



Journal of Frontiers in Multidisciplinary Research

Explainable Machine Learning for Predicting Semen Quality Classes in Artificial Insemination Bulls Using Longitudinal Feature Engineering: A Decision-Support Framework for Smart Livestock Reproduction

Dorothy Assan ^{1*}, Luke Mapiliyao ², Abbegail Dube ³, Never Assan ⁴

¹ Department of Biotechnology and Bioinformatics, Faculty of Science and Technology, Midlands State University, Private Bag 9055, Senga Road, Gweru, Zimbabwe

² Ministry of Lands and Rural Development, Mangwe District, Box 87, Plumtree, Zimbabwe

³ Mechanization and Irrigation Development, Esigodini Agricultural College, Ministry of Agriculture, Private Bag 5808, Bulawayo, Zimbabwe

⁴ Faculty of Agriculture, Zimbabwe Open University, Bulawayo Regional Campus, Bulawayo, Zimbabwe

* Corresponding Author: **Dorothy Assan**

Article Info

E-ISSN: 3050-9726

P-ISSN: 3050-9718

Volume: 07

Issue: 02

Received: 12-07-2026

Accepted: 14-08-2026

Published: 16-09-2026

Page No: 298-305

Abstract

Background: Routine semen records contain repeated observations from the same bull, but their temporal structure is often underused when predicting subsequent semen quality. This study developed an explainability-oriented longitudinal prediction framework in which each predictor was traceable to reproductive information available before the target collection and compared transparent historical features with fixed machine-learning models.

Methods: A retrospective longitudinal dataset comprising 57 semen collection records from seven artificial insemination bulls was subjected to predefined quality control. Eight observations were excluded and the first collection for each bull was unavailable for lag-based prediction, yielding 42 modelling observations. Semen quality was defined as high quality when total motility was $\geq 70\%$ and lower quality when total motility was $< 70\%$. Longitudinal feature engineering generated immediate, rolling and cumulative historical reproductive predictors using only information available before each target collection. Simple historical indicators and Logistic Regression, Random Forest and XGBoost models were evaluated using Leave-One-Bull-Out Cross-Validation (LOBOCV). Explainability was addressed through transparent feature definitions, temporal traceability and comparison with parsimonious models; no post-hoc explainability method was used. Matthews correlation coefficient (MCC) was prespecified as the primary model-selection criterion.

Results: The historical mean of previous total motility produced the strongest verified simple prediction, with 71.4% accuracy, 67.5% balanced accuracy and $MCC = 0.391$. Previous total motility achieved 66.7% accuracy and $MCC = 0.296$, whereas the three-collection rolling mean achieved 64.3% accuracy, 59.4% balanced accuracy and $MCC = 0.211$. Logistic Regression using previous total motility achieved $MCC = 0.352$, and adding previous progressive motility did not improve MCC. The combination of previous total motility and historical mean total motility produced $MCC = 0.060$. Under the fixed longitudinal machine-learning specifications, enhanced Random Forest and XGBoost models achieved MCC values of -0.247 and -0.224 , respectively, while enhanced Logistic Regression achieved $MCC = -0.407$. **Conclusion:** Longitudinal feature engineering extracted a measurable predictive signal from historical reproductive records, but the strongest verified performance was obtained from transparent cumulative historical information rather than from increasingly complex machine-learning models. The findings support an interpretable decision-support pathway in which routinely accumulated semen records are transformed into temporally valid bull-level performance profiles, while emphasising that validation across substantially larger numbers of independent bulls is required before operational deployment.

DOI: <https://doi.org/10.54660/JFMR.2026.7.2.298-305>

Keywords: explainable machine learning, artificial insemination bulls, semen quality, longitudinal feature engineering, decision support, smart livestock reproduction, total motility

1. Introduction

Artificial insemination (AI) is an important reproductive technology in livestock production because it enables the dissemination of genetic material from selected males and places semen quality at the centre of breeding-programme efficiency. Routine semen evaluation provides information on characteristics such as total motility, progressive motility, semen volume and sperm concentration, which are used to assess ejaculates before processing and use. Variation among ejaculates from the same bull means that a single collection may not fully represent the bull's longitudinal semen-performance profile (Harstine *et al.*, 2018;

Ngoda *et al.*, 2023; Koçyiğit *et al.*, 2026) [5, 8, 10].

Routine AI operations generate repeated semen observations from the same bull, creating longitudinal records that can potentially be transformed into predictive features for reproductive decision support. Longitudinal feature engineering provides a structured way of converting past records into immediate, rolling and cumulative representations of historical performance. Such features are attractive because their biological and temporal origins can be explicitly traced, allowing prediction to remain linked to information that would actually have been available at the time of decision making.

Longitudinal prediction nevertheless creates two important methodological challenges. First, successive ejaculates from an individual bull are not statistically independent. If observations from the same bull are divided between model-development and validation datasets, a model may exploit animal-specific characteristics and produce apparently strong performance when evaluated on another collection from the same bull. Second, temporal information leakage can occur when predictors incorporate measurements from the target collection or observations recorded after the prediction point. Both problems can lead to optimistic estimates of predictive performance (Roberts *et al.*, 2017) [12].

A credible smart-livestock decision-support framework therefore requires both temporal validity and model transparency. Predictors should be constructed exclusively from information available before the target collection, while observations from the same bull should remain within either the training or validation portion of each fold. Leave-One-Bull-Out Cross-Validation (LOBOCV) provides a stringent approach because each bull is completely withheld in turn, making the evaluation more closely aligned with generalisation to previously unseen animals.

Explainability is particularly important in small livestock datasets, where apparent gains from algorithmic complexity may be difficult to distinguish from model instability. In the present framework, explainability is approached by design: historical features are explicitly defined, their temporal provenance is known, and their predictive contribution can be compared with parsimonious models. This framing differs from post-hoc explainability approaches such as SHAP or LIME and does not claim that such post-hoc methods were applied in the present analysis.

Historical semen performance can be represented at different levels of temporal aggregation. Previous total motility may capture the most recent state of performance, a rolling mean may represent recent collection history, and a cumulative historical mean may represent longer-term performance. Comparing these transparent indicators provides an interpretable benchmark before introducing more complex statistical or machine-learning models.

Machine-learning methods such as Random Forest and Extreme Gradient Boosting (XGBoost) can represent nonlinear relationships and interactions among predictors. However, increased flexibility does not necessarily improve generalisation when the number of independent animals is small and repeated observations dominate the dataset. A central practical question is therefore not whether a complex algorithm can fit the available records, but whether longitudinally engineered features provide demonstrable predictive value and whether the resulting approach remains sufficiently transparent for reproductive decision support (Chen & Guestrin, 2016; Hürland *et al.*, 2023) [2, 7].

Against this background, the present study developed an explainability-oriented longitudinal prediction framework for subsequent semen-quality classification in artificial-insemination bulls. The objectives were to: (i) engineer temporally valid longitudinal features from historical semen records while preventing information leakage; (ii) compare transparent historical indicators with Logistic Regression, Random Forest and XGBoost under Leave-One-Bull-Out Cross-Validation; (iii) determine whether increasing algorithmic complexity provided additional predictive value beyond parsimonious historical representations of semen performance; and (iv) translate the verified findings into a conceptual decision-support framework for future smart livestock reproduction.

2. Materials and Methods

2.1. Ethical Approval

All procedures involving animals were conducted in accordance with the ethical standards for animal experimentation universal standards. Animal handling and data collection procedures were performed to minimise stress and ensure animal welfare throughout the study.

2.2. Study Site

This research was conducted at the Matopos Research Station in Bulawayo, Zimbabwe, located at 22.23°S latitude and 31.30°E longitude. The region experiences a dry season from April to October and a rainy season from November to March, with an average annual rainfall of less than 446.8 mm (Assan 2023) [1]. The area is characterized by high temperatures, ranging from 21.6°C to 11.4°C during the hottest months, and low rainfall (<450 mm) (Hagreveas *et al.*, 2004; Homann *et al.*, 2007) [4, 6]. The research site comprises rangelands with sweet veld vegetation, which provides high nutritional value suitable for sustaining ruminants (Ward *et al.*, 1979; Ncube, 2005; Van Rooyen *et al.*, 2007) [9, 13, 14].

2.3. Study Dataset and Longitudinal Data Structure

The original dataset comprised 57 semen collection records from seven bulls. Each record represented an individual semen collection event and included bull descriptors, collection information and routinely recorded semen variables. Records were ordered chronologically within bull so that observations preceding a target collection could be distinguished from the target and subsequent observations. Bull identity was retained only as the grouping unit for animal-level validation.

2.4. Data Quality Control and Modelling Population

A predefined quality-control procedure was applied to the original 57 records. Eight observations were excluded, leaving 49 biologically valid records. Because the prediction task required at least one previous semen collection, the first available collection for each bull could not contribute to lag-based prediction. The final longitudinal modelling population therefore comprised 42 observations from seven bulls. The modelling outcome contained 25 high-quality observations (59.5%) and 17 lower-quality observations (40.5%).

2.5. Definition of the Prediction Outcome

Semen quality was defined from total sperm motility (TM). Collections with TM $\geq 70\%$ were classified as higher total-motility class (Class 1), while collections with TM $< 70\%$ were classified as lower total-motility class (Class 0). The

target TM measurement was used only to define the observed outcome and was never used as a predictor of its own classification.

2.6. Temporal Ordering and Leakage Prevention

All records were chronologically ordered within bull before feature construction. For a target collection at time t , predictors were restricted to observations at $t-1$, $t-2$ and earlier. No current-collection semen measurement and no future observation was permitted to contribute to predictor construction. This temporal boundary was applied consistently to simple benchmarks and machine-learning models.

2.7. Longitudinal Predictor Construction

The longitudinal predictors represented immediate, short-term and cumulative historical reproductive performance. Previous total motility (Previous TM) and previous progressive motility (Previous PM) were obtained from the immediately preceding collection. Rolling3_TM represented the mean total motility across the three most recent eligible previous collections. HistMean_TM represented the mean total motility across all eligible previous collections. Additional historical variables included previous semen volume, previous sperm concentration, rolling and cumulative progressive-motility measures, historical standard deviations and collection-history variables. These variables were constructed only from observations preceding the target collection. The feature-engineering process was therefore temporally traceable: each engineered predictor could be linked to a clearly defined set of preceding reproductive observations.

2.8. Simple Longitudinal Prediction Benchmarks

Three transparent threshold-based benchmarks were evaluated: Previous TM $\geq 70\%$; Rolling3_TM $\geq 70\%$; and HistMean_TM $\geq 70\%$. The benchmarks were evaluated under the same LOBOCV prediction structure used for learned models. Their purpose was to establish whether a reproducible historical signal existed before introducing model complexity.

2.9. Logistic Regression Models

The final learned longitudinal specifications were restricted to the prespecified predictor combinations reported in the Results. All predictors were calculated using information available before the target collection, and model fitting and evaluation followed the fixed LOBOCV protocol.

2.10. Baseline and Enhanced Machine-Learning Models

Random Forest and XGBoost were evaluated alongside Logistic Regression. Baseline models used only information available before the target collection, including bull age, breed, ejaculate number, collection day, collection year and collection-interval information. Current semen measurements that define or describe the target ejaculate were excluded. Enhanced models incorporated strictly historical reproductive predictors. The fixed Random Forest specification used 500 trees, minimum samples per leaf of 2, square-root feature selection, balanced class weights and random seed 2026. XGBoost used 300 estimators, maximum depth 2, learning rate 0.03, subsample 0.8, column

subsampling 0.8, regularisation parameter 2.0, minimum child weight 2 and random seed 2026. No hyperparameter tuning was performed in the final reconstruction (Hürland *et al.*, 2023; Chen & Guestrin, 2016) ^[2, 7].

The explainability component of the study was operationalised through transparent feature engineering and direct comparison of predictor strategies rather than through a post-hoc explainability algorithm. Accordingly, no SHAP, LIME or equivalent post-hoc feature-attribution analysis was conducted. The interpretation focused on whether clearly defined historical features and parsimonious models could provide useful predictive information relative to more complex fixed machine-learning specifications.

2.11. Leave-One-Bull-Out Cross-Validation

Because observations were repeated within bulls, Leave-One-Bull-Out Cross-Validation (LOBOCV) was used as the primary validation framework. In each fold, all eligible observations from one bull were held out while the remaining bulls were used for model development. This procedure was repeated until each bull had served as the held-out validation unit. The approach prevents observations from the same bull appearing in both training and validation data and therefore provides a more demanding assessment of generalisation to previously unseen animals.

2.12. Performance Evaluation and Model Selection

Performance was evaluated using accuracy, balanced accuracy, precision, recall, specificity, F1-score, ROC-AUC, MCC and Cohen's kappa where applicable. Matthews correlation coefficient was prespecified as the primary model-selection criterion because it summarises the four confusion-matrix components and remains informative when class frequencies are unequal (Chicco *et al.*, 2021) ^[3]. Secondary metrics were used to provide complementary information rather than to override the primary criterion. (Chicco *et al.*, 2021) ^[3].

2.13. Statistical and Computational Reproducibility

The final computational reconstruction used a fixed, leakage-aware analysis protocol. The raw dataset was preserved unchanged; longitudinal predictors were generated from preceding observations only; validation was performed at bull level; and numerical results were retained from the computational output rather than reconstructed from earlier manuscript tables. The computational workflow was implemented in Python using standard machine-learning tools, including scikit-learn and XGBoost (Pedregosa *et al.*, 2011; Chen & Guestrin, 2016) ^[2, 11]. (Roberts *et al.*, 2017) ^[12].

3. Results

3.1. Dataset Validation and Longitudinal Modelling Population

The original dataset comprised 57 semen collection records from seven artificial insemination bulls. Following predefined quality control, eight observations were excluded, leaving 49 biologically valid records. The first available collection for each bull could not contribute to lag-based prediction, resulting in 42 longitudinal modelling observations from seven bulls. Of these, 25 (59.5%) were classified as higher total-motility class and 17 (40.5%) as lower total-motility class.

3.2. Performance of Simple Longitudinal Prediction Indicators

The simple longitudinal benchmarks demonstrated a measurable predictive signal in historical semen performance. As shown in Table 1, previous total motility at the 70% threshold achieved 66.7% accuracy, 64.5% balanced accuracy and an MCC of 0.296. The three-collection rolling

mean produced 64.3% accuracy, 59.4% balanced accuracy and an MCC of 0.211. The historical mean of previous total motility provided the strongest simple prediction, achieving 71.4% accuracy, 67.5% balanced accuracy and an MCC of 0.391. Thus, cumulative historical information outperformed both the immediately preceding observation and the short rolling window under the same LOBOCV framework.

Table 1: Performance of simple longitudinal prediction indicators under Leave-One-Bull-Out Cross-Validation.

Prediction strategy	Accuracy (%)	Balanced accuracy (%)	MCC
Previous TM $\geq 70\%$	66.7	64.5	0.296
Rolling 3-collection mean TM $\geq 70\%$	64.3	59.4	0.211
Historical mean TM $\geq 70\%$	71.4	67.5	0.391

Note: Accuracy and balanced accuracy are expressed as percentages; MCC is unitless. The classification threshold for all strategies was TM $\geq 70\%$.

Figure 1 provides a visual comparison of the relative performance of the three simple longitudinal prediction

indicators and similarly shows the superior overall performance of the historical mean of previous total motility.

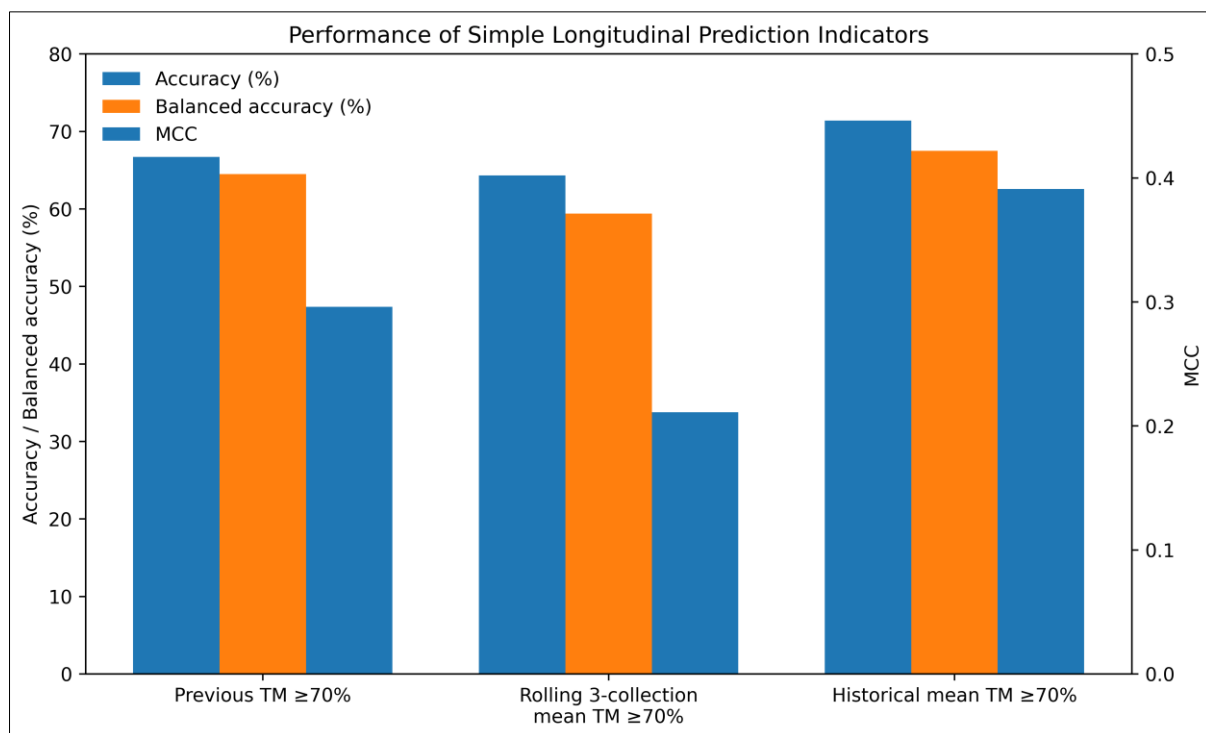


Fig 1: Performance of simple longitudinal prediction indicators under Leave-One-Bull-Out Cross-Validation.

3.3. Performance of Parsimonious Logistic Regression Models

The performance of the parsimonious longitudinal Logistic Regression models is summarised in Table 2. Logistic Regression based on Previous TM achieved 69.0% accuracy, 67.4% balanced accuracy, an ROC-AUC of 0.692 and an MCC of 0.352. Adding Previous PM did not change accuracy, balanced accuracy or MCC, although ROC-AUC increased slightly to 0.699. In contrast, combining Previous TM with HistMean_TM reduced predictive performance,

producing 57.1% accuracy, 52.7% balanced accuracy, an ROC-AUC of 0.584 and an MCC of 0.060.

Overall, Table 2 shows that increasing the number of biologically related historical predictors did not necessarily improve model generalisation in this small longitudinal dataset. The strongest Logistic Regression performance was obtained using Previous TM alone, while the addition of Previous PM did not improve the primary performance criterion, MCC.

Table 2: Performance of parsimonious longitudinal Logistic Regression models under Leave-One-Bull-Out Cross-Validation.

Predictor specification	Accuracy (%)	Balanced accuracy (%)	ROC-AUC	MCC
Previous TM	69.0	67.4	0.692	0.352
Previous TM + Previous PM	69.0	67.4	0.699	0.352
Previous TM + Historical mean TM	57.1	52.7	0.584	0.060

Note: Accuracy and balanced accuracy are expressed as percentages; ROC-AUC and MCC are unitless. MCC was the prespecified primary evaluation criterion.

3.4. Baseline and Enhanced Machine-Learning Performance

The comparative performance of the baseline and enhanced machine-learning models under the fixed LOBOCV framework is presented in Table 3. Under the fixed model specifications, baseline machine-learning models using only temporally available bull and collection descriptors showed limited predictive performance. Baseline Logistic Regression achieved an MCC of -0.515 , followed by baseline Random Forest (MCC = -0.343) and baseline XGBoost (MCC = -0.446). Incorporation of longitudinal reproductive predictors did not reverse this overall pattern. Enhanced Logistic Regression achieved an MCC of -0.407 , whereas enhanced Random

Forest and enhanced XGBoost achieved MCC values of -0.247 and -0.224 , respectively. Although the enhanced Random Forest and XGBoost models achieved higher accuracy, balanced accuracy, F1 scores and ROC-AUC values than their corresponding baseline models (Table 3), their MCC values remained negative. Overall, Table 3 indicates that incorporating longitudinal reproductive predictors improved several secondary performance measures for the Random Forest and XGBoost models, but did not produce positive MCC values under the prespecified primary evaluation criterion. Thus, the additional algorithmic complexity and expanded predictor set did not provide useful generalisation according to MCC in this small longitudinal dataset.

Table 3: Performance of baseline and enhanced machine-learning models under the fixed Leave-One-Bull-Out Cross-Validation framework, with Matthews correlation coefficient as the prespecified primary evaluation criterion.

Model	Accuracy	Balanced accuracy	Precision	Recall	Specificity	F1	ROC-AUC	MCC
Baseline Logistic Regression	23.8	28.5	11.1	4.0	52.9	5.9	0.181	-0.515
Baseline Random Forest	31.0	33.5	35.7	20.0	47.1	25.6	0.301	-0.343
Baseline XGBoost	26.2	28.6	28.6	16.0	41.2	20.5	0.244	-0.446
Enhanced Logistic Regression	28.6	29.6	35.3	24.0	35.3	28.6	0.141	-0.407
Enhanced Random Forest	40.5	37.8	50.0	52.0	23.5	51.0	0.379	-0.247
Enhanced XGBoost	40.5	38.7	50.0	48.0	29.4	49.0	0.391	-0.224

Note: Accuracy, balanced accuracy, precision, recall, specificity, and F1 are expressed as percentages. ROC-AUC and MCC are unitless. MCC was the prespecified primary evaluation criterion.

3.5. Comparative Predictive Performance

The comparative performance of the verified prediction strategies is summarised in Figure 2. The analysis showed a clear separation between the strongest simple longitudinal predictors and the fixed machine-learning models. The historical mean TM threshold achieved the highest MCC (0.391), followed by Logistic Regression based on Previous TM and Previous TM + Previous PM, both of which achieved an MCC of 0.352. Previous TM threshold and the three-collection rolling mean achieved MCC values of 0.296 and 0.211, respectively. As illustrated in Figure 2, the strongest predictive performance was therefore obtained from simple temporally valid historical indicators and parsimonious Logistic Regression models rather than from the more complex fixed machine-learning models. In contrast, all enhanced machine-learning models produced negative MCC values under LOBOCV. These findings indicate that, in this dataset, increasing model complexity did not provide additional predictive value under stringent animal-level validation.

3.6 Robustness and Uncertainty

Because only seven bulls constituted the independent validation units, the apparent performance of the strongest historical indicator should be interpreted cautiously. Descriptive observation-level bootstrap analysis of the historical-mean threshold produced a median MCC of 0.403 with a descriptive 95% interval of 0.090–0.666 and a median balanced accuracy of 0.679 with a descriptive interval of 0.536–0.811. These intervals are not independent-bull confidence intervals because repeated observations are clustered within only seven bulls. They are therefore presented as descriptive uncertainty estimates rather than evidence of population-level generalisability.

4. Discussion

4.1. Historical Semen Performance as a Predictive Signal

The principal finding was that historical semen performance contained a measurable predictive signal for the total-motility class of a subsequent collection. The historical mean of previous total motility provided the strongest verified simple performance, achieving 71.4% accuracy, 67.5% balanced accuracy and MCC = 0.391. The cumulative historical measure outperformed both the immediately preceding total motility and the three-collection rolling mean. This suggests that, within the present dataset, the predictive information was not restricted to the most recent collection. A cumulative measure may provide a more stable representation of reproductive performance than a single observation, although this mechanism was not directly tested.

4.2. Parsimony Versus Additional Predictors

The Logistic Regression results reinforce the value of parsimony. Previous TM alone achieved MCC = 0.352, and addition of Previous PM did not improve the primary criterion. Combining Previous TM with HistMean_TM reduced MCC to 0.060. The result does not imply that historical mean TM is biologically unimportant; rather, it shows that combining correlated historical summaries did not improve prediction in this small dataset. Predictor retention should therefore be based on demonstrated incremental value under independent validation rather than on the assumption that more biologically related variables will necessarily improve performance.

4.3. Algorithmic Complexity and Generalisation Across Bulls

The fixed machine-learning results provide an important counterpoint to the assumption that more flexible algorithms

necessarily produce better prediction. All baseline models had negative MCC values, and the enhanced models also remained below zero. Enhanced Random Forest and XGBoost had higher conventional accuracy than the corresponding baseline models, but their MCC values remained negative. Because MCC was prespecified as the primary criterion, these results do not support selecting the complex models over the historical-mean threshold or the parsimonious Logistic Regression models. Instead, they suggest that the dominant predictive signal in the present data was captured by simple historical reproductive information.

4.4. Importance of Leave-One-Bull-Out Validation

LOBOCV was essential because repeated semen observations from the same bull are not independent. Random record-level splitting could place collections from the same animal in both training and validation datasets and allow models to exploit bull-specific characteristics. LOBOCV avoids this problem by holding out all observations from one bull at a time. The resulting estimates therefore represent a more demanding test of generalisation to previously unseen bulls. This is consistent with broader recommendations that cross-validation should respect temporal, hierarchical and other dependence structures in the data (Roberts *et al.*, 2017) ^[12].

4.5. Implications for Artificial-Insemination Bull Monitoring

The historical mean of previous total motility is attractive as a candidate decision-support variable for artificial-insemination bull monitoring because it is transparent, simple to calculate and can be updated whenever a new semen collection is recorded. It should complement rather than replace direct laboratory evaluation. The current ejaculate remains the object of direct assessment, whereas the historical predictor provides contextual information about prior semen performance (Harstine *et al.*, 2018; Koçyigit *et al.*, 2026) ^[5, 8].

The results support a conceptual decision-support pathway for smart livestock reproduction: (1) routinely capture semen and collection records; (2) preserve temporal ordering and apply data-quality checks; (3) update transparent historical features after each eligible collection; (4) generate a subsequent semen-quality risk or class indication using a validated parsimonious model; and (5) present the output as contextual support for laboratory and management decisions rather than as an autonomous replacement for semen assessment. This framework remains conceptual and should not be interpreted as an operationally validated software system.

4.6. Implications for Small Longitudinal Livestock Datasets

The study illustrates an important methodological issue in livestock machine learning: repeated observations can increase the apparent dataset size without increasing the number of independent animals to the same extent. The present modelling population contained 42 observations but only seven independent bulls. Consequently, algorithmic flexibility can be large relative to the effective independent sample size. A staged strategy is therefore appropriate: preserve temporal structure, evaluate simple biological predictors, validate at the animal level, and only then assess

whether more complex algorithms provide demonstrable incremental benefit. (Roberts *et al.*, 2017) ^[12].

4.7. Limitations and Future Research

The principal limitation was the small number of independent bulls. The study also used routinely collected observational records and therefore lacked comprehensive information on nutrition, health events, environmental conditions, season, thermal stress, body condition and management. The binary 70% TM threshold is an operational simplification of semen quality, which is multidimensional. Future research should evaluate multiple semen-quality attributes and reproductive outcomes, include substantially larger numbers of bulls, and conduct external validation across breeds, seasons, management systems and AI centres (Harstine *et al.*, 2018; Hürland *et al.*, 2023) ^[5, 7].

4.8. Overall Interpretation

The central contribution of the study is the demonstration that longitudinal semen history contains a measurable predictive signal under strict temporal and animal-level validation. Longitudinal feature engineering made that signal explicit through temporally traceable representations of immediate, rolling and cumulative reproductive history. The strongest verified result came from a simple historical-mean threshold, while the parsimonious Previous TM Logistic Regression model also performed well relative to the more complex machine-learning models.

From an explainability perspective, the findings suggest that transparency can be achieved through disciplined feature construction and comparison with parsimonious models rather than through algorithmic complexity alone. The results do not establish a ready-to-deploy prediction system or demonstrate the use of post-hoc explainable-AI algorithms. Rather, they provide a reproducible and interpretable basis for future smart-livestock decision-support research in which larger independent datasets can determine whether the observed longitudinal signal is stable and transportable (Roberts *et al.*, 2017) ^[12].

5. Conclusion

This study demonstrated that longitudinal feature engineering can extract a measurable predictive signal from historical semen records for subsequent total-motility classification in artificial insemination bulls. The strongest verified performance was obtained from the historical mean of previous total motility, which achieved 71.4% accuracy, 67.5% balanced accuracy and MCC = 0.391 under LOBOCV. Previous total motility alone also provided useful predictive information, with the corresponding Logistic Regression model achieving MCC = 0.352, whereas addition of previous progressive motility did not improve the primary criterion.

The findings further indicate that predictive performance was not proportional to algorithmic complexity. Under the fixed computational specifications, enhanced Random Forest and XGBoost models produced negative MCC values and did not outperform the strongest parsimonious longitudinal strategies. The explainability contribution of the study therefore lies in transparent, temporally traceable feature engineering and parsimonious comparative modelling rather than in post-hoc explainable-AI algorithms.

Collectively, the results support a conceptual decision-support framework for smart livestock reproduction in which

routinely accumulated semen records are transformed into continuously updated historical performance profiles that complement, rather than replace, direct laboratory evaluation. Substantially larger independent datasets, external validation and explicit post-hoc explainability analyses will be required before operational deployment.

5.1. Practical and Research Implications

Routinely accumulated semen records could potentially be transformed into continuously updated historical semen-performance profiles for individual bulls. In a future smart-livestock implementation, these profiles could provide a transparent contextual layer alongside conventional laboratory assessment. The decision-support value would arise from making historical information explicit, temporally valid and easy to interpret rather than from algorithmic complexity alone.

Future studies should prioritise independent animal replication, broader biological and environmental predictors, external validation and evaluation of clinically or operationally meaningful reproductive outcomes. They should also test explicit explainable-AI methods, including post-hoc feature-attribution approaches, in sufficiently large independent datasets before any operational decision-support system is deployed (Hürland *et al.*, 2023; Harstine *et al.*, 2018) [5, 7].

Acknowledgements

The authors acknowledge Matopos Research Station for providing access to the facilities and historical semen records used in this study.

Funding

This research received no specific funding from any public, commercial, or not-for-profit funding agency.

Declaration of Competing Interests

The authors declare that they have no conflict of interest.

Data Availability

The data supporting the findings of this study are available from the corresponding author on reasonable request, subject to applicable institutional permissions and data-governance requirements.

Author Contributions

Dorothy Assan: Data curation, formal analysis, methodology, and writing – original draft.

Abbegal Dube: Investigation, data collection, and writing – review and editing.

Luke Mapiliyao: Investigation, data collection, and writing – review and editing.

Never Assan: Conceptualization, methodology, supervision, validation, and writing – review and editing.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this manuscript, the authors used PaperPal to assist with language refinement, manuscript organization, and editorial preparation. Following the use of this tool, the authors reviewed and edited the content as

necessary and take full responsibility for the accuracy, integrity, and originality of the published work.

References

- Assan, N., Musasira, M., Mpofu, M., Mwareya, N., 2023. Species dependent correlation analysis and regression models of body weight on linear body measures in indigenous sheep and goats of Zimbabwe. *Adv. Mod. Agric.* 4(2), 2388. <https://doi.org/10.54517/ama.v4i2.2388>.
- Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- Chicco, D., Tötsch, N., Jurman, G., 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining* 14, 13. <https://doi.org/10.1186/s13040-021-00244-z>.
- Hagereaves, S.K., Bruce, D., Beffa, L.M., 2004. Disaster mitigation options for livestock production in communal farming systems in Zimbabwe. ICRISAT & FAO, Bulawayo, Zimbabwe.
- Harstine, B.R., Utt, M.D., DeJarnette, J.M., 2018. Review: Integrating a semen quality control program and sire fertility at a large artificial insemination organization. *Animal* 12 (Suppl. 1), s63–s74. <https://doi.org/10.1017/S1751731118000319>.
- Homann, S., Van Rooyen, A.F., Moyo, T., Nengomasha, Z., 2007. Goat production and marketing: baseline information for semi-arid Zimbabwe. ICRISAT, Zimbabwe.
- Hürland, M., Kuhlitz, D.A., Kuhlitz, C., Osmers, J.H., Jung, M., Schulze, M., 2023. The use of machine learning methods to predict sperm quality in Holstein bulls. *Theriogenology* 197, 16–25. <https://doi.org/10.1016/j.theriogenology.2022.11.032>.
- Koçyiğit, A., Ün, H., Kaya, C., Esin, B., Tanrikulu, M.D., Çevik, M., 2026. Linking laboratory findings to field fertility: A comparative study of frozen semen from dairy and beef bulls. *Front. Vet. Sci.* 13, 1794630. <https://doi.org/10.3389/fvets.2026.1794630>.
- Ncube, S., 2005. The role of proanthocyanidins and related flavonoids in the utilization of sorghum grain varieties as feed for ruminants in the semi-arid areas of Zimbabwe [doctoral dissertation]. University of Zimbabwe, Harare.
- Ngoda, P., Kashoma, I., Msalya, G., 2023. Semen quality of progeny-tested breeding bulls maintained at the National Artificial Insemination Centre, Arusha, Tanzania. *East African Journal of Science, Technology and Innovation* 4(3). <https://doi.org/10.37425/eajsti.v4i3.612>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, R., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

12. Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>.
13. Van Rooyen, A.F., Freeman, A., Moyo, S., Rohrbach, D., 2007. Livestock development in Southern Africa: Research priorities. ICRISAT.
14. Ward, H.K., Richardson, F.D., Denny, R.P., Dye, P.J., 1979. Matopos Research Station: a perspective. *Rhod. Agric. J.* 76(1), 5–18.

How to Cite This Article

Assan D, Dube LMA, Assan N. Explainable Machine learning for predicting semen quality classes in artificial insemination bulls using longitudinal feature engineering: a Decision-Support framework for smart livestock reproduction. *J Front Multidiscip Res.* 2026;7(2):298-305. doi:10.54660/JFMR.2026.7.2.298-305

Creative Commons (CC) License

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.