.3168/jds.S0022-0302(68)87317-5.

Ferguson, J. D., D. T. Galligan, and N. Thomsen. 1994. Principal 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.

Friggens, N. C., P. Berg, P. Theilgaard, I. R. Korsgaard, K. L. Ingvartsen, P. Løvendahl, and J. Jensen. 2007. Breed and parity effects on energy balance profiles through lactation: Evidence of genetically driven body energy change. J. Dairy Sci. 90:5291-5305. https://doi.org/10.3168/jds.2007-0173.

Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M.-L. Vanrobays, A. Lainé, S. Vanderick, and C. Grelet. 2016. Capitalizing on fine milk composition for breeding and management of dairy cows. J. Dairy Sci. 99:4071-4079. https://doi.org/10.3168/jds.2015-10140.

Graham, P., Young, J. and Penny, R., 2009. Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models. Journal of Official Statistics, 25(2):245-268.

Graves, A. and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 18:602-610. https://doi.org/10.1016/j.neunet.2005.06.042.

Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. Neural Comput. 9:1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

ICAR. 2020. Procedure 2 of Section 2 of ICAR Guidelines computing of accumulated lactation yield. Accessed March. 26, 2023. https://www.icar.org/Guidelines/02-Procedure-2-Computing-Lactation-Yield.pdf.

 29

ICAR. 2022. Appendix 1 of Section 5 of ICAR Standard trait definition for dairy cattle. Accessed March. 26, 2023. https://www.icar.org/Guidelines/05-Conformation-recording-Appendix-1.pdf

Kristensen, E., L. Dueholm, D. Vink, J. Andersen, E. Jakobsen, S. Illum-Nielsen, F. Petersen, and C. Enevoldsen. 2006. Within-and across-person uniformity of body condition scoring in Danish Holstein cattle. J. Dairy Sci. 89:3721-3728.

Kuzuhara, Y., K. Kawamura, R. Yoshitoshi, T. Tamaki, S. Sugai, M. Ikegami, Y. Kurokawa, T. Obitsu, M. Okita, and T. Sugino. 2015. A preliminarily study for predicting body weight and milk properties in lactating Holstein cows using a three-dimensional camera system. Comput. Electron. Agric. 111:186-193. https://doi.org/10.1016/j.compag.2014.12.020.

Leary, N.O., Leso, L., Buckley, F., Kenneally, J., McSweeney, D. and Shalloo, L., 2020. Validation of an automated body condition scoring system using 3D imaging. Agriculture, 10(6):246. https://doi.org/10.3390/agriculture10060246

Lobato, F., C. Sales, I. Araujo, V. Tadaiesky, L. Dias, L. Ramos, and A. Santana. 2015. Multi-objective genetic algorithm for missing data imputation. Pattern Recognit. Lett. 68:126-131. https://doi.org/10.1016/j.patrec.2015.08.023.

Luo, X., Hu, Y., Gao, Z., Guo, H., & Su, Y. (2023). Automated measurement of livestock body based on pose normalisation using statistical shape model. Biosystems Engineering, 227, 36–51. https://doi.org/10.1016/j.biosystemseng.2023.01.016

Macciotta, N. P. P., D. Vicario, and A. Cappio-Borlino. 2005. Detection of different shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models. J. Dairy Sci. 88:1178-1191. https://doi.org/10.3168/jds.S0022-0302(05)72784-3.

Mao, I., K. Sloniewski, P. Madsen, and J. Jensen. 2004. Changes in body condition score and in its genetic variation during lactation. Livest. Prod. Sci. 89:55-65. https://doi.org/10.1016/j.livprodsci.2003.12.005.

30

Martins, B. M., Mendes, A. L. C., Silva, L. F., Moreira, T. R., Costa, J. H. C., Rotta, P. P., Chizzotti M. L. and Marcondes, M. I. (2020). Estimating body weight, body condition score, and type traits in dairy cows using three dimensional cameras and manual body measurements. Livestock science, 236, 104054. https://doi.org/10.1016/j.livsci.2020.104054.

Mayeres, P., J. Stoll, J. Bormann, R. Reents, and N. Gengler (2004). Prediction of daily milk, fat, and protein production by a random regression test-day model. J. Dairy Sci. 87:1925-1933. https://doi.org/10.3168/jds.S0022-0302(04)73351-2.

McParland, S., G. Banos, E. Wall, M. Coffey, H. Soyeurt, R. Veerkamp, and D. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94:3651-3661. https://doi.org/10.3168/jds.2010-3965.

Meng, Z. and Z. Shi. 2012. Extended rough set-based attribute reduction in inconsistent incomplete decision systems. Inf. Sci. 204:44-69. https://doi.org/10.1016/j.ins.2012.04.004.

Mota, L. F. M., S. Pegolo, T. Baba, F. Peñagaricano, G. Morota, G. Bittante, and A. Cecchinato. 2021. Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. J. Dairy Sci. 104:8107-8121. https://doi.org/10.3168/jds.2020-19861.

Mullins, I.L., Truman, C.M., Campler, M.R., Bewley, J.M. and Costa, J.H., 2019. Validation of a commercial automated body condition scoring system on a commercial dairy farm. Animals, 9(6):287. https://doi.org/10.3390/ani9060287

Poppe, M., G. Bonekamp, M. L. van Pelt, and H. A. Mulder. 2021. Genetic analysis of resilience indicators based on milk yield records in different lactations and at different lactation stages. J. Dairy Sci. 104:1967-1981. https://doi.org/10.3168/jds.2020-19245.

 31

Poppe, M., R. F. Veerkamp, M. L. van Pelt, and H. A. Mulder. 2020. Exploration of variance, autocorrelation, and skewness of deviations from lactation curves as resilience indicators for breeding. J. Dairy Sci. 103:1667-1684. https://doi.org/10.3168/jds.2019-17290.

Pryce, J. E., M. P. Coffey, and G. Simm. 2001. The relationship between body condition score and reproductive performance. J. Dairy Sci. 84:1508-1515. https://doi.org/10.3168/jds.S0022-0302(01)70184-1.

Roche, J. R., P. G. Dillon, C. R. Stockdale, L. H. Baumgard, and M. J. VanBaale. 2004. Relationships among international body condition scoring systems. J. Dairy Sci. 87:3076-3079. https://doi.org/10.3168/jds.s0022-0302(04)73441-4.

Roche, J. R., N. C. Friggens, J. K. Kay, M. W. Fisher, K. J. Stafford, and D. P. Berry. 2009. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. J. Dairy Sci. 92:5769-5801. https://doi.org/10.3168/jds.2009-2431.

Rubin, D. B. 1976. Inference and missing data. Biometrika 63:581-592. https://doi.org/10.1093/biomet/63.3.581.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1985. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, San Diego, CA. https://apps.dtic.mil/sti/pdfs/ADA164453.pdf

Sainani, K.L., 2015. Dealing with missing data. PM&R, 7(9):990-994.

Sakaguchi, M. 2009. Differences between body condition scores and body weight changes in postpartum dairy cows in relation to parity and reproductive indices. Can. Vet. J. 50:649-656.

Sargent, F., V. Lytton, and O. Wall Jr. 1968. Test interval method of calculating dairy herd improvement association records. J. Dairy Sci. 51:170-179. https://doi.org/10.3168/jds.S0022-0302(68)86943-7.

32

Schaeffer, L.R. and J. Jamrozik. 1996. Multiple-trait prediction of lactation yields for dairy cows. J. Dairy Sci. 79:2044-2055. https://doi.org/10.3168/jds.S0022-0302(96)76578-5.

Schröder, U.J. and R. Staufenbiel. 2006. Invited review: Methods to determine body fat reserves in the dairy cow with special regard to ultrasonographic measurement of backfat thickness. J. Dairy Sci. 89:1-14. https://doi.org/10.3168/jds.S0022-0302(06)72064-1.

Song, X., E.A.M. Bokkers, S. van Mourik, P.W.G. Groot Koerkamp, and P. van der Tol. 2019. Automated body condition scoring of dairy cows using 3-dimensional feature extraction from multiple body regions. J. Dairy Sci. 102:4294-4308. https://doi.org/10.3168/jds.2018-15238.

Spoliansky, R., Y. Edan, Y. Parmet, and I. Halachmi. 2016. Development of automatic body condition scoring using a low-cost 3-dimensional Kinect camera. J. Dairy Sci. 99:7714-7725. https://doi.org/10.3168/jds.2015-10607.

Truman, C. M., M. R. Campler, and J. H. Costa. 2022. Body condition score change throughout lactation utilizing an automated BCS System: A descriptive study. Animals 12:601. https://doi.org/10.3390/ani12050601.

Van Buuren, S. 2018. Flexible imputation of missing data. CRC press. New-York, NY.

Walsh, S., F. Buckley, K. Pierce, N. Byrne, J. Patton, and P. Dillon. 2008. Effects of breed and feeding system on milk production, body weight, body condition score, reproductive performance, and postpartum ovarian function. J. Dairy Sci. 91:4401-4413. https://doi.org/10.3168/jds.2007-0818.

Wathes, D.C., Z. Cheng, N. Bourne, V. J. Taylor, M. P. Coffey, and S. Brotherstone. 2007. Differences between primiparous and multiparous dairy cows in the inter-relationships between metabolic traits, milk yield and body condition score in the periparturient period. Domest. Anim. Endocrinol. 33:203-225. https://doi.org/10.1016/j.domaniend.2006.05.004.

33

Wildman, E., G. Jones, P. Wagner, R. Boman, H. Troutt Jr, and T. Lesch. 1982. A dairy cow body condition scoring system and its relationship to selected production characteristics. J. Dairy Sci. 65:495-501. https://doi.org/10.3168/jds.S0022-0302(82)82223-6.

Wilmink, J. 1987. Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. Livest. Prod. Sci. 16:335-348. https://doi.org/10.1016/0301-6226(87)90003-0.

Woolson, R. F. 2007. Wilcoxon signed rank test. Wiley Encyclopedia of Clinical Trials:1-3.

Zhang, J., Lei, J., Wu, J., Lu, H., Guo, H., Pezzuolo, A., et al. (2023). Automatic method for quantitatively analyzing the body condition of livestock from 3D shape. Computers and Electronics in Agriculture, 214, Article 108307. https://doi.org/10.1016/j.compag.2023.108307

ORCID

J. Chelotti - 0000-0003-2365-1298

H. Atashi - 0000-0002-6853-6608

M. Ferrero - 0000-0002-4298-6440

C. Grelet - 0000-0003-3313-485X

H. Soyeurt - 0000-0001-9883-9047

L. Giovanini - 0000-0002-6391-7671

H. L. Rufiner - 0000-0002-1083-5891

N. Gengler - 0000-0002-5981-5509

34

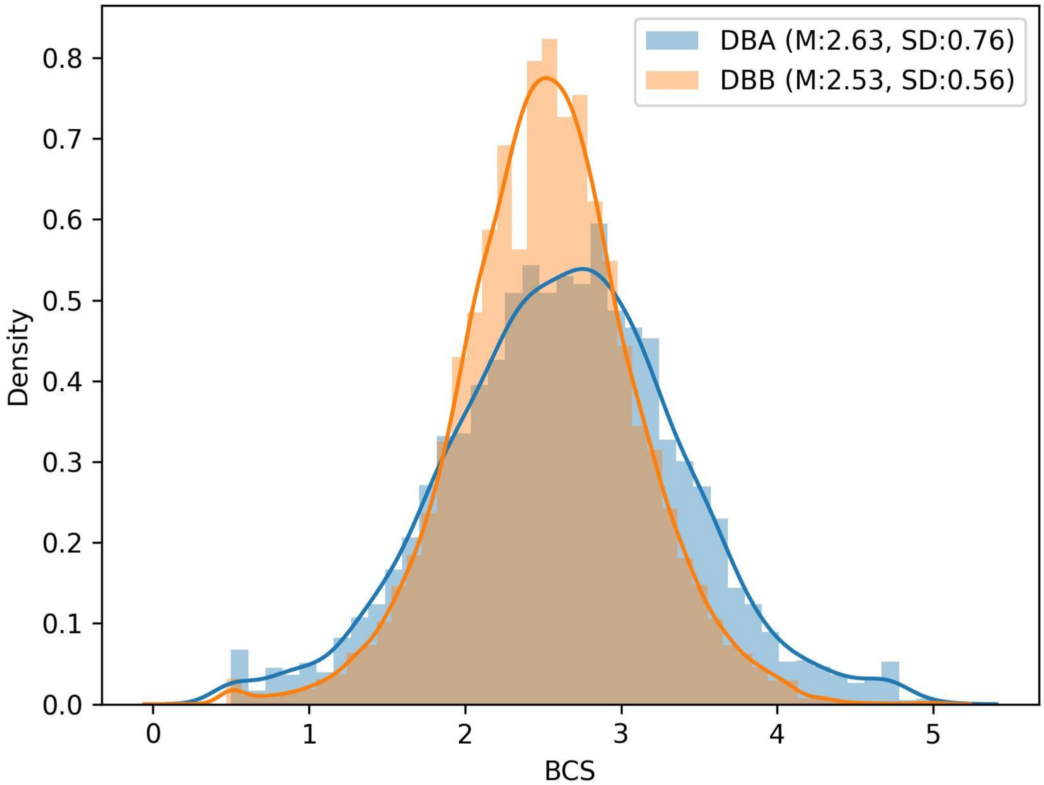

Figure 1. Distribution of BCS data for each database. DBA = data collected in the Alsace region in France, DBB = data collected in the Bourgogne-Franche-Comté region in France

35

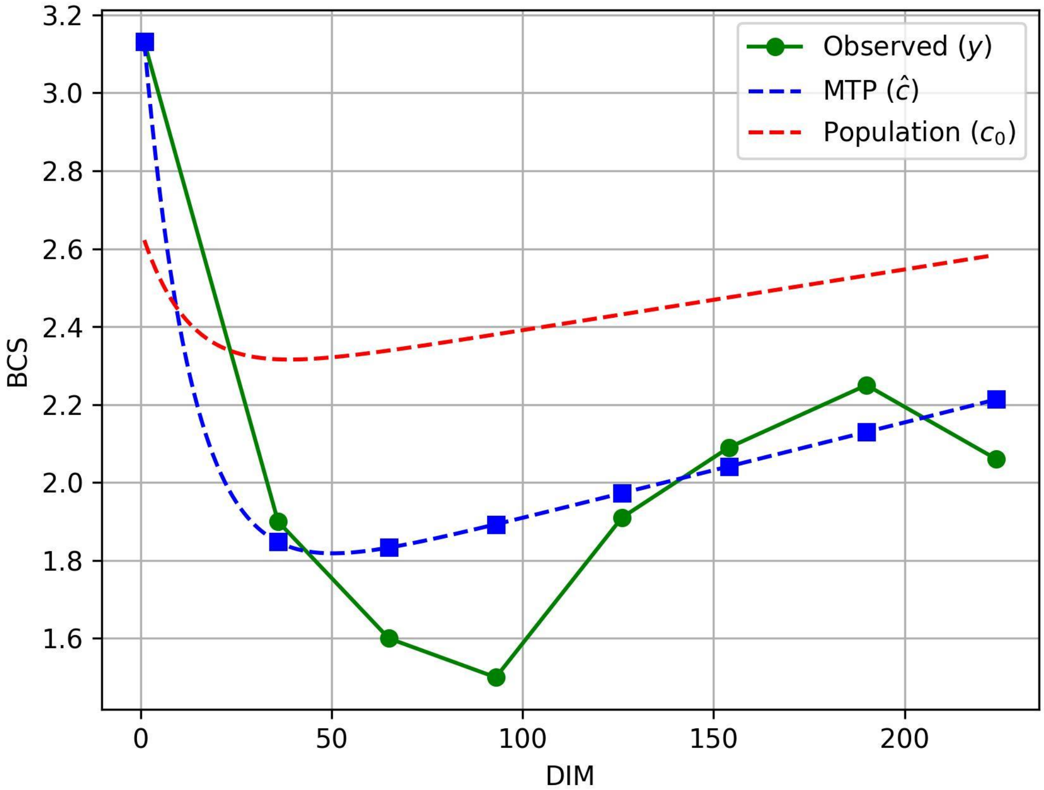

Figure 2. The multiple-trait prediction procedure curve (blue squared) represents the estimated lactation curve parameters as a combination of the population curve (red dashed) representing the population curve parameters and the observed BCS values (green dotted) for each specific cow and lactation combination.

36

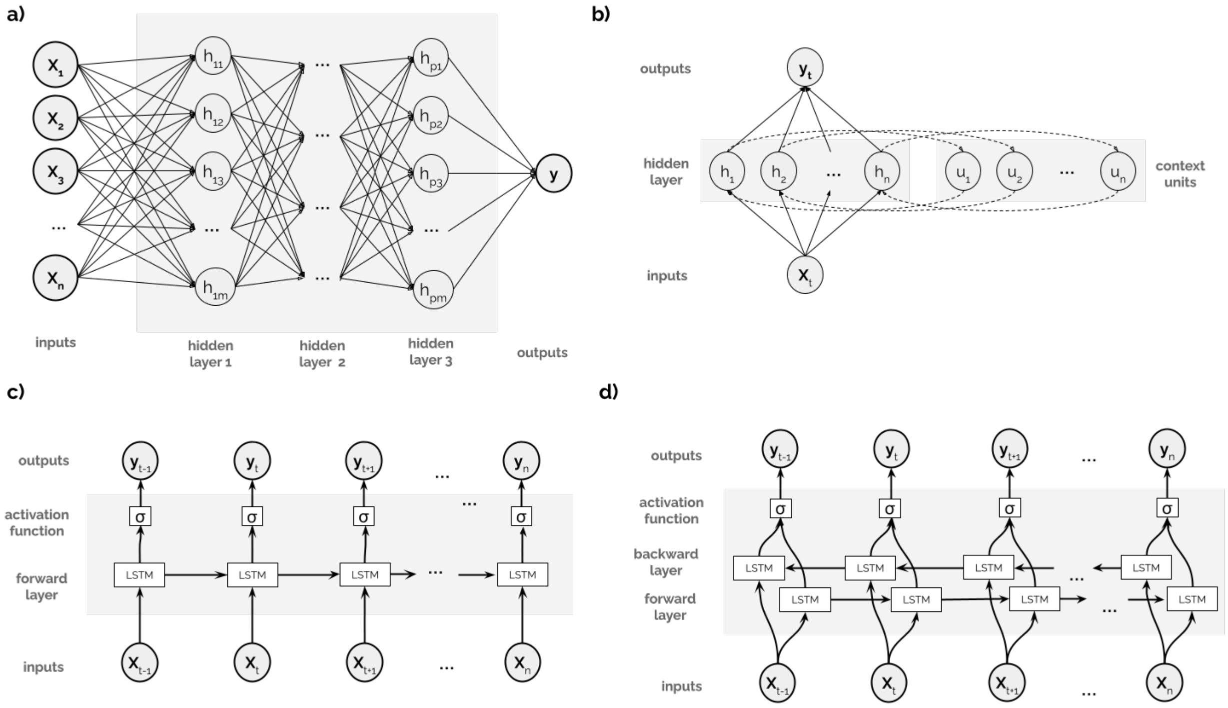

Figure 3. General architecture of the used Machine Learning methods: Multi-layer perceptron (MLP) (a), Elman network (Elman) (b), long-short term memories (LSTM) (c), and Bi-directional LSTM (BiLSTM) (d).

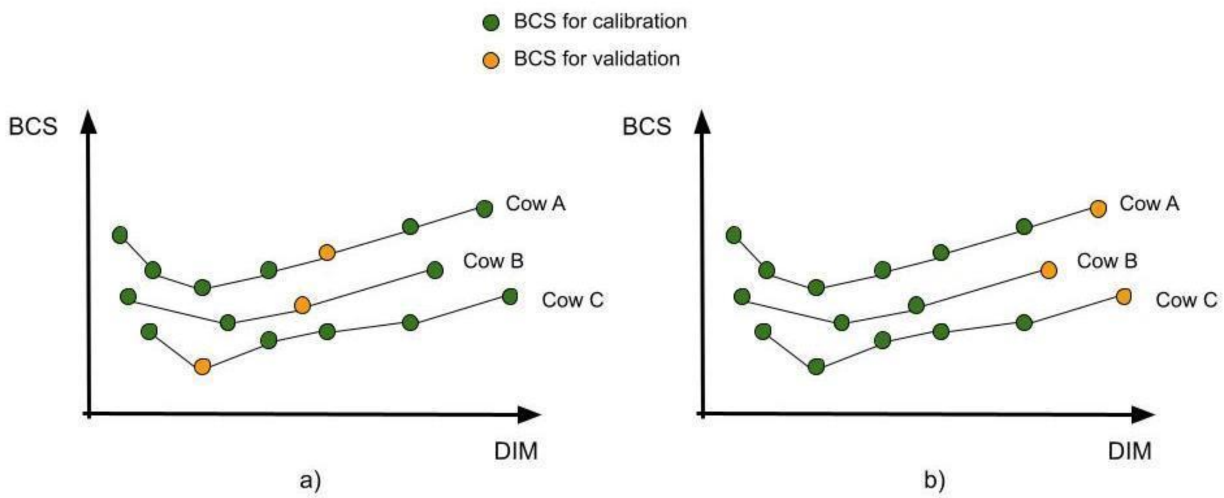

Figure 4. Exemplification of lactation curves composed of discrete BCS observations for both settings: intermediate data imputation (a) and forecasting (b). In (a), BCS values in-between were randomly selected to build the validation set (orange points), while the remaining points were kept for the calibration set (green points). In (b), only the last values of each sequence were selected to build the validation set (orange points), while the remaining points were kept for the calibration set (green points).

38

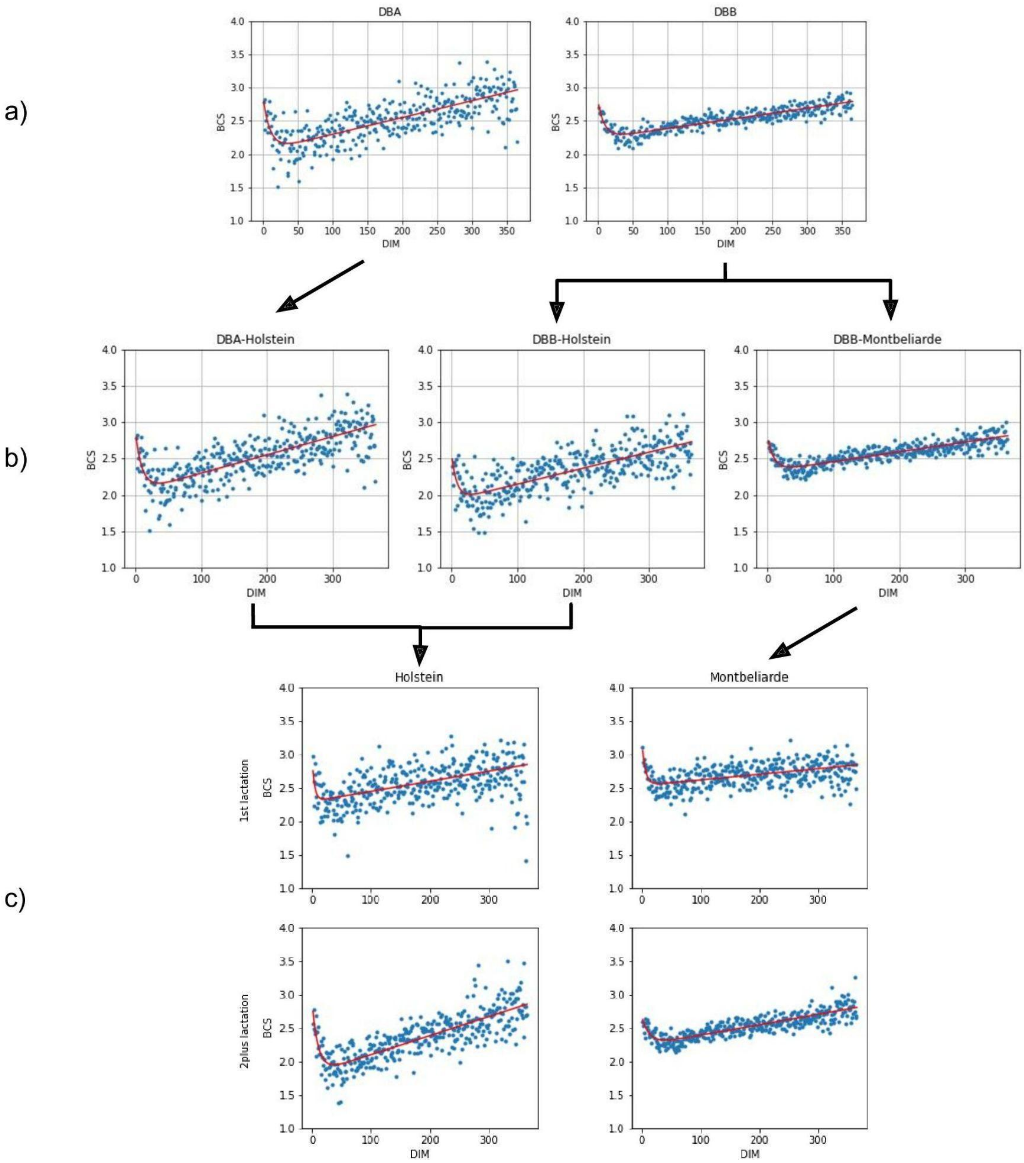

Figure 5. Evolution of average BCS by days in milk (DIM) (blue dots) and its corresponding mean curve (red curves) through the lactation for each database (a), population defined as breeds inside databases (b) and parities and breeds (c).

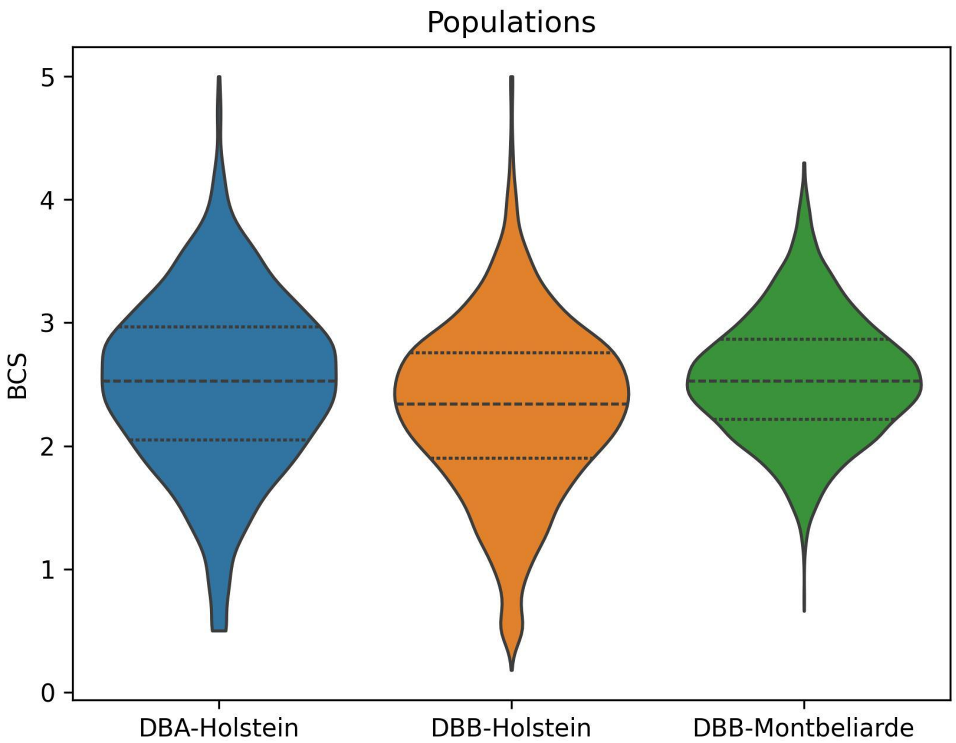

Figure 6. Global distribution of BCS values for each population. Median (dashed lines) and quartiles (dotted lines) of the populations are included in the Figure.

 40

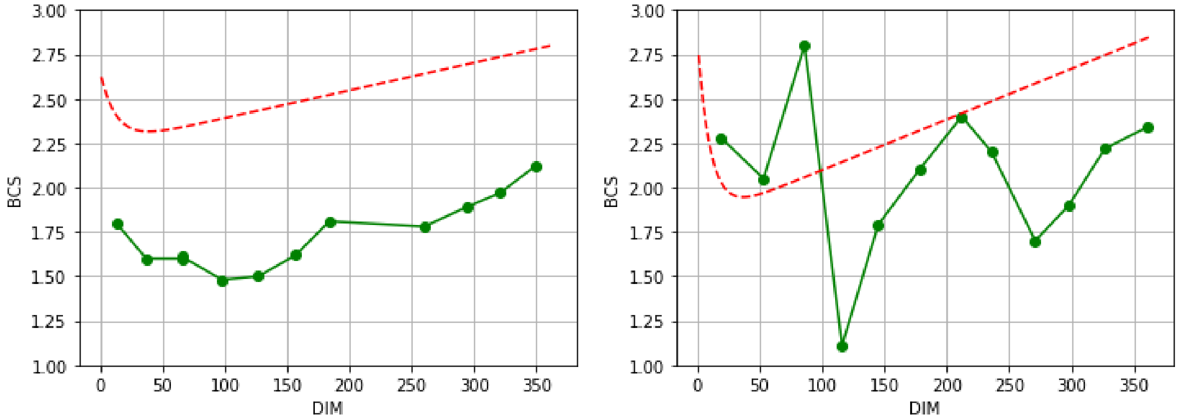

Figure 7. Examples for a typical curve (left) and an atypical curve (right). The observed (green dotted) and the expected (red dashed) curves for the population are shown for each cow and lactation.

41

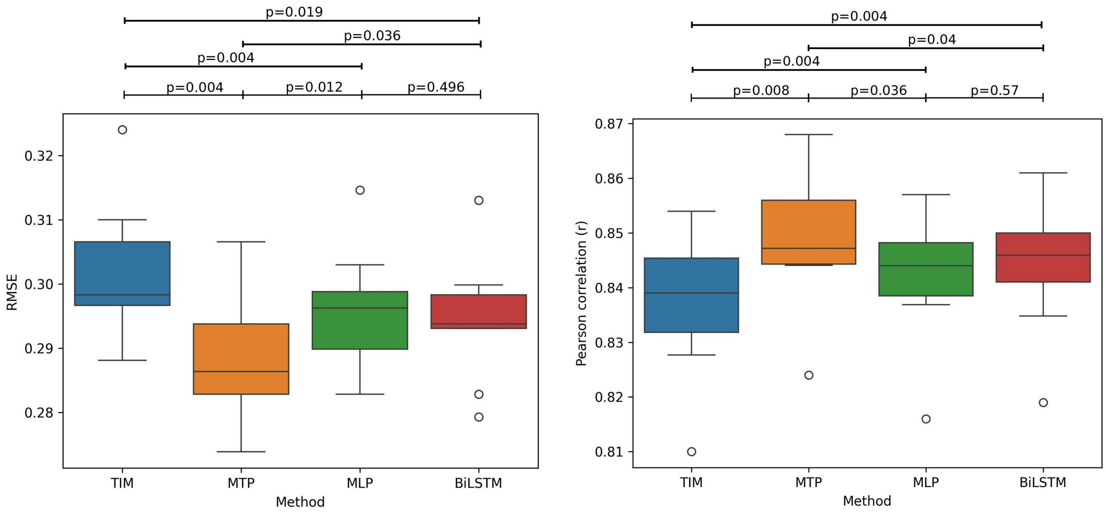

Figure 8. Distribution of root mean squared error (RMSE) and the Pearson correlation (r) for each method over 10 random executions. P-values (Wilcoxon test) are at the top of the Figure.

Table 1. Details of the raw BCS databases used in this study.

	DBA[1]	DBB[2]
Number of BCS records	5,629	20,578
Primiparous	37.8%	28.6%
Multiparous	62.2%	71.4%
Number of animals	932	2,106
Primiparous	52.7%	50.0%
Multiparous	47.3%	50.0%
Number of herds	8	18
Majority breeds[3]		
Holstein	86%	22%
Montbéliarde	-	77%
Other[4]	14%	1%
Number of groups cow and parity (BCS curves)	1,367	3,380
Number of BCS records by cow and parity (SD)	4.03 (2.14)	5.61 (3.54)
Recording period	Jan. 2019 - Dec. 2020	Nov. 2018 - Oct. 2020

[1] DBA = data from the Alsace region in France.

[2] DBB = data from the Bourgogne-Franche-Comté region in France.

[3] Expressed as a percentage of the total number of animals.

[4] Includes crossbreeds and other minority breeds.

Table 2. Body condition score details for studied populations.

	Montbéliarde		Holstein	
	Primiparous	Multiparous	Primiparous	Multiparous
Mean (SD)	2.68 (0.49)	2.51 (0.49)	2.66 (0.63)	2.32 (0.73)
Median	2.65	2.50	2.57	2.34
Nadir	2.56	2.32	2.33	1.95
Nadir DIM	28	39	19	37
Delta (δ)[1]	0.15	0.28	0.28	0.52

[1] The parameter of the linear term in equation (3), indicating the general slope after nadir and strongly related with the recovery of the body condition.

Table 3. Macro-average over 10 random executions for the imputation of intermediate BCS values. Under the Wilcoxon test, MTP achieved significantly better results than the rest of the methods ($p<0.05$).

	Traditional		ML[1]	
	TIM[2]	MTP[3]	MLP[4]	BiLSTM[5]
Root Mean Squared Error (RMSE) ↓	0.302	**0.288**	0.297	0.295
Pearson Correlation (r) ↑	0.837	**0.849**	0.843	0.845

1. ML = Machine learning.
2. TIM = Test Interval Method.
3. MTP = Multiple-Trait Procedure.
4. MLP = Multi-Layer Perceptron.
5. BiLSTM = Bi-directional Long-Short Term Memories.

Table 4. Methods evaluated for the forecasting of BCS values.

	Traditional	ML[1]	
	MTP[2]	Elman[3]	LSTM[4]
Root Mean Squared Error (RMSE) ↓	**0.348**	0.373	0.356
Pearson Correlation (r) ↑	**0.760**	0.728	0.751

1. ML = Machine learning.
2. MTP = Multiple-Trait Procedure.
3. Elman = Elman network.
4. LSTM = Long-Short Term Memories.