Computers and Electronics in Agriculture 219 (2024) 108777

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Predicting bovine daily milk yield by leveraging genomic breeding values

Andrea Mario Vergani [a,*], Alessandro Bagnato [b], Marco Masseroli [a]

[a] Politecnico di Milano, Department of Electronics, Information and Bioengineering, Via Ponzio 34/5, 20133 Milano, Italy
[b] Università degli Studi di Milano, Department of Veterinary Medicine and Animal Sciences, Via dell'Università 6, 26900 Lodi, Italy

ARTICLE INFO

Keywords:
Milk production forecasting
Genomic breeding value
Machine learning
Individual prediction
Phenomics

ABSTRACT

The main goal of this work, conducted on a herd of 502 Holstein cows situated in Italy, is to propose a machine learning-based approach to forecast the individual bovine daily milk production by explicitly leveraging genotypic information. As part of our study, we also evaluated the importance in the prediction of genotypic and phenotypic variables usually available within herd. The methodology we propose is based on two consecutive models: a genomic prediction one to calculate the animal's genomic breeding value from marker data, followed by a feed forward neural network combining such additive genetic effect and the environmental features (parity, days in milk, age at calving in months, month of calving) for milk yield forecasting. In particular, we both assess the inclusion of genomic breeding values calculated within herd or provided by the breeders' association, discovering that the latter ones allow for better final predictions. The results of our model outperform the ones by a linear mixed model with the same inputs on average, day by day and at the individual level. Moreover, we propose a problem formulation that also leverages additional factors partially controllable by breeders: in this case, features such as the number of milkings and the concentrate consumption inside the automatic milking system prove to highly impact on the final prediction, and hence on milk production. To the best of our knowledge, the proposed problem formulation based on genomic breeding values is a novelty in the individual bovine milk yield machine learning forecasting literature. Given the low genotyping costs and the availability of a larger number of environmental features in farms equipped with a wide range of sensors, as automatic milking systems, our solution can support breeders' herd management and animal monitoring, thanks to the possibility to forecast the full lactation curve in advance even for primiparous bovines and newborn calves. With this work, we successfully achieve our objectives of including genomic information in bovine milk yield machine learning-based forecasting, thus improving the performance on this task, and of evaluating the impact on prediction of common genotypic and phenotypic information available to breeders.

1. Introduction

Milk yield forecasting is a constantly developing discipline in the dairy cattle breeding sector, allowing for improved animal monitoring, herd management and decision-support for farmers. For these reasons, some models based on machine and deep learning for predicting the lactation curve at the individual level have been recently proposed (Liseune et al., 2021; Nguyen et al., 2020; Zhang et al., 2022, 2020), also leveraging the availability of considerable quantities of data collected within herd (for instance, by automatic milking systems - AMSs). In particular, Nguyen et al. (2020) showed the importance of diet parameters (feed composition and quantity) to predict the weekly milk production of Holstein Friesian dairy cows. The approach by Zhang et al. (2020) explicitly considered the time series nature of the problem, through a recurrent neural network architecture to forecast the lactation curve. Also Liseune et al. (2021) implemented a deep learning solution exploiting previous cycle records in order to predict the milk production curve of bovines in subsequent lactations. Conversely, the study by Zhang et al. (2022) focused on forecasting the milk yield of primiparous Holstein cows, through a framework combining clustering, classification and regression and leveraging monthly production data of relatives.

However, despite the growing availability of bovine genotypic data to breeders and the theoretical potential of integrating them in the machine learning-based milk yield forecasting task, we are not aware of any study implementing such kind of solution. For these reasons, the main goal of this work is to combine phenotypic and genotypic data for the prediction of the bovine daily milk production trait at the individual level; indeed, as discussed by Zhang et al. (2022), a proper use of the genomic information is likely to improve the state-of-the-art forecasting performance. Moreover, we also aim at integrating in the predictive task some variables that are both usually collected by AMSs

* Corresponding author.
E-mail address: andreamario.vergani@polimi.it (A.M. Vergani).

https://doi.org/10.1016/j.compag.2024.108777
Received 14 April 2023; Received in revised form 21 January 2024; Accepted 20 February 2024
Available online 27 February 2024

Table 1
Data description for the variables used in the study, with means, standard deviations (std), quantiles (q), minimum (min) and maximum (max) values. The statistics about genomic breeding values by the breeders' association, fixed by bovine, were computed considering the 502 animals; those about parity, age at calving (in months) and month of calving, fixed by lactation, considering the 865 lactations; the ones about days in milk, mean number of milkings, daily concentrate consumption and daily milk production, varying record by record, considering the 154,580 observations.

Variable	mean (std)	min	0.25 q	0.5 q	0.75 q	max
Genomic breeding value (kg)	741.7 (506.8)	−984.0	419.0	765.0	1102.8	2114.0
Parity	1.76 (0.95)	1	1	1	2	6
Age at calving (in months)	35.05 (12.86)	21	24	33	41	96
Month of calving	7.13 (3.73)	1	3	8	10	12
Days in milk	157.1 (102.3)	1	71	145	230	450
Mean number of milkings (/day)	2.74 (0.75)	1.00	2.10	2.70	3.30	6.90
Concentrate consumption (kg)	4.27 (1.33)	0.00	3.54	4.05	5.29	9.55
Milk production (kg)	39.7 (9.1)	1.9	33.7	39.5	45.7	77.3

and partially controllable by breeders: in particular, we believe that the analysis of the importance of such input features in the final predictions could better support dairy farmers in their herd management decisions and lactation planning. Summarizing, our objective is to propose a machine learning method to forecast the bovine daily milk yield at the individual level by explicitly leveraging genotypic information, as well as partially controllable phenotypic variables collected by AMSs.

2. Materials and methods

2.1. Data description

The data analyzed in this work came from a single Holstein Friesian herd equipped with AMSs, situated in Lombardy (in the northern flat of Italy) and monitored for 26 consecutive months (between February 2020 and March 2022). A total of 9 AMSs were installed in the farm, model Lely Astronaut. The considered 502 animals were genotyped with the Neogen GeneSeek Genomic Profiler Bovine 100k SNP chip, so their genotypes on 95,256 single nucleotide polymorphisms (SNPs) were available; the cows, under a unique management system, were in a freestall barn structure with natural ventilation subsidized by fans, in order to guarantee the most uniform conditions as a response to changes in weather (i.e., temperature and humidity). AMSs collected 154,580 valid daily milk production records, also comprising date, parity, days in milk, rumination time, eating time, mean number of milkings (averaged on the last seven days) and concentrate consumption during milking in the AMS. Other available data included the dates of birth of the bovines and their milk yield genomic estimated breeding values (gEBVs) calculated by the reference breeders' association (ANAFIBJ[1]) based on a wide set of animals. Overall, the available phenotypic data came from 865 different lactations, with an average value of about 1.72 lactations per bovine. The length of the lactation period was variable, depending on the animal, the environmental conditions and the herd management strategy, resulting in fewer observations for high values of days in milk. In the farm under analysis, the bovines voluntarily entered the AMS, so the number of daily milkings was not fixed. Finally, the animals were dried-off when their milk yield was durably below 25 kg. Table 1 reports some statistics about the available data used in this study.

Despite the availability of gEBVs by the breeders' association, in this study we also assessed the value of information of breeding values calculated within the herd to the task of milk yield forecasting, as a reference for comparison. In order to clean the genotype data, we performed the quality control steps proposed by Liang et al. (2021) (10% threshold on missingness per SNP and per individual, 5% threshold on minor allele frequency and 10^{-5} Hardy-Weinberg p-value threshold): in this way, 82,495 SNPs and all the bovines were retained. Missing genotype values were imputed using AlphaImpute software (Antolín et al., 2017; Hickey et al., 2012).

2.2. Data stratification

One of the advantages of integrating the genotypic information in the task of milk yield forecasting is the opportunity to generalize on new animals. Indeed, genotyping a bovine allows to predict its expected performance as parent of next generation even before observing any phenotypic records from it. For this reason, we directly split the collected daily observations on the individuals (not on the records themselves): 102 (about 20%) of the bovines were selected to constitute the test set, while the other 400 were distributed in 10 folds to perform a cross-validation for model selection; in this way, all the daily records of an animal were grouped in the same fold, and hence the generalization performance on new bovines could be assessed.

Moreover, in order to avoid strong unbalances in the number of samples and environmental feature values within each fold, we stratified the phenotypic data. In particular, the stratification considered only the longest lactation of every bovine (since less than 1% of them had records for two lactations of at least 270 days), and it has been performed on the combination of the following effect levels: first or subsequent lactation (2 levels); expected lactation peak (two months after calving, according to Hutjens, 2002, 2016 and confirmed by Hudson et al., 2019) in "warm" (from May to September, with average minimum temperatures above 10 °C and maximum temperatures above 20 °C) or "cold" (from October to April, with average minimum temperatures below 10 °C and maximum temperatures below 20 °C) months (2 levels), according to the historical mean temperatures by month in the Provincial Seat (Milano) of the analyzed dairy farm[2]; lactation with more than 270 daily records, or not (2 levels); presence of other lactations for the considered animal, or not (2 levels). Specifically, the first two described stratification effects are meant to balance the environmental feature values in the folds, while the last two ensure that the number of daily records inside every fold is approximately stable. We chose 270 as the number of records to identify almost full lactations (observations from at least nine out of ten 30-days lactation stages, according to Adamczyk et al., 2021).

Following the proposed methodology, the bovines were stratified according to the 16 described combinations of effect levels. Table 2 reports information about the number of stratification combinations present in each cross-validation fold and in the test set.

2.3. Modeling

2.3.1. Linear mixed model

We introduce the following linear mixed model with repeatability for daily milk yield observations, defined with the help of domain experts:

$$y_{ij} = \mu + lac_{ij} + day_{ij} + ageCalv_{ij} + monthCalv_{ij} + a_i + p_i + e_{ij} = b_{ij} + a_i + p_i + e_{ij} \quad (1)$$

[1] National breeders' association for Italian Friesian, Brown and Jersey cattle: http://www.anafi.it/en (Accessed April 1, 2023).

[2] Source: Il Meteo, Milano average temperatures, https://www.ilmeteo.it/portale/medie-climatiche/Milano (Accessed October 9, 2023).

Table 2
Cardinality of stratification combinations in each cross-validation fold and in the test set. Combinations of effect levels for stratification are reported in the rows, with the following notation: first bit (First) equal to 1 for first lactations, 0 for subsequent lactations; second bit (Cold) equal to 1 for expected lactation peak in "cold" months, 0 for "warm" months; third bit (Short) equal to 1 for lactations lasting less than 270 daily records, 0 otherwise; last bit (Other) equal to 1 for animals having multiple lactations, 0 otherwise. Fold numbers (1-10) and test set are reported in the columns.

First, Cold, Short, Other	1	2	3	4	5	6	7	8	9	10	Test	Total
1111	12	12	12	12	12	12	12	12	12	11	30	149
1110	4	4	4	4	4	5	5	5	5	4	11	55
1101	2	2	2	1	1	1	1	1	1	1	4	17
1100	5	5	5	5	5	4	5	5	5	5	13	62
1011	4	4	4	4	4	4	4	4	4	4	10	50
1010	1	1	1	1	1	1	0	1	1	1	2	11
1001	0	0	0	0	0	0	0	1	1	1	1	4
1000	2	2	2	2	2	2	2	2	2	3	5	26
0111	0	0	0	0	0	0	1	1	1	0	1	4
0110	1	1	1	2	2	2	2	1	1	1	3	17
0101	0	0	0	0	0	0	0	0	0	0	0	0
0100	6	6	6	6	6	6	5	5	5	6	15	72
0011	0	0	0	0	0	0	0	0	0	0	0	0
0010	1	1	1	1	1	1	1	0	0	1	2	10
0001	0	0	0	0	0	0	0	0	0	0	0	0
0000	2	2	2	2	2	2	2	2	2	2	5	25
Total	40	40	40	40	40	40	40	40	40	40	102	502

In this model, the daily milk production record j of animal i (y_{ij}) is written as the sum of the following components: the fixed effects related to the record (b_{ij}, comprising the historical daily milk yield μ, the parity lac_{ij}, the days in milk day_{ij}, the age at calving in months $ageCalv_{ij}$ and the month of calving $monthCalv_{ij}$); the additive genetic effect of animal i (a_i); the permanent environmental effect associated with animal i (p_i); and a residual error (e_{ij}). Details about random effects and covariance matrices are presented together with Eq. (2), constituting an extended matrix formulation of Eq. (1).

In particular, the additive genetic effects (a) are directly derived from the SNP effects (g):

$$y = Xb + WZg + Wp + e = Xb + Wa + Wp + e \tag{2}$$

Here, y represents the $(n_{pheno}, 1)$ vector of daily milk production observations (n_{pheno} is the total number of records), b the $(n_{fixed}, 1)$ vector of fixed effects (n_{fixed} is the number of levels of the fixed effects), g the $(n_{snp}, 1)$ vector of additive SNP effects (n_{snp} is the number of available SNPs), p the $(n_{animal}, 1)$ vector of permanent environmental effects (n_{animal} is the number of animals) with 0 mean and covariance matrix $G\sigma_p^2$ (G is defined in Eq. (3)), e the $(n_{pheno}, 1)$ vector of residuals with 0 mean and covariance matrix $I\sigma_e^2$ (I is the identity matrix); in addition, Z is the normalized (n_{animal}, n_{snp}) incidence matrix (relating an animal to its SNPs values) proposed by VanRaden (2008); X and W, instead, are incidence matrices ((n_{pheno}, n_{fixed}) and (n_{pheno}, n_{animal}), respectively) associating fixed and animal-related effects to the specific trait records. Moreover, as hinted at above, the $(n_{animal}, 1)$ vector of additive genetic effects a is given by $a = Zg$; a has 0 mean and covariance matrix equal to $G\sigma_a^2$.

As proposed by VanRaden (2008), the genomic relationship matrix G is defined as follows:

$$G = \frac{ZZ'}{2\sum_j p_j(1-p_j)} \tag{3}$$

In Eq. (3), the matrix Z' is the transpose of Z, while p_j represents the minor allele frequency of SNP j.

The proposed model was solved with the BLUPF90 software (Misztal et al., 2002); its solutions were used to integrate the additive genetic effects of animals (calculated within herd) in the machine learning predictions as genotypic features. Moreover, the linear mixed model also constituted a baseline reference model for forecasting the bovine daily milk yield.

2.3.2. Machine learning algorithms

In order to pursue the main goal of this work, we propose a machine learning approach for predicting the bovine daily milk production trait leveraging genomic information. In particular, we integrated the milk yield additive genetic effect calculated on the individual animal level, also known as gEBV, into the forecasting task; in practice, we built our dataset considering: the daily milk yield records as samples; the four fixed effects also considered in the linear mixed model presented in Eq. (1) (i.e., parity, days in milk, age at calving in months, month of calving) as features, together with the genomic breeding value; the mean number of milkings on the last seven days and the daily concentrate consumption as optional covariates; the daily milk production as target. Input features were chosen with the help of domain experts: in particular, the fixed effects were selected in accordance with the linear model of Eq. (1); the additional factors, instead, were chosen by experts among the information recorded by AMSs and considered to be mostly controllable by breeders. Given that the concentrate consumption was daily recorded, but farmers are used to periodically taking feed intake decisions, we also considered the feature averaged weekly, monthly, quarterly or over the full lactation.

Machine learning algorithms were trained for predicting the daily milk production (in kg) of a bovine given a set of environmental and genotypic features; indeed, the phenotypic target is both influenced by fixed and genetic effects. In particular, the only genotypic feature that we considered is the gEBV of the bovine, which is a single number quantifying the individual's additive genetic effects for the analyzed trait; such a quantity has been calculated in a genomic prediction step preceding the milk production forecasting one, as illustrated in Fig. 1. Specifically, we explored two options for the derivation of the additive genetic effects: a first one was to derive them within the considered herd with the model expressed by Eqs. (1) and (2), while another one consisted in directly taking advantage of the gEBVs calculated by the breeders' association, which are based on a much wider set of animals. The gEBV was used as an input of the phenomic forecasting model.

Regarding the environmental effects, we also assessed the importance for the final prediction of two additional features, with respect to the ones already considered in Eq. (1): the mean daily number of milkings (on the last seven days) and the concentrate consumption.

As described in Section 2.2, the dataset was stratified and split, with 102 (about 20%) bovines in the test set and 10 folds for cross-validation; we performed a grid search for hyperparameter tuning and model selection, according to the minimization of the mean squared error in cross-validation. Moreover, the feature values were scaled through

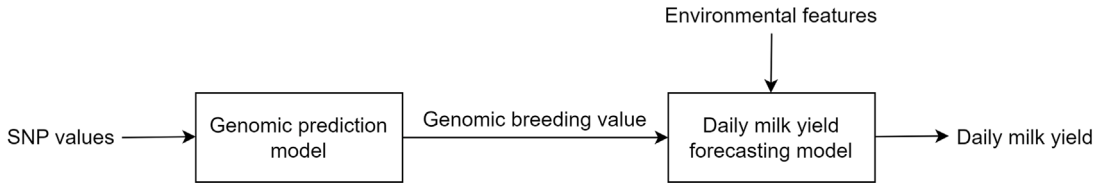

Fig. 1. Complete model architecture at inference time: first, the genomic breeding value is calculated from single nucleotide polymorphisms (SNPs) through a genomic prediction step; then, the genomic breeding value is combined with the environmental features in a machine learning model that predicts the bovine daily milk yield.

the Min-Max normalization. The proposed steps for feature scaling and model selection are quite common in the machine learning field.

Finally, we trained the following regression algorithms, and assessed their performance, for the proposed forecasting task: AdaBoost regression (Freund and Schapire, 1997) with decision trees as base estimators, Feed Forward Neural Networks (FFNN) (Rosenblatt, 1958), K-Nearest Neighbors regression (KNN) (Cover and Hart, 1967), Linear Regression, Random Forest regression (Breiman, 2001) and linear Support Vector Regression (SVR) (Smola and Schölkopf, 2004).[3]

2.4. Prediction assessment

2.4.1. Metrics

In order to assess the performance of our phenomic forecasting model, we primarily considered the mean squared error between actual and predicted daily milk yield values: the parameters of machine learning models were optimized according to mean squared error loss, and the minimization of cross-validation mean squared error was used for hyperparameter tuning and model selection. Moreover, we also monitored the Pearson correlation coefficient, the concordance correlation coefficient and the mean absolute error between predicted milk production values and the ground truth, as additional metrics to assess the model quality and the added value of information when including the gEBV, the mean number of milkings and the concentrate consumption as features of the forecasting task.

2.4.2. Models for performance comparison

As a baseline for prediction performance comparison against machine learning models, we selected the linear mixed model presented in Eqs. (1) and (2), similarly to the approach proposed by Nguyen et al. (2020) on the same task (i.e., machine learning-based milk yield prediction). Indeed, given the model architecture in Fig. 1, the phenomic forecasting performance of the linear mixed model was useful to assess the added value of information of our machine learning model (i.e., the enhancement in prediction given by the latter); analogously, we also evaluated the importance of single features (e.g., gEBV, mean number of milkings, concentrate consumption), as well as their combinations. Therefore, we compared the forecasting pipeline in Fig. 1 with every possible combination of single steps: only the linear model for daily milk yield forecasting; only the machine learning model (potentially with optional environmental features), without genomic information; genomic prediction followed by machine learning-based phenomic forecasting, potentially with optional environmental features. To ensure a fair and complete comparison between the machine learning model and the linear mixed one, we also assessed the predictive performance of the latter with the mean number of milkings and the concentrate consumption as additional fixed effects (i.e., extended version of the model presented in Eq. (2), with the optional environmental features included in the vector of fixed effects b).

[3] Feed Forward Neural Networks were implemented with the *Keras* Python library. All the other models were implemented with the *scikit-learn* Python library, in particular with the following classes: *AdaBoostRegressor* for AdaBoost, *KNeighborsRegressor* for K-Nearest Neighbors, *LinearRegression* for Linear Regression, *RandomForestRegressor* for Random Forest, *LinearSVR* for linear Support Vector Regression.

Given our primary goal to introduce a machine learning model with genomic information for the prediction of daily milk production, the comparison with other models not explicitly part of the pipeline in Fig. 1 (e.g., mechanistic models) was beyond the scope of this work; applications of machine learning-based milk yield forecasting without genomic information (Liseune et al., 2021; Nguyen et al., 2020; Zhang et al., 2022, 2020) and comparisons with other models (Murphy et al., 2018) are already extensively present in the literature.

3. Results

3.1. Prediction performance of machine learning algorithms

3.1.1. Genomic breeding values by the breeders' association

As described in Section 2.3.2, in order to include in the problem formulation the additive genetic effect as a feature, we leveraged the gEBV that the breeders' association (ANAFIBJ) calculated through a genomic prediction model based on the availability of a large number of over 360,000 genotypes.[4] Table 3 reports (in decreasing performance order) the average 10-fold cross-validation results of the application of the tested machine learning algorithms on our dataset. In particular, the presented results were obtained without considering the additional environmental features (mean number of milkings and concentrate consumption), whose impact is directly discussed in Section 3.3.

We noticed that all the machine learning models were able to obtain average milk yield prediction values very close to the ground truth. Moreover, we observed that the feed forward neural network models outperformed the other analyzed machine learning algorithms in cross-validation, in terms of mean squared error, mean absolute error, Pearson correlation coefficient and concordance correlation coefficient. Overall, adding more than one hidden layer did not seem to be substantially beneficial, also taking into account the larger number of hyperparameters and the longer training times: for this reason, we selected the neural network with one hidden layer as the forecasting model to be considered in subsequent analyses.

Obviously, in order to understand whether the gEBV is beneficial to the milk yield prediction, we also assessed the performance of a feed forward neural network with one hidden layer when considering only the mandatory environmental features (parity, days in milk, age at calving in months, month of calving): in such a situation, we observed a cross-validation mean squared error equal to $50.76\,\mathrm{kg}^2$ (standard deviation of $5.75\,\mathrm{kg}^2$), mean absolute error equal to $5.63\,\mathrm{kg}$ (standard deviation of $0.32\,\mathrm{kg}$), Pearson correlation coefficient equal to 0.63 (standard deviation of 0.04) and concordance correlation coefficient equal to 0.57 (standard deviation of 0.04). Comparing these results with those presented in Table 3, we concluded that the integration of the gEBV sensibly contributed to a better forecasting performance for the daily milk yield trait.

[4] "Genetic evaluation notes Holstein and Jersey" on ANAFIBJ website: http://server01.anafi.it/IndiciGenetici/Schede-calcolo-Indici-ING-2021.pdf (Accessed April 1, 2023).

Table 3
Machine learning model performances in cross-validation (Mean Squared Error - MSE, Mean Absolute Error - MAE, Pearson correlation coefficient - r, Concordance Correlation Coefficient - CCC), obtained considering the genomic breeding values by the breeders' association, compared to the ground truth in terms of average milk yield. FFNN stands for Feed Forward Neural Network, HL for hidden layer(s), KNN for K-Nearest Neighbors, SVR for Support Vector Regression; σ indicates the standard deviation of the performance measure across the 10 cross-validation folds.

Model	MSE [kg^2]	MAE [kg]	r	CCC	Average yield [kg]
Ground truth	–	–	–	–	39.59
FFNN 2 HL	41.71 ($\sigma = 4.27$)	5.06 ($\sigma = 0.27$)	0.71 ($\sigma = 0.03$)	0.66 ($\sigma = 0.04$)	39.66
FFNN 1 HL	41.97 ($\sigma = 4.61$)	5.08 ($\sigma = 0.31$)	0.71 ($\sigma = 0.03$)	0.66 ($\sigma = 0.03$)	39.72
Random Forest	45.82 ($\sigma = 4.96$)	5.35 ($\sigma = 0.33$)	0.68 ($\sigma = 0.03$)	0.62 ($\sigma = 0.03$)	39.63
AdaBoost	49.90 ($\sigma = 5.69$)	5.60 ($\sigma = 0.28$)	0.65 ($\sigma = 0.03$)	0.54 ($\sigma = 0.03$)	39.20
Linear SVR	64.36 ($\sigma = 4.92$)	6.27 ($\sigma = 0.26$)	0.49 ($\sigma = 0.04$)	0.38 ($\sigma = 0.04$)	39.58
Linear Regression	64.40 ($\sigma = 4.97$)	6.27 ($\sigma = 0.26$)	0.49 ($\sigma = 0.04$)	0.38 ($\sigma = 0.04$)	39.60
KNN	70.63 ($\sigma = 7.54$)	6.62 ($\sigma = 0.37$)	0.53 ($\sigma = 0.06$)	0.53 ($\sigma = 0.06$)	39.59

3.1.2. Genomic breeding values calculated within the herd

Alternatively to the gEBV calculated by the breeders' association, we used as a genotypic feature the additive genetic effect deriving from the SNPs, through the within herd application of the linear mixed model presented in Eq. (2). Running the model, we observed that the heritability for the daily milk production trait on the considered dataset (without observing the test set records) was equal to 0.47 (with variances of the additive genetic effect, permanent environmental effect and residual respectively equal to $33.28\,\text{kg}^2$, $10.47\,\text{kg}^2$ and $27.58\,\text{kg}^2$, as estimated by restricted maximum likelihood (Corbeil and Searle, 1976)). Moreover, the gEBVs ranged from $-16.54\,\text{kg}$ to $12.22\,\text{kg}$, with a mean value close to 0 ($1.81 \times 10^{-4}\,\text{kg}$) and a standard deviation equal to $4.58\,\text{kg}$. The Pearson correlation coefficient between the gEBVs calculated within the herd and from the breeders' association on test set bovines was equal to 0.42, thus underlying the limited accuracy of additive genetic effects calculated in a small size animal population.

Regarding the considered milk production forecasting model, we directly worked with a feed forward neural network with one hidden layer; indeed, such algorithm is chosen as the reference model for integrating the gEBV feature into the predictive task, as described by the numerical results and the discussion presented in Section 3.1.1. Moreover, once again we do not consider the additional environmental features (mean number of milkings and concentrate consumption) in the problem formulation at this stage of the analysis.

As already discussed in Section 3.1.1, in order to analyze the impact of the breeding value in the prediction, we first assessed the forecasting performance of the model considering only the fixed effects: in such a situation, as already reported in Section 3.1.1, we obtained a cross-validation mean squared error equal to $50.76\,\text{kg}^2$ (standard deviation of $5.75\,\text{kg}^2$) and Pearson correlation coefficient of 0.63 (standard deviation of 0.04). Of course, since the analyzed trait is also influenced by genetics, we may expect a higher performance when including also the gEBV in the problem formulation; however, this actually happened only when a random noise was added to the breeding values themselves. In such a scenario, we performed a grid search for finding the best (according to the cross-validation mean squared error) standard deviation of the random Gaussian noise with mean equal to 0, which resulted to be equal to 0.15 (we remind that the breeding values were scaled through Min-Max normalization, as reported in Section 2.3.2). In this way, the prediction performance in cross-validation showed a mean squared error of $48.77\,\text{kg}^2$ (standard deviation of $5.17\,\text{kg}^2$) and a Pearson correlation coefficient of 0.65 (standard deviation of 0.04). The limited impact of the gEBVs calculated within the herd, compared to the ones provided by the breeders' association, is discussed in Section 4, together with the need to add a random noise in this scenario. As a matter of fact, deciding not to include the noise led to even detrimental results, already evident by the fact that the cross-validation loss was always bigger than the train one, since the first training epochs.

Concluding, the addition of the breeding value calculated within the herd in the bovine milk yield forecasting process was still beneficial, but with a lower impact and a needed higher care to be taken during model training.

3.2. Performance comparison: machine learning versus linear mixed model

The results presented in Section 3.1 highlight the importance of the gEBV as a feature in the task of bovine daily milk yield forecasting. In particular, including the additive genetic effect by the breeders' association proved to be more beneficial than using the value calculated within the herd. For this reason, in this section we compare the test set forecasting results obtained with the linear mixed model of Eqs. (1) and (2) with the ones obtained with the top performing feed forward neural network (with one hidden layer) in cross-validation (the one leveraging the gEBV calculated by the breeders' association). As usual, the environmental features considered in this analysis were the parity, the days in milk, the age at calving (in months) and the month of calving.

Regarding the average forecasting performance on the test set, with a mean absolute error equal to $4.83\,\text{kg}$ and a Pearson correlation coefficient of 0.71, the neural network sensibly outperformed the analyzed linear mixed model (mean absolute error of $7.10\,\text{kg}$, Pearson correlation coefficient equal to 0.48). Moreover, the neural network did not only perform better at the average overall level, but also day by day, as shown in Fig. 2: unsurprisingly, a Wilcoxon signed-rank test (Wilcoxon, 1945) accepted, at reasonable p-value thresholds, the alternative hypothesis stating that the distribution of mean absolute errors by the neural network had a lower median than the one by the linear mixed model. Finally, we observe that the performance gap is present even when not considering the genotypic feature in the machine learning formulation: as expected, this fact confirms that the relation between the considered input factors and the daily milk yield is intrinsically non-linear, so it is better captured by a machine learning approach with a non linear model. However, even if the linear mixed model is outperformed by the inclusion of non-linearities, it is still able to capture the average shape and behavior of the lactation curve, as shown in Fig. 3. Besides the average performance, it is also interesting to assess and compare the predictions at the individual level, in order to understand whether the neural network outperforms the linear mixed model also on single lactation curves. In particular, our test set comprised 179 lactations (by 102 bovines), in which we observed that the mean absolute error by the machine learning model was lower than the one by the linear mixed model in 131 milk production curves (73.2% of cases); moreover, in 39 of these lactations (21.8% of the total cases), the mean absolute error by the linear mixed model was at least twice the one by the neural network. These results indeed suggest that the machine learning approach outperforms the linear one also at the individual level.

3.3. Feature importance

As described in Section 2.3.2, we added the mean daily number of milkings (on the last week) and the concentrate consumption during milking in the AMS to the machine learning problem formulation. As

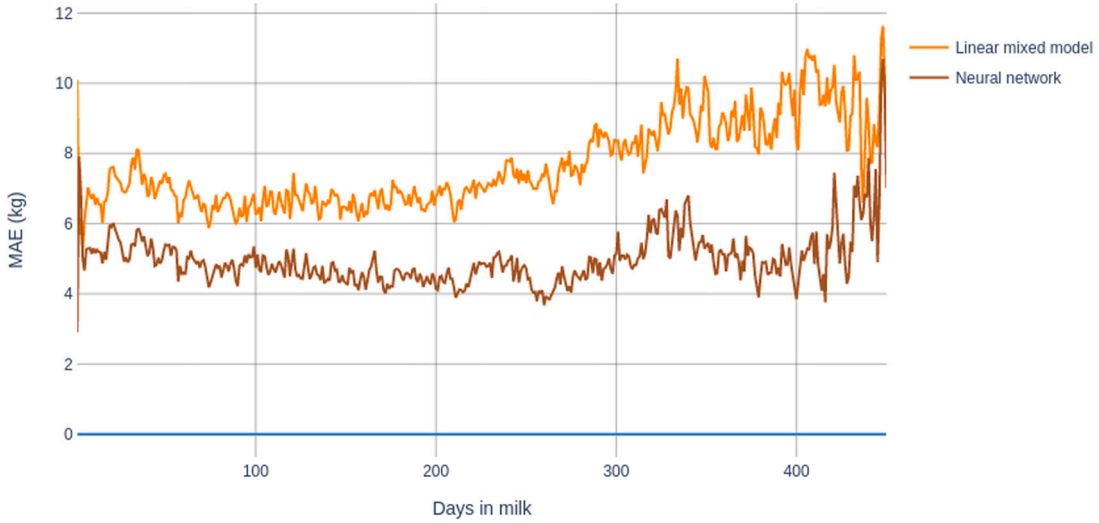

Fig. 2. Mean absolute error (MAE) on test set lactation curves, computed for every value of days in milk: mean absolute error associated with predictions by the linear mixed model (in orange) and by the feed forward neural network with one hidden layer (in brown).

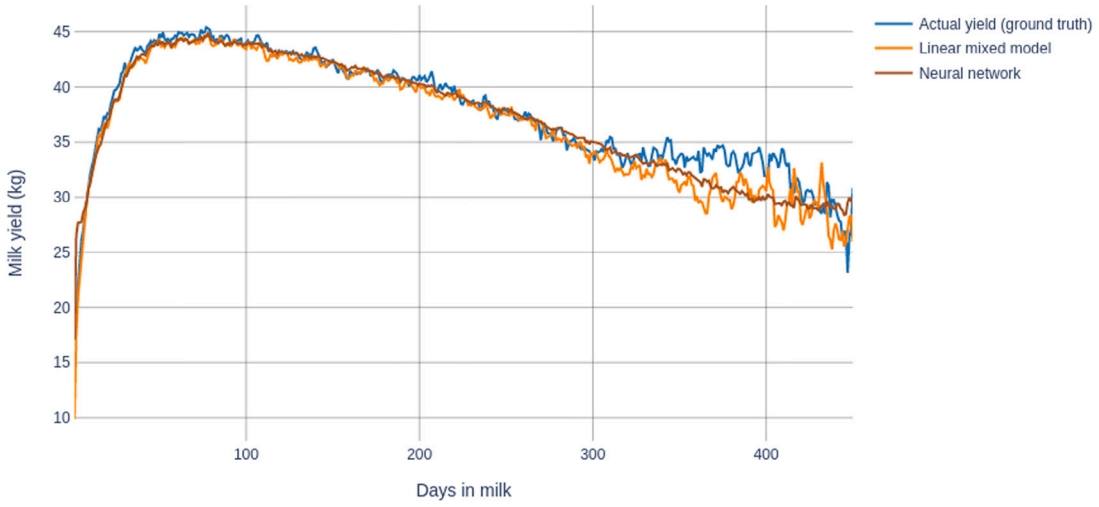

Fig. 3. Predictions of the average lactation curve: average lactation curve on the test set (in blue) and its predictions by the linear mixed model (in orange) and by the feed forward neural network with one hidden layer (in brown).

a matter of fact, the results of the inclusion of such features were very promising; always considering feed forward neural networks with one hidden layer, we observed a substantial performance improvement in the task of daily milk yield forecasting: the cross-validation mean squared error decreased from $41.97\,kg^2$ (standard deviation of $4.61\,kg^2$) to $26.57\,kg^2$ (standard deviation of $4.17\,kg^2$), the mean absolute error improved from $5.08\,kg$ (standard deviation of $0.31\,kg$) to $3.96\,kg$ (standard deviation of $0.30\,kg$), while the Pearson and the concordance correlation coefficients increased from 0.71 (standard deviation of 0.03) to 0.83 (standard deviation of 0.02) and from 0.66 (standard deviation of 0.03) to 0.81 (standard deviation of 0.02), respectively. In addition to these average performance values, we also observed lower test set mean absolute errors both day by day (as shown in Fig. 4) and at the individual level (adding the optional factors was beneficial in 120 lactations over 179, i.e., in more than two-thirds of the analyzed situations). As described in Section 2.4.2, we also evaluated the performance of the linear mixed model with the mean daily number of milkings and the concentrate consumption as additional fixed effects: this model achieved a mean absolute error equal to $5.38\,kg$ and a Pearson correlation coefficient of 0.69 on the test set, showing a sensible improvement with respect to the baseline linear mixed model without the optional factors, but being noticeably outperformed by machine learning (even without the additional covariates). Tables 4 and 5 respectively report average and individual level results on the test set also for primiparous and multiparous bovines, comparing the linear mixed models and the neural networks with and without optional features: in both the cases of first and subsequent lactations, on average and at the individual level, the best model configuration included also the mean number of milkings and the concentrate consumption.

Moreover, as explained in Section 2.4.2, we assessed the forecasting performance of our model with different combinations of input features, in order to understand the impact of each of them on the final predictions: Table 6 reports, in increasing order, the cross-validation performances obtained with the various input configurations.

Another interesting point is constituted by the averaging of the concentrate consumption feature over some predefined time intervals, as explained in Section 2.3.2, also in order to try to remove the potential noise associated with a daily measure. To this end, Table 7 shows the forecasting performance obtained by our neural network model when averaging the concentrate consumption over one week, one month, three months, or the full lactation: we observe that, for all the considered input configurations, the best solution seems to consist in considering concentrate consumption values averaged week by week.

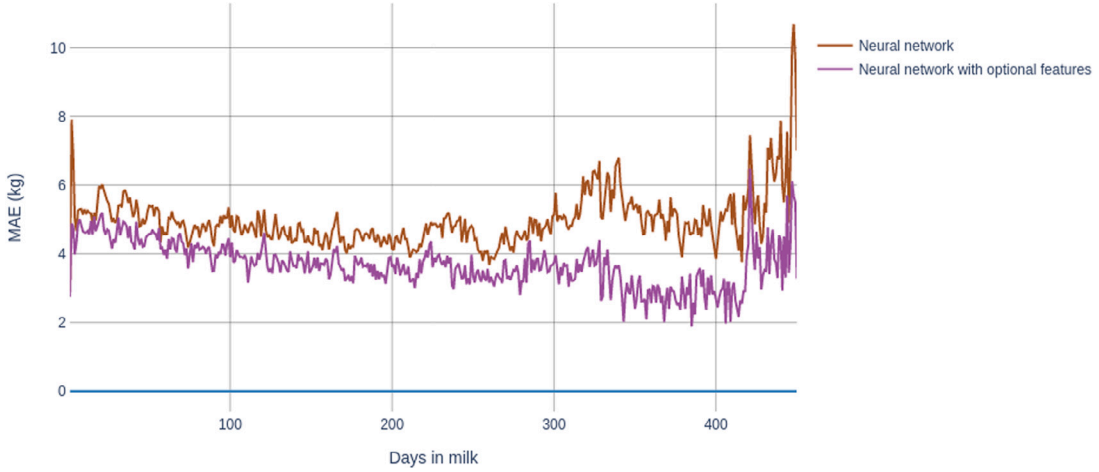

Fig. 4. Further mean absolute error (MAE) on test set lactation curves, computed for every value of days in milk: mean absolute error associated with predictions by the feed forward neural network with one hidden layer (in brown, see also Fig. 2) and by this latter one when also considering the mean daily number of milkings (on the last seven days) and the concentrate consumption during milking in the automatic milking system (in purple). A Wilcoxon signed-rank test confirmed, at any reasonable p-value threshold, that the distribution associated with the purple curve had a significantly lower median than the one of the distribution associated with the brown curve.

Table 4
Comparison of mean absolute error (in kg) performance on the test set by the linear mixed models (LMM) and the neural networks (FFNN) without ((4) for LMM, (4+a) for FFNN) and with ((6) for LMM, (4+milk+conc+a) for FFNN) optional features: results for lactations from primiparous bovines, multiparous bovines and the complete test set. "*4*" represents the mandatory environmental factors (parity, days in milk, age at calving in months, month of calving), "*6*" the mandatory (parity, days in milk, age at calving in months, month of calving) and additional (mean daily number of milkings, concentrate consumption) environmental factors, *a* refers to the milk yield genomic breeding value by the breeders' association, *milk* to the mean daily number of milkings (on the last seven days), *conc* to the daily concentrate consumption during milking in the automatic milking system.

Model	Primiparous	Multiparous	Complete test set
LMM(4)	6.39	7.91	7.10
LMM(6)	4.68	6.18	5.38
FFNN(4+a)	4.42	5.27	4.83
FFNN(4+milk+conc+a)	3.36	4.42	3.87

For this reason, our best input configuration according to the minimization of the mean squared error in cross-validation is the one including the gEBV by the breeders' association, parity, days in milk, age at calving (in months), month of calving, mean daily number of milkings (on the last seven days) and average concentrate consumption (week by week) during milking in the AMS: on the test set, the model achieved a mean squared error equal to 21.62 kg^2, a mean absolute error of 3.49 kg, a Pearson correlation coefficient of 0.85 and a concordance correlation coefficient equal to 0.85.

4. Discussion

The results presented in Section 3.1 show that the inclusion of the gEBVs by the breeders' association in the milk yield forecasting problem formulation was more beneficial than using the ones calculated within the herd. This evidence may be explained by the fact that the additive genetic effects estimated within the herd were less accurate than the ones calculated by the breeders' association, because the considered population was sensibly smaller (502 individuals, compared to the more than 360,000 ones considered by ANAFIBJ); indeed, as shown in Fig. 1, the forecasting model architecture is based on two consecutive predictions, so the error of the first one reflects on the daily milk yield estimation. As a support, we remind that the gEBVs calculated within herd and from the breeders' association were not highly correlated (Pearson correlation equal to 0.42), as reported in Section 3.1.2.

An interesting point about this aspect is the fact that, in the case of calculations entirely within the herd, the gEBVs were beneficial only when a random noise was added. The most likely reason is that, as mentioned, the additive genetic effects estimated within a herd of about 500 animals could not be as accurate as those calculated using the prediction equation obtained on the whole ANAFIBJ population; by adding a random noise, instead, our neural network was forced to only partially rely on the genetic feature, or equivalently to account for its uncertainty. On the contrary, the inclusion of a random noise to the additive genetic effects calculated by the breeders' association had a negligible impact, because such values are more stable.

In any case, the integration of a genomic selection model prior to the final prediction allows embedding animal genomics knowledge into the forecasting task. The main advantage of using the gEBV is to summarize in a single feature all the genotypic information connected to the milk yield trait. Indeed, problem formulations directly based on SNPs (instead of the gEBV) were not able to achieve promising results, at least in our scenario. In particular, feed forward neural networks with environmental features and SNPs identified by Lasso (Tibshirani, 1996) feature selection (in order to avoid the curse of dimensionality problem) showed results similar to the ones by the model only considering the fixed effects; we obtained analogous results when, instead of selecting SNPs using Lasso, we reduced the dimensionality with Principal Component Analysis (Hotelling, 1933; Pearson, 1901), prior to the regression task. These results are clearly in contrast with the ones obtained when considering the gEBV (see Table 6) instead of SNPs. As a matter of fact, the problem of directly working with SNPs is related to their dimensionality and the small impact that each of them has in influencing the studied quantitative complex trait.

For this reason, the proposed solution is composed of two steps: the genomic prediction one, which is an established method to calculate gEBV from marker data (Meuwissen et al., 2001), followed by the milk yield forecasting one, which is able to capture the non-linearities of the problem. To the best of our knowledge, the integration of genotypic information in the machine learning-based prediction of the bovine milk production trait, as well as the combination of genomic prediction and phenomic forecasting in the task, is a novelty in the literature. Only Zhang et al. (2022) tried to couple environmental features with inheritance-related features (yield data from relatives), but without explicitly integrating genomic information; in particular, the latter study achieved a test set mean absolute error equal to 5.0 kg, thus being outperformed by our model (mean absolute error of 4.83 kg on the test set, even without considering the additional environmental features of number of milkings and concentrate consumption).

Regarding the additional environmental features, instead, Section 3.3 shows that they have a considerable impact in the predictions: indeed, adding information about the number of milkings or

Table 5
Comparison of individual level Mean Absolute Error (MAE) performance on the test set by the linear mixed models (LMM) and the neural networks (FFNN) without ((4) for LMM, (4+a) for FFNN) and with ((6) for LMM, (4+milk+conc+a) for FFNN) optional features: results for lactations from primiparous bovines (89 total curves), multiparous bovines (90 total curves) and the complete test set (179 total curves). "*4*" represents the mandatory environmental factors (parity, days in milk, age at calving in months, month of calving), "*6*" the mandatory (parity, days in milk, age at calving in months, month of calving) and additional (mean daily number of milkings, concentrate consumption) environmental factors, *a* refers to the milk yield genomic breeding value by the breeders' association, *milk* to the mean daily number of milkings (on the last seven days), *conc* to the daily concentrate consumption during milking in the automatic milking system.

MAE configuration	Primiparous	Multiparous	Complete test set
LMM(4)<LMM(6)<FFNN(4+a)<FFNN(4+milk+conc+a)	5	0	5
LMM(4)<LMM(6)<FFNN(4+milk+conc+a)<FFNN(4+a)	2	1	3
LMM(4)<FFNN(4+a)<LMM(6)<FFNN(4+milk+conc+a)	2	0	2
LMM(4)<FFNN(4+a)<FFNN(4+milk+conc+a)<LMM(6)	1	2	3
LMM(4)<FFNN(4+milk+conc+a)<LMM(6)<FFNN(4+a)	0	2	2
LMM(4)<FFNN(4+milk+conc+a)<FFNN(4+a)<LMM(6)	2	0	2
LMM(6)<LMM(4)<FFNN(4+a)<FFNN(4+milk+conc+a)	1	2	3
LMM(6)<LMM(4)<FFNN(4+milk+conc+a)<FFNN(4+a)	5	3	8
LMM(6)<FFNN(4+a)<LMM(4)<FFNN(4+milk+conc+a)	3	0	3
LMM(6)<FFNN(4+a)<FFNN(4+milk+conc+a)<LMM(4)	2	1	3
LMM(6)<FFNN(4+milk+conc+a)<LMM(4)<FFNN(4+a)	5	5	10
LMM(6)<FFNN(4+milk+conc+a)<FFNN(4+a)<LMM(4)	4	5	9
FFNN(4+a)<LMM(4)<LMM(6)<FFNN(4+milk+conc+a)	3	1	4
FFNN(4+a)<LMM(4)<FFNN(4+milk+conc+a)<LMM(6)	2	1	3
FFNN(4+a)<LMM(6)<LMM(4)<FFNN(4+milk+conc+a)	0	4	4
FFNN(4+a)<LMM(6)<FFNN(4+milk+conc+a)<LMM(4)	1	2	3
FFNN(4+a)<FFNN(4+milk+conc+a)<LMM(4)<LMM(6)	3	6	9
FFNN(4+a)<FFNN(4+milk+conc+a)<LMM(6)<LMM(4)	8	9	17
FFNN(4+milk+conc+a)<LMM(4)<LMM(6)<FFNN(4+a)	0	2	2
FFNN(4+milk+conc+a)<LMM(4)<FFNN(4+a)<LMM(6)	1	0	1
FFNN(4+milk+conc+a)<LMM(6)<LMM(4)<FFNN(4+a)	5	2	7
FFNN(4+milk+conc+a)<LMM(6)<FFNN(4+a)<LMM(4)	9	9	18
FFNN(4+milk+conc+a)<FFNN(4+a)<LMM(4)<LMM(6)	2	7	9
FFNN(4+milk+conc+a)<FFNN(4+a)<LMM(6)<LMM(4)	23	26	49

Table 6
Features impact in cross-validation: predictions performance (Mean Squared Error - MSE, Mean Absolute Error - MAE, Pearson correlation coefficient - r, Concordance Correlation Coefficient - CCC) by feed forward neural networks with one hidden layer. Among the features, "*4*" represents the mandatory environmental factors (parity, days in milk, age at calving in months, month of calving), *a* refers to the milk yield genomic breeding value by the breeders' association, *milk* to the mean daily number of milkings (on the last seven days), *conc* to the daily concentrate consumption during milking in the automatic milking system. Instead, σ indicates the standard deviation of the performance measure across the 10 cross-validation folds.

Features	MSE [kg^2]	MAE [kg]	r	CCC
4	50.76 ($\sigma = 5.75$)	5.63 ($\sigma = 0.32$)	0.63 ($\sigma = 0.04$)	0.57 ($\sigma = 0.04$)
4+a	41.97 ($\sigma = 4.61$)	5.08 ($\sigma = 0.31$)	0.71 ($\sigma = 0.03$)	0.66 ($\sigma = 0.03$)
4+milk	37.44 ($\sigma = 4.82$)	4.81 ($\sigma = 0.31$)	0.75 ($\sigma = 0.03$)	0.71 ($\sigma = 0.04$)
4+conc	34.66 ($\sigma = 3.76$)	4.54 ($\sigma = 0.23$)	0.77 ($\sigma = 0.02$)	0.74 ($\sigma = 0.02$)
4+milk+conc	31.79 ($\sigma = 3.99$)	4.36 ($\sigma = 0.27$)	0.79 ($\sigma = 0.03$)	0.76 ($\sigma = 0.03$)
4+milk+a	30.23 ($\sigma = 5.00$)	4.28 ($\sigma = 0.35$)	0.80 ($\sigma = 0.03$)	0.78 ($\sigma = 0.03$)
4+conc+a	29.78 ($\sigma = 3.52$)	4.21 ($\sigma = 0.24$)	0.80 ($\sigma = 0.02$)	0.78 ($\sigma = 0.02$)
4+milk+conc+a	26.57 ($\sigma = 4.17$)	3.96 ($\sigma = 0.30$)	0.83 ($\sigma = 0.02$)	0.81 ($\sigma = 0.02$)

the concentrate consumption was even more beneficial than simply considering the genetic effects, as reported by Table 6. As a matter of fact, also the results by Nguyen et al. (2020) underlined the importance of diet parameters and milkings per day in the Holstein milk yield predictions. The impact of the additional environmental variables is related to the fact that they are able to monitor the actual individual conditions of the bovines, besides their genetic predispositions for the analyzed trait. Moreover, we also observed that, in the case of the concentrate consumption, averaging the value week by week led to the best results: the considered amount of time is likely to allow both reducing the noise connected to the high variability of the daily measures and capturing temporary conditions about an animal; in addition, a weekly measure also has the advantage of being more easily controllable by breeders, who are likely to periodically take feed intake decisions.

Analogously to the solution proposed by Liseune et al. (2021), one of the advantages of our model is the possibility to forecast the daily milk production even before the beginning of the lactation: in particular, our predictions may leverage only the parity, the days in milk, the age at calving (in months), the month of calving and the gEBV. Moreover, both the number of milkings and the concentrate consumption are features partially controllable by breeders; thus, including them in the problem formulation still allows predicting in advance how they may influence the daily milk production (with the forecasting exactitude depending on how much the breeder is actually able to control them). Of course, controlling (even partially) the number of milkings is not possible in all the herd management configurations: depending on the milking system, the farmer may be able to plan the number of milkings or to limit its maximum; moreover, this number varies by animal and cannot be known with certainty when the bovine

Table 7
Impact of average values of daily concentrate consumption on the cross-validation mean squared error: predictions by feed forward neural networks with one hidden layer. Among the features, "*4*" represents the mandatory environmental factors (parity, days in milk, age at calving in months, month of calving), *a* refers to the milk yield genomic breeding value by the breeders' association, *milk* to the mean daily number of milkings (on the last seven days), *conc* to the concentrate consumption during milking in the automatic milking system. Moreover, every row denotes a different setting for the concentrate consumption: either considered day by day, or averaged on one week, one month, three months, or the full lactation. The values reported inside the cells are the cross-validation mean squared errors (in kg^2) for the different configurations; σ, instead, indicates the standard deviation (in kg^2) of the performance measure across the 10 cross-validation folds.

Concentrate on	4+conc	4+milk+conc	4+milk+conc+a
1 day	34.66 ($\sigma = 3.76$)	31.79 ($\sigma = 3.99$)	26.57 ($\sigma = 4.17$)
1 week	26.45 ($\sigma = 3.52$)	25.38 ($\sigma = 3.17$)	22.55 ($\sigma = 3.31$)
1 month	27.08 ($\sigma = 3.45$)	25.91 ($\sigma = 3.46$)	23.15 ($\sigma = 3.34$)
3 months	28.45 ($\sigma = 3.57$)	27.49 ($\sigma = 3.73$)	24.41 ($\sigma = 3.62$)
Full lactation	30.68 ($\sigma = 3.17$)	28.15 ($\sigma = 3.58$)	25.63 ($\sigma = 3.87$)

is only a calf; however, our framework is flexible and works even if the number of milkings is missing. Furthermore, while the model by Liseune et al. (2021) was only able to forecast lactation curves of non-primiparous bovines, our solution also allows predicting daily milk yield records even for the first lactation cycle (before or within the latter). The test set results by our model, even considering only the four mandatory environmental features and the gEBV, quite sensibly outperform the ones by Liseune et al. (2021) both in terms of root mean squared error (6.25 kg, lower than 7.38 kg by Liseune et al., 2021) and mean absolute error (4.83 kg, lower than 5.58 kg by Liseune et al., 2021); instead, the test set Pearson correlation coefficient achieved by Liseune et al. (2021) (0.75) is slightly better than ours (0.71).

Summarizing, we believe that the applicability of the model architecture we propose is concrete, also considering the low genotyping costs and the fact that no additional investment for phenotypic and environmental data collection is required in farms equipped with AMSs. Indeed, our solution may help breeders in their management decisions, through machine learning-based milk yield forecasting entirely relying on data usually available within the herd (i.e., gEBVs and AMS data). Thanks to the presence of controllable features (at least partially), farmers can use the model to predict individual lactation curves under different configurations (e.g., different months of birth, ages at calving, weekly concentrate values, ...), and in turn ground their decisions on the one that best fits the farm's needs (indeed, also according to expected milk yield); the approach works even in the case of primiparous bovines, because the model inputs do not include individual historical information from previous lactations, as explained in Section 2.3.2. Moreover, our solution is clearly more flexible than relying only on some milk curves based on genetic and environmental features: first, a machine learning approach allows working in a larger space of input configurations; additionally, the model architecture proposed in Fig. 1 may be effortlessly extended to include other environmental covariates, if available and useful for the farmer. Concluding, the availability of milk yield predictions to breeders may be an efficient decision support to help them in better planning herd management and reproduction, as well as in monitoring bovines, thus potentially contributing to better sustainability and advancements in the dairy sector.

Limitations of our approach include the lack of time series modeling considering previous records (in an autoregressive way or even from past lactations), which would enhance the predictive power of the solution. The main limitation of this work is connected to the fact that data came from a single herd and 26 consecutive months; our model was trained and tested on different animals but of the same herd; so, it is likely to perform worse if applied to other farms. However, the approach and architecture proposed in Fig. 1 are general, thus they may be easily replicated in other single herds, or in more generic situations with data jointly coming from multiple farms. Lastly, we did not compare our model architecture with other models not part of the pipeline in Fig. 1, but this is beyond the scope of this work.

Finally, we remind that our best model, according to the cross-validation results presented in Tables 6 and 7, is the feed forward neural network considering the gEBV by the breeders' association and all the environmental features (both mandatory and optional ones), with concentrate consumption weekly average values; its performance on the test set is reported at the end of Section 3.3.

5. Conclusions

This work, conducted on a herd of 502 Holstein cows situated in northern Italy, presents a novel solution to combine environmental and genotypic information for the prediction of the bovine daily milk yield, using machine learning techniques. In particular, the main objective of our study has been met by defining a model architecture that includes a genomic prediction model to calculate additive genetic effects from SNP data, followed by an artificial neural network with one hidden layer considering as features the gEBV from the initial step, the parity, the days in milk, the age at calving (in months) and the month of calving. In this configuration, the inclusion of the genotypic feature under the form of additive genetic effect proved to be especially impactful when the latter one was calculated starting from a consistently large training population; indeed, our best results were obtained when considering gEBV by the breeders' association. Moreover, we observed the importance of additional environmental features that can be included in the problem formulation, besides that of the additive genetic effect: in this case, the integration of the weekly average values of number of milkings and of the concentrate consumption allowed for a better performance in the forecasting task.

Funding

This work was supported by EAFRD Rural Development Program 2014–2020, Management Authority Regione Lombardia - OP. 16.1.01 Project ID n. 201801062430 - "Operational Group EIP AGRI" https://ec.europa.eu/eip/agriculture/en/eip-agri-projects/projects/operational-groups.

CRediT authorship contribution statement

Andrea Mario Vergani: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Alessandro Bagnato:** Conceptualization, Resources, Supervision, Writing – review & editing. **Marco Masseroli:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Adamczyk, K., Grzesiak, W., Zaborski, D., 2021. The use of artificial neural networks and a general discriminant analysis for predicting culling reasons in holstein-friesian cows based on first-lactation performance records. Animals 11 (3), http://dx.doi.org/10.3390/ani11030721.

Antolín, R., Nettelblad, C., Gorjanc, G., Money, D., Hickey, J.M., 2017. A hybrid method for the imputation of genomic data in livestock populations. Genet. Select. Evol. 49 (1), http://dx.doi.org/10.1186/s12711-017-0300-y.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Corbeil, R.R., Searle, S.R., 1976. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics 18 (1), 31–38. http://dx.doi.org/10.2307/1267913.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 13 (1), 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. System Sci. 55 (1), 119–139. http://dx.doi.org/10.1006/jcss.1997.1504.

Hickey, J.M., Kinghorn, B.P., Tier, B., van der Werf, J.H., Cleveland, M.A., 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Select. Evol. 44 (9), 11. http://dx.doi.org/10.1186/1297-9686-44-9.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educat. Psychol. 24, 498–520. http://dx.doi.org/10.1037/h0071325.

Hudson, C., Cook, J.G., Laven, R., 2019. 25 - Veterinary control of herd fertility in intensively managed dairy herds. In: Noakes, D.E., Parkinson, T.J., England, G.C. (Eds.), Veterinary Reproduction and Obstetrics, tenth ed. W.B. Saunders, St. Louis (MO), pp. 467–484. http://dx.doi.org/10.1016/B978-0-7020-7233-8.00025-2.

Hutjens, M., 2002. Dairy farm management systems | dry lot – Dairy cow breeds. In: Roginski, H. (Ed.), Encyclopedia of Dairy Sciences. Elsevier, Oxford, pp. 693–699. http://dx.doi.org/10.1016/B0-12-227235-8/00125-5.

Hutjens, M., 2016. Dry-lot dairy cow breeds. In: McSweeney, P.L., McNamara, J.P. (Eds.), Encyclopedia of Dairy Sciences, third ed. Academic Press, Oxford, pp. 234–241. http://dx.doi.org/10.1016/B978-0-08-100596-5.00706-X.

Liang, M., Miao, J., Wang, X., Chang, T., An, B., Duan, X., Xu, L., Gao, X., Zhang, L., Li, J., Gao, H., 2021. Application of ensemble learning to genomic selection in chinese simmental beef cattle. J. Anim. Breed. Genet. 138 (3), 291–299. http://dx.doi.org/10.1111/jbg.12514.

Liseune, A., Salamone, M., Van den Poel, D., van Ranst, B., Hostens, M., 2021. Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning. Comput. Electron. Agric. 180, 105904. http://dx.doi.org/10.1016/j.compag.2020.105904.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157 (4), 1819–1829. http://dx.doi.org/10.1093/genetics/157.4.1819.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D., 2002. BLUPF90 and related programs (BGF90), in: CD-ROM communication, proceedings of the 7th world congress on genetics applied to livestock production. Montpellier 2002, 7–28.

Murphy, M., Zhang, F., Upton, J., Shine, P., Shalloo, L., 2018. A Review of Milk Production Forecasting Models. pp. 14–61.

Nguyen, Q.-T., Fouchereau, R., Frénod, E., Gerard, C., Sincholle, V., 2020. Comparison of forecast models of production of dairy cows combining animal and diet parameters. Comput. Electron. Agric. 170, http://dx.doi.org/10.1016/j.compag.2020.105258.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11), 559–572. http://dx.doi.org/10.1080/14786440109462720.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65 (6), 386–408. http://dx.doi.org/10.1037/h0042519.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222. http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

VanRaden, P., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91 (11), 4414–4423. http://dx.doi.org/10.3168/jds.2007-0980.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1 (6), 80–83. http://dx.doi.org/10.2307/3001968.

Zhang, F., Weigel, K.A., Cabrera, V.E., 2022. Predicting daily milk yield for primiparous cows using data of within-herd relatives to capture genotype-by-environment interactions. J. Dairy Sci. 105 (8), 6739–6748. http://dx.doi.org/10.3168/jds.2021-21559, URL https://www.sciencedirect.com/science/article/pii/S0022030222003307.

Zhang, W., Yang, K., Yu, N., Cheng, T., Liu, J., 2020. Daily milk yield prediction of dairy cows based on the GA-LSTM algorithm. In: 2020 15th IEEE International Conference on Signal Processing. ICSP, 1, pp. 664–668. http://dx.doi.org/10.1109/ICSP48669.2020.9320926.