International Journal of Biological and Biomedical Research

Predictive Value of Historical Semen Collection Records for Semen Quality Classification in Artificial Insemination Bulls

Dorothy Assan [1*], Teresa Rukuni [2], Abbegal Dube [3], Never Assan [4]
[1] Department of Biotechnology and Bioinformatics, Faculty of Science and Technology, Midlands State University, Gweru, Zimbabwe
[2] Matopos Research Institute, Department of Research and Extension, Ministry of Lands and Agriculture, Bulawayo, Zimbabwe
[3] Department of Agricultural Education, Esigodini Agricultural College, Ministry of Lands, Agriculture, Fisheries, Water and Rural Development, Esigodini, Zimbabwe
[4] Faculty of Agriculture, Zimbabwe Open University, Bulawayo Regional Campus, Bulawayo, Zimbabwe

* Corresponding Author: **Dorothy Assan**

Article Info

E-ISSN: 3107-7137
Volume: 02
Issue: 06
Received: 06-09-2026
Accepted: 04-10-2026
Published: 02-11-2026
Page No: 17-24

Abstract
Background: Routine semen collection records contain repeated observations from individual artificial insemination (AI) bulls and may provide information about the quality of subsequent ejaculates. However, prediction requires strict separation between historical information and the target collection.
Objective: To evaluate whether historical semen records predict subsequent ejaculate total-motility class and whether multivariable classifiers improve on transparent historical indicators.
Methods: Source workbooks were reconciled into 94 collection records from 10 bulls. One conflicting bull 832 event was held out, leaving a provisional primary population of 93 records. High-quality semen was defined as total motility (TM) ≥70%. Predictors for each target date were constructed only from records on earlier collection dates. Seventy-three eligible ejaculate observations from nine bulls, distributed across 44 target sessions, were evaluated using leave-one-bull-out cross-validation. Historical threshold benchmarks were compared with logistic regression, random forest and XGBoost.
Results: The cumulative mean of previous collection-date TM means achieved 72.6% accuracy, 70.8% balanced accuracy and Matthews correlation coefficient (MCC) of 0.422. Among multivariable classifiers, random forest achieved the highest pooled performance (75.3% accuracy, 71.9% balanced accuracy, ROC-AUC 0.770 and MCC 0.474), compared with logistic regression (MCC 0.382) and XGBoost (MCC 0.313). Excluding eight legacy-flagged records changed the relative model ranking.
Conclusion: Earlier-date semen history contained a measurable but exploratory predictive signal. The small number of bulls, sensitivity to record handling and unresolved source checks preclude claims of operational readiness; historical indicators should be evaluated as complements to, not substitutes for, direct semen assessment.

DOI: https://doi.org/10.54660/IJBBR.2026.2.6.17-24

Keywords: artificial insemination, bull, semen quality, total motility, historical records, leave-one-bull-out cross-validation, machine learning

1. Introduction

Semen quality is a critical determinant of the efficiency of artificial insemination (AI) programmes because ejaculate characteristics influence semen processing, cryopreservation, storage, and subsequent fertilisation potential. Total sperm motility is routinely evaluated during semen assessment and remains an important operational indicator of ejaculate quality. However, semen characteristics can vary among successive ejaculates from the same bull, creating uncertainty about the quality of a subsequent collection and highlighting the potential value of incorporating reproductive history into predictive assessment.

Variation associated with bull age, ejaculate number and collection frequency has been documented in commercial AI settings, although effects may differ between fresh and post-thaw measurements [1, 2, 3, 4, 5].

The increasing availability of routinely collected reproductive records provides an opportunity to complement conventional laboratory evaluation with data-driven prediction. Machine-learning approaches have been applied to predict bull semen characteristics, while collection-history and management data have also been used to forecast sperm production [6, 7]. However, predictive performance alone does not establish practical utility. For a predictive model to provide meaningful information at the time of decision-making, its predictors must be restricted to information that would genuinely be available before the outcome being predicted. This requirement is particularly important in reproductive datasets containing repeated observations from the same animal.

Longitudinal semen records present two related challenges. First, successive ejaculates from an individual bull are biologically and statistically related, and semen production and motility may vary with ejaculate number and collection frequency [4, 5]. Treating individual collections as independent observations can therefore produce overly optimistic estimates of predictive performance. A model trained on some ejaculates from a bull may learn animal-specific characteristics and subsequently appear accurate when predicting additional ejaculates from the same animal. Second, temporal information leakage can occur when predictors incorporate measurements from the target ejaculate or information recorded after the prediction point. Such leakage can generate apparently strong predictive performance without demonstrating the ability to predict a genuinely subsequent semen-quality outcome. [8].

A rigorous longitudinal prediction framework should therefore establish an explicit temporal information boundary. Predictors should be constructed exclusively from information available before the target semen collection, while repeated observations from the same bull should be kept together during model validation. Under this framework, prediction becomes a test of whether historical reproductive information can generalise beyond the animals used for model development rather than a contemporaneous classification of semen quality using measurements obtained from the target ejaculate. [8].

Historical semen records may contain several forms of potentially informative reproductive history. Previous studies have used collection history and management information to predict sperm production, and machine-learning models have been evaluated for bull semen-quality traits [6, 7]. These studies establish the feasibility of prediction from routine AI-centre records, but their outcome definitions and predictor sets differ from a task that classifies the total motility of a subsequent ejaculate using only earlier-date semen measurements. The mean from the immediately preceding collection date may capture short-term persistence in reproductive performance, whereas rolling measures can characterise recent performance across several distinct collection dates. Cumulative historical measures may provide information about longer-term reproductive performance and consistency. Constructing these summaries from earlier collection dates, rather than allowing ejaculates collected on the same date to predict one another, provides a temporally interpretable representation of each bull's prior semen performance.

Whether increasing algorithmic complexity improves prediction under these conditions remains uncertain. Random Forest and Extreme Gradient Boosting (XGBoost) can represent nonlinear relationships and interactions among predictors, but their flexibility does not necessarily translate into improved generalisation when the number of independent animals is small and repeated observations dominate the dataset. A parsimonious longitudinal predictor may therefore perform as well as, or better than, a more complex machine-learning model. Establishing this requires validation at the animal level rather than random partitioning of individual semen records [6, 7, 8, 9].

Against this background, the present study developed a date-restricted longitudinal framework for predicting semen-quality classes of subsequent ejaculates from artificial insemination bulls using routinely collected reproductive records. Predictors for each target collection date were constructed exclusively from records belonging to earlier dates for the same bull. The study evaluated simple historical benchmarks and compared them with Logistic Regression, Random Forest and XGBoost under Leave-One-Bull-Out Cross-Validation. The specific objectives were to: (i) construct historical predictors that respect a strict earlier-date information boundary; (ii) compare transparent historical benchmarks with statistical and machine-learning classifiers under animal-level validation; and (iii) determine whether multivariable algorithmic approaches provided additional predictive value over cumulative historical semen summaries. [6, 8].

2. Materials and Methods

2.1. Ethical Approval

All procedures involving animals were conducted in accordance with the ethical standards for animal experimentation approved by the Zimbabwe Open University Animal Research Ethics Committee (Approval No. PG/2026/64.2). Animal handling and data collection procedures were performed to minimise stress and ensure animal welfare throughout the study.

2.2. Study Site

The study used semen-production records from artificial insemination bulls maintained at Matopos Research Station, Zimbabwe. The site is located in a semi-arid production environment characterised by a marked dry season and variable rainfall. Site-level environmental characteristics were not used as predictors in the present analysis.

2.3. Study Dataset, Source Reconciliation and Longitudinal Structure

The source workbooks were reconciled into an event-level database containing 94 semen collection records from 10 artificial insemination bulls. Each row represented an individual ejaculate record and included bull identity, collection date, age, breed code, ejaculate number and routinely recorded semen measurements, including total motility (TM), progressive motility (PM), semen volume, sperm concentration and the collection descriptor. Collection interval information was retained where recorded. The original source values were preserved during reconciliation.

Records were matched across the available workbooks using bull identifier, collection date and ejaculate number, with

measurement values compared where matching events appeared in more than one file. One event for bull 832 (9 December 2024, ejaculate 1) had two conflicting sets of measurements. The record with TM 92.89% and PM 91.99% was corroborated by other available files and was provisionally retained; the competing record was held outside the primary dataset pending verification against the original collection register. The resulting provisional primary dataset contained 93 records. All eight collection events labelled 21019 in the expanded ALL sheet matched, on collection date, ejaculate number and all recorded semen measurements, corresponding records labelled 21029 or 21-029 in four other source workbooks. The bull was also consistently described as year of birth 2021, age four years and breed code 2. These records were therefore treated as one animal for analysis, using 21029 as the harmonised analysis identifier while retaining 21019 as the identifier appearing in the expanded ALL sheet. No separate animal master register was available, but the event-level crosswalk showed complete agreement across all eight records.

The eight observations previously labelled as flagged in a legacy quality-control file were retained in the provisional primary population. Agreement among spreadsheets was treated as evidence of transcription consistency, not as proof of original laboratory measurement accuracy. Statistical outlier flags were used as review prompts and were not, by themselves, grounds for deleting observations.

2.4. Data Quality Control and Analysis Populations

The provisional primary analysis used the 93 reconciled records after the single conflicting bull 832 event had been held out. Basic plausibility checks were applied to the measured variables, and duplicated bull-date-ejaculate keys were reviewed. The eight legacy-flagged observations remained in the primary analysis because the available source files did not establish that their values were erroneous. Their influence was examined in a separate sensitivity analysis that excluded these eight observations and reconstructed all historical predictors from the remaining records.

The primary population included 56 high-quality records and 37 lower-quality records using the TM classification threshold described below. Because prediction required at least one previous distinct collection date for the same bull, all ejaculates on a bull's first recorded collection date were ineligible as target prediction observations. The first date was treated as a single session even when more than one ejaculate was recorded on that date. Under this rule, 73 ejaculate-level observations from nine bulls were eligible for prediction. These observations arose from 44 bull-by-date target collection sessions. One bull with a single source record did not contribute an eligible prediction observation.

2.5. Definition of the Prediction Outcome

Semen quality was defined from total sperm motility. Each target ejaculate with TM ≥70% was classified as high-quality semen (Class 1), whereas an ejaculate with TM <70% was classified as lower-quality semen (Class 0). The 70% cut-point was used as an operational classification threshold for this study and should not be interpreted as a universal biological threshold for fertility or as a substitute for a full semen-quality assessment. The target ejaculate's TM value was used only to define the observed outcome and was not included among the predictors for its own classification.

2.6. Temporal Ordering and Prevention of Information Leakage

Records were ordered within bull by collection date and, within a date, by ejaculate number. The prediction time boundary was defined at the collection-date level. For a target ejaculate collected on date t, only records from the same bull with collection dates strictly earlier than t were eligible to construct predictors. No ejaculate collected on the target date, including an earlier-numbered ejaculate from that same date, was allowed to contribute to the target's historical predictors. Consequently, ejaculates collected on the same date shared the same historical predictor values but remained separate ejaculate-level outcomes in the primary analysis. The first distinct collection date for each bull could not be used as a prediction target because no earlier-date history was available.

2.7. Construction of Date-Restricted Historical Predictors

Historical variables were calculated separately for each bull and target collection date. Where several ejaculates were recorded on a previous date, measurements were first averaged within that date (Table 1). Historical summaries across dates were then constructed from these date-level means. This approach prevented dates with more recorded ejaculates from automatically receiving greater weight in cumulative and rolling historical summaries.

Table 1: Date-restricted historical predictors used in the multivariable models.

Predictor	Definition
Previous-date mean TM	Arithmetic mean TM across all ejaculates on the immediately preceding distinct collection date.
Previous-date mean PM	Arithmetic mean PM across all ejaculates on the immediately preceding distinct collection date.
Cumulative historical mean TM	Arithmetic mean of the date-specific TM means across all prior distinct collection dates; each date receives equal weight.
Rolling three-date mean TM	Arithmetic mean of date-specific TM means for the three most recent prior distinct collection dates, or fewer where less history is available.
Previous-date mean semen volume	Mean recorded semen volume across ejaculates on the immediately preceding date; zero values were treated as missing for this predictor only.
Previous-date mean sperm concentration	Mean sperm concentration across ejaculates on the immediately preceding date, using the units recorded in the source dataset.

The historical cumulative mean was therefore a mean of prior collection-date means rather than a pooled mean across every previous ejaculate. The rolling-three-date measure likewise represented three distinct prior collection dates, not the three most recent ejaculate records. Predictor values were recalculated independently for every target date. A zero semen-volume value was treated as missing only when

constructing the previous-date volume predictor; the original source value was not overwritten.

2.8. Simple Historical Benchmarks

Three transparent threshold benchmarks were evaluated: (i) previous-date mean TM ≥70%; (ii) mean TM across the three most recent previous collection dates ≥70%; and (iii) cumulative historical mean TM across all previous collection dates ≥70%. These benchmarks were applied to the same eligible target observations used for the learned models. Their purpose was to quantify the predictive information available from transparent historical summaries before considering multivariable classification.

2.9. Logistic Regression and Machine-Learning Models

Logistic regression was evaluated using four specifications: previous-date mean TM alone; previous-date mean TM together with previous-date mean PM; cumulative historical mean TM together with previous-date mean TM; and a multivariable specification containing all six date-restricted historical predictors listed in Table 1. In the multivariable model, the six inputs were previous-date mean TM, previous-date mean PM, cumulative historical mean TM, rolling three-date mean TM, previous-date mean semen volume and previous-date mean sperm concentration. Logistic regression was fitted using scikit-learn with median imputation, within-fold standardisation and a maximum of 4,000 iterations. No class weighting was applied to logistic regression in this reconstruction.

Random Forest and XGBoost were evaluated using the same six-predictor multivariable specification. The Random Forest model used 500 trees, a minimum leaf size of two observations, square-root feature selection, balanced class weights and random seed 2026. XGBoost used 300 estimators, maximum tree depth of 2, learning rate of 0.03, row subsampling of 0.8, column subsampling of 0.8, regularisation parameter $\lambda=2$, minimum child weight of 2, scale_pos_weight of 1 and random seed 2026. Median imputation was performed within each training fold for both tree-based models; standardisation was not applied. No hyperparameter tuning was performed in the final reconstruction. [6, 9, 10]

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

Leave-One-Bull-Out Cross-Validation (LOBO-CV) was used to evaluate generalisation to bulls not used for model fitting. In each fold, all eligible target observations from one bull were held out, while eligible observations from the remaining bulls were used to train the model. The procedure was repeated until each of the nine prediction-contributing bulls had served as the held-out unit. The held-out bull's historical predictor values were calculated from that bull's own records dated before each target collection date, as would be available for a future prediction; however, no outcome record from the held-out bull was used to fit the model. This grouping prevented records from the same bull appearing in both the training and validation partitions. [8]

2.11. Performance Evaluation

Out-of-fold predictions from all nine held-out bulls were pooled to calculate accuracy, balanced accuracy, precision, recall, specificity, F1-score, ROC-AUC and Matthews correlation coefficient (MCC). For probability-based classifiers, predicted probabilities ≥0.50 were classified as high-quality. MCC was treated as the primary summary measure because it incorporates all four confusion-matrix components and is informative when class frequencies are unequal. Other metrics were used to describe complementary aspects of classification performance. For deterministic threshold benchmarks, ROC-AUC was not calculated because the benchmark supplied a binary classification rather than a continuous prediction score. [11]

2.12. Session-Level and Data-Quality Sensitivity Analyses

A secondary session-level analysis examined the effect of grouping ejaculates collected on the same bull-date. For this analysis, TM values were averaged across target ejaculates within each bull-by-date session, and the session was classified as high-quality when its mean TM was ≥70%. The resulting session-level classification was compared with the cumulative historical-mean threshold and with out-of-fold probabilities from the multivariable classifiers. Because all ejaculates in a target bull-date session shared the same historical predictors, one session-level probability was retained for each session. This was a secondary robustness analysis with a different outcome unit from the primary ejaculate-level prediction task.

A separate data-quality sensitivity analysis excluded the eight observations labelled "Previously flagged" in the legacy quality-control file. All historical predictors were reconstructed after exclusion so that removed observations did not contribute to the history of later target collections. The resulting sensitivity population contained 85 source records, 67 eligible ejaculate-level prediction observations, 42 target sessions and nine bulls contributing predictions. Results were compared with those from the provisional 93-record primary population. The conflicting bull 832 record remained outside both populations pending original-record verification.

2.13. Statistical Software and Reproducibility

Data reconciliation and feature construction were conducted in Python using pandas and NumPy. Model fitting, preprocessing and performance evaluation used scikit-learn; XGBoost was implemented using the XGBoost Python interface. Random seeds were fixed at 2026 for the stochastic tree-based models. Predictor construction was performed before model fitting according to the strict earlier-date rule, and all learned preprocessing steps were fitted within each LOBO-CV training fold. The reconciled source database and out-of-fold prediction outputs were retained to support auditability and reproducibility.

3. Results and Discussion

3.1. Characteristics of the Semen Collection Records and Prediction Population

The reconciled source database contained 94 collection records (Table 2). One conflicting record for bull 832 (9 December 2024, ejaculate 1) was held outside the provisional primary dataset because two incompatible sets of semen measurements were recorded for the same collection event. The provisional primary dataset therefore comprised 93 records from 10 bulls. Total motility (TM) was available for all records and averaged 67.58% (SD 21.65; median 73.61%; range 6.32–93.19%). Progressive motility (PM) averaged 64.99% (SD 22.69; median 71.01%; range 4.22–92.90%). Mean semen volume was 4.32 mL (SD 3.29), while mean

sperm concentration was 423.96 in the units recorded in the source dataset (SD 260.16).

Using the classification threshold of TM ≥70%, 56 of the 93 source records (60.2%) were classified as high-quality and 37 (39.8%) as lower-quality. Under the revised temporal design, the first collection date for each bull could not contribute prediction observations because no earlier collection history was available. This yielded 73 eligible ejaculate-level prediction observations from nine bulls. These comprised 44 high-quality and 29 lower-quality outcomes. The 73 observations were distributed across 44 bull-by-date target collection sessions, with multiple ejaculates represented in some sessions. Bull 818 contributed only one source record and therefore no eligible prediction observation.

Table 2: Descriptive characteristics of the 93-record provisional dataset.

Variable	n	Mean ± SD	Median	Range
Total motility (%)	93	67.58 ± 21.65	73.61	6.32–93.19
Progressive motility (%)	93	64.99 ± 22.69	71.01	4.22–92.90
Semen volume (mL)	93	4.32 ± 3.29	4.00	0.00–14.50
Sperm concentration (source units)	93	423.96 ± 260.16	375.08	69.41–1656.76
Collection descriptor (CD)	93	74.24 ± 81.31	52.00	0–496

3.2. Temporal Construction of Historical Predictors and Benchmark Performance

To prevent information from the target collection session entering its own predictors, all historical variables were reconstructed using records from dates strictly earlier than the date of the target ejaculate. Ejaculates collected on the same date were therefore not used to predict one another. For the immediately preceding collection date, semen measurements were averaged across the ejaculates recorded on that date. The cumulative historical TM indicator was calculated by first obtaining a mean TM for each previous collection date and then averaging those date-specific means, giving each previous date equal weight regardless of the number of ejaculates recorded. The rolling indicator was the mean of the three most recent previous date-specific TM means, or fewer dates where less history was available (Table 3).

Three transparent threshold benchmarks were evaluated using the same 73 eligible observations. The cumulative mean of previous collection-date means achieved the strongest benchmark performance (accuracy 72.6%, balanced accuracy 70.8%, MCC 0.422). The mean TM from the three most recent previous collection dates produced accuracy of 69.9%, balanced accuracy of 68.5% and MCC of 0.371. The mean TM from only the immediately preceding collection date was less informative (accuracy 61.6%, balanced accuracy 59.4%, MCC 0.190).

Table 3: Threshold-based historical indicators under leave-one-bull-out validation.

Historical benchmark (threshold ≥70%)	Accuracy (%)	Balanced accuracy (%)	MCC
Immediately preceding collection-date mean TM	61.6	59.4	0.190
Mean TM across previous three collection dates	69.9	68.5	0.371
Cumulative mean of previous collection-date means	72.6	70.8	0.422

The ordering of these benchmarks suggests that information distributed across several prior collection dates was more useful than relying on the most recent date alone. Averaging date-specific means may reduce the influence of short-term variation within individual ejaculates or collection sessions. However, the cumulative indicator is still a relatively simple historical descriptor, and its performance should not be interpreted as evidence that historical TM can replace direct assessment of the target ejaculate.

3.3. Comparison of Logistic Regression and Machine-Learning Classifiers

Multivariable classifiers were evaluated using the same leave-one-bull-out folds. The six predictors were the mean TM and PM on the immediately preceding collection date, the cumulative historical mean TM across previous dates, the rolling mean TM across up to three previous dates, and the mean semen volume and sperm concentration on the immediately preceding date. Zero semen-volume entries were treated as missing for predictor construction only; the original recorded values were retained in the source database. Median imputation was fitted within each training fold. Predictors were standardised within the logistic-regression pipeline; tree-based models used median imputation without standardisation.

Random forest achieved the strongest pooled record-level performance among the three multivariable classifiers, with accuracy of 75.3%, balanced accuracy of 71.9%, ROC-AUC of 0.770 and MCC of 0.474. Logistic regression achieved accuracy of 71.2%, balanced accuracy of 67.9%, ROC-AUC of 0.688 and MCC of 0.382. XGBoost produced accuracy of 67.1%, balanced accuracy of 65.7%, ROC-AUC of 0.730 and MCC of 0.313 (Table 4). Random forest had the highest recall (88.6%), but specificity was moderate (55.2%), indicating that a substantial proportion of lower-quality outcomes were classified as high-quality.

Table 4: Pooled out-of-fold performance of multivariable classifiers using leave-one-bull-out cross-validation.

Model	Accuracy (%)	Balanced accuracy (%)	ROC-AUC	MCC
Logistic regression	71.2	67.9	0.688	0.382
Random forest	75.3	71.9	0.770	0.474
XGBoost	67.1	65.7	0.730	0.313

The random-forest result was numerically higher than the logistic-regression and XGBoost results on the principal pooled metrics. Nevertheless, the difference between random forest and the cumulative historical-mean threshold benchmark was modest (MCC 0.474 versus 0.422). The logistic-regression model also showed only moderate discrimination. The results do not establish that a more complex model is consistently superior to a cumulative historical summary. They instead indicate that multiple historical features may contain useful information, while the apparent gains from model complexity remain uncertain in a dataset with only nine validation bulls.

3.4. Session-Level Sensitivity Analysis

Because all ejaculates collected on the same date share identical historical predictors under the revised temporal design, an additional session-level sensitivity analysis was conducted. For this secondary analysis, target ejaculates were aggregated within each bull-by-date session, and the session was classified as high-quality when its mean TM was at least 70% (Table 5). This produced 44 session-level observations: 26 sessions with mean TM ≥70% and 18 sessions below the threshold. This session-level outcome differs from the primary ejaculate-level classification and is therefore presented as a robustness check rather than a replacement endpoint.

Table 5: Secondary session-level sensitivity analysis using the mean TM of ejaculates within each target collection date.

Model / benchmark	n sessions	Accuracy (%)	Balanced accuracy (%)	ROC-AUC	MCC
Cumulative historical-mean threshold	44	75.0	74.6	—	0.488
Logistic regression	44	68.2	66.2	0.746	0.332
Random forest	44	75.0	72.9	0.786	0.474
XGBoost	44	70.5	69.9	0.761	0.394

At the session level, the cumulative historical-mean threshold and random forest both achieved 75.0% accuracy, with MCC values of 0.488 and 0.474, respectively. Random forest had ROC-AUC of 0.786. The similarity between the benchmark and random-forest results reinforces the conclusion that a simple cumulative summary is competitive with multivariable classification. Because the number of sessions is small and sessions remain clustered within bulls, these estimates should be regarded as exploratory.

3.5. Bull-Level Variation and Sensitivity to Legacy-Flagged Records

Pooled cross-validated performance concealed differences between bulls. The number of eligible predictions per bull ranged from one to 15, and some bulls contributed predictions from only one outcome class. For example, bull 20010 contributed seven lower-quality outcomes and no high-quality outcomes, while bull 832 contributed 11 high-quality and four lower-quality outcomes (Table ^). Accuracy values for bulls with one or very few prediction observations are not stable estimates of individual-level predictive performance. These differences mean that the pooled metrics summarise performance across nine held-out bulls but do not establish uniform predictive performance across animals.

Table 6: Prediction-observation distribution and random-forest accuracy by held-out bull.

Bull ID	Prediction records	Sessions	High TM	Lower TM	RF accuracy (%)
20010	7	3	0	7	100.0
21029	6	5	2	4	66.7
2112	1	1	1	0	100.0
2148	1	1	0	1	0.0
777	9	5	7	2	55.6
832	15	8	11	4	73.3
861	9	8	3	6	77.8
875	12	6	9	3	75.0
888	13	7	11	2	84.6

A sensitivity analysis excluding the eight records labelled “Previously flagged” in the legacy quality-control file retained 85 source records and produced 67 eligible prediction observations from nine bulls across 42 target sessions. The cumulative historical-mean benchmark decreased from an MCC of 0.422 to 0.327. Random-forest MCC decreased from 0.474 to 0.336, whereas logistic-regression MCC changed from 0.382 to 0.418. XGBoost MCC changed from 0.313 to 0.320 (Table 7). The sensitivity results therefore show that the apparent advantage of the cumulative benchmark and random forest depends on the handling of the legacy-flagged observations; the ranking of the algorithms is not stable across the two data populations.

Table 7: Sensitivity of threshold and model performance to exclusion of eight legacy-flagged observations.

Model / benchmark	n	Accuracy (%)	Balanced accuracy (%)	ROC-AUC	MCC
Cumulative historical-mean threshold	73	72.6	70.8	—	0.422
Cumulative historical-mean threshold (exclude flags)	67	68.7	65.9	—	0.327
Logistic regression (primary)	73	71.2	67.9	0.688	0.382
Logistic regression (exclude flags)	67	73.1	69.6	0.648	0.418
Random forest (primary)	73	75.3	71.9	0.770	0.474
Random forest (exclude flags)	67	68.7	66.7	0.711	0.336
XGBoost (primary)	73	67.1	65.7	0.730	0.313
XGBoost (exclude flags)	67	68.7	65.2	0.689	0.320

The eight flagged observations were not excluded from the primary analysis merely because they were extreme. Agreement across available spreadsheets supports consistency of transcription for several entries, but does not confirm the accuracy of original laboratory measurements. Cross-source checking showed that all eight events labelled 21019 in the expanded ALL sheet matched records labelled 21029/21-029 in four other workbooks; these were therefore treated as one animal under the harmonised analysis identifier 21029, while the original source identifier was preserved. The zero semen volume recorded for bull 20010 on 19 November 2024 (ejaculate 2) was repeated identically across all five checked sources. It was retained in the source database but treated as missing only when constructing the previous-date volume predictor; the original laboratory register is still needed to establish the meaning of the zero entry. The low-motility records for bull 20010 on 3 December 2024 (TM 6.32%; PM 4.22%) and 23 December 2024 (TM 16.15%; PM 13.59%) were also repeated identically across all five sources and were retained in the primary analysis, with their influence examined through the legacy-flag exclusion sensitivity analysis. Their laboratory accuracy cannot be confirmed without the original records. For the bull 832 collision, the 92.89% TM / 91.99% PM record appeared in all five checked sources, whereas the competing 89.67% TM / 87.97% PM record appeared only in the expanded ALL sheet and its expanded working extract; the former was retained provisionally and the latter remained held out pending original-register verification.

3.6. Implications and Limitations

The revised analysis indicates that historical semen records contain information associated with the TM classification of subsequent ejaculates when the predictors are restricted to earlier collection dates. The cumulative mean of date-specific historical TM values provided a useful transparent benchmark, and random forest produced the strongest pooled performance among the multivariable classifiers in the primary provisional dataset. However, the benchmark remained competitive, the algorithm ranking changed when legacy-flagged records were excluded, and bull-level performance varied. Historical records may therefore have value as an adjunct for identifying bulls whose recent semen quality warrants closer review, but the current evidence does not justify automated acceptance or rejection of ejaculates.

Several limitations constrain interpretation. First, only nine bulls contributed eligible predictions, and the number of observations per bull was uneven. Second, the 73 ejaculate-level outcomes arose from 44 target collection-date sessions, so outcomes sharing the same historical predictors are clustered within session. Third, validation was internal and leave-one-bull-out; no independent AI centre or external dataset was available. Fourth, although the 21019/21029 identifier discrepancy was reconciled through complete eight-event cross-source agreement, the original animal master register was unavailable. The meaning of bull 20010's zero semen-volume entry and the laboratory accuracy of its two very low motility records remain unverified against original laboratory records; the competing bull 832 event also remains held out pending register verification. Fifth, the binary outcome was based on TM ≥70% and does not directly measure fertilising capacity, conception rate or field fertility. Finally, the modest sample size and small number of independent bulls make estimates sensitive to individual animals and data-handling decisions. Larger multi-centre datasets, pre-specified temporal validation, and linkage to fertility outcomes are required before practical deployment can be considered.

4. Conclusion

Under a temporally restricted design in which predictors were drawn only from earlier collection dates, historical semen records showed measurable but moderate predictive value for classifying subsequent ejaculates. The cumulative mean of previous collection-date TM values achieved an MCC of 0.422, while random forest achieved an MCC of 0.474 among multivariable classifiers. These results are exploratory: model rankings changed when the eight legacy-flagged observations were excluded, and performance varied across the nine bulls contributing predictions. Historical collection records may complement routine semen evaluation, but the present evidence is insufficient to establish generalisable or operationally ready prediction.

The next research priority is replication in larger multi-centre bull populations with verified individual identities and collection registers, followed by external validation and linkage to fertility outcomes. Until such evidence is available, historical semen indicators should be used only as supplementary context for routine laboratory assessment.

Declarations

Declaration of AI and AI-assisted technologies was used in the manuscript preparation process

During preparation of this manuscript, the authors used Grammarly to assist with language editing, structural revision and manuscript preparation. The authors reviewed and edited the output, verified the reported results against the project records, and take full responsibility for the final content.

CRediT Authorship Contribution Statement

Dorothy Assan: Conceptualization, Methodology, Formal analysis, Software, Validation, Visualization, Writing – original draft, Writing – review and editing; Teresa Rukuni: Investigation, Data curation, Animal data collection, Writing – review and editing; Abbegal Dube & Never Assan:

Investigation, Data curation, Animal data collection, Writing – review and editing.

Declaration of Competing Interest

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

Funding

No specific funding was received for this study.

Data Availability

The underlying dataset is not publicly deposited because it contains institutionally held animal reproductive records. Data availability can be considered by the corresponding author on reasonable request, subject to institutional permissions.

Ethics Statement

All procedures involving animals were conducted in accordance with the ethical standards for animal experimentation approved by the Zimbabwe Open University Animal Research Ethics Committee (Approval No. PG/2026/64.2). Animal handling and data collection procedures were performed to minimise stress and ensure animal welfare throughout the study.

Acknowledgements

The authors gratefully acknowledge Matopos Research Institute for providing the data used in this study. The study originated from an undergraduate research project. The first author also extends sincere appreciation to the institutional supervisor, Teresa Rukuni, for her guidance, support and valuable contributions throughout the research process.

References

1. B.R. Harstine, M.D. Utt, J.M. DeJarnette, Review: Integrating a semen quality control program and sire fertility at a large artificial insemination organization, Animal 12 (Suppl. 1) (2018) s63–s74. https://doi.org/10.1017/S1751731118000319.
2. P. Ngoda, I. Kashoma, G. Msalya, 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) (2023). https://doi.org/10.37425/eajsti.v4i3.612.
3. A. Koçyiğit, H. Ün, C. Kaya, B. Esin, M.D. Tanrıkulu, M. Çevik, Linking laboratory findings to field fertility: A comparative study of frozen semen from dairy and beef bulls, Frontiers in Veterinary Science 13 (2026) 1794630. https://doi.org/10.3389/fvets.2026.1794630.
4. E.M. Murphy, A.K. Kelly, C. O'Meara, B. Eivers, P. Lonergan, S. Fair, Influence of bull age, ejaculate number, and season of collection on semen production and sperm motility parameters in Holstein Friesian bulls in a commercial artificial insemination centre, Journal of Animal Science 96(6) (2018) 2408–2418. https://doi.org/10.1093/jas/sky130.
5. P. Taaffe, C.M. O'Meara, M. Stiavnicka, C.J. Byrne, B. Eivers, P. Lonergan, S. Fair, Increasing the frequency of ejaculate collection in young dairy bulls increases semen production and field fertility, Theriogenology 182 (2022) 45–52. https://doi.org/10.1016/j.theriogenology.2022.01.030.
6. M. Hürland, D.A. Kuhlgatz, C. Kuhlgatz, J.H. Osmers, M. Jung, M. Schulze, The use of machine learning methods to predict sperm quality in Holstein bulls, Theriogenology 197 (2023) 16–25. https://doi.org/10.1016/j.theriogenology.2022.11.032.
7. A.E. Quick, J. Meronek, K. Amburn, K. Rozeboom, K.A. Weigel, Predicting sperm production of young dairy bulls using collection history and management factors, Journal of Dairy Science 104(5) (2021) 5817–5826. https://doi.org/10.3168/jds.2020-19617.
8. D.R. Roberts, V. Bahn, S. Ciuti, M.S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J.J. Lahoz-Monfort, B. Schröder, W. Thuiller, D.I. Warton, B.A. Wintle, F. Hartig, C.F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40(8) (2017) 913–929. https://doi.org/10.1111/ecog.02881.
9. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
10. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
11. D. Chicco, N. Tötsch, G. Jurman, The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Mining 14 (2021) 13. https://doi.org/10.1186/s13040-021-00244-z.

How to Cite This Article

Assan D, Rukuni T, Dube A, Assan N. Predictive value of historical semen collection records for semen quality classification in artificial insemination bulls. Int J Biol Biomed Res. 2026;2(6):17–24. doi:10.54660/IJBBR.2026.2.6.17-24.

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-NCSA 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.