Journal of Agricultural Science and Technology (JAST), 29(4)
In Press, Pre-Proof Version

# Weighted Milk Yield Feature for Daily Milk Yield Prediction in Dairy Cows

**Abdullah Ammar Karcioglu[1], Muhammed Berat Ozgungordu[2], and Mete Yağanoğlu[3]***

**Abstract**

Milk production lies at the intersection of food science, agricultural economics and animal science. Work across these fields concerns milk quality, yield, sustainability and animal welfare. This study aimed to estimate the daily milk yield of cows on dairy farms in the Eastern Anatolia region using routinely collected production and reproductive variables. The final dataset comprised 8,128 individual cows and included breed, lactation number, age, fertility status, days in milk, days pregnant, and insemination count. Seven regression algorithms were applied: Decision Tree Regression (DTR), Linear Regression (LR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR), Lasso Regression, Ridge Regression (RR), and Support Vector Regression (SVR). Age Category, Pregnancy Status, and Weighted Milk Yield (WMY) were evaluated as engineered features. On the primary test split, GBR produced the lowest errors, with an MAE of 3.985 L/day, an RMSE of 4.992 L/day, a MAPE of 22.741%, and an $R^2$ of 0.438. Across 100 additional predefined 80:20 partitions, GBR achieved a mean MAE of 4.100 ± 0.072 L/day and ranked for MAE, RMSE, and $R^2$ in all 100 partitions. Although WMY ranked highly in the model-based feature-importance analysis, its incremental contribution was very small: MAE changed from 3.989 L/day without WMY to 3.985 L/day with WMY, and paired-bootstrap confidence intervals for the performance differences included zero. These findings indicate that GBR provided the best predictive performance among the evaluated algorithms, while WMY should be interpreted as a study-specific engineered predictor rather than as a feature with a demonstrated independent improvement in predictive performance.

**Keywords***:* Machine Learning, Milk Yield Depending on Breed, Milk Yield Prediction, Pregnancy and Milk Yield in Cows.

## 1. Introduction

Milk production has an important place both economically and nutritionally throughout the world, and accurately analyzing the factors affecting milk yield is critical for farm efficiency

[1] Department of Software Engineering, Faculty of Engineering, Ataturk University, Erzurum, Türkiye.
[2] Department of Artificial Intelligence and Data Engineering, Faculty of Engineering, Ataturk University, Erzurum, Türkiye.
[3] Department of Computer Engineering, Faculty of Engineering, Ataturk University, Erzurum, Türkiye.
*Corresponding author; e-mail: yaganoglu@atauni.edu.tr

and financial gain (Shahinfar et al., 2014). Lactation curves are generally used for this purpose: empirical mathematical models describe daily milk production as a function of time since calving (Macciotta et al., 2005; Wood, 1967), and cows have an average milking period of 305 days and a resting period of 60 days that affects milk yield (Wilmink, 1987). However, the lactation curve alone is not sufficient. Reproductive performance is also associated with milk production and the productive lifespan of dairy cows (De Vries and Marcondes, 2020; Pinedo et al., 2020). These factors may jointly contribute to variation in milk yield and may also contain overlapping information.

In this study, a machine learning-based approach is presented to estimate the milk yield (MY) of Simmental, Holstein, Montofon and Jersey cows raised on milk farms in the Eastern Anatolia region. Breed, age, days in milk, lactation number, insemination count, fertility status and pregnancy-related variables were used together within a common machine-learning framework, and additional features were derived from them, including an empirical composite feature termed Weighted Milk Yield (WMY). The contributions of this study are listed below:

- The original dataset obtained from milk farms in the Eastern Anatolia region forms the basis of this study and is an important resource for understanding milk yield in the region.
- The study jointly evaluates breed, age, days in milk, lactation number, insemination count, fertility status, and pregnancy-related variables within a common machine-learning framework.
- Age Category and Pregnancy Status were derived from existing variables and evaluated together with the original predictors.
- WMY was developed as an empirical composite feature from days in milk, age, lactation number and Holstein breed status, and its contribution was evaluated by feature-importance and ablation analyses.

Among the seven regression algorithms evaluated, GBR produced the lowest prediction errors on the test set, and the contribution of the engineered features was examined through ablation analysis rather than feature importance alone.

## 2. Related Works

Data science applications are widely used in milk production facilities, and decision-support systems based on regression and other machine-learning algorithms are increasingly developed. However, previous studies have generally examined subsets of animal, reproductive, genomic, behavioral, environmental or historical production variables rather than combining routinely available farm variables within a common modelling framework.

Several studies have applied machine-learning algorithms to milk production data. Ji et al. (2022) predicted the following month's daily milk yield, milk composition and milking frequency on a robotic dairy farm. Suseendran and Duraisamy (2021) reported that ANN and SVR achieved the highest accuracy, Zhang et al. (2020) showed that GA-LSTM was more accurate than the traditional LSTM algorithm, and Tarjan et al. (2023) used a DNN with phenotypic and pedigree data for Holstein-Friesian cattle in Serbia. Time-series approaches have been applied at the national scale for milk (Mishra et al., 2020) and buffalo milk (Subbanna et al., 2021) production, and Amin (2003) found the lactation stages to be significant for Egyptian buffaloes. Salamone et al. (2022) estimated first-test-day milk yield using Random Forest models, and random regression models have been used to estimate genetic parameters for test-day milk yield (Santos et al., 2021; Uribe and Lembeye, 2020). Vergani et al. (2024) combined genomic prediction with a feedforward neural network, and Dallago et al. (2019) found an artificial neural network to outperform linear regression and random forest.

Recent animal-science studies provide further evidence for the biological relevance of days in milk, parity, age-related production stage and breed. Milk yield has been reported to be lower in first-parity Holstein cows, to increase until approximately the third parity and then decline, with peak production between 61 and 90 days in milk (Timkovičová Lacková et al., 2022), and breed, parity and lactation stage significantly affected daily production under heat-stress conditions (Mićić et al., 2022). Comparable relationships were identified in 659 dairy cows (Vrhel et al., 2024), and early-lactation daily yields of 19.6, 24.1 and 26.3 kg were reported for first-, second- and third-parity cows (Walsh et al., 2024).

Recent machine-learning studies have emphasized the contribution of production history, sensor variables, genomic information and appropriate validation strategies. Hooker et al. (2025) found that previous daily milk-yield records produced the largest predictive improvement in 1,312 commercial dairy cows, Ferrari et al. (2026) evaluated a feedforward neural network across generations using 67,010 daily observations, and Darabighane and Atzori (2026) found gradient boosting to perform best under walk-forward validation. These results are not directly comparable because prediction horizons, input variables, populations and validation strategies differ.

In this study, parameters collected from commercial dairy farms operating under routine production conditions were used together, and milk yield in the Eastern Anatolia region was analyzed with a unique dataset. Seven regression models were compared using MAE, RMSE, MAPE and $R^2$ under a common training and test protocol, and WMY was evaluated through feature-importance analysis, ablation and paired bootstrap resampling. Table 1 summarizes selected recent studies; because they use different targets, variables, units and validation protocols, the reported values should be interpreted within their original settings rather than as a ranking.

**Table 1.** Literature summary for milk yield estimation.

| Study | Method | Year | Metric | Result |
|---|---|---|---|---|
| Ji et al. (2022) | Time Series | 2022 | Accuracy | 80 |
| Tarjan et al. (2023) | DNN | 2023 | Accuracy | 76.1 |
| Salamone et al. (2022) | RFR | 2022 | RMSE | 6.08 |
| Vergani et al. (2024) | Linear Mixed Model | 2024 | MAE | 3.96 |
| Dallago et al. (2019) | Neural Network Model | 2019 | MAE | Less than 4 kg |
| Hooker et al. (2025) | Ridge Regression, Gradient Boosting, and Random Forest | 2025 | MAE | 2.69 kg |
| Ferrari et al. (2026) | Feedforward Neural Network | 2026 | RMSE | 5.98 kg |
| Darabighane and Atzori (2026) | Gradient Boosting, Random Forest, and Linear Mixed Model | 2026 | MAE<br>RMSE | 1.39 kg<br>2.07 kg |

## 3. Materials and Methods

### 3.1 Dataset

In this study, daily milk yield was estimated through data preprocessing, feature engineering, feature scaling, model training and performance evaluation. Dry-cow records, records with missing daily milk-yield values, records outside the farm-specific 8–40 L/day data-quality eligibility range and records with invalid negative insemination counts were removed. No class-balancing procedure was applied because the study addressed a regression problem. MinMax and Standard scaling were both evaluated and MinMaxScaler was retained, fitted only to the training data to prevent information leakage.

The dataset was collected from commercial dairy farms in the Eastern Anatolia region under routine production conditions and included Holstein, Jersey, Simmental and Montofon cows. The 8–40 L/day range was defined as a farm-specific data-quality eligibility criterion in consultation with the data providers. According to the data providers from the participating farms, values below 8 L/day for cows recorded as actively milked were considered potentially unreliable, particularly because decimal-separator or misplaced-decimal errors could occur during data entry. Daily yields above 40 L/day were reported to be uncommon under the production conditions of these farms and were therefore also treated as potentially unreliable records. These limits were used as operational data-quality criteria and should not be interpreted as universal biological cut-off values. The findings should therefore be interpreted as population-specific and may not generalize to farms operating under different environmental, feeding or management conditions.

The complete raw dataset contained 13,357 records, each representing a different cow; repeated observations from the same cow were not present. All preprocessing and exclusion steps described below were applied starting from this complete raw dataset. Collar identifiers were assigned locally within each farm and were therefore excluded from the predictor set. Because each record represented a separate animal, the same cow could not occur simultaneously in the training and test datasets.

The data-cleaning procedure first excluded 3,038 dry-cow records, reducing the dataset to 10,319 records. A further 317 records with missing daily milk yield were removed. Application of this farm-specific 8–40 L/day data-quality eligibility range removed 1,810 records below 8 L/day and 62 records above 40 L/day. Finally, two records containing negative insemination counts were excluded. The resulting analytical dataset consisted of 8,128 individual cow records.

The analytical dataset contained seven original predictors, namely lactation number, age in months, breed, fertility status, days in milk, days pregnant and insemination count, and daily milk yield in L/day as the regression target. Categorical variables were converted into binary indicator variables.

Table 2 presents five anonymized example records; the row number was not used as a predictor.

**Table 2.** Example records from the final analytical dataset**.**

| Lactation No | Age (months) | Breed | Fertility | Days in milk | Days pregnant | Insemination count | Milk yield |
|---|---|---|---|---|---|---|---|
| 3 | 107.90 | Holstein | Not pregnant | 77 | 0 | 1 | 32.10 |
| 2 | 97.00 | Jersey | Inseminated | 55 | 0 | 1 | 21.60 |
| 4 | 59.60 | Simmental | Inseminated | 55 | 0 | 2 | 18.90 |
| 5 | 94.60 | Holstein | Pregnant | 70 | 74 | 5 | 22.70 |
| 4 | 97.00 | Montofon | Postpartum | 4 | 0 | 7 | 18.75 |

Figure 1 summarizes the analytical workflow. After deterministic data cleaning and feature engineering, the dataset was divided into training and test subsets using an 80:20 ratio. A fixed random split with seed 297 was used for the primary seven-model comparison, resulting in 6,502 training and 1,626 test records. To assess the robustness of the model comparison to random data partitioning, an additional repeated holdout validation was performed using 100 predefined random 80:20 partitions (seeds 1000–1099). All 100 partitions were evaluated without performance-based selection. For each partition, the seven regression algorithms were refitted on the corresponding training subset and evaluated on the corresponding holdout subset.

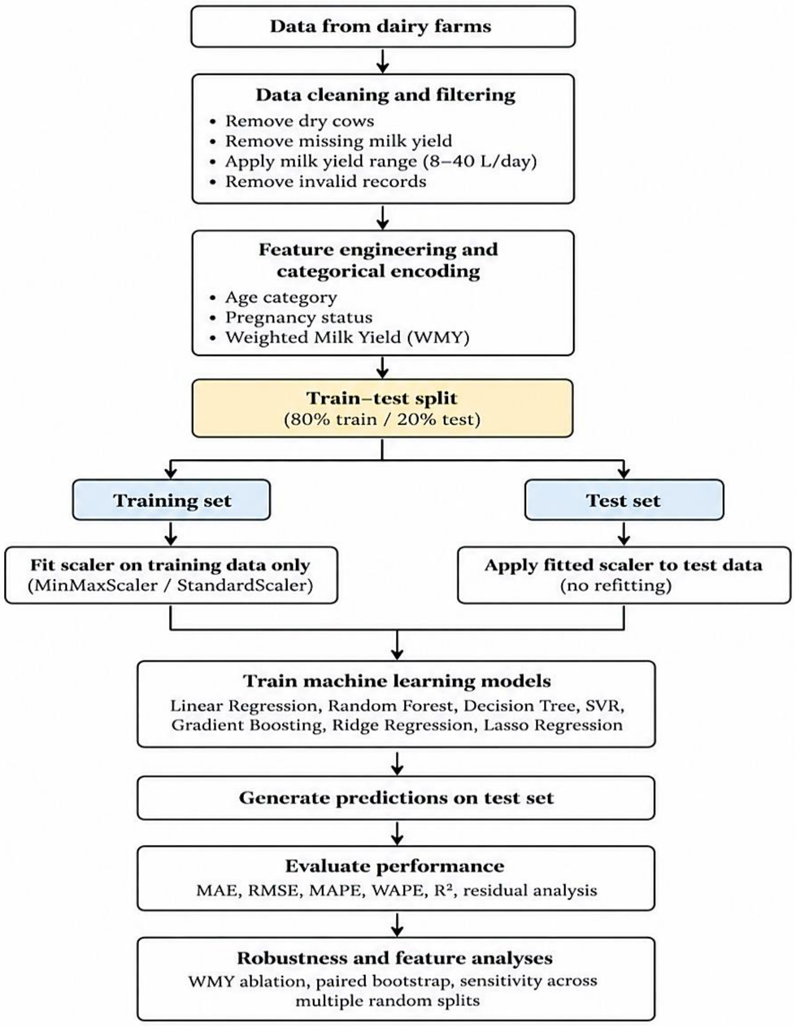


**Fig. 1.** General framework of the study.

### 3.2 Data Pre-Processing

In the pre-processing phase, features in the string data structure were converted into numerical data using encoder methods, one-hot encoding being an example of individually encoding categorical variables into feature vectors (Cerda et al., 2018). Fertility, breed, age-category and pregnancy-status variables were therefore represented by binary indicator variables.

The target variable, MY, was converted from liters to milliliters during model computation, and MAE and RMSE were converted back to liters when reporting the results. Missing insemination counts were encoded as 0 because these records represented cows that had not yet been inseminated, and comma entries in the number of days pregnant were interpreted as 0.

Two features, named "Age Category" and "Pregnancy Status", were derived from existing variables and one-hot encoded. Age Category used three intervals: Young (<36 months), Middle Aged (36-72 months) and Old (>72 months). Pregnancy Status was derived from days pregnant: None (0), Early Pregnancy (1-89), Mid Pregnancy (90-180) and Late Pregnancy (>180). Fertility status was represented by four mutually exclusive categories and breed by four categories. Because both engineered categories are deterministic transformations of existing variables, their incremental contribution was assessed through ablation analysis.

Visualizations were used descriptively and not as evidence of predictor independence; dependence and multicollinearity were evaluated using correlation and variance inflation factor analyses.

The resulting features and their definitions are presented in Table 3.

**Table 3.** Encoded variables used in model development.

| Variable | Explanation |
|---|---|
| Lactation No. | Number of completed lactations |
| Age (Months) | Cow's age in months |
| NDM | Number of days in milk |
| NDP | Number of days pregnant |
| Insemination | Number of inseminations |
| Fertility_Pregnant | Pregnant cows |
| Fertility_Postpartum | Cows in the postpartum category |
| Fertility_Insemination | Cows that have been inseminated |
| Fertility_NotPregnant | Cows classified as not pregnant/returned to cycle |
| Breed_Holstein | Holstein cows |
| Breed_Jersey | Jersey cows |
| Breed_Montofon | Montofon cows |
| Breed_Simmental | Simmental cows |
| Age Category_Young | Cows younger than 36 months |
| Age Category_Middle Aged | Cows aged 36–72 months |
| Age Category_Old | Cows older than 72 months |
| Pregnancy Status_Early Pregnancy | Cows with 1–89 days of pregnancy |
| Pregnancy Status_Mid Pregnancy | Cows with 90–180 days of pregnancy |
| Pregnancy Status_Late Pregnancy | Cows with more than 180 days of pregnancy |
| Pregnancy Status_None | Cows with 0 days of pregnancy |
| MY | Daily milk yield; expressed in mL during model computation and reported as L/day |

After encoding, the dataset contained 8,128 records, 20 numerical predictors and one target variable.

Mean daily yield varied little with age, staying between roughly 19 and 21 L/day across most 12-month intervals (Figure 2). Because the youngest and the oldest intervals were sparsely populated, the apparent differences at either extreme carry little weight.

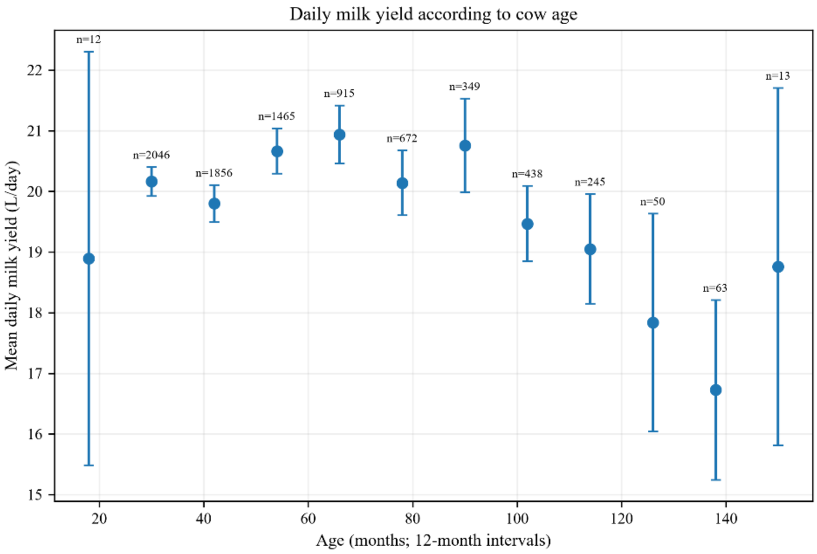


**Fig. 2.** Daily milk yield by 12-month age interval (mean ± 95% CI; n per group).

Breed separated the animals far more clearly (Figure 3). Holstein cows averaged 22.61 L/day and Simmental cows 17.12 L/day, and the overall effect was significant (Welch's ANOVA, $F(3,1271.67)= 421.87$, $p<0.001$). Holm-adjusted pairwise Welch tests placed Holstein ahead of every other breed; Jersey and Montofon were indistinguishable from each other ($p=0.789$) but both outproduced Simmental ($p< 0.001$).

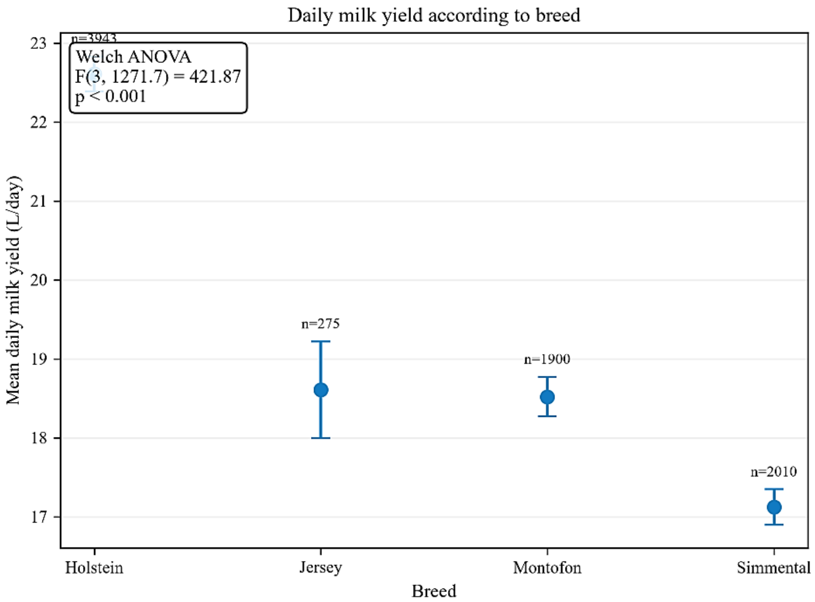


**Fig. 3.** Daily milk yield by breed (mean ± 95% CI; n per group).

Examining age within each breed (Figure 4) shows that the Holstein advantage held across the whole age range rather than being confined to particular ages; the wider intervals at older ages reflect thinner data.

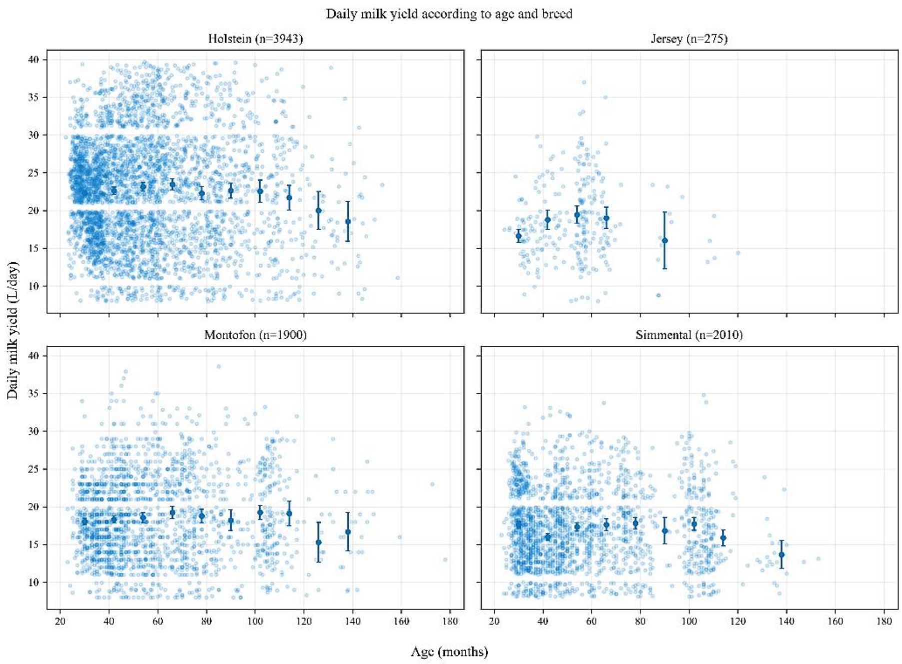


**Fig. 4.** Daily milk yield by age and breed; points are individual cows and overlaid markers are 12-month interval means ± 95% CI.

Grouping age into the three categories used in the models gives the same picture (Figure 5): Holstein cows led in every category, and for Holstein, Jersey and Montofon cows the Middle-Aged group generally outproduced both the Young and the Old groups.

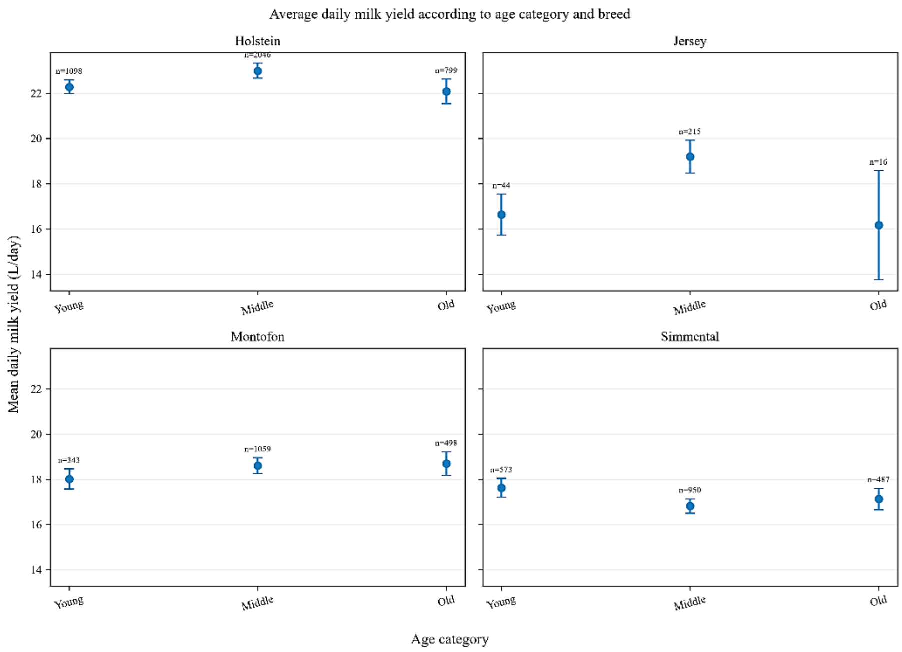


**Fig. 5.** Daily milk yield by age category and breed (mean ± 95% CI; n per group).

Yield followed the expected parity curve (Figure 6), rising from the first lactation to about the fourth and declining thereafter, in line with earlier reports (Ray et al., 1992). The widening intervals in later lactations again follow from the smaller number of animals in those groups.

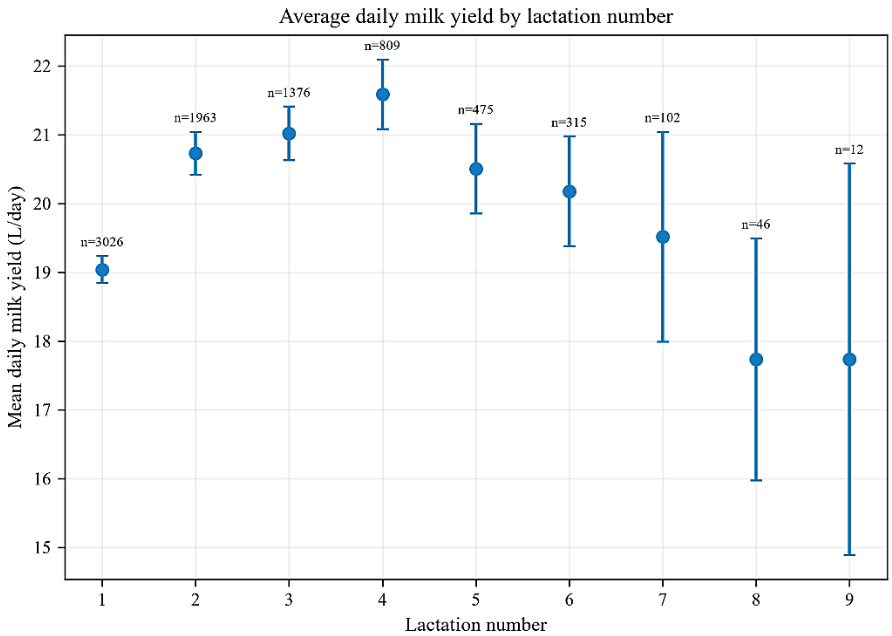


**Fig. 6.** Daily milk yield by lactation number (mean ± 95% CI; n per group).

Figures 7 and 8 compare mean daily milk yield across lactation numbers and breeds for pregnant and non-pregnant cows. Non-pregnant cows generally showed higher mean milk yield within corresponding groups, although the magnitude varied; previous studies have similarly reported a reduction in milk yield with advancing pregnancy (Bar-Anan and Genizi, 1981; Coulon et al., 1995). These patterns do not imply that the predictors were independent, and correlation and multicollinearity analyses indicated substantial overlap.

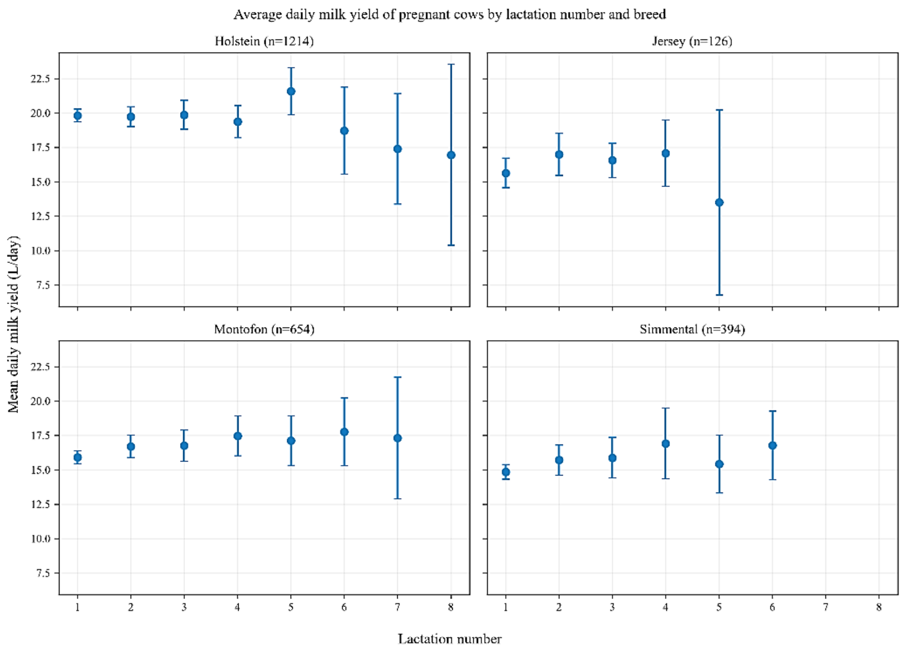


**Fig. 7.** Mean daily milk yield of pregnant cows by lactation number and breed. Error bars: 95% CI.

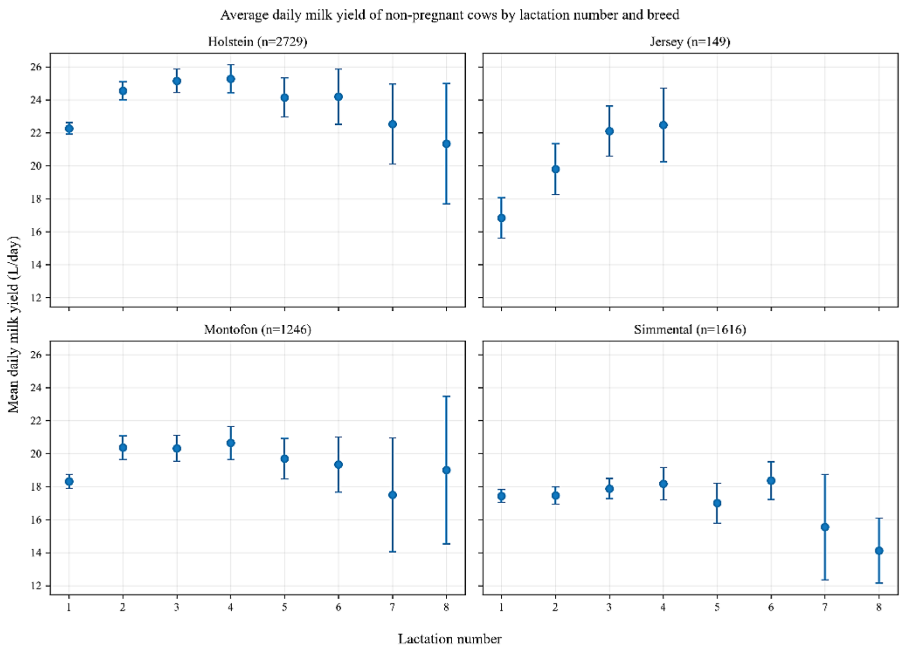


**Fig. 8.** Mean daily milk yield of non-pregnant cows by lactation number and breed. Error bars: 95% CI.

### 3.3 Feature Extraction

Beyond the age and pregnancy categories described in Section 3.2, Weighted Milk Yield (WMY) was developed as an empirical composite feature from the number of days in milk (NDM), age in months, lactation number (LacNo) and Holstein breed status, because lactation stage, parity, age-related production stage and breed are associated with milk yield (Mićić et al., 2022; Timkovičová Lacková et al., 2022; Vrhel et al., 2024; Walsh et al., 2024). WMY is therefore a study-specific construct rather than an established biological index, and was calculated using Equation (1):

$$\text{Weighted Milk Yield} = \frac{\text{Number of Days in Milk x Age(Month)}}{\text{Lac\_No}}\, x\ Breed_{HOLS} \tag{1}$$

Where NDM is the number of days in milk, Age is the cow's age in months, LacNo is the lactation number and HOLS is a binary indicator equal to 1 for Holstein cows and 0 otherwise, so WMY is zero for non-Holstein cows. The operations were defined by the authors as part of the feature-engineering design and are not biologically validated coefficients.

Because WMY is deterministically derived from variables that were also retained as predictors, its incremental contribution was evaluated by comparing otherwise identical GBR models with and without WMY.

The model-level importance of the predictors after inclusion of WMY was examined using the impurity-based feature-importance values of the fitted GBR model (Figure 9). WMY appeared among the most influential predictors; however, feature importance is not evidence of an independent improvement, which was therefore evaluated by ablation.

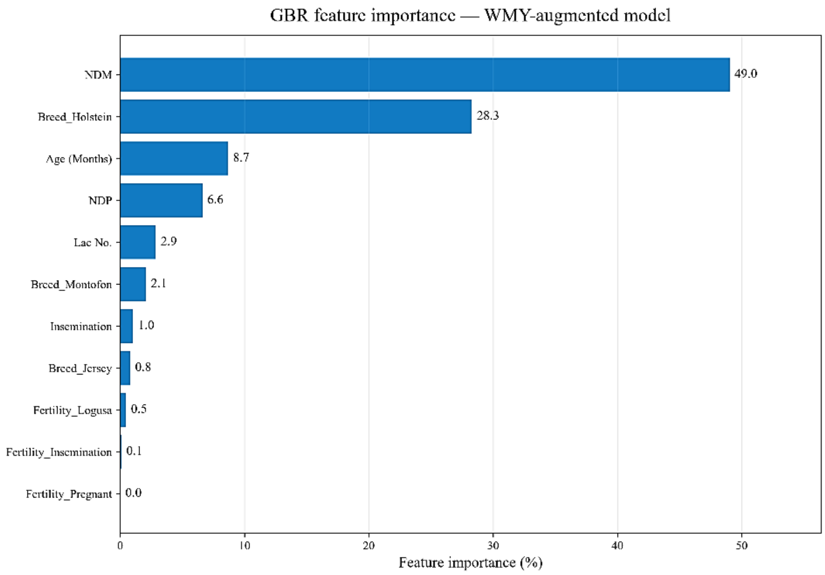


**Fig. 9.** Impurity-based feature-importance ranking of the WMY-augmented GBR model

3.4 Normalization.

Standard scaling transforms the distribution of data by subtracting the mean and dividing by the standard deviation, centering the data around zero with a standard deviation of 1, and allows features with different scales to be processed more effectively (Rashmi and Shantala, 2024).

Min-Max normalization converts data to the range [0-1] by ranking how much each field value exceeds the minimum (Yağanoğlu, 2022). Both methods were evaluated and Min-Max normalization was retained; the scaler was fitted using the training data and applied to the test data using the same transformation.

### 3.5 GBR Configuration and Hyperparameter Sensitivity

Hyperparameters are external parameters that are not directly estimated during model fitting and can be configured to affect predictive performance (Alibrahim and Ludwig, 2021). Grid-search procedures provide a systematic approach for evaluating predefined combinations of hyperparameter values (Shekar and Dagnew, 2019). In the present study, Table 4 presents the fixed GBR configuration used in the primary model comparison and WMY analyses. To assess the sensitivity of GBR performance to alternative hyperparameter settings without using the test data, GridSearchCV was applied only to the 6,502-record training subset using five-fold cross-validation. The search evaluated 48 combinations: n_estimators = {100, 200}, learning_rate = {0.05, 0.10}, max_depth = {2, 3, 4}, min_samples_split = {2, 4}, and min_samples_leaf = {1, 3}. Mean cross-validated MAE was used as the selection criterion. MinMaxScaler and GBR were implemented within a single scikit-learn Pipeline, so scaling parameters were estimated separately from the training portion of each cross-validation fold. The held-out test subset was not used during scaling, hyperparameter evaluation, or model selection.

**Table 4.** Fixed GBR configuration used in the primary model comparison.

| | |
|---|---|
| Learning_rate | 0.1 |
| Max_depth | 3 |
| Min_samples_leaf | 3 |
| Min_samples_split | 4 |
| N_estimators | 100 |

### 3.6 Performance Evaluation Metrics

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and the coefficient of determination ($R^2$). MAE and RMSE were the primary error measures because they are expressed on the same scale as daily milk yield, RMSE giving greater weight to larger errors; $R^2$ quantifies the explained variability and MAPE is a supplementary relative-error measure.

MAE was calculated as follows (Bozkurt, 2022):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (2)$$

RMSE was calculated as follows (Singla et al., 2022):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

MAPE was calculated as follows (Chen et al., 2018):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (4)$$

The coefficient of determination was calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

Where n is the number of observations, y is the observed daily milk yield, ŷ is the corresponding predicted value and ȳ is the mean observed milk yield. Lower MAE, RMSE and MAPE values indicate smaller prediction errors, whereas higher $R^2$ values indicate greater explained variance.

### 3.7 Machine Learning Algorithms

Support Vector Regression (SVR) carries the support vector formulation over to continuous targets. Kernel functions let it represent both linear and nonlinear relationships, and its loss function ignores errors that fall inside a predefined tolerance band while penalizing those outside it (Karcioğlu and Aydin, 2019).

Decision Tree Regression (DTR) represents the predictor space as a tree of nodes and branches (Pekel, 2020). Candidate split points are evaluated for every independent variable, and the split that minimizes the chosen error criterion is retained as the decision point.

Lasso Regression selects predictors and estimates their coefficients at the same time. A constraint keeps the sum of the absolute coefficients below a specified value (λ), which shrinks the less informative coefficients towards zero and minimizes prediction error (Ranstam and Cook, 2018).

Ridge Regression (RR) is a parameter estimation method for the multiple linear regression problem (McDonald, 2009). A regularization term added to the loss shrinks the size of the coefficients and thereby limits overfitting.

Linear Regression (LR) fits the line, or the hyperplane in the multivariate case, that best describes the relationship between the numerical predictors and the response (Okur and Cetin, 2019).

Random Forest Regression (RFR) combines the predictions of many decision trees rather than relying on a single one (Rodriguez-Galiano et al., 2015). The diversity among the trees suppresses noise in the data and improves the generalizability of the model.

Gradient Boosting Regression (GBR) is an ensemble technique in which weak predictive models, usually decision trees, are combined stepwise into a strong predictive model (Zhang and Haghani, 2015). At each step a new model is trained on the differences between actual and predicted values, and the process continues for the specified number of boosting iterations.

## 4. Experimental Results

The analyses were conducted using Python 3.14.2 and scikit-learn 1.9.0. The dataset was split into 80% training and 20% testing, all preprocessing transformations were fitted using the training data, and seven regression algorithms were compared using MAE, RMSE, MAPE and $R^2$.

Table 5 summarizes the performance of the seven regression algorithms on the primary test split. GBR achieved the lowest prediction errors (MAE 3.985 L/day, RMSE 4.992 L/day, MAPE 22.741%) and the highest $R^2$ (0.438), whereas DTR produced the highest errors (MAE 5.668 L/day, RMSE 7.336 L/day, MAPE 31.236%). Lasso, Ridge and Linear Regression performed very similarly, and RFR produced slightly lower MAE and MAPE values.

Repeated holdout validation across 100 additional predefined 80:20 partitions produced a mean ± SD GBR MAE of 4.100 ± 0.072 L/day, RMSE of 5.173 ± 0.088 L/day, MAPE of 23.519 ± 0.529%, and $R^2$ of 0.399 ± 0.018. GBR ranked first among the seven evaluated algorithms for MAE, RMSE, and $R^2$ in all 100 partitions. The repeated-validation mean MAE was slightly higher than the 3.985 L/day obtained on the primary test split, indicating that absolute performance varied across random partitions, whereas the relative ranking of GBR remained stable. Observed and predicted values are shown in Figures 10 and 11 and the residuals in Figure 12.

**Table 5.** Performance of the regression algorithms on the primary seed-297 test split.

| Model | MAE (L/day) | RMSE (L/day) | MAPE (%) | $R^2$ |
|---|---|---|---|---|
| SVR | 5.393 | 6.669 | 29.491 | -0.003 |
| DTR | 5.668 | 7.336 | 31.236 | -0.214 |
| Lasso | 4.218 | 5.288 | 24.270 | 0.370 |
| Ridge | 4.220 | 5.290 | 24.284 | 0.369 |
| LR | 4.221 | 5.292 | 24.289 | 0.369 |
| RFR | 4.189 | 5.335 | 23.830 | 0.358 |
| **GBR** | **3.985** | **4.992** | **22.741** | **0.438** |

Figure 10 compares observed and predicted values across the test observations. SVR predictions were concentrated within a narrow range, whereas the tree-based models produced a wider range; GBR followed the observed variation most closely.

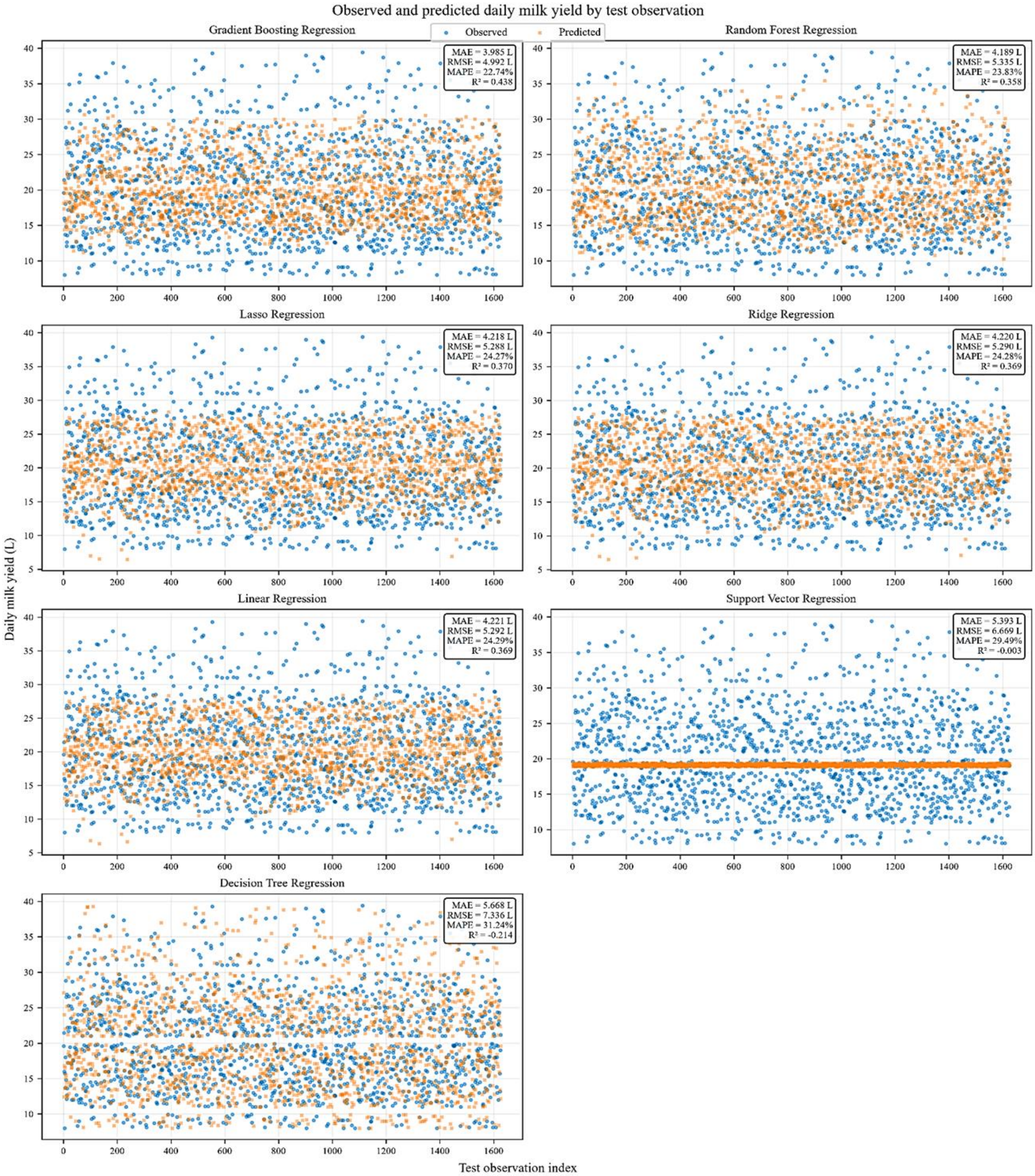


**Fig. 10.** Observed and predicted daily milk yield for the seven models on the test set.

Figure 11 shows the agreement between observed and predicted values. GBR showed comparatively closer agreement with the identity line, although dispersion remained toward the lower and higher ends of the milk-yield range.

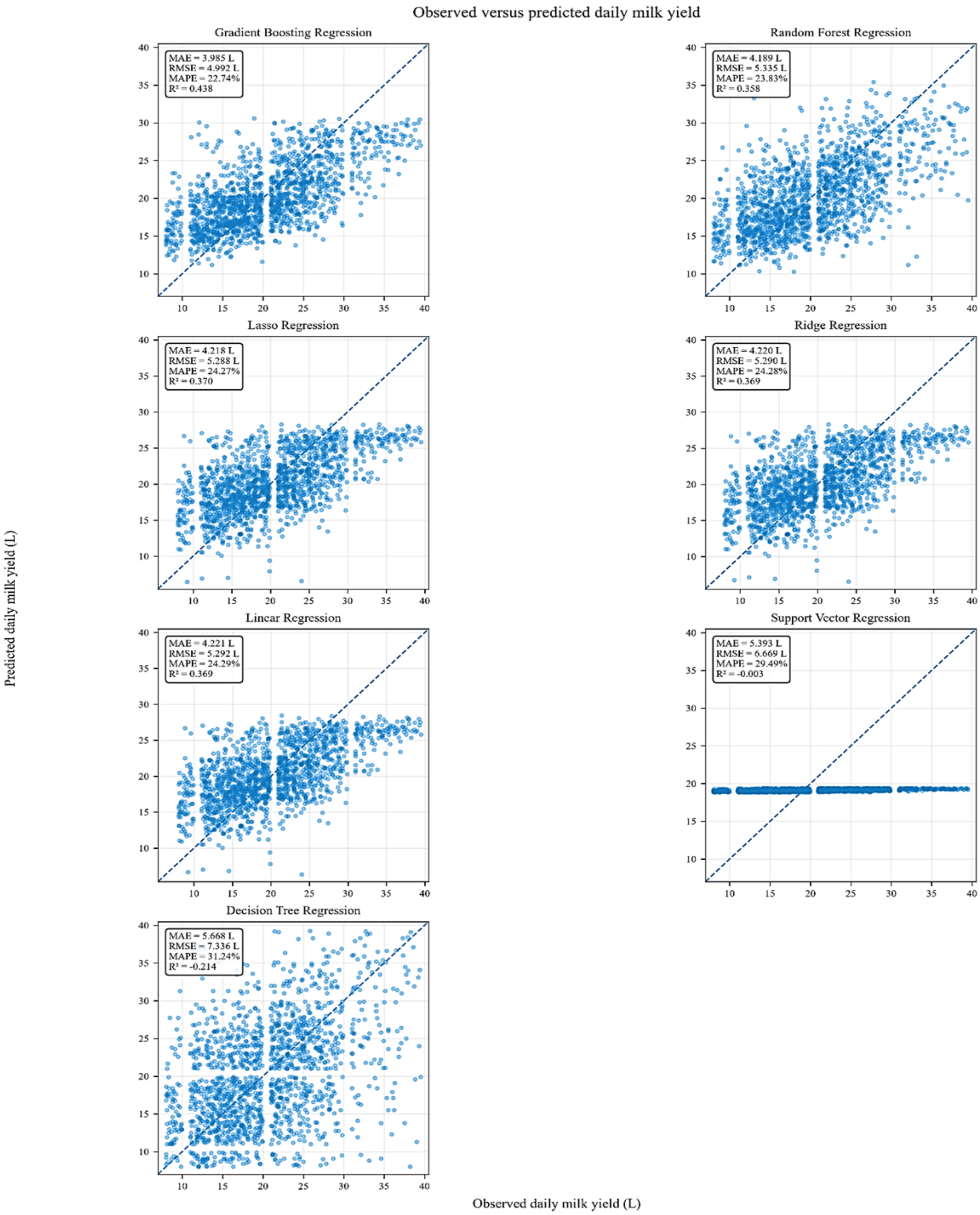


**Fig. 11.** Observed versus predicted daily milk yield for the seven models. The identity line represents perfect agreement.

Figure 12 presents the residual distributions. GBR showed a smaller residual spread than the other algorithms, although relatively large residuals remained for some observations, so performance was not uniform across the full milk-yield range.

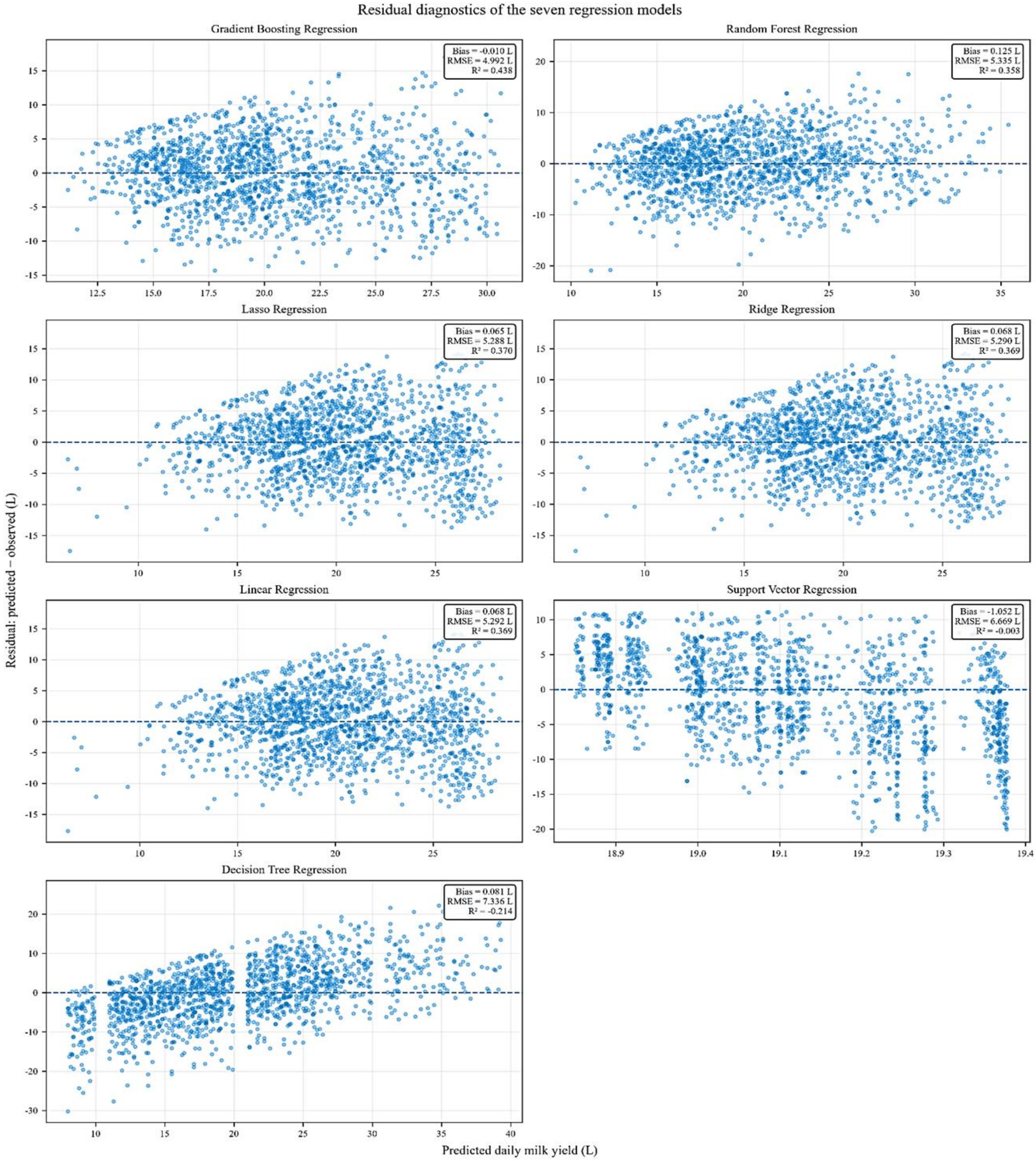


**Fig. 12.** Residuals versus predicted daily milk yield for the seven models on the test set.

The incremental contribution of WMY was evaluated by comparing otherwise identical GBR models with and without the engineered feature. Without WMY the model yielded an MAE of 3.989 L/day, an RMSE of 4.993 L/day, a MAPE of 22.787% and an $R^2$ of 0.438; with WMY the values were 3.985 L/day, 4.992 L/day, 22.741% and 0.438. Paired bootstrap analyses showed 95% confidence intervals for the differences that included zero.

Correlation and multicollinearity analyses showed substantial overlap among several predictors. The strongest Pearson correlations were between age and lactation number (r=0.932), days pregnant and the pregnant fertility indicator (r=0.865), and age and the Old age-category indicator (r=0.832). WMY had a VIF of 3.422 in the reference-coded analysis, so it was not the primary source of multicollinearity.

## 5. Discussion

Among the seven regression models evaluated, GBR produced the lowest prediction errors (MAE 3.985 L/day, RMSE 4.992 L/day, MAPE 22.741%, $R^2$ 0.438), RFR and the linear models showed intermediate performance, and SVR and DTR produced higher errors. However, the $R^2$ value indicates that a substantial proportion of the variability in daily milk yield remained unexplained, and several predictors contained overlapping information.

The predictive performance observed here can be considered together with previous milk-yield prediction studies, although direct numerical comparison requires caution because populations, predictor sets, target definitions, prediction horizons, units and validation procedures differ. In a recent study conducted in Türkiye, Çanga Boğa and Boğa (2026) evaluated six machine-learning algorithms using 830 records from a commercial dairy farm in Niğde with partially overlapping predictors; Random Forest and Gradient Boosting performed best ($R^2$ approximately 0.88, RMSE approximately 5.2 L), and multicollinearity between age and lactation number was also considered. Salamone et al. (2022) reported RMSE values of 6.08-6.24 kg, Vergani et al. (2024) reduced MAE from 5.63 to 3.96 kg when genomic and management information was incorporated, and Dallago et al. (2019) reported an MAE below 4 kg. Hooker et al. (2025) obtained an MAE of 2.69 kg when previous milk-yield records were available, and Darabighane and Atzori (2026) an MAE of 1.39±0.12 kg. These differences should not be read as a direct ranking, because the available information and experimental designs differ substantially.

WMY was developed as an empirical composite predictor derived from days in milk, age, lactation number and Holstein breed status. Although it ranked among the most influential predictors in the feature-importance analysis, ablation showed only a very small improvement: MAE decreased from 3.989 to 3.985 L/day and the paired-bootstrap confidence intervals included zero. High model-internal feature importance therefore does not necessarily correspond to an independent improvement in predictive performance. Because WMY is derived from predictors already available to the model and equals zero for non-Holstein cows, it is a study-specific engineered predictor rather than a validated milk-yield index.

Several limitations should be considered. The dataset was obtained from commercial dairy farms in Eastern Anatolia and no independent farm-level dataset was available for external validation, so the results may not generalize to different regions, breeds, feeding systems or management conditions. The analytical cohort was restricted to the farm-specific 8–40 L/day data-quality eligibility range, which may limit generalizability to cows with milk yields outside this interval. Because absolute model performance can vary according to the train–test partition, the primary single-split results were supplemented by repeated holdout validation across 100 additional predefined partitions. GBR ranked first among the seven algorithms for MAE, RMSE, and $R^2$ in all 100 partitions, supporting the stability of the relative model-comparison result. The moderate $R^2$ value also indicates that variables not available here, including feeding, health, environmental, management, sensor or genomic information, may explain further variability.

## 6. Conclusions

Daily milk yield was modelled from records routinely kept on dairy farms in the Eastern Anatolia region, using breed, age, days in milk, lactation number, pregnancy-related characteristics and insemination information. Of the seven regression algorithms compared, GBR gave the smallest prediction errors (MAE 3.985 L/day, RMSE 4.992 L/day, MAPE 22.741%, $R^2$ 0.438), although the moderate $R^2$ falls short of supporting a claim of high predictive accuracy. Weighted Milk Yield ranked highly in the feature-importance analysis, yet its ablation gain was marginal (MAE 3.989 versus 3.985 L/day, with paired-bootstrap intervals including zero) and it cannot be credited with a statistically reliable improvement of its own. These conclusions are constrained by a single-region dataset without external validation, a cohort confined to the farm-specific 8-40 L/day data-quality eligibility range. Repeated validation across 100 additional predefined partitions supported the stability of GBR as the best-performing algorithm among the models evaluated, although absolute performance varied across partitions. Independent or prospective validation, together with feeding, health, environmental, management, sensor or genomic predictors, remains the necessary next step.

## References

1. Alibrahim, H. and Ludwig, S. A. 2021. Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. 2021 IEEE Congress on Evolutionary Computation (CEC), PP. 1551-1559.
2. Amin, A. A. 2003. Test-Day Model of Daily Milk Yield Prediction across Stages of Lactation in Egyptian Buffaloes. Arch. Anim. Breed., 46(1): 35-45.
3. Bar-Anan, R. and Genizi, A. 1981. The Effects of Lactation, Pregnancy and Calendar Month on Milk Records. Anim. Sci., 33(3): 281-290.
4. Bozkurt, F. 2022. A Comparative Study on Classifying Human Activities Using Classical Machine and Deep Learning Methods. Arab. J. Sci. Eng., 47(2): 1507-1521.
5. Çanga Boğa, D. and Boğa, M. 2026. Milk Yield Prediction Using Animal and Production Features: Optimal Combinations and Machine Learning Approaches. Turk. J. Agric. Food Sci. Technol., 14(6): 1476-1482.
6. Cerda, P., Varoquaux, G. and Kégl, B. 2018. Similarity Encoding for Learning with Dirty Categorical Variables. Mach. Learn., 107(8-10): 1477-1494.
7. Chen, J. F., Do, Q. H., Nguyen, T. V. A. and Doan, T. T. H. 2018. Forecasting Monthly Electricity Demands by Wavelet Neuro-Fuzzy System Optimized by Heuristic Algorithms. Information, 9(3): 51.
8. Coulon, J. B., Pérochon, L. and Lescourret, F. 1995. Modelling the Effect of the Stage of Pregnancy on Dairy Cows' Milk Yield. Anim. Sci., 60(3): 401-408.
9. Dallago, G. M., de Figueiredo, D. M., de Resende Andrade, P. C., dos Santos, R. A., Lacroix, R., Santschi, D. E. and Lefebvre, D. M. 2019. Predicting First Test Day Milk Yield of Dairy Heifers. Comput. Electron. Agric., 166: 105032.
10. Darabighane, B. and Atzori, A. S. 2026. Comparison of a Linear Mixed Model and Tree-Based Machine Learning Models for Daily Milk Yield Prediction in Dairy Cows during Summer. Information, 17(5): 415.
11. De Vries, A. and Marcondes, M. I. 2020. Overview of Factors Affecting Productive Lifespan of Dairy Cows. Animal, 14(S1): 155-164.

12. Ferrari, C., Punturiero, C., Delledonne, A., Vergani, A. M., Masseroli, M., Strillacci, M. G. and Bagnato, A. 2026. Cross-Generational Validation of a Feedforward Neural Network for Milk Yield Prediction in Dairy Cattle. Animals, 16(5): 707.

13. Hooker, J., de Medeiros, B. B., Saha, C., Abdulrahman, T. and Alves, A. A. C. 2025. Evaluating the Contribution of Behavioral, Milking System, and Environmental Data to Short-Term Milk Yield Prediction in Commercial Dairy Cows Using Machine Learning. J. Dairy Sci., 108(10): 11133-11150.

14. Ji, B., Banhazi, T., Phillips, C. J., Wang, C. and Li, B. 2022. A Machine Learning Framework to Predict the Next Month's Daily Milk Yield, Milk Composition and Milking Frequency for Cows in a Robotic Dairy Farm. Biosyst. Eng., 216: 186-197.

15. Karcioğlu, A. A. and Aydin, T. 2019. Sentiment Analysis of Turkish and English Twitter Feeds Using Word2Vec Model. 2019 27th Signal Processing and Communications Applications Conference (SIU), PP. 1-4.

16. Macciotta, N. P. P., Vicario, D. and Cappio-Borlino, A. 2005. Detection of Different Shapes of Lactation Curve for Milk Yield in Dairy Cattle by Empirical Mathematical Models. J. Dairy Sci., 88(3): 1178-1191.

17. McDonald, G. C. 2009. Ridge Regression. Wiley Interdiscip. Rev. Comput. Stat., 1(1): 93-100.

18. Mićić, N., Stanojević, D., Samolovac, L., Petričević, V., Stojiljković, N., Gantner, V. and Bogdanović, V. 2022. The Effect of Animal-Related and Some Environmental Effects on Daily Milk Production of Dairy Cows under the Heat Stress Conditions. Mljekarstvo, 72(4): 250-260.

19. Mishra, P., Fatih, C., Niranjan, H. K., Tiwari, S., Devi, M. and Dubey, A. 2020. Modeling and Forecasting of Milk Production in Chhattisgarh and India. Indian J. Anim. Res., 54(7): 912-917.

20. Okur, H. and Cetin, A. 2019. Credit Risk Estimation with Machine Learning. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), PP. 1-6.

21. Pekel, E. 2020. Estimation of Soil Moisture Using Decision Tree Regression. Theor. Appl. Climatol., 139(3-4): 1111-1119.

22. Pinedo, P., Santos, J. E. P., Chebel, R. C., Galvão, K. N., Schuenemann, G. M., Bicalho, R. C., Gilbert, R. O., Rodriguez-Zas, S. L., Seabury, C. M., Rosa, G. and Thatcher, W. 2020.

Associations of Reproductive Indices with Fertility Outcomes, Milk Yield, and Survival in Holstein Cows. J. Dairy Sci., 103(7): 6647-6660.

23. Ranstam, J. and Cook, J. A. 2018. LASSO Regression. Br. J. Surg., 105(10): 1348.

24. Rashmi, C. R. and Shantala, C. P. 2024. Evaluating Deep Learning with Different Feature Scaling Techniques for EEG-Based Music Entrainment Brain Computer Interface. e-Prime Adv. Electr. Eng. Electron. Energy, 100448.

25. Ray, D. E., Halbach, T. J. and Armstrong, D. V. 1992. Season and Lactation Number Effects on Milk Production and Reproduction of Dairy Cattle in Arizona. J. Dairy Sci., 75(11): 2976-2983.

26. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M. 2015. Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines. Ore Geol. Rev., 71: 804-818.

27. Salamone, M., Adriaens, I., Vervaet, A., Opsomer, G., Atashi, H., Fievez, V., Aernouts, B. and Hostens, M. 2022. Prediction of First Test Day Milk Yield Using Historical Records in Dairy Cows. Animal, 16(11): 100658.

28. Santos, E. P. B., Feltes, G. L., Negri, R., Cobuci, J. A. and Silva, M. V. G. B. 2021. Estimation of Genetic Parameters for Test-Day Milk Yield in Girolando Cows Using a Random Regression Model. Arq. Bras. Med. Vet. Zootec., 73.

29. Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P. and Weigel, K. 2014. Prediction of Insemination Outcomes in Holstein Dairy Cattle Using Alternative Machine Learning Algorithms. J. Dairy Sci., 97(2): 731-742.

30. Shekar, B. H. and Dagnew, G. 2019. Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), PP. 1-8.

31. Singla, P., Duhan, M. and Saroha, S. 2022. Different Normalization Techniques as Data Preprocessing for One Step Ahead Forecasting of Solar Global Horizontal Irradiance. In: "Artificial Intelligence for Renewable Energy Systems", Woodhead Publishing.

32. Subbanna, Y. B., Kumar, S. and Puttaraju, S. K. M. 2021. Forecasting Buffalo Milk Production in India: Time Series Approach. Buffalo Bull., 40(2): 335-343.

33. Suseendran, G. and Duraisamy, B. 2021. Predication of Dairy Milk Production Using Machine Learning Techniques. In: "Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021", Springer Singapore, PP. 579-588.

34. Tarjan, L., Šenk, I., Pracner, D., Štrbac, L., Šaran, M., Ivković, M. and Dedović, N. 2023. Application of Machine Learning in Estimating Milk Yield According to the Phenotypic and Pedigree Data of Holstein-Friesian Cattle in Serbia. Contemp. Agric., 72(4): 181-187.

35. Timkovičová Lacková, P., Maskaľová, I. and Vajda, V. 2022. Effect of Parity and Days in Milk on Milk Urea Concentration and Milk Components in Holstein Dairy Cows. Folia Vet., 66(2): 1-8.

36. Uribe, H. and Lembeye, F. 2020. Estimation of Genetic Parameters for Milk Yield Using a Random Regression Test-Day Model in First Parity Dairy Cows under Pasture-Based Systems of Los Ríos Region in Chile. Austral J. Vet. Sci., 52(3): 10.

37. Vergani, A. M., Bagnato, A. and Masseroli, M. 2024. Predicting Bovine Daily Milk Yield by Leveraging Genomic Breeding Values. Comput. Electron. Agric., 219: 108777.

38. Vrhel, M., Ducháček, J., Gašparík, M., Vacek, M., Codl, R. and Pytlík, J. 2024. Association between Production and Reproduction Parameters Based on Parity and Breed of Dairy Cows in the Czech Republic. Arch. Anim. Breed., 67: 197-205.

39. Walsh, S., Delaby, L., Kennedy, M., Galvin, N., McKay, Z. C. and Egan, M. 2024. Intake Profile, Milk Production, and Energy Balance of Early-Lactation Spring-Calving Holstein Friesian and Jersey × Holstein Friesian Dairy Cows in High-Utilization Pasture-Based Systems. J. Dairy Sci., 107(10): 8058-8071.

40. Wilmink, J. B. M. 1987. Adjustment of Test-Day Milk, Fat and Protein Yield for Age, Season and Stage of Lactation. Livest. Prod. Sci., 16(4): 335-348.

41. Wood, P. D. P. 1967. Algebraic Model of the Lactation Curve in Cattle. Nature, 216(5111): 164-165.

42. Yağanoğlu, M. 2022. Hepatitis C Virus Data Analysis and Prediction Using Machine Learning. Data Knowl. Eng., 142: 102087.

43. Zhang, W., Yang, K., Yu, N., Cheng, T. and Liu, J. 2020. Daily Milk Yield Prediction of Dairy Cows Based on the GA-LSTM Algorithm. 2020 15th IEEE International Conference on Signal Processing (ICSP), Vol. 1, PP. 664-668.

44. Zhang, Y. and Haghani, A. 2015. A Gradient Boosting Method to Improve Travel Time Prediction. Transp. Res. Part C Emerg. Technol., 58: 308-324.