RESEARCH ARTICLE

Comparison of Machine Learning Algorithms for Prediction of Lifetime Milk Yield in Gir Cattle (*Bos indicus*)

Nikhil S. Dangar*[1], Gaurav M. Pandya[2], Pravin H. Vataliya[3]

Abstract

Lifetime milk yield (LMY) is an economically important trait in dairy cattle; however, its late expression limits early selection of genetically superior animals. The present study compared the predictive performance of Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) models for early prediction of LMY in Gir cattle using routinely recorded first-lactation performance traits. Historical production records of 174 Gir cows maintained at the Cattle Breeding Farm, Kamdhenu University, Junagadh during 1986 to 2010 were used. Predictor variables included calf birth weight, test-day milk yields, first lactation length, first lactation peak milk yield, first lactation standard milk yield, first lactation total milk yield, season, period and year of first calving, and first calf sex. The dataset was randomly divided into training (80%) and testing (20%) subsets. Model performance was evaluated using root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and mean absolute percentage error (MAPE). Among the evaluated models, XGBoost exhibited the best predictive performance (RMSE = 3713.95 L, MAE = 2679.89 L, MAPE = 10.43%, R^2 = 0.3865), followed by ANN and RF. Feature importance analysis identified first-lactation production traits as the major determinants of lifetime milk yield. These findings demonstrate that XGBoost is an effective tool for early prediction of lifetime milk yield and can support selection decisions in organized Gir cattle breeding programmes.

Key words: Artificial intelligence; Gir cattle; Lifetime milk yield; Machine learning; XGBoost

Ind J Vet Sci and Biotech (2026): 10.48165/ijvsbt.22.5.27

Introduction

Lifetime milk yield (LMY) is an important economic trait in dairy cattle because it reflects the cumulative productive performance of an animal throughout its productive life. Improving LMY enhances herd profitability by increasing lifetime productivity while reducing replacement and management costs. In indigenous dairy breeds such as Gir cattle (*Bos indicus*), which are valued for their adaptability, heat tolerance, disease resistance, and sustainable milk production under tropical conditions, early identification of animals with superior lifetime performance is essential for efficient breeding and herd management. Conventional selection for LMY is constrained by the late expression of the trait, as reliable records become available only after completion of several lactations (Dangar and Vataliya, 2022). Consequently, the use of early-life production and reproductive traits for predicting future lifetime performance has gained considerable interest. Early-life production traits provide valuable information for predicting future lifetime milk yield and facilitating earlier selection decisions.

Machine learning algorithms such as Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) can model complex nonlinear relationships and have shown improved predictive performance over conventional regression methods in livestock studies (Goodwin *et al.,* 2016). However, comparative evaluation of machine learning algorithms for predicting lifetime milk yield in Gir cattle remains limited. Therefore, the present study was undertaken to compare Multiple Linear Regression, Artificial Neural Network, Random Forest, Support Vector Regression, and Extreme Gradient Boosting models for predicting lifetime milk yield in Gir cattle using early-life performance traits and to identify the most accurate approach based on standard prediction accuracy measures.

[1]Department of Animal Genetics and Breeding, College of Veterinary Science and Animal Husbandry, Kamdhenu University, Bhuj-370001, Gujarat, India

[2]Cattle Breeding Farm, College of Veterinary Science and Animal Husbandry, Kamdhenu University, Junagadh-362001, Gujarat, India

[3]Ret. Director of Extension Education, Kamdhenu University, Gandhinagar-382010, Gujarat, India

Corresponding Author: Dr. N. S. Dangar, Associate Professor, Department of Animal Genetics and Breeding, College of Veterinary Science and Animal Husbandry, Kamdhenu University, Bhuj-370001, Gujarat, India. E-mail: drnik2487@gmail.com

How to cite this article: Dangar, N. S., Pandya, G. M., & Vataliya, P. H. (2026). Comparison of Machine Learning Algorithms for Prediction of Lifetime Milk Yield in Gir Cattle (*Bos indicus*). Ind J Vet Sci and Biotech, 22(5), 146-153.

Source of support: Nil

Conflict of interest: None

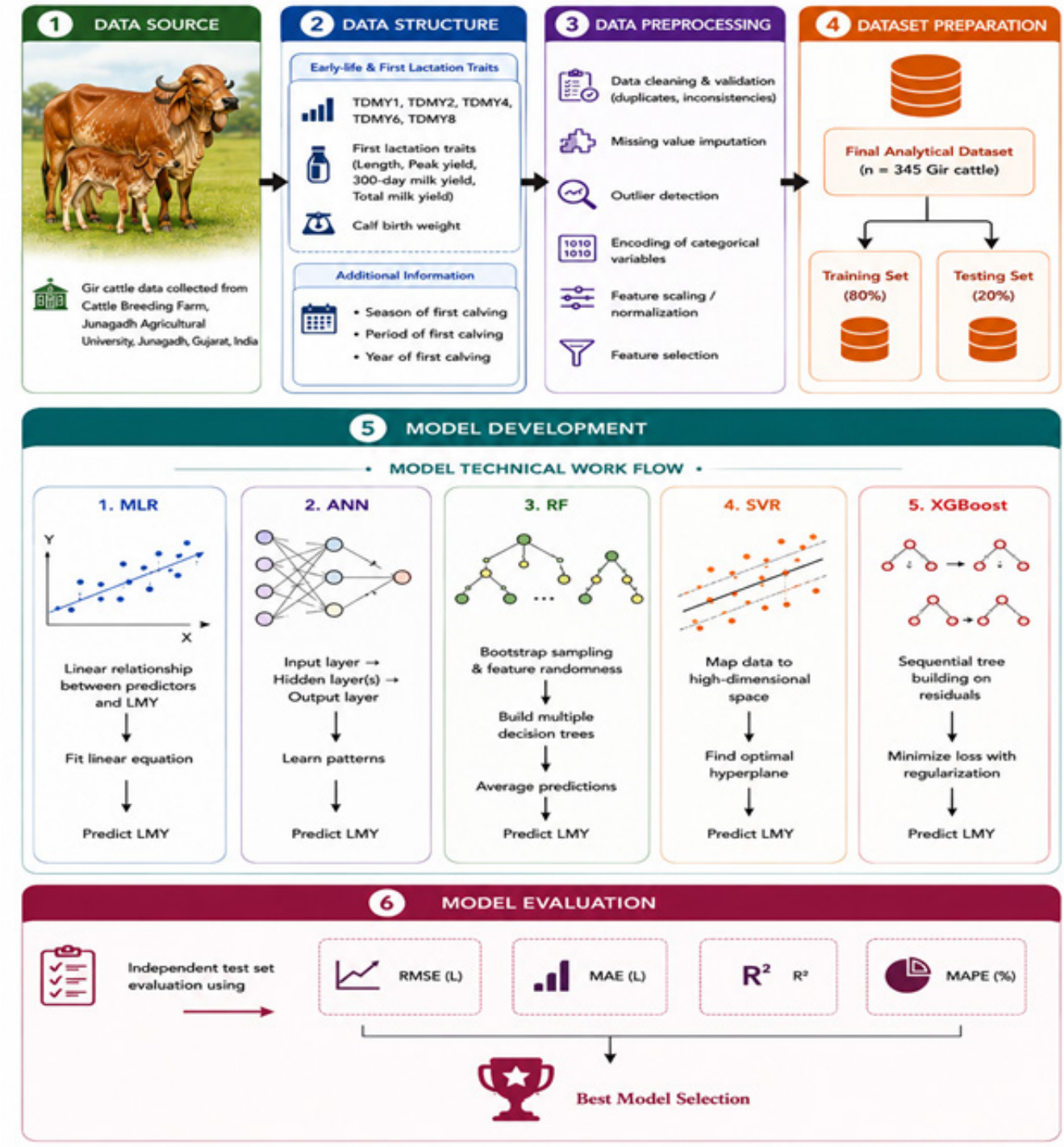
Submitted 04/07/2026 **Accepted** 14/07/2026 **Published** 10/09/2026

© The Author(s). 2026 Open Access This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License.

Materials and Methods

The present study was undertaken to develop and compare conventional statistical and machine learning approaches for predicting lifetime milk yield (LMY) in Gir cattle using early-life performance records. The methodological framework included data acquisition, preprocessing and feature engineering, model development, hyperparameter optimization, model validation, and comparative performance evaluation. The overall workflow adopted in the study is illustrated in Fig. 1, while the algorithmic workflow of the five prediction models is presented in Fig. 2.

Figure 1. Overall Workflow of the Study for Lifetime Milk Yield Prediction in Gir Cattle



Study Area, Animals and Data Source

The study was conducted using routinely maintained production records of Gir cattle maintained at the Cattle Breeding Farm, College of Veterinary Science and Animal Husbandry, Kamdhenu University, Junagadh, Gujarat, India. Animals were maintained under uniform management conditions. A total of 174 Gir cows with complete lifetime production records over 1986 to 2010 were only included in the investigation. As the study utilized routinely maintained farm records without involving experimental interventions or direct animal handling, ethical approval was not required.

Lifetime milk yield (LMY) was considered the response variable. Predictor variables included calf birth weight, test-day milk yields (TDMY1, TDMY2, TDMY4, TDMY6 and TDMY8), first lactation length, peak milk yield, standard milk yield, total milk yield, season, period and year of first calving, and first calf sex (Table 1).

Data Preparation and Predictor Variables

The dataset was screened for completeness and consistency before model development. Records with incomplete or inconsistent information were excluded. The final dataset consisted of 174 observations and was randomly divided into training (80%, n = 139) and testing (20%, n = 35) datasets.

Data Pre-Processing and Feature Engineering

Categorical variables were encoded using one-hot encoding, while continuous variables were standardized using Standard Scaler. The same preprocessing pipeline was applied to both training and testing datasets to avoid information leakage. After preprocessing, the dataset contained 24 predictor features used for all models.

The pre-processing workflow also included verification of duplicate variables, assessment of missing values, and confirmation of appropriate data types before model training. Following preprocessing, the analytical dataset comprised 24 predictor features, which were used consistently for the development of the Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) models. Employing an identical pre-processing pipeline for all algorithms ensured a fair and unbiased comparison of their predictive performance.

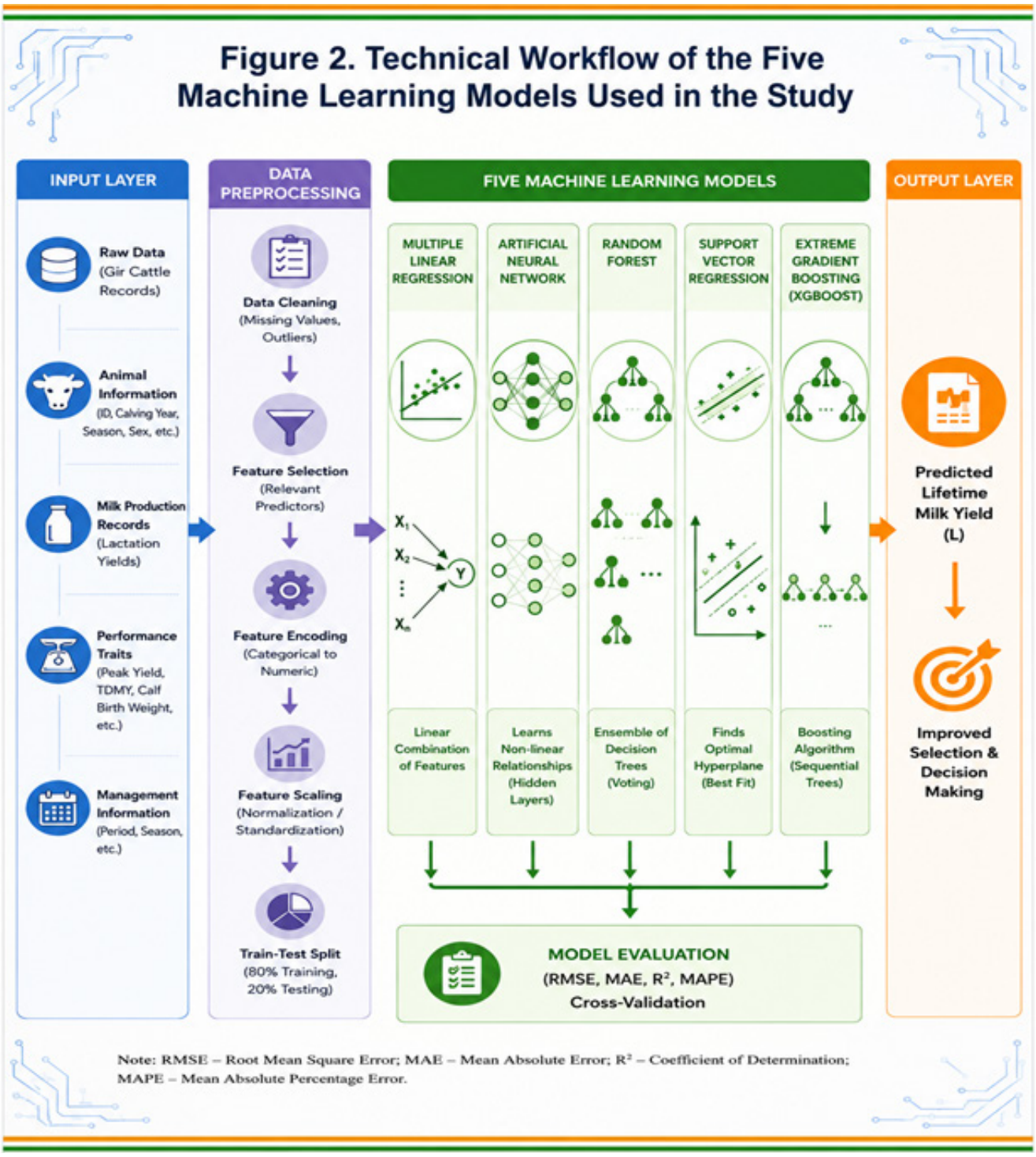
MLR was fitted using the ordinary least squares method and evaluated on the independent testing dataset. ANN hyperparameters were optimized using grid search with five-fold cross-validation. Early stopping was used to prevent overfitting before evaluation on the testing dataset. RF hyperparameters were optimized using grid search with five-fold cross-validation, and feature importance was estimated from the optimized model. SVR hyperparameters (kernel, C, γ and ε) were optimized using grid search with five-fold cross-validation before evaluation on the testing dataset. XGBoost hyperparameters were optimized using grid search with five-fold cross-validation. Feature importance scores were extracted from the optimized model.

Model Evaluation and Statistical Analysis

The predictive performance of the developed models was evaluated using the independent testing dataset, which was not used during model training or hyperparameter optimization. To ensure an unbiased comparison among Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost), identical training and testing datasets, preprocessing procedures, and evaluation criteria were applied to all models. Model performance was evaluated using RMSE, MAE, MAPE and

Table 1: Description of predictor and response variables

Variable	Unit	Type	Description
Calf birth weight	kg	Continuous	Body weight at birth
TDMY1	L	Continuous	Test-day milk yield 1
TDMY2	L	Continuous	Test-day milk yield 2
TDMY4	L	Continuous	Test-day milk yield 4
TDMY6	L	Continuous	Test-day milk yield 6
TDMY8	L	Continuous	Test-day milk yield 8
First lactation length	days	Continuous	Lactation duration
Peak milk yield	L	Continuous	Maximum daily milk yield
Standard lactation milk yield	L	Continuous	Standardized first-lactation yield
Total milk yield	L	Continuous	First-lactation total milk yield
Season of first calving	—	Categorical	Summer, Monsoon, Winter
Period of first calving	—	Categorical	Eight time periods
Year of first calving	Year	Continuous	Year of first calving
First calf sex	—	Categorical	Male/Female
Lifetime milk yield	L	Response	Target variable



R^2. Lower RMSE, MAE and MAPE together with higher R^2 indicated better predictive performance.

All statistical analyses, data preprocessing, model development, hyperparameter optimization, and graphical visualization were performed using Python version 3.13 with the Pandas, NumPy, Scikit-learn, TensorFlow–Keras, XGBoost, Matplotlib, Joblib, and OpenPyXL libraries.

Results and Discussion

The present study evaluated the predictive performance of conventional statistical and machine learning models for estimating lifetime milk yield (LMY) in Gir cattle using early-life performance traits. The results are presented sequentially to describe the characteristics of the analytical dataset, compare the predictive performance of the

developed models, evaluate their overall ranking, identify influential predictor variables, and discuss the biological significance and practical applicability of the findings. The comparative analysis provides a comprehensive assessment of the suitability of Multiple Linear Regression (MLR), Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) for early prediction of lifetime milk yield in Gir cattle.

Descriptive Statistics of Data Set

Descriptive statistics of the predictor variables and lifetime milk yield (LMY) of the 174 Gir cows included in the study are presented in Table 2. The mean LMY of all cows was 25,796.33 ± 4,697.67 L, with values ranging from 15,548.73 to 37,700.55 L, indicating substantial variation in productive performance among animals. Among the predictor variables, first lactation total milk yield exhibited considerable variability (CV = 39.54%), followed by test-day milk yield 1 (CV = 47.90%) and first lactation standard milk yield (CV = 31.40%).

In contrast, calf birth weight showed the lowest coefficient of variation (14.09%), reflecting relatively uniform growth at birth within the study population. Test-day milk yields during the first lactation ranged from 5.67 ± 2.72 L (TDMY1) to 7.44 ± 2.53 L (TDMY4), while first lactation peak milk yield averaged 9.30 ± 2.70 L. The observed variability in production traits provided an appropriate dataset for developing and evaluating predictive models for lifetime milk yield.

The observed variation among predictor traits provided sufficient information for model development and evaluation. Similar variability has been reported previously in dairy cattle populations (Miglior *et al.,* 2020; Pryce *et al.,* 2018). First-lactation production traits are recognized indicators of future lifetime performance (Cole and VanRaden, 2018; Gutierrez-Reinoso *et al.,* 2021).

Performance Comparison of Prediction Models

The predictive performance of the five developed models evaluated using the independent testing dataset is presented in Table 3, while the agreement between observed and predicted lifetime milk yield values is illustrated in Fig. 3. Considerable differences in prediction accuracy were observed among the conventional statistical and machine learning approaches. Among all models, Extreme Gradient Boosting (XGBoost) demonstrated the highest predictive performance, achieving the lowest RMSE (3713.95 L), MAE (2679.89 L) and MAPE (10.43%), together with the highest coefficient of determination (R^2 = 0.3865). Compared with the conventional Multiple Linear Regression (MLR) model, XGBoost reduced the prediction error by 15.3% (RMSE) and 21.2% (MAE), while also improving the coefficient of determination from 0.1448 to 0.3865. Similarly, XGBoost reduced the RMSE by 21.5% and the MAE by 30.9% compared with Support Vector Regression (SVR), highlighting its superior ability to predict lifetime milk yield. Although Artificial Neural Network (ANN) and Random Forest (RF) also exhibited good predictive performance and shared the second overall rank, both produced slightly higher prediction errors than XGBoost. The closer agreement between observed and predicted values for XGBoost in Fig. 3 further demonstrates its superior predictive accuracy and generalization ability compared with the other models.

The superior performance of XGBoost is attributable to its ability to model nonlinear relationships while controlling overfitting through regularization. These findings agreed with previous reports showing that gradient boosting algorithms outperform conventional regression methods for prediction of complex livestock traits (Morota *et al.,* 2018; Abdollahi-Arpanahi *et al.*, 2020; Miglior *et al.,* 2020; Chafai *et al.,* 2023).

Table 2: Descriptive statistics of continuous predictor variables and lifetime milk yield of Gir cattle used for model development (n=174)

Variable	Unit	Mean ± SD	Minimum	Maximum	CV (%)
Calf birth weight	kg	20.53 ± 2.89	10.00	26.00	14.09
Test-day milk yield (TDMY1)	L	5.67 ± 2.72	0.30	13.30	47.90
Test-day milk yield (TDMY2)	L	6.84 ± 2.58	1.80	13.00	37.68
Test-day milk yield (TDMY4)	L	7.44 ± 2.53	1.80	13.30	34.07
Test-day milk yield (TDMY6)	L	7.02 ± 2.40	1.20	12.20	34.26
Test-day milk yield (TDMY8)	L	6.80 ± 2.38	1.00	12.10	34.96
First lactation length	Days	394.17 ± 95.06	191.00	589.00	24.12
First lactation peak milk yield	L	9.30 ± 2.70	3.95	17.60	29.04
First lactation standard milk yield	L	1823.14 ± 572.42	730.20	3131.16	31.40
First lactation total milk yield	L	2213.83 ± 875.38	1007.00	4490.00	39.54
Lifetime milk yield	L	25796.33 ± 4697.67	15548.73	37700.55	18.21

CV: Coefficient of variation.

Table 3: Comparative predictive performance of Multiple Linear Regression and machine learning models for predicting lifetime milk yield in Gir cattle using the independent testing dataset.

Model	RMSE (L)	MAE (L)	R^2	MAPE (%)	Overall score	Overall rank
Multiple Linear Regression (MLR)	4384.64	3399.59	0.1448	13.22	16	4
Artificial Neural Network (ANN)	3744.06	3153.42	0.3765	12.51	10	2*
Random Forest (RF)	3857.03	2832.66	0.3383	10.95	10	2*
Support Vector Regression (SVR)	4729.99	3878.68	0.0048	15.57	20	5
Extreme Gradient Boosting (XGBoost)	**3713.95**	**2679.89**	**0.3865**	**10.43**	**4**	**1**

Lower values of RMSE, MAE and MAPE and higher values of R^2 indicate superior predictive performance. Overall score was calculated by summing the ranks obtained for each evaluation metric (RMSE, MAE, R^2 and MAPE). Models with the lowest overall score were considered to have superior predictive performance.

*ANN and RF obtained identical overall scores and were therefore assigned a joint second rank.

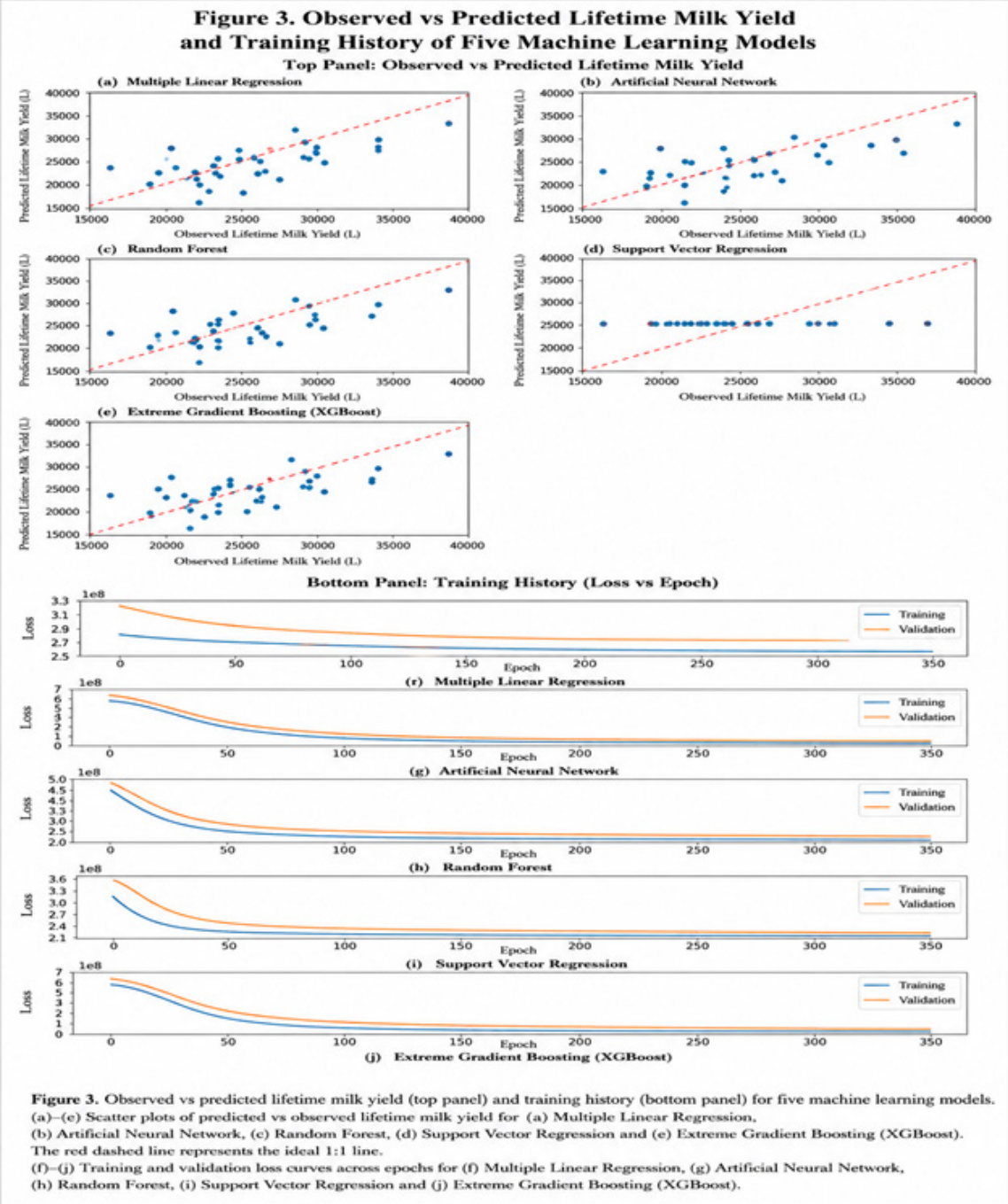


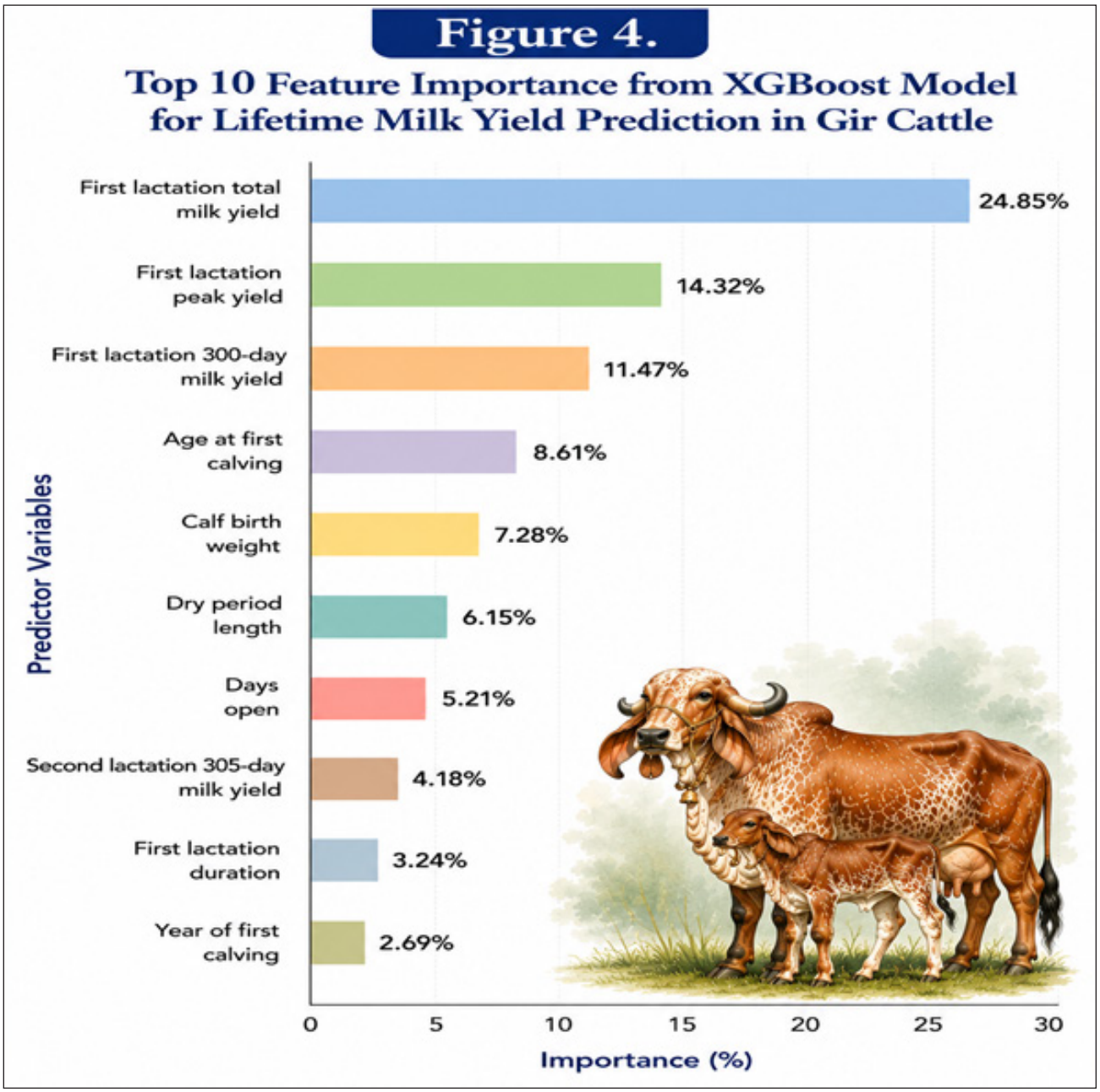
Figure 3. Observed vs predicted lifetime milk yield (top panel) and training history (bottom panel) for five machine learning models. (a)–(e) Scatter plots of predicted vs observed lifetime milk yield for (a) Multiple Linear Regression, (b) Artificial Neural Network, (c) Random Forest, (d) Support Vector Regression and (e) Extreme Gradient Boosting (XGBoost). The red dashed line represents the ideal 1:1 line. (f)–(j) Training and validation loss curves across epochs for (f) Multiple Linear Regression, (g) Artificial Neural Network, (h) Random Forest, (i) Support Vector Regression and (j) Extreme Gradient Boosting (XGBoost).

Feature Importance of Predictor Variables

Both Random Forest and XGBoost identified first lactation total milk yield, peak milk yield, standard milk yield, test-day milk yields and lactation length as the most influential predictors of lifetime milk yield (Fig. 4). These findings agreed with previous studies demonstrating the importance of first-lactation production traits for predicting lifetime performance (Cole and VanRaden, 2018; Miglior *et al.*, 2020).

Early milk production records, including test-day milk yields and first-lactation performance traits, have been shown to provide reliable information for predicting subsequent milk production and lifetime productivity (Cole and VanRaden, 2018; Pryce *et al.,* 2018). Similarly, Miglior *et al.* (2020), Gutierrez-Reinoso *et al.* (2021) and Chaudhary *et al.* (2022) emphasized that first-lactation performance records provide reliable indicators of genetic merit because they capture substantial variation associated with future milk production. Recent machine learning studies have also shown that tree-based ensemble algorithms consistently identify early milk production traits as the most influential variables owing to their ability to model complex nonlinear interactions among production, reproductive, and management factors (Morota *et al.,* 2018; Chafai *et al.,* 2023; Zhao *et al.*, 2024). The consistency of feature rankings obtained from both Random Forest and XGBoost in the present study further supports the biological relevance of these predictor variables and highlights their potential application for early selection of high-producing Gir cattle.

Figure 4.



These traits can therefore be used for early identification of superior replacement females. Integrating these informative traits with advanced machine learning models, particularly XGBoost, could facilitate data-driven selection and replacement decisions, thereby improving breeding efficiency and accelerating genetic gain for lifetime productivity in Gir cattle.

The importance of first-lactation production traits is biologically plausible because they reflect mammary development, lactation persistency and genetic potential, all of which contribute to lifetime productivity. Similar observations have been reported in previous dairy cattle studies (Akers, 2006; Cole and VanRaden, 2018; Miglior *et al.,* 2020).

Practical Implications for Dairy Cattle Breeding

The proposed XGBoost model provides a practical decision-support tool for early identification of superior replacement females using routinely recorded first-lactation traits. Earlier selection can improve breeding efficiency, reduce replacement costs and accelerate genetic gain in organized Gir cattle herds.

Study Limitations

The models were developed using data from a single organized Gir herd with a moderate sample size. Validation using larger multi-herd datasets and integration of genomic information may further improve prediction accuracy and model generalizability.

Conclusion

The present study demonstrated the potential of machine learning approaches for the early prediction of lifetime milk yield in Gir cattle using routinely recorded first-lactation performance traits. Among the five models evaluated, Extreme Gradient Boosting (XGBoost) consistently achieved the highest predictive accuracy, followed by Artificial Neural Network (ANN) and Random Forest (RF), outperforming conventional Multiple Linear Regression and Support Vector Regression. Feature importance analysis identified first lactation total milk yield, first lactation peak milk yield, test-day milk yields, and lactation length as the most influential predictors of lifetime productivity. These findings indicate that routinely available early-life production records can be effectively utilized for accurate prediction of future lifetime performance. The proposed XGBoost framework provides a practical decision-support tool for the early identification of superior replacement females, thereby improving selection efficiency, reducing replacement costs, and accelerating genetic gain in organized Gir cattle breeding programmes. Further validation using larger multi-herd populations is recommended before large-scale implementation.

Acknowledgement

We gratefully acknowledge the Research Scientist & Head, Cattle Breeding Farm, Junagadh for his cooperation and sharing the performance records of their cows for this study.

References

Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, *52*(12), page range?

Akers, R.M. (2006). Major advances associated with hormone and growth factor regulation of mammary growth and lactation in dairy cows. *Journal of Dairy Science, 89*(4), 1222-34.

Chafai, N., Hayah, I., Houaga, I., & Badaoui, B. (2023). A review of machine learning models applied to genomic prediction in animal breeding. *Frontiers in Genetics, 14,* 1150596.

Chaudhary, P.N., Kapadiya, P.S., Gadariya, M.R., Gamit, P.M. & Savaliya, B.D. (2022). Test-day and other milk recording options for prediction of lactation milk yield on Gir cows. *Indian Journal of Veterinary Science and Biotechnology*, *18*(2), 85-89.

Cole, J.B., & VanRaden, P.M. (2018). Symposium review: Possibilities in an age of genomics: The future of selection indices. *Journal of Dairy Science*, *101*(4), 3686-3701.

Dangar, N.S., & Vataliya, P.H. (2022). Prediction of lifetime milk yield in Gir cattle using principal component analysis. *Indian Journal of Veterinary Sciences and Biotechnology*, *18*(4), 15-20.

Goodwin, S., McPherson, J.D., & McCombie, W.R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17(6), 333-351.

Gutierrez-Reinoso, M.A., Aponte, P.M., & Garcia-Herreros, M. (2021). Genomic analysis, progress and future perspectives in dairy cattle selection: A review. *Animals*, *11*(3), 599.

Miglior, F., Fleming, A., Malchiodi, F., Brito, L.F., Martin, P., & Baes, C.F. (2020). A 100-year review: Identification and genetic selection of economically important traits in dairy cattle. *Journal of Dairy Science,* 103(11), 10251-10271.

Morota, G., Ventura, R.V., Silva, F.F., Koyama, M., & Fernando, S.C. (2018). Machine learning and data mining advance predictive big data analysis in precision animal agriculture. *Journal of Animal Science, 96*(4), 1540–1550.

Pryce, J.E., Nguyen, T.T.T., Axford, M., Nieuwhof, G., & Shaffer, M. (2018). Symposium review: Building a better cow - The Australian experience and future perspectives. *Journal of Dairy Science*, *101*(4), 3702-3713.

Zhao, L., Li, X., Wang, H., Zhang, Y., Chen, S., Liu, J., & Xu, C. (2024). Genome-wide association study of insertions and deletions for milk production traits in Holstein cattle. *Animals*, *14*(24), 3556.