



Assessing the Production Potential and Lactation Dynamics of Karan Fries Cattle Utilizing Machine Learning based Predictive Modelling

Ishmeet Kumar¹, Ritik Kumar Singh¹, Dibyasha Kar¹, R.K. Deepikarani¹, Pooja Beniwal¹, Deepti Sinha¹, Gargi Ramola¹, Sabyasachi Mukherjee¹

10.18805/IJAR.B-5763

ABSTRACT

Background: This study evaluates machine learning (ML) algorithms as a robust alternative to traditional linear models, which often overlook complex biological patterns for predicting lactation performance in Karan Fries (KF) cattle. Using phenotypic records from 473 lactating animals, the analysis focuses on predicting 305-day milk yield (305 DMY) and total milk yield (TMY). The population exhibited an average peak yield of 13.28 ± 0.25 kg and a persistent average fat concentration of 4.14%.

Methods: This study was done by integrating Wood's incomplete gamma function to model lactation curves, compressing longitudinal test-day data into mathematically significant parameters. These parameters, alongside fixed environmental factors, served as predictors for three models: Multiple linear regression (MLR), random forest (RF) and extreme gradient boosting (XGBoost).

Result: Results revealed that the RF model significantly outperformed both MLR and XGBoost. For 305DMY, RF achieved a high coefficient of determination ($R^2 = 0.88$) compared to the poor performance of the MLR baseline ($R^2 = 0.21$). A similar hierarchy persisted for TMY, where RF attained an R^2 of 0.87, effectively reducing the root mean square error (RMSE) by over 50% compared to MLR. The study concludes that integrating Wood's parameters with ensemble-based ML models offers a transformative tool for dairy breeding by facilitating early selection and optimized herd management.

Key words: Karan fries cattle, Lactation curve, Multiple linear regression, Random forest, Wood's model, XGBoost.

INTRODUCTION

Historically, Indian cattle were characterized by low yields and delayed calving intervals (Bhardwaj *et al.*, 2023). To address this, the Karan Fries (KF) breed was developed at ICAR-NDRI, Karnal, by crossing majorly Holstein-Friesian with Tharparkar cattle starting in 1971 (Gurnani *et al.*, 1986). The goal was to combine high milk yield with indigenous heat tolerance and disease resistance. Anatomically, the breed features a wide rump and balanced udder for calving ease and productivity (Kumar *et al.*, 2025). While monitoring TMY and 305-DMY is vital for genetic assessment (Singh *et al.*, 2016; Vyas *et al.*, 2025), traditional linear models and ssGBLUP often struggle with the non-linear “noise” of biological data (Brito *et al.*, 2021; Ilayaraja *et al.*, 2025; Nayeri *et al.*, 2019; Bayer *et al.*, 2021). Machine Learning (ML) offers a flexible alternative, extracting patterns without rigid statistical assumptions (Chafai *et al.*, 2023). Ensemble algorithms like random forest (RF) and xgBoost excel at managing high-dimensional data and quantifying feature importance, often outperforming multiple linear regression (MLR) (Topuz *et al.*, 2025; Streefland *et al.*, 2023). By capturing complex environmental and parity interactions, ML models enhance precision breeding (Khan *et al.*, 2025). This study compares RF, xgBoost and MLR to demonstrate how advanced modeling can refine lactation predictions and support earlier genetic selection in Karan Fries herds.

¹Animal Genetics and Breeding Division, ICAR-National Dairy Research Institute, Karnal-132 001, Haryana, India.

Corresponding Author: Sabyasachi Mukherjee, Animal Genetics and Breeding Division, ICAR-National Dairy Research Institute, Karnal-132 001, Haryana, India.

Email: sabayasachimukherjee@gmail.com

ORCIDs: <https://orcid.org/0009-0002-4291-4292>, <https://orcid.org/0009-0009-0182-6190>, <https://orcid.org/0009-0004-4834-8193>, <https://orcid.org/0009-0009-4699-6134>, <https://orcid.org/0009-0006-3524-7654>, <https://orcid.org/0009-0006-6442-8921>, <https://orcid.org/0009-0007-6714-0162>, <https://orcid.org/0000-0002-1711-6582>

How to cite this article: Kumar, I., Singh, R.K., Kar, D., Deepikarani, R.K., Beniwal, P., Sinha, D., Ramola, G. and Mukherjee, S. (2026). Assessing the Production Potential and Lactation Dynamics of Karan Fries Cattle Utilizing Machine Learning based Predictive Modelling. *Indian Journal of Animal Research*. **60(7)**: 1307-1317. doi: 10.18805/IJAR.B-5763.

Submitted: 20-01-2026 Accepted: 09-03-2026 Online: 08-06-2026

MATERIALS AND METHODS

Study area and data description

The study utilized Karan Fries cattle records from ICAR-NDRI, Karnal. Data pre-processing involved removing incomplete pedigrees, inconsistent intervals and outliers (TMY < 500 kg and ± 3 SD from mean were removed). The primary dataset included 10 sequential monthly test-day

records (T1-T10) tracking milk yield (TDMY) and fat yield (TDFY). To evaluate model performance, total milk yield (TMY) and 305-day milk yield (305-DMY) were defined as target variables, resulting in a final set of 473 complete lactations. The coefficient of variation (CV) was used to assess data variability.

Statistical analysis

Descriptive statistics were computed for TMY, TDFY and lactation length (LL). To ensure the biological plausibility and statistical integrity of the dataset, outliers were identified and removed using the interquartile range (IQR) method (Meinert *et al.*, 1998). The IQR method was preferred over conventional mean-based approaches because extreme outliers in raw data can significantly inflate the sample mean and standard deviation. Consequently, records falling outside these calculated boundaries were excluded prior to evaluating phenotypic relationships. To address missing values, linear interpolation via the *imputeTS* R package was used instead of mean substitution, which fails to capture temporal lactation dynamics Moritz and Bartz-Beielstein (2017). Additionally, the *lubridate* package standardized DOB and DOC variables Moritz and Bartz-Beielstein (2017), ensuring precise calculation of lactation intervals and derived traits.

Adjustment for fixed effects and yield calculations

To account for non-genetic environmental variance, first lactation records were corrected for fixed factors season of birth (SOB), Season of calving (SOC) and year of calving (YOC) using IBM SPSS (v25.0). These adjustments (Table 3) mitigate seasonal bias in predictions. Additionally, lactation and fat yields were calculated via the Test Interval Method (TIM), International Committee for Animal Recording (ICAR) guidelines (ICAR, 2020).

Non-parametric lactation descriptors

This approach allowed for a flexible analysis of the lactation cycle without relying on complex mathematical assumptions. Whereas peak yield (Y_{max}), as the highest amount of milk produced on any single test day. Initial yield (Y_1), as the milk yield recorded during the very first test day of the lactation period and slope to peak (S_{asc}), as the value measures how quickly milk production increases during early lactation as the animal reaches its maximum output. It represents the rate of change from the first test day to the day of peak production as:

$$S_{asc} = \frac{Y_{max} - Y_1}{\Delta t}$$

Where,

Δt = Standardized to 60 days which represents the approximate average time reported for the KF population to attain peak yield.

Furthermore, lactation persistency (P) was conceptualized as the ability of the animal to maintain production after the

peak yield, calculated as the ratio of the last recorded yield (Y_{last}) to the peak yield (Wood, 1967).

$$P = \frac{Y_{last}}{Y_{max}}$$

Lactation length distribution was visualized in R using the *ggplot2* package (Wickham, 2016). Instead of frequency counts, we utilized kernel density estimation (KDE) and a density-scaled y-axis to provide a continuous probability density function (PDF) alongside empirical histogram bars (Silverman, 2018).

Parametric lactation curve modelling (Wood's model)

The shape of the lactation curve was modelled for each animal from the test day records using Wood's incomplete gamma function (Wood, 1967). Unlike calculating simple means, this parametric approach captures the biological underlying mechanism of lactation using the equation:

$$Y_t = at^be^{-ct}$$

Where,

Y_t = The milk yield at day t .

a (Scaling factor) = The initial milk production capacity.

b (Rate of rise) = The incline phase dominated by cell proliferation.

c (Rate of decline) = The decline phase.

By fitting this model using non-linear least squares (NLS) regression, the complex, high-dimensional longitudinal data (10 test days) was compressed into three biologically interpretable parameters (a , b , c). These parameters were then used as static predictors in the machine learning models, effectively teaching the algorithms the trajectory of the curve (Graeboll *et al.*, 2015).

Statistical modelling and machine learning algorithms

Three distinct regression algorithms were selected to model the relationship between the lactation features and the target yield traits (305-DMY and TMY), from a parametric baseline (Multiple linear regression) to non-parametric Random Forest (RF) and advanced boosting frameworks (XGBoost). To ensure a rigorous and unbiased comparison between the three modelling approaches, a consistent 5-fold cross-validation (CV) protocol was implemented using the '*caret package*' in R, (Adnan *et al.*, 2022). The dataset was partitioned into a training set (80%) and an independent testing set (20%) using stratified random sampling to ensure that the distribution of milk yields was preserved in both subsets.

Multiple linear regression (MLR)

Multiple linear regression (MLR) served as the parametric baseline for this study. It posits a linear, additive relationship between the dependent variable (Y) and the set of independent predictors $X_1, X_2, \dots, X_p$. The model estimates coefficients (β) that minimize the sum of squared residuals:

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i$$

Where,

β_0 = The intercept.

β_j = the partial change in yield for a unit change in predictor j .

$\epsilon_i \sim N(0, \sigma^2)$ = Stated as the residual error.

The MLR model was implemented using the method 'lm' function within 'caret' package. While MLR was reported as widely used in dairy science for its interpretability (Dongre *et al.*, 2012), it relies on strict assumptions of linearity, homoscedasticity and normality of errors. In the context of this study, by comparing MLR against tree-based models, we can explicitly quantify the non-linear gain, the improvement in accuracy achieved by capturing complex interactions that MLR fails to model (Shmueli, 2011).

Descriptive predictor variables

The predictive modelling framework utilized a multi-dimensional feature set comprising variables. To capture the biological complexity of the lactation cycle, parameters from Wood's incomplete gamma function (a, b, c) Initial milk yield, ascending slope and descending slope respectively, were integrated alongside non-parametric descriptors such as peak yield and persistency. Environmental variance was accounted for through fixed effects (SOB, SOC and YOC) as Season of Birth, Season of Calving and Year of Calving, while longitudinal test-day records (TD-1 to TD-10) provided the temporal basis for the 305DMY and TMY forecasts.

Random forest regression (RF)

Random forest (RF) was defined as the non-parametric ensemble learning method based on bootstrap aggregating. It constructs a multitude of decorrelated decision trees during training and outputs the average prediction of the individual trees (Breiman, 2001).

Mathematically, for an input vector x , the RF prediction $\hat{f}_{rf}(x)$ was the average of B trees:

$$\hat{f}_{rf}(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

Where each tree T_b was created on a bootstrap sample of the data. Crucially, at each split in the tree, the algorithm considers only a random subset of predictors. This randomness decorrelates the trees, reducing the variance of the final model without increasing bias.

The model was implemented using the 'random forest' package (Liaw and Wiener, 2002). Furthermore, the model explicitly enabled variable importance calculations to identify the primary physiological and environmental contributors to production (Liaw and Wiener, 2002). RF was selected for its inherent robustness to multicollinearity and its capacity to handle mixed complex biological data.

Extreme gradient boosting (XGBoost)

XGBoost (eXtreme gradient boosting) was selected for its ability to enhance prediction accuracy by iteratively focusing

on residual errors. Unlike random forest, which constructs multiple trees in parallel, XGBoost followed a sequential process where each new tree was designed to correct the mistakes of its predecessor. This boosting approach allowed the model to "learn" from its own errors, making it highly effective at capturing subtle patterns in lactation data that other models might have overlooked (Chen and Guestrin, 2016). The model operated by minimizing an objective function at each step t , which combined a loss function L (Squared error) and a regularization term $\Omega(f_t)$:

$$\text{Obj}^{(t)} = \sum_{i=1}^n L[y_i, \hat{y}_i^{t-1} + f_t(x_i)] + \Omega(f_t)$$

Hyperparameter tuning was employed in order to maximize predictive performance, a grid search strategy was employed to systematically identify the most effective hyperparameters (Chen and Guestrin, 2016).

Performance metrics

The predictive performance of the final models was rigorously evaluated on the test set using three complementary statistical metrics. Firstly, Coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE), (Willmott and Matsuura, 2005).

RESULTS AND DISCUSSION

Descriptive statistics and lactation profile

The average total milk yield (TMY) was recorded as 4414.58±69.75 kg, with high coefficient of variation (CV) of 31.64%, (Table 1). Whereas 305DMY was 3476.28±39.12 kg which were found to have comparatively low variability (CV = 22.53%). Peak yield was reported as with the mean of 13.28±0.25 kg, with the mean TDMY was 8.32±0.16 kg. The mean total fat yield per lactation was 142.61 kg. The average lactation length was 426.11 days, indicating a prolonged production cycle. Furthermore, the population exhibited a high average fat concentration of 4.14%. The Persistency found to have mean of 0.61±0.01 with a high CV of 36.54%, indicating a wide disparity in the animals' ability to maintain milk production levels post-peak (Table 1).

Trends in test-day milk and fat yields

The early lactation phase (TD1 to TD3) was marked by a characteristic ascent, initiating at 8.60±0.23 kg and reaching a maximum of 9.56±0.24 kg by the third test day. This period corresponded with the highest observed fat yield, which also peaked at TD3 (0.51±0.01 kg). Following this peak, the mid-lactation phase (TD4 to TD6) exhibited a gradual decline, with milk yield decreasing from 8.97±0.23 kg at TD4 to 8.24±0.21 kg by TD6. The late lactation stage (TD7 to TD10) demonstrated a persistent downward trajectory (Rana *et al.*, 2024). Yields dropped significantly to a minimum of 6.93±0.18 kg at TD8 before stabilizing slightly to conclude at 7.41±0.20 kg on the tenth test day. Fat yields depict same pattern, generally declining to 0.36±0.06 kg by the end of the lactation period (Table 2).

Table 1: Descriptive statistics of production traits and lactation curve parameters.

Variable	Mean±SE	Range	CV (%)
TMY (kg)	4414.58±69.75	1663.00-9182.00	31.64
305 DMY (kg)	3476.28±39.12	507.50-5636.50	22.53
Peak yield (kg)	13.28±0.25	3.99-25.00	36.98
Total fat yield (kg)	142.61±3.28	86.23-222.68	19.69
Mean TDMY (kg)	8.32±0.16	3.84-18.75	38.07
Lactation length (days)	426.11±6.98	320-573	13.99
Average fat (%)	4.14±0.16	2.75-4.78	33.02
Persistency	0.61±0.01	0.09-1.00	36.54

Table 2: Descriptive statistics of test-day milk and fat yields (Mean±SE).

Test day	Milk yield (kg)	Fat yield (kg)
TD1	8.60±0.23	0.50±0.21
TD2	8.95±0.26	0.40±0.14
TD3	9.56±0.24	0.51±0.12
TD4	8.97±0.23	0.41±0.10
TD5	8.43±0.22	0.45±0.17
TD6	8.24±0.21	0.38±0.01
TD7	8.40±0.20	0.44±0.02
TD8	6.93±0.18	0.33±0.04
TD9	7.72±0.18	0.42±0.01
TD10	7.41±0.20	0.36±0.06

Effect of non-genetic factors

The analysis of variance revealed that the SOB had a statistically significant effect on the 305DMY ($p < 0.05$). Cows born in Season 3 exhibited the highest standardized yield 3614.02 ± 73.88 kg, which was higher as comparable to those born in Season 4 (3551.51 ± 66.64) kg but significantly superior to animals born in seasons 1 and 2. In contrast, the season of birth did not significantly influence other production parameters, including LL, TMY, TFY and FP ($p > 0.05$). The SOC significantly influenced the TFY, with the highest fat yield observed in animals calving during Season 1 (147.54 ± 2.59) kg. This was significantly higher than the yield recorded for Season 3 (136.73 ± 2.75) kg, while Seasons 2 and 4 showed intermediate values. Furthermore, the YOC showed a significant effect on milk composition, specifically FP. It was observed that YOC-4 recorded the highest fat percentage ($7.34 \pm 0.10\%$), differing significantly from the lower percentages observed in YOC-1 ($6.92 \pm 0.11\%$) and YOC-3. No statistically significant differences were found for LL and TMY across the different seasons and years of calving (Table 3).

Non-parametric lactation and parametric lactation curve modelling

Non-parametric modelling revealed an initial yield of 8.78 ± 5.08 kg and a peak of 13.2 ± 5.49 kg. A persistency of 0.620 ± 0.28 indicated that KF cows maintained ~62% of peak production. As shown in (Fig 1), the mean lactation length was 426.11 days (Solid blue line). The KDE curve (dashed line) peaked between 380-480 days, with standard

deviation markers (red dotted lines) confirming most cows exceeded the standard 305-day lactation period.

The Karan Fries lactation trajectory was modeled using Wood's function (Wood, 1967) via non-linear regression (nls in R). The scaling parameter (a) was 55.1, representing initial production; the inclining rate (b) was 0.102; and the decay rate (c) was 0.00113. Together, these three parameters standardize the population's production pattern and serve as fundamental predictors for total lactation yield.

Phenotypic correlations

A strong positive correlation exists between total milk yield and total fat yield ($r = 0.716$), indicating that higher volumes yield greater cumulative fat. Similarly, lactation length and total fat yield are positively correlated ($r = 0.68$), highlighting the role of persistency in maximizing nutrient harvest. Conversely, total milk yield and average fat percentage share a moderate negative correlation ($r = -0.56$).

Multiple linear regression

Building on the previous analysis, the predictive performance for both 305DMY and TMY was evaluated across the testing dataset (Table 4), the MLR model demonstrated moderate predictive capability, achieving an R^2 of 0.21, resulting in an RMSE of 729.71 kg and an MAE of 570.56 kg. A similar trend was observed for TMY prediction (Table 4), yielding a lower R^2 of 0.1012. The associated error metrics were notably higher than those for 305DMY, with an RMSE of 1405.87 kg and an MAE of 1082.63 kg, highlighting the increased complexity of predicting TMY compared to the standardized 305DMY (Fig 2 and 5).

Random forest

Random forest (RF) model demonstrated a substantial leap in predictive accuracy. For 305 DMY, the RF model achieved an R^2 of 0.88. This performance represents a dramatic improvement over the MLR baseline, reducing the RMSE to 315.98 kg and the MAE to 209.61 kg. A similarly robust pattern emerged for TMY prediction, where the RF model yielded an R^2 of 0.8688, an RMSE of 634.12 kg and an MAE of 432.20 kg. Remarkably, the algorithm maintained consistent efficacy across both 305DMY and TMY with R^2 of (0.88 vs. 0.87), suggesting that Random Forest effectively captures the underlying biological determinants of milk

production regardless of the specific trait definition or scale. (Fig 3 and 5; Table 4).

Extreme gradient boosting

In addition to the MLR and Random Forest approaches, the Extreme Gradient Boosting (XGBoost) model was evaluated. For 305DMY, XGBoost achieved an R^2 of 0.82, with an RMSE of 359.63 kg and an MAE of 264.01 kg. While this represents a substantial improvement over the linear model, it fell short of the accuracy achieved by the Random Forest. This hierarchical trend persisted in the TMY predictions, where XGBoost again occupied the middle ground between MLR and RF. For TMY, the model yielded an R^2 of 0.7818, an RMSE of 755.03 kg and an MAE of 551.84 kg (Fig 4 and 5; Table 4).

Variable importance and biological drivers

Variable importance rankings (Fig 6) identify mean test-day milk yield (Mean TDMY) as the primary predictor across all models. For TMY and 305DMY, persistency (67.0%) and slope early peak (62.8%) were the leading secondary predictors. Although included, raw Wood's function coefficients (a, b, c) showed low importance (<5%), suggesting that Random Forest models prioritize derived biological indicators over raw mathematical parameters to describe the lactation curve.

Residual distribution

The residual analysis reveals a near-Gaussian distribution of errors for the ensemble models, particularly random forest. The concentration of residuals around the zero-line and the lack of skewness in the error histograms (Fig 7) indicate that the model assumptions hold true for this population.

This study interprets key production metrics including fat yield, fat percentage, Total Milk Yield (TMY), 305-day milk yield (305DAY) and extended lactation length to establish a comprehensive performance profile of the KF cattle. By blending traditional biometric frameworks, such as Wood's Gamma Function, with modern machine learning techniques like Random Forest (RF) and XGBoost.

Comparative analysis of milk production traits

Average total milk yield (TMY) was 4414.58 ± 69.75 kg, exceeding estimates by Worku *et al.* (2021). The 305-day milk yield (305-DMY) of 3476.28 ± 39.12 kg also surpassed findings by Kumar *et al.* (2025) (Table 2). A high TMY coefficient of variation (CV) of 31.64% reflected significant herd heterogeneity. The average lactation length (LL) of 426.11 days exceeded the 305-day ICAR (2023) standard, indicating high persistence while maintaining 4.14% fat. Despite a peak yield of 13.2 ± 5.49 kg and mean persistency of 0.61 ± 0.01 , a high persistency CV (36.54%) suggested greater diversity than noted by Dash *et al.* (2016), signalling substantial potential for future genetic improvement in KF cattle.

Table 3: Least square means of fixed factors.

Factor	No. of obs.	TMY (kg)	305 DMY (kg)	LL (days)	TFY (kg)	FP (%)
Season of calving (SOC)	473	-	-	-	-	-
1 (Winter)	178	4396.45 ± 131.51^a	3454.33 ± 72.62^a	405.25 ± 12.53^a	147.54 ± 2.59^b	7.23 ± 0.10^a
2 (Summer)	154	4546.54 ± 120.09^a	3510.96 ± 66.32^a	405.56 ± 11.44^a	141.56 ± 2.36^{ab}	7.01 ± 0.10^a
3 (Rainy)	64	4244.40 ± 139.85^a	3428.55 ± 77.23^a	392.01 ± 13.32^a	136.73 ± 2.75^a	6.97 ± 0.11^a
4 (Autumn)	76	4579.93 ± 136.75^a	3562.74 ± 75.52^a	424.87 ± 13.02^a	141.95 ± 2.69^{ab}	7.09 ± 0.11^a
Season of birth (SOB)	473	-	*	-	-	-
1 (Winter)	161	4374.58 ± 126.53^a	3471.80 ± 69.87^a	401.04 ± 12.05^a	141.96 ± 2.49^a	7.18 ± 0.10^a
2 (Summer)	128	4246.04 ± 148.64^a	3319.25 ± 82.09^a	415.40 ± 14.16^a	140.10 ± 2.92^a	6.92 ± 0.12^a
3 (Rainy)	99	4652.12 ± 133.78^a	3614.02 ± 73.88^b	412.28 ± 12.74^a	145.63 ± 2.63^a	6.98 ± 0.11^a
4 (Autumn)	85	4494.58 ± 120.67	3551.51 ± 66.64^b	398.98 ± 11.49^a	140.09 ± 2.37^a	7.23 ± 0.10^a
Year of calving (YOC)	473	-	**	-	-	*
1 (2015-2017)	112	4565.45 ± 142.29^a	3618.39 ± 78.58^b	409.57 ± 13.55^a	142.32 ± 2.80^a	6.92 ± 0.11^a
2 (2018-2020)	82	4412.10 ± 126.27^a	3389.03 ± 69.73^a	417.98 ± 12.03^a	140.77 ± 2.48^a	7.08 ± 0.10^{ab}
3 (2021-2023)	186	4256.71 ± 131.40^a	3429.67 ± 72.56^{ab}	399.94 ± 12.51^a	142.67 ± 2.58^a	6.95 ± 0.10^a
4 (2024-2025)	93	4533.06 ± 128.85^a	3519.49 ± 71.16^{ab}	400.20 ± 12.27^a	142.02 ± 2.53^a	7.34 ± 0.10^b

Least square means (LSM) along with standard error; NS: Non-significant ($P > 0.05$); *: Significant at $P < 0.05$; **: Significant at $P < 0.01$. Means bearing different superscripts (a, b) within the same column differ significantly. TMY: Total milk yield, 305 DMY: 305-day milk yield, LL: Lactation length, TFY: Total fat yield, FP: Fat percentage.

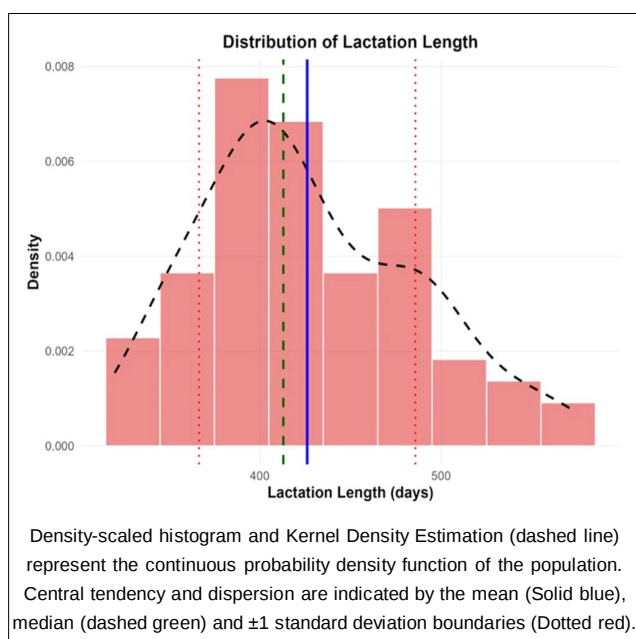


Fig 1: Distribution of lactation length.

Trends in test day milk and fat yields

Karan Fries (KF) lactation followed the standard biological trajectory: incline, stabilization and decline (Wood, 1967). Milk yield rose from 8.60 ± 0.23 kg (TD1) to 8.95 ± 0.26 kg (TD2), driven by mammary cell proliferation and fat mobilization (Bauman and Currie, 1980). This ascending trend aligns with Kokate *et al.* (2012) and Singh *et al.* (2016). Peak fat yield (0.51 ± 0.12 kg) confirmed successful homeorhetic adaptation to early metabolic demands (Bauman, 2000). The decline phase began at TD5 (8.43 ± 0.22 kg), reached a low at TD8 (6.93 ± 0.18 kg) and ended at 7.41 ± 0.20 kg by TD10, with fat yield dropping to 0.36 kg due to the gradual loss of secretory cells (Knight and Wilde, 1993).

Influence of fixed factors on production traits

In this study, SOC did not significantly influence TMY, 305-DMY, LL, or FP ($p > 0.05$), suggesting the KF breed's adaptability to environmental fluctuations aligning with Ilayaraja *et al.* (2025) but contradicting Worku *et al.* (2021). However, SOC significantly impacted total fat yield ($p = 0.041$), with summer calvers (Season 1) yielding the highest. SOB significantly

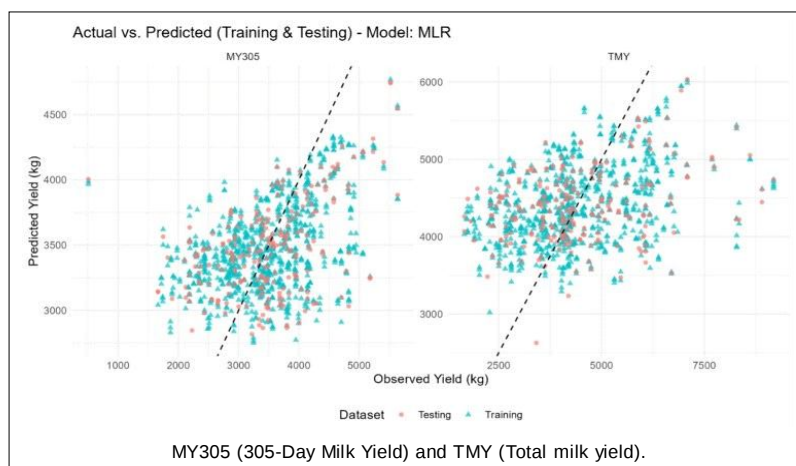


Fig 2: MLR based modelling: Training vs testing.

Table 4: Performance metrics for 305DMY and TMY prediction models: Testing set performance with comparison to training set metrics, showing R^2 , RMSE and MAE for multiple linear regression (MLR), random forest (RF) and extreme gradient boosting (XGBoost) models.

Model	R^2 (Training)	RMSE_Train (kg)	MAE_Train (kg)	R^2 (Testing)	RMSE (kg)	MAE (kg)
Performance metrics for 305 DMY prediction models						
MLR	0.18	703.83	557.48	0.22	729.71	570.56
RF	0.98	140.93	88.99	0.89	315.98	209.61
XGB	0.95	187.38	141.16	0.82	359.63	264.01
Performance metrics for TMY prediction models						
MLR	0.13	1282.27	1002.76	0.10	1405.87	1082.63
RF	0.98	270.56	173.45	0.87	634.12	432.20
XGB	0.94	378.71	275.36	0.78	755.03	551.84

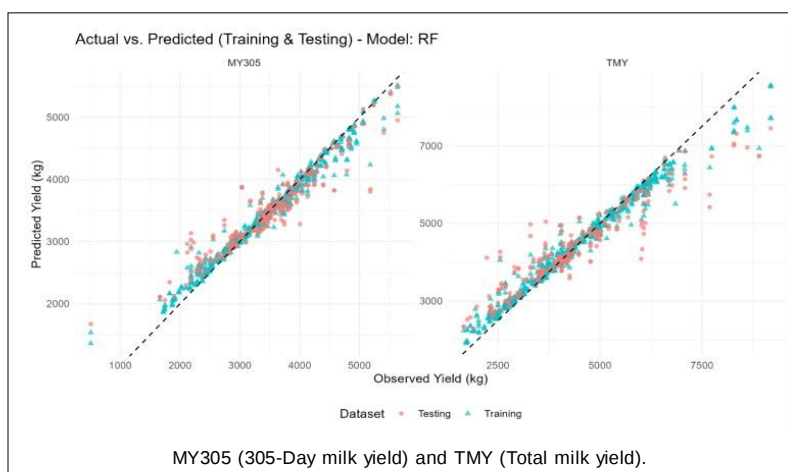


Fig 3: RF based modelling: Training vs testing.

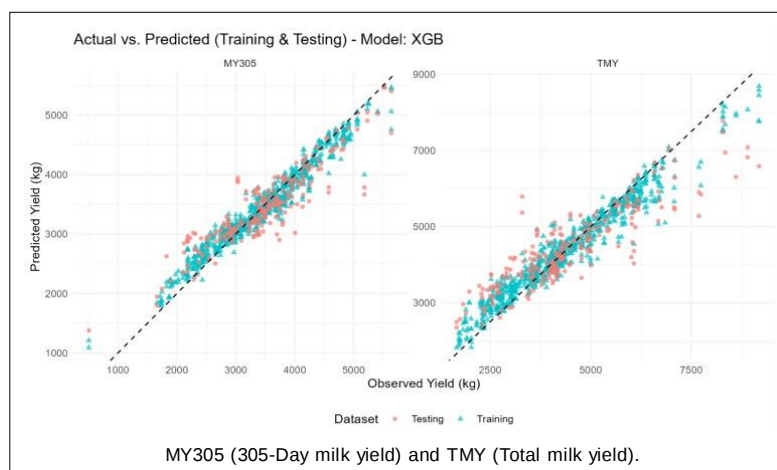


Fig 4: XGBoost based modelling: Training vs testing.

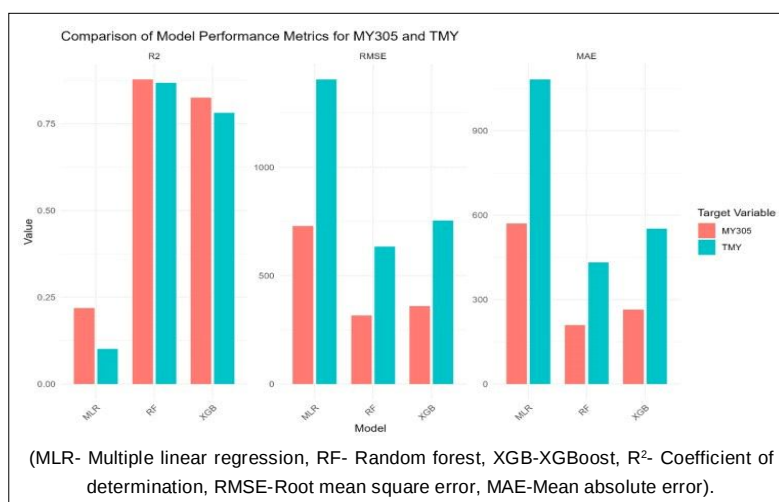


Fig 5: Comparison of model performance metrics for MY305 (305-day milk yield) and TMY (Total milk yield).

influenced 305-DMY ($p = 0.048$), with Season 3-born cows performing best, suggesting that early-life conditions affect long-term productivity. YOC significantly affected 305-DMY and FP ($p = 0.02$), likely due to annual variations in climate and fodder availability, consistent with Worku *et al.* (2021) but differing from Yadav *et al.* (2017).

Phenotypic correlations

A primary finding was the strong positive correlation between total milk yield (TMY) and total fat yield (TFY) ($r = 0.716$), similarly reported by Upadhyay *et al.* (2024). Furthermore, the positive correlation between lactation length and fat yield ($r = 0.685$) underscores the physiological importance of production persistency (ICAR-NDRI, 2015). Conversely, a moderate negative correlation was observed between total milk yield and average fat percentage ($r = -0.564$). This illustrates the dilution effect, a biological trade-off where increased fluid volume leads to a decreased concentration of solids.

Comparative evaluation of statistical and machine learning models

The RF algorithm significantly outperformed MLR and XGBoost in predicting 305 DMY and TMY. For the test set, RF achieved high values (0.878 for 305DMY; 0.869 for TMY), while XGBoost showed intermediate performance and MLR performed poorly ($R^2 < 0.22$). RF's superiority is validated by its lower error metrics; for 305DMY, RF yielded the lowest RMSE (315.98 kg) and MAE (209.61 kg). In contrast, MLR's RMSE was more than double (729.71 kg). The >360 kg reduction in MAE compared to MLR confirms RF as a more reliable tool for lactation forecasting. Although a high training R^2 (0.98) versus test R^2 (0.88) suggests slight overfitting typical of ensemble learning, the robust test performance demonstrates that RF effectively captured key biological signals. Future validation using multi-herd data and tree-pruning could further refine this gap.

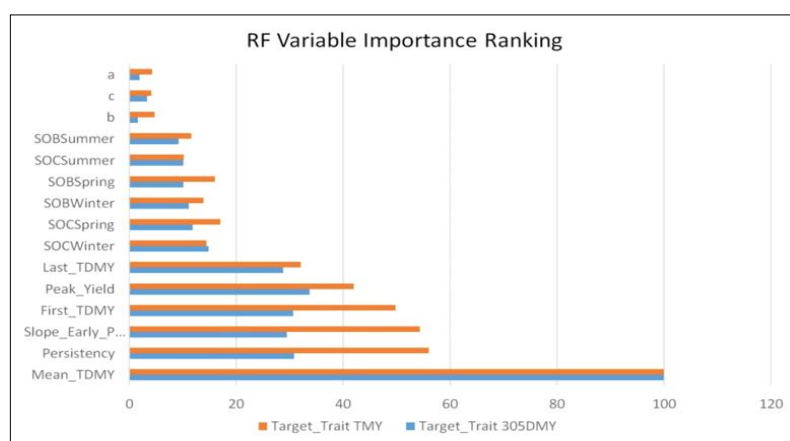


Fig 6: Variable importance rankings for random forest (RF).

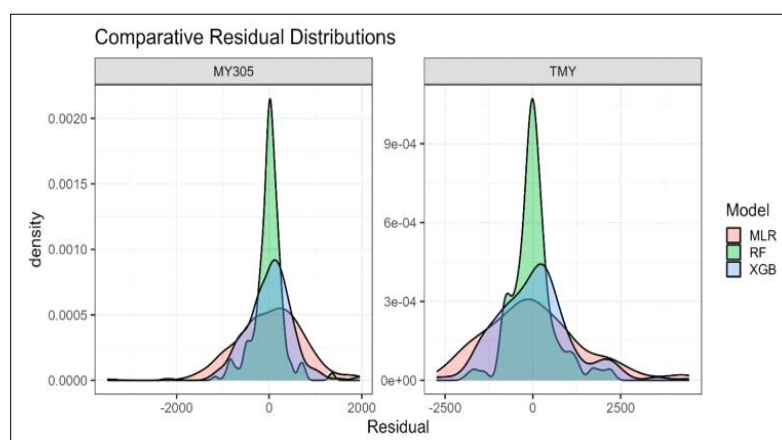


Fig 7: Comparative residual distribution for 305-day milk yield (305DMY) and total milk yield (TMY) across multiple linear regression (MLR), random forest (RF) and XGBoost models.

By assuming linear relationships, MLR fails to capture the inherently non-linear lactation curve, which involves a sharp rise followed by a gradual decline (Dongre *et al.*, 2012). Its high residual errors (RMSE = 729.71) confirm an inability to model intricate physiological patterns and environmental interactions (Sahoo *et al.*, 2019). Consequently, MLR's lack of adaptability to non-linear data limits its precision despite its high interpretability (Singh *et al.*, 2021).

Conversely, RF has emerged as a superior tool for predicting TMY and 305DMY, consistently outperforming linear methods (Worku *et al.*, 2021; Singh *et al.*, 2021). Its low MAE (432.20 kg) suggests better handling of missing data and physiological fluctuations than MLR (Tirink *et al.*, 2023). While XGBoost uses sequential error minimization (Chen and Guestrin, 2016), its 305DMY accuracy ($R^2 = 0.825$) trailed RF, likely due to data deviations. Although RF showed a tendency to overfit (R^2 0.98 training vs. test) a common trait in small biological datasets ($n=473$) it achieved the highest predictive power. These results align with findings in Iranian Holsteins = R^2 0.91; (Abdollahi-Arpanahi *et al.*, 2020) and Italian Holsteins = R^2 0.88; (Bovo *et al.*, 2021). Future multi-herd studies using nested cross-validation could further refine these estimates.

A key finding is the hierarchical preference of ML models for specific lactation features. While Wood's function provided the essential framework, its raw parameters (a, b, c) showed low relative importance in the RF model. This suggests an information-routing effect: because Persistency and Early Slope were derived from the Wood's curve, they can serve as biological summaries for the lactation profile in KF cattle. Since ML algorithms seek the most efficient predictive paths, they prioritize these integrated biological features over individual mathematical constants. This highlights Wood's function's value as a powerful tool for feature engineering rather than just a standalone fitting method.

CONCLUSION

This study demonstrates the superior productive potential of KF cattle, characterized by a TMY of 4414.58 kg, 305-day yield of 3476.28 kg, extended lactation length of 426 days and stable 4.14% fat content, indicating strong persistency. The breed's adaptability is evidenced by the non-significant effect of calving season on yield, while the significant impact of birth season underscores the importance of early-life management for lifetime performance. Test-day yields followed a biologically plausible curve, peaking at TD3 before a gradual decline. Among the predictive models, Random Forest proved most robust ($R^2 = 0.878$), followed by XGBoost; both significantly outperformed Multiple Linear Regression, which failed to capture non-linear lactation dynamics. Integrating Wood's parameters with advanced machine learning provides an robust framework for precision dairy

management. Implementing high-accuracy models like Random Forest can optimize early selection, nutritional strategies and culling policies to enhance the economic efficiency of Karan Fries herds.

ACKNOWLEDGEMENT

The authors thank ICAR, New Delhi and the Director, ICAR-NDRI, Karnal for all the facilities and permission to carry out this study. Sincere gratitude was expressed to Dr. Anupama Mukherjee for providing mentorship, data and computational assistance.

Informed consent

The data used in this study were obtained from record room of AGB Division of the Institute. Ethical clearance of the research work was obtained vide File No. 53/IAEC/24/01 dated 30.9.2024.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this article. No specific funding is available for the publication of this manuscript or funding the APC.

REFERENCES

- Abdollahi-Arpanahi, R., Gianola, D. and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics, Selection, Evolution: GSE*. **52(1)**: 12.
- Adnan, M., Alarood, A.A.S., Uddin, M.I. and Rehman, I.U. (2022). Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *Peer J. Computer Science*. **8**: e803.
- Bauman, D.E. (2000). Regulation of Nutrient Partitioning During Lactation: Homeostasis and Homeorhesis Revisited. In *Ruminant Physiology: Digestion, Metabolism, Growth and Reproduction*. Wallingford UK: CABI. pp. 311-328.
- Bauman, D.E. and Currie, W.B. (1980). Partitioning of nutrients during pregnancy and lactation: A review of mechanisms involving homeostasis and homeorhesis. *Journal of Dairy Science*. **63(9)**: 1514-1529.
- Bayer, P.E., Petereit, J., Danilevicz, M.F. anderson, R., Batley, J. and Edwards, D. (2021). The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome*. **14(3)**: e20112. <https://doi.org/10.1002/tpg2.20112>.
- Bhardwaj, S., Togla, O., Mumtaz, S., Yadav, N., Tiwari, J., Muansangi, L., Illa, S.K., Wani, Y.M., Mukherjee, S. and Mukherjee, A. (2023). Comparative assessment of the effective population size and linkage disequilibrium of Karan Fries cattle revealed viable population dynamics. *Animal Bioscience*. **37(5)**: 795-806.
- Bovo, M., Miki, A., Stefano, B., Daniele, T. and Patrizia, T. (2021). Random forest modelling of milk yield of dairy cows under heat stress conditions. *Animals*. **11(5)**: 1314.
- Breiman, L. (2001). Random forests. *Machine Learning*. **45(1)**: 5-32.

- Brito, L.F., Bédère, N., Douhard, F., Oliveira, H.R., Arnal, M., Peñagaricano, F. and Miglior, F. (2021). Genetic selection of high-yielding dairy cattle toward sustainable farming systems in a rapidly changing world. *Animal*. **15**: 100292.
- Chafai, N., Hayah, I., Houaga, I. and Badaoui, B. (2023). A review of machine learning models applied to genomic prediction in animal breeding. *Frontiers in Genetics*. **14**: 1150596.
- Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785-794.
- Dash, S.K., Gupta, A.K., Avtar, S., Chakravarty, A.K., Mohanty, T.K., Panmei, A. and Pushp, R.S. (2016). Genetic analysis of first lactation production and fertility traits in Karan Fries cattle. *Indian Journal of Animal Sciences*. **86**(10): 1159-1164.
- Dongre, V.B., Gandhi, R.S., Singh, A. and Ruhil, A.P. (2012). Comparative efficiency of artificial neural networks and multiple linear regression analysis for prediction of first lactation 305-day milk yield in Sahiwal cattle. *Livestock Science*. **147**(1-3): 192-197.
- Graesboll, K., Kirkeby, C., Nielsen, S.S. and Christiansen, L.E. (2015). Danish holsteins favor bull offspring: Biased Milk production as a function of fetal sex and calving difficulty. *Plos One*. **10**(4): e0124051.
- Grolemund, G. and Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*. **40**(3): 1-25.
- Gurnani, M., Sethi, R.K. and Nagarcenkar, R. (1986). Development of Karan Fries cattle at NDRI, Karnal. *Dairy Information Bulletin*. **3**(9): 1-2.
- Ilayaraja, I., Chitra, A., Vyas, J., Kumar, I., Muansangi, L., Singh, S.P., Gowane, G.R., Mukherjee, A. and Mukherjee, S. (2025). Improved genetic evaluation in Karan Fries cattle using multitrait single-step genomic best linear unbiased prediction method. *Animal*. **19**(12): 101692.
- International Committee for Animal Recording. (2020). Procedure 2-Computing of Accumulated Lactation Yield (Version January 2020).
- International Committee for Animal Recording. (2023). Section 2-Guidelines for Dairy Cattle Milk Recording (Version June 2023). <https://www.icar.org/Guidelines/02-Overview-Cattle-Milk-Recording.pdf>.
- Khan, S.A., Shukla, A.K., Yadav, S.K. and Vishwakarma, G.K. (2025). Machine learning models for analysis and prediction of optimal egg production. *Quality and Quantity*. pp 1-21.
- Knight, C.H. and Wilde, C.J. (1993). Mammary cell changes during pregnancy and lactation. *Livestock Production Science*. **35**(1-2): 3-19.
- Kokate, L.S., Singh, A., Banu, R., Gandhi, R.S., Chakravarty, A.K., Gupta, A.K. and Sachdeva, G.K. (2012). Sire evaluation based on predicted first lactation 305-day milk yield using monthly test day milk yield values of Karan Fries cattle. *Indian Journal of Animal Sciences*. **82**(12): 1539.
- Kumar, I., Mukherjee, A., Gowane, G.R., Kamboj, M.L., Malhotra, R.K. and Mukherjee, S. (2025). Assessing the importance of linear type traits and their association with functional, production and reproduction traits in Karan Fries cattle through multi-trait bayesian method. *Acta Agriculturae Scandinavica, Section A-Animal Science*. **74**(1): 1-12.
- Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. *R News*. **2**(3): 18-22.
- Meinert, T.R. and Norman, H.D. (1998). Merit of outliers for milk yield as indicators of accuracy of genetic evaluations of sires. *Journal of Dairy Science*. **81**(11): 2951-2955.
- Moritz, S. and Bartz-Beielstein, T. (2017). Impute TS: Time series missing value imputation in R. *The R Journal*. **9**: 207.
- Nayeri, S., Sargolzaei, M. and Tulpan, D. (2019). A review of traditional and machine learning methods applied to animal breeding. *Animal Health Research Reviews*. **20**(1): 31-46.
- Rana, E., Gupta, A.K., Ruhil, A.P., Mall, S. and Ashokan, M. (2024). Evaluating peak milk yield as a predictor for first lactation 305-day milk yield in murrah buffaloes. *Indian Journal of Animal Research*. **60**(5): 775-780. doi: 10.18805/IJAR.B-5311.
- Sahoo, S.K., Singh, A., Ambhore, G.S., Dash, S.K. and Dubey, P.P. (2019). Comparative efficiency of different multiple linear regression prediction equations of first lactation 305-day milk yield for sire evaluation in Murrah buffaloes. *Indian Journal of Animal Research*. **53**(10): 1287-1291. doi: 10.18805/ijar.B-3661.
- Shmueli, G. and Koppius, O.R. (2011). Predictive analytics in information systems research. *MIS Quarterly*. **35**: 55-572.
- Silverman, B.W. (2018). Density Estimation for Statistics and Data Analysis. Routledge, New York.
- Singh, A., Singh, A., Singh, M., Gupta, A.K. and Dash, S.K. (2016). Estimation of genetic parameters of first lactation 305-day and monthly test-day milk yields in Karan Fries cattle. *Indian Journal of Animal Sciences*. **86**(4): 436-440.
- Singh, M., Prakash, V., Dash, S., Dixit, S. and Gupta, A.K. (2021). Multiple linear regression analysis using monthly test day milk yield predicting the first lactation production performance for sire evaluation in murrah buffaloes. *Indian Journal of Dairy Science*. **74**(1): 89-94.
- Streefland, G.J., Herrema, F. and Martini, M. (2023). A Gradient Boosting model to predict the milk production. *Smart Agricultural Technology*. **6**: 100302.
- Tırınk, C., Piwczynski, D., Kolenda, M. and Önder, H. (2023). Estimation of body weight based on biometric measurements by using random forest regression, support vector regression and CART algorithms. *Animals*. **13**(5): 798.
- Topuz, D. and Tekgöz, S. (2025). Hybrid ensemble model for lactation milk yield prediction of holstein cows. *Kafkas Üniversitesi Veteriner Fakültesi Dergisi*. **31**(5): 603-611.
- Upadhyay, A., Chakravarty, A.K., Dash, S., Shivahre, P.R. and Gonge, D.S. (2024). Effect of environmental factors on milk production traits and energy-corrected milk yield in karan fries cattle. *International Journal of Bio-Resource and Stress Management*. **15**(11): 1-8.
- Vyas, J., Kumar, I., Chitra, A. et al. (2025). Longitudinal genomic and pedigree inbreeding metrics reveal inbreeding depression in milk quality traits of Sahiwal cattle maintained in organised herd. *Nucleus*. doi: 10.1007/s13237-025-00611-9.
- Wickham, H. (2016). Data Analysis. In ggplot2: Elegant Graphics for Data Analysis. Cham: Springer international publishing. pp. 189-201.

- Willmott, C.J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*. **30**: 79-82.
- Wood, P.D.P. (1967). Algebraic model of the lactation curve in cattle. *Nature*. **216(5111)**: 164-165.
- Worku, D., Gowane, G.R., Kumar, R., Joshi, P., Gupta, I.D. and Verma, A. (2021). Estimation of genetic parameters for production and reproductive traits in indian Karan-Fries cattle using multi-trait Bayesian approach. *Tropical Animal Health and Production*. **53(3)**: 369.
- Yadav, A.K., Mukherjee, A., Mumtaz, S., Kumar, S. and Pathak, P. (2017). Effect of nongenetic factors on milk yield and milk constituents in Karan Fries cattle. *Ruminant Science*. **6(2)**: 225-230.