

Accepted Manuscript

Published online: October 04, 2026

Doi: 10.34172/japid.4369

Submitted: June 18, 2026

Revised: September 17, 2026

Accepted: September 19, 2026

Development and internal validation of reduced-feature machine-learning models for identifying severe periodontitis from routine clinical records

Shourjo Sarkar Munsil¹ • Banashree Baishya¹ • Kashish Agrawal¹ • Ritam Chowdhury¹ • Santosh Rayagouda Patil² • Shikhar Daniel³ • Aysha Jebin Alikkal^{4*}

¹Department of Periodontology and Oral Implantology, Kalinga Institute of Dental Sciences, KIIT University, Bhubaneswar, India

²Research Faculty, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India

³Department of Oral Medicine and Radiology, Chhattisgarh Dental College and Research Institute, Rajnandgaon, India

⁴Department of Periodontics, Krishnadevaraya College of Dental Sciences and Hospital, Rajiv Gandhi University of Health Sciences, Bangalore, India

ORCID and Email: Shourjo Sarkar Munsil: 0009-0004-9112-3416, shourjosarkar250@gmail.com

Email: Banashree Baishya: banashreebaishya24@gmail.com

Email: Kashish Agrawal: kashishagrawal2468@gmail.com

Email: Ritam Chowdhury: chowdhury.ritam1997@gmail.com

Email: Santosh Rayagouda Patil: drpsantosh@gmail.com

Email: Shikhar Daniel: shikhardaniel555@gmail.com

ORCID: Aysha Jebin Alikkal: 0000-0003-1036-242X

***Corresponding Author; Email:** ayshpathu3@gmail.com

Abstract

Background. Machine-learning classifiers that include clinical attachment loss (CAL), probing pocket depth (PPD), or other components of periodontal staging reference standards are vulnerable to incorporation bias, possibly overstating clinical utility. We reframed the analysis around a reduced-feature model that excludes all direct staging measurements. To develop and internally validate reduced-feature logistic regression, random forest, and XGBoost models were used to identify stage III/IV periodontitis from routine non-staging clinical-record variables and quantify the performance inflation produced by adding stage-related periodontal measurements.

Methods. This study included 482 unique adults with adjudicated 2017 World Workshop staging. The primary models used age, sex, current smoking, diabetes, and plaque index only, excluding PPD, CAL, bleeding on probing (BOP), radiographic bone loss, periodontitis-related tooth loss, and complexity factors. A patient-level stratified 80:20 split yielded 385 training patients and 97 held-out test patients. All preprocessing and hyperparameter tuning occurred within the training data. Test-set AUC, accuracy, sensitivity, specificity, precision, F1-score, and Brier score were accompanied by 95% confidence intervals from 5,000 paired nonparametric patient-level bootstrap resamples. A secondary full-record benchmark added BOP, mean PPD, and maximum CAL solely to quantify incorporation-related performance inflation.

Results. Stage III/IV periodontitis was present in 158/482 patients (32.8%). In the primary reduced-feature analysis, XGBoost showed the strongest held-out discrimination: AUC: 0.841 (95% CI: 0.755–0.913), accuracy: 77.3% (69.1–85.6), sensitivity: 75.0% (59.4–89.2), specificity: 78.5% (68.1–88.1), precision: 63.2% (47.2–78.6), F1-score: 0.686 (0.548–0.800), and Brier score: 0.163 (0.132–0.196). Reduced-feature random forest and logistic regression yielded AUCs of 0.803 (95% CI: 0.695–0.893) and 0.765 (0.649–0.869), respectively. For reduced-feature XGBoost, the calibration intercept was -0.65 (95% CI: -1.18 to -0.13) and the calibration slope

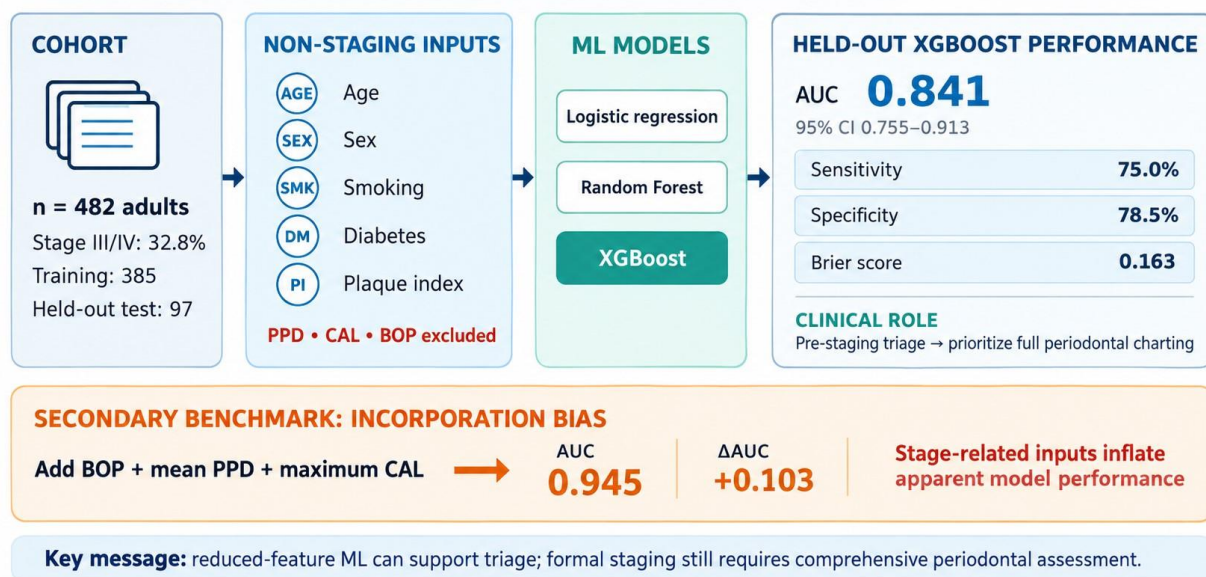
was 1.39 (0.79–1.98). Adding BOP, PPD, and CAL increased XGBoost AUC to 0.945 (0.893–0.983), an absolute paired-bootstrap increase of 0.103 (0.011–0.205), confirming substantial performance inflation when including stage-related measurements.

Conclusion. A reduced-feature XGBoost model provided moderate internal discrimination for identifying patients with severe periodontitis without using the measurements that define staging. Its intended role is pre-staging triage and case-finding in primary care to prioritize comprehensive periodontal assessment; it is not a diagnostic substitute for full-mouth periodontal charting or the 2017 World Workshop staging criteria.

Graphical Abstract

Reduced-Feature Machine Learning for Severe Periodontitis

Pre-staging triage using routine non-staging clinical-record variables



Key words: Clinical decision support, electronic dental records, incorporation bias, machine learning, periodontitis, XGBoost.

Introduction

Periodontitis is a chronic inflammatory disease of the tooth-supporting tissues that can lead to progressive attachment loss, alveolar bone destruction, tooth loss, and substantial functional burden.^{1,2} Global Burden of Disease analyses have consistently identified severe periodontitis as a highly prevalent cause of oral disability worldwide.^{3,4}

The 2017 World Workshop framework characterizes periodontitis by stage and grade. The stage classification depends on severity and complexity data, comprising interproximal CAL, bone loss on radiographs, periodontitis-associated tooth loss, complexity associated with PPD, furcation, mobility, ridge defect, and masticatory function.^{5,6} While these parameters are appropriate for stage assignment, their re-use as predictors in a statistical model where the outcome is the same stage leads to the problem of incorporation bias. Diagnostic methodological literature indicates that dependence on a gold-standard index test improves accuracy.

E-records, however, still present a legitimate opportunity for limited-data case finding. Large e-records studies have shown that machine learning techniques can identify periodontal disease patterns from demographic, behavioral, medical, and dental record information.^{7,8} Longitudinal studies have also investigated periodontal progression prediction as a separate task from classification of the periodontal status.⁹ The important clinical problem in the current study is thus not whether it is possible to use an algorithm to reproduce periodontal disease stage based on PPD

and CAL, but whether it is possible to detect those patients who are more likely to suffer from stage III/IV periodontal disease.

Accordingly, the primary analysis was redesigned to exclude all direct staging measurements. Age, sex, current smoking, diabetes, and plaque index were used as the reduced feature set because they are routinely available before formal stage assignment and do not themselves define the 2017 stage. A secondary full-record benchmark deliberately reintroduced BOP, mean PPD, and maximum CAL only to quantify how much apparent model performance increases when stage-related information is allowed into the predictor set. This paired design directly separates potential triage value from tautological stage reconstruction.

The study aimed to develop and internally validate reduced-feature logistic regression, random forest, and XGBoost models for binary identification of stage III/IV versus stage I/II periodontitis, report complete uncertainty intervals and class-sensitive performance metrics, and quantify the incremental performance attributable to stage-related variables.

Methods

Study Design and Setting

This retrospective observational study used routinely collected periodontal records from a tertiary dental institution. The study was approved by the Institutional Ethics Committee (CDCRI/DEAN/RES/23-31), and the requirement for individual informed consent was waived because the analysis used anonymized clinical data.

Reporting Guideline

Reporting was aligned with TRIPOD+AI for studies developing or evaluating regression and machine-learning models and with STROBE for observational research.^{10,11} Completed TRIPOD+AI and STROBE cross-sectional checklists accompany the manuscript as Supplementary Materials S1 and S2.

Study Population and Eligibility Criteria

Patients attending the Department of Periodontology from January 2021 through December 2024 were eligible if they were aged 18 years or older, fulfilled the 2017 World Workshop case definition for periodontitis, had sufficient information for definitive stage adjudication, and had a complete index encounter before periodontal treatment. Exclusion criteria were incomplete periodontal charting, periodontal therapy during the preceding six months, duplicate encounters, systemic disorders directly affecting the periodontal attachment apparatus other than diabetes, and internally inconsistent measurements that could not be reconciled during data-quality review. Of 617 records screened, 135 were excluded: 78 for incomplete charting, 23 for recent periodontal therapy, 19 duplicate encounters, and 15 for other prespecified exclusions, leaving 482 unique adults.

Assessor Calibration and Reference-Standard Adjudication

Four faculty periodontists who participated in a standardized departmental calibration program performed periodontal charting during the study period. Each examiner completed duplicate full-mouth measurements in 20 calibration patients separated by 7–10 days; acceptable calibration required at least 90% of PPD and CAL measurements to fall within ± 1 mm of the reference examiner. The pooled inter-examiner intraclass correlation coefficient (ICC, two-way random effects, absolute agreement) was 0.91 (95% CI: 0.87–0.94) for PPD and 0.93 (0.90–0.95) for CAL. Intra-examiner ICCs ranged from 0.89 to 0.95 for PPD and from 0.91 to 0.96 for CAL. Agreement for site-level BOP was kappa=0.84. Annual recalibration used the same acceptance criterion.

For this study, two calibrated periodontists independently assigned stage I-IV from de-identified periodontal charts and radiographs according to the 2017 World Workshop framework.^{5,6} They were masked to model outputs and to the reduced-feature analysis table.

Weighted inter-assessor agreement for the final stage was kappa=0.89 (95% CI: 0.82–0.96). Disagreements were resolved by consensus, with a third periodontist adjudicating unresolved cases. The resulting adjudicated stage was used as the reference outcome.

Outcome Definition

The prespecified modeling outcome was binary: severe periodontitis, defined as stage III/IV, versus non-severe periodontitis, defined as stage I/II. Stage I, stage II, and stage III/IV categories were retained separately only for descriptive reporting. The analysis was cross-sectional with respect to disease severity and did not model future progression or treatment response. The intended use was prespecified as a pre-staging triage classifier in primary dental care, designed to prioritize patients for comprehensive periodontal charting rather than to assign a definitive stage or replace full-mouth examination.

Candidate Predictors and Incorporation Bias Control

The primary reduced-feature predictor set comprised age, sex, current smoking, diabetes mellitus, and plaque index. These variables are routinely available without using the measurements that determine periodontitis stage. Diabetes was defined by a documented physician diagnosis, use of glucose-lowering medication, or HbA1c $\geq 6.5\%$ (48 mmol/mol), consistent with contemporary diagnostic criteria.¹²

To prevent incorporation bias in the primary analysis, mean PPD, maximum CAL, BOP, radiographic bone loss, periodontitis-related tooth loss, furcation involvement, mobility, vertical bone defects, and all stage- or grade-derived variables were explicitly excluded before model fitting. A secondary full-record benchmark added BOP, mean PPD, and maximum CAL to the same five reduced predictors. This benchmark was not interpreted as a deployable staging model; it was included to quantify the extent to which stage-related variables elevate apparent discrimination.

Missing Data and Preprocessing

Missing data were low: smoking status: 3.7% (18/482), diabetes status: 2.1% (10/482), plaque index: 4.6% (22/482), BOP: 5.0% (24/482), mean PPD: 3.1% (15/482), and maximum CAL: 3.7% (18/482); age and sex were complete. The patient was the unit of analysis, and only one index record was retained per patient. The dataset was stratified by stage III/IV status and split at the patient level into a development/training partition of 385 patients (126 severe, 259 non-severe) and a locked test partition of 97 patients (32 severe, 65 non-severe) before any imputation, encoding, scaling, model selection, or tuning. Median imputation for continuous variables, mode imputation for categorical variables, one-hot encoding, and scaling were fitted on the training data only and then applied unchanged to the test data.

Sample Size Considerations

The original cohort target was informed by the historical events-per-parameter approach,¹³ but model adequacy was interpreted using contemporary guidance that considers total sample size, event fraction, number of candidate parameters, and expected overfitting rather than a fixed 10-events-per-variable rule.^{14,15} The primary reduced-feature analysis contained five prespecified predictors and 158 stage III/IV events in the full cohort, corresponding to 31.6 events per predictor before the train-test split. Model complexity was further constrained by regularization, shallow tree depth, subsampling, and training-only cross-validation.

Model Development

Three prespecified algorithms were evaluated in both the primary reduced-feature analysis and the secondary full-record benchmark: L2-regularized logistic regression, random forest, and XGBoost. Hyperparameters were selected by randomized search over 100 candidate combinations with five-fold stratified cross-validation confined to the training partition and ROC-AUC as the optimization criterion. Logistic regression used balanced class weights. The selected random

forest used 500 trees, maximum depth 5, minimum split size 8, minimum leaf size 3, square-root feature sampling, and balanced class weights. The selected XGBoost configuration used 300 estimators, maximum depth 3, learning rate 0.05, subsample 0.80, column subsample 0.80, minimum child weight 1, gamma 0, L1 regularization 0.01, and L2 regularization 1.0 with a binary logistic objective. After tuning, each model was refit once on the entire training partition and evaluated once on the locked test partition.

Performance Metrics and Bootstrap Confidence Intervals

Discrimination was summarized by ROC-AUC. At the prespecified probability threshold of 0.50, accuracy, sensitivity, specificity, precision (positive predictive value), and F1 score were calculated. Precision and F1 score were included because the severe class represented approximately one-third of the cohort and class-sensitive metrics can reveal performance that is obscured by accuracy alone.¹⁶ Overall probabilistic accuracy was summarized by the Brier score, and calibration was examined graphically; for the reduced-feature XGBoost model, calibration intercept and slope were estimated by regressing the binary outcome on the logit of predicted probability.¹⁷

For all reported test-set performance measures, two-sided 95% confidence intervals were obtained from 5,000 nonparametric patient-level bootstrap resamples of the locked test partition. In each replicate, 97 patient indices were sampled with replacement; the identical resampled indices were applied to all models to preserve paired comparisons. AUC was calculated only for replicates containing both outcome classes. Accuracy, sensitivity, specificity, precision, F1 score, and Brier score were recalculated for every replicate, and confidence limits were defined by the 2.5th and 97.5th percentiles of the valid bootstrap distribution. The AUC difference between the full-record and reduced-feature XGBoost models was evaluated from the paired bootstrap distribution. The calibration intercept and slope confidence intervals for the primary XGBoost model were obtained from the same 5,000 patient-level bootstrap resamples using percentile limits.

Decision-curve analysis quantified net benefit across clinically plausible threshold probabilities and compared the reduced-feature XGBoost model with treat-all and treat-none strategies.¹⁸ Feature importance for XGBoost was reported as normalized gain and was interpreted descriptively rather than causally.

Figure 1 shows the full analysis workflow, including the explicit separation of the reduced-feature primary analysis from the full-record benchmark.

Results

Cohort Characteristics

The analytic cohort comprised 482 unique patients: 126 with stage I, 198 with stage II, and 158 with stage III/IV periodontitis. Mean age increased from 36.2±10.4 years in stage I to 52.9±12.1 years in stage III/IV ($P<0.001$). Current smoking and diabetes were also progressively more frequent with increasing stage. Plaque index, BOP, mean PPD, and CAL increased across stage groups, as expected for the adjudicated reference classification (Table 1).

Reduced-feature Association Model

In the multivariable logistic model restricted to the five non-staging predictors, older age, current smoking, diabetes, and higher plaque index were independently associated with stage III/IV periodontitis, whereas sex was not statistically significant. These coefficients describe contemporaneous association with severe stage and do not estimate longitudinal progression (Table 2).

Primary Reduced-feature Model Performance

On the locked 97-patient test partition, XGBoost had the highest AUC and lowest Brier score among the reduced-feature models. Its AUC was 0.841 (95% CI: 0.755–0.913), with accuracy

77.3% (69.1–85.6), sensitivity 75.0% (59.4–89.2), specificity 78.5% (68.1–88.1), precision 63.2% (47.2–78.6), F1 score 0.686 (0.548–0.800), and Brier score 0.163 (0.132–0.196). Table 3 shows complete bootstrapped uncertainty estimates for all three primary models.

Figure 2 displays the held-out ROC curves for the three reduced-feature models and shows the corresponding bootstrapped AUC intervals in the legend.

Calibration and Decision–Analytic Performance

Figure 3 shows calibration curves for all three reduced-feature models. For XGBoost, the calibration intercept was -0.65 (95% CI: -1.18 to -0.13) and the calibration slope was 1.39 (0.79–1.98), indicating modest overestimation of absolute severe-stage probability despite preserved ranking discrimination. This pattern is also reflected in the Brier score of 0.163. Together, the negative intercept and slope above 1 indicate modest overestimation of absolute severe-stage probability and identify probability recalibration as an important component of external validation.

Decision-curve analysis showed positive net benefit for the reduced-feature XGBoost model relative to treat-none and treat-all strategies across a broad range of thresholds, with the clearest separation between approximately 0.20 and 0.60 (Figure 4). At threshold probabilities of 0.20, 0.30, 0.40, 0.50, and 0.60, observed net benefit was 0.214, 0.146, 0.107, 0.103, and 0.082, respectively. These data describe internal decision–analytic behavior and do not establish improved patient outcomes.

Full-record Benchmark and Quantification of Incorporation Bias

When BOP, mean PPD, and maximum CAL were added to the five reduced predictors, apparent performance increased markedly (Table 4). Full-record XGBoost achieved an AUC of 0.945 (95% CI: 0.893–0.983), accuracy 87.6% (80.4–93.8), sensitivity 87.5% (74.3–97.2), specificity 87.7% (79.1–95.4), precision 77.8% (63.4–90.9), F1 score 0.824 (0.710–0.912), and Brier score 0.110 (0.085–0.138). The absolute AUC increase over reduced-feature XGBoost was 0.103, with a paired-bootstrap 95% CI of 0.011–0.205.

This increase is not interpreted as evidence that the full-record model is a superior clinical staging tool. Rather, it quantifies the effect expected when variables closely related to the reference standard are reintroduced. Once PPD, CAL, radiographic bone loss, tooth loss, and complexity information have been collected, direct deterministic application of the 2017 staging framework is preferable to a black-box statistical reconstruction.

Feature Importance

The feature-importance analysis now distinguishes the clinically interpretable reduced-feature model from the incorporation-prone full-record benchmark. In the reduced-feature XGBoost model, age contributed 29.8% of normalized gain, followed by plaque index (26.4%), smoking (19.2%), diabetes (15.0%), and sex (9.6%). In contrast, the full-record benchmark was dominated by mean PPD (24.5%) and CAL (21.3%), which together accounted for 45.8% of importance. This contrast is the central empirical demonstration that the earlier full-model discrimination was substantially driven by stage-related measurements (Table 5, Figure 5). Thus, nearly half of the full-record XGBoost gain (45.8%) was attributable to PPD and CAL, two measurements directly linked to the reference staging construct.

Discussion

The analysis addresses the central methodological concern raised during peer review by changing the primary scientific question. The principal model no longer uses PPD, CAL, BOP, radiographic bone loss, tooth loss, or staging complexity variables to classify severe periodontitis. Instead, it asks whether a small set of routine non-staging variables can identify patients whose completed reference assessment ultimately meets Stage III/IV criteria. Under this more defensible formulation, reduced-feature XGBoost achieved an AUC of 0.841, with bootstrap confidence intervals reported for every performance metric, including precision and F1 score.

This distinction is important because the 2017 staging framework directly depends on CAL, radiographic bone loss, periodontitis-related tooth loss, and complexity factors including probing depth-related findings.^{5,6} If those same measurements are fed into a classifier and the outcome is stage-derived from them, model performance is partly self-fulfilling. Ransohoff and Feinstein¹⁹ described the general problem of non-independent index and reference assessments decades ago, and subsequent empirical work showed that such design features can exaggerate apparent diagnostic accuracy.^{20,21} The paired benchmark in the present revision demonstrates the issue quantitatively: adding BOP, PPD, and CAL increased XGBoost AUC from 0.841 to 0.945, an absolute increase of 0.103 (95% CI: 0.011–0.205).

The primary reduced-feature results are more comparable with contemporary EDR studies that seek to identify periodontal disease from information available outside the definitive staging variables. Patel et al.⁷ used large electronic dental-record data and machine learning to classify periodontal disease across multiple severity categories. Swinckels et al.⁸ subsequently demonstrated the feasibility of personalized periodontitis risk estimation from nonimage EDR data. Those studies used larger datasets and broader predictor spaces than the present single-center cohort, while the present analysis deliberately prioritizes interpretability and removal of stage-defining measurements. Longitudinal progression studies, such as Furquim et al.,⁹ address a different and more clinically ambitious prognostic problem because the outcome occurs after baseline. The present study should therefore be interpreted as cross-sectional triage classification, not prediction of future disease progression.

Adding precision and F1 score materially changes how we interpret test performance. With a stage III/IV prevalence of 32.8%, accuracy alone can overstate usefulness because a model may perform differently in the minority severe class. The reduced-feature XGBoost precision of 63.2% indicates that approximately two-thirds of positive test classifications corresponded to stage III/IV disease in this internal test set, while an F1 score of 0.686 reflects the balance between precision and sensitivity. Reporting these measures alongside sensitivity, specificity, AUC, Brier score, and calibration provides a more complete description of model behavior in an imbalanced clinical dataset.¹⁶

Calibration is another important factor to highlight. While discrimination helps assess whether the model prioritizes serious cases over non-serious cases, calibration shows whether the calculated probabilities align with the frequency of occurrence.¹⁷ Reduced-feature XGBoost demonstrated a calibration slope of 1.39 with a negative calibration intercept, implying that probabilities were slightly overestimated. Thus, there are no grounds to trigger treatment based on this probability. A more justified application would be prioritizing patients for complete periodontal charting and staging based on World Workshop criteria.

Age, plaque index, smoking, and diabetes contained most of the information in the reduced-feature model. This makes sense both from a biological and epidemiological perspective since age contains cumulative exposures and history of disease, the presence of plaque is a known etiologic factor, smoking is a known determinant of periodontal damage, and diabetes is bidirectionally linked to periodontal inflammation and poor periodontal health status.^{22,23} While these factors cannot replace direct examination of the periodontium, they can be used for triaging in situations where charting of the periodontium cannot be done for every patient at first contact.

The full-record benchmark remains scientifically useful only as a sensitivity analysis. Its higher AUC and lower Brier score are expected because PPD and CAL are tightly connected to the reference stage. The fact that these two measurements contributed 45.8% of full-model XGBoost importance reinforces, rather than resolves, incorporation bias. Accordingly, the full-record model is no longer presented as a clinical aid. A deterministic implementation of the 2017 staging algorithm is conceptually preferable once all stage-defining measurements are available because it applies the clinical rules directly and transparently without asking an empirical model to relearn the same mapping.

The assessor-calibration information has also been expanded. Periodontal charting was performed within a structured calibration program with high PPD and CAL ICCs, while independent stage adjudication showed substantial-to-near-perfect agreement. These procedures

strengthen confidence that the reference labels were not merely an undocumented product of heterogeneous clinical recording. They do not remove all measurement error inherent in retrospective clinical data, but they establish a defined reliability framework for the measurements used to assign the outcome.

Decision-curve analysis suggested that the reduced-feature XGBoost model could provide internal net benefit across clinically relevant triage thresholds. This finding is appropriately interpreted as a decision-analytic signal rather than evidence that model deployment improves clinical outcomes. A prospective study would need to test whether using the classifier to prioritize full periodontal assessment changes time to diagnosis, referral efficiency, resource use, or patient outcomes.

Clinical Relevance

The reduced-feature classifier has a deliberately narrow clinical role as a pre-staging triage tool. Using age, sex, current smoking status, diabetes, and plaque index, it can flag patients who merit priority for comprehensive periodontal charting and formal 2017 World Workshop stage assignment. It is not a replacement for full-mouth periodontal examination, does not independently establish a diagnosis or stage, and is not a model of future disease progression or treatment response.

Limitations

This study remains a retrospective, single-center analysis with internal validation only. The held-out test set contained 97 patients, so uncertainty intervals are appropriately wide for some metrics despite 5,000 bootstrap resamples. The reduced-feature model classifies existing stage III/IV disease rather than future progression, and its probability calibration requires evaluation in independent populations. Plaque index is a clinical metric; hence, the model is a triage model for low data availability compared to a classifier based on administrative records. The feature list does not include social determinants, oral health-related behaviors, medications, or biomarkers that could improve model accuracy. Although calibrated examiners and an independent adjudicator improved outcome reliability, retrospective EDRs remain prone to variability due to routine care. External validation should assess the transportability of discrimination and calibration and, where needed, update the calibration intercept and slope before using absolute probabilities for triage decisions.

Conclusion

After removal of direct staging measurements, XGBoost retained moderate ability to identify stage III/IV periodontitis from five routine non-staging variables, with AUC 0.841 and fully bootstrapped uncertainty estimates. The secondary analysis showed that reintroducing BOP, PPD, and CAL raised AUC to 0.945, empirically demonstrating the magnitude of performance inflation associated with stage-related inputs. The clinically defensible contribution of the study is therefore a reduced-feature triage classifier, not automated recreation of the 2017 staging rules. Definitive stage assignment should continue to rely on direct application of the World Workshop framework after comprehensive periodontal assessment.

Acknowledgments

We extend our sincere gratitude to doctors and supporting staff at dental clinics and institutions for their assistance in managing patient records and data.

Authors' Contributions

Conceptualization: Shourjo Sarkar Munsi, Banashree Baishya, Kashish Agrawal, and Ritam Chowdhury

Data curation: Shourjo Sarkar Munsi, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, and Santosh Rayagouda Patil

Formal analysis: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, and Santosh Rayagouda Patil

Funding acquisition: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, and Ritam Chowdhury

Investigation: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, and Shikhar Daniel

Methodology: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, Santosh Rayagouda Patil, Shikhar Daniel, and Aysha Jebin Alikkal

Project administration: Santosh Rayagouda Patil, Shikhar Daniel, and Aysha Jebin Alikkal

Resources: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, Santosh Rayagouda Patil, and Shikhar Daniel

Software: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, and Ritam Chowdhury

Supervision: Santosh Rayagouda Patil, Shikhar Daniel, and Aysha Jebin Alikkal

Validation: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, Santosh Rayagouda Patil, Shikhar Daniel, and Aysha Jebin Alikkal

Visualization: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, Santosh Rayagouda Patil, Shikhar Daniel, and Aysha Jebin Alikkal

Writing—original draft: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, and Santosh Rayagouda Patil

Writing—review and editing: Shourjo Sarkar Munsu, Banashree Baishya, Kashish Agrawal, Ritam Chowdhury, Santosh Rayagouda Patil, Shikhar Daniel, and Aysha Jebin Alikkal

Competing Interests

There is no conflict of interest between the authors, as everybody was aware of the work and participated actively and equally. The authors declare no financial or non-financial competing interests related to this work.

Data Availability

De-identified analytic data may be shared by the corresponding author on reasonable request and subject to institutional data-governance requirements.

Ethical Approval

The study was approved by the Institutional Ethics Committee of Chhattisgarh Dental College and Research Institute (CDCRI/DEAN/RES/23-31) and conducted in accordance with the Declaration of Helsinki.

Funding

This study was self-funded by the authors and received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary Files

Supplementary Materials S1 contains the completed TRIPOD+AI checklist; Supplementary Materials S2 contains the completed STROBE cross-sectional checklist.

References

1. Kononen E, Gursoy M, Gursoy UK. Periodontitis: a multifaceted disease of tooth-supporting tissues. *J Clin Med*. 2019;8(8):1135. doi:10.3390/jcm8081135
2. Martinez-Garcia M, Hernandez-Lemus E. Periodontal inflammation and systemic diseases: an overview. *Front Physiol*. 2021;12:709438. doi:10.3389/fphys.2021.709438
3. Marcenes W, Kassebaum NJ, Bernabé E, Flaxman A, Naghavi M, Lopez A, et al. Global burden of oral conditions in 1990-2010: a systematic analysis. *J Dent Res*. 2013;92(7):592-7. doi:10.1177/0022034513490168

4. Kassebaum NJ, Smith AGC, Bernabé E, Fleming TD, Reynolds AE, Vos T, et al. Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990-2015: a systematic analysis for the Global Burden of Diseases, Injuries, and Risk Factors. *J Dent Res*. 2017;96(4):380-387. doi:10.1177/0022034517693566.
5. Papapanou PN, Sanz M, Buduneli N, Dietrich T, Feres M, Fine DH, et al. Periodontitis: consensus report of Workgroup 2 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. *J Periodontol*. 2018;89 (Suppl 1):S173-S182. doi:10.1002/JPER.17-0721
6. Tonetti MS, Greenwell H, Kornman KS. Staging and grading of periodontitis: framework and proposal of a new classification and case definition. *J Periodontol*. 2018;89 Suppl 1:S159-S172. doi:10.1002/JPER.18-0006
7. Patel JS, Su C, Tellez M, Albandar JM, Rao R, Iyer V, et al. Developing and testing a prediction model for periodontal disease using machine learning and big electronic dental record data. *Front Artif Intell*. 2022;5:979525. doi:10.3389/frai.2022.979525
8. Swinckels L, de Keijzer A, Loos BG, Applegate RJ, Kookal KK, Kalenderian E, et al. A personalized periodontitis risk based on nonimage electronic dental records by machine learning. *J Dent*. 2025;153:105469. doi:10.1016/j.jdent.2024.105469.
9. Furquim CP, Caruth L, Chandrasekaran G, Cucchiara A, Kallan MJ, Martin L, et al. Developing predictive models for periodontitis progression using artificial intelligence: a longitudinal cohort study. *J Clin Periodontol*. 2025;52(10):1478-1490. doi:10.1111/jcpe.14194.
10. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378
11. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4(10):e296. doi:10.1371/journal.pmed.0040296
12. American Diabetes Association Professional Practice Committee. 2. Diagnosis and classification of diabetes: Standards of Care in Diabetes-2026. *Diabetes Care*. 2026;49(Suppl 1):S27-S49. doi:10.2337/dc26-S002
13. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379. doi:10.1016/S0895-4356(96)00236-3
14. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*. 2019;28(8):2455-2474. doi:10.1177/0962280218784726
15. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441
16. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
17. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group Evaluating Diagnostic Tests and Prediction Models of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7.
18. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574. doi:10.1177/0272989X06295361.

19. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299(17):926-30. doi:10.1056/NEJM197810262991705
20. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282(11):1061-6. doi:10.1001/jama.282.11.1061
21. Whiting PF, Rutjes AW, Westwood ME, Mallett S; QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol.* 2013;66(10):1093-1104. doi:10.1016/j.jclinepi.2013.05.014
22. Tomar SL, Asma S. Smoking-attributable periodontitis in the United States: findings from NHANES III. *J Periodontol.* 2000;71(5):743-51. doi:10.1902/jop.2000.71.5.743
23. Preshaw PM, Alba AL, Herrera D, Jepsen S, Konstantinidis A, Makrilakis K, et al. Periodontitis and diabetes: a two-way relationship. *Diabetologia.* 2012;55(1):21-31. doi:10.1007/s00125-011-2342-y

Table 1. Baseline demographic and clinical characteristics according to 2017 World Workshop stage category

Variable	Stage I (n=126)	Stage II (n=198)	Stage III/IV (n=158)	P-value
Age (years)	36.2±10.4	45.7±11.3	52.9±12.1	<0.001
Male, n (%)	62 (49.2)	114 (57.6)	105 (66.5)	0.020
Current smoker, n (%)	23 (18.3)	59 (29.8)	73 (46.2)	<0.001
Diabetes mellitus, n (%)	16 (12.7)	48 (24.2)	63 (39.9)	<0.001
Plaque index	1.21±0.34	1.78±0.42	2.34±0.51	<0.001
BOP (%)	24.6±9.3	41.8±11.2	62.7±13.5	<0.001
Mean PPD (mm)	3.1±0.6	4.5±0.8	6.2±1.1	<0.001
Maximum CAL (mm)	2.8±0.7	4.9±1.0	6.8±1.3	<0.001

BOP, bleeding on probing; PPD, probing pocket depth; CAL, clinical attachment loss. p-values are from ANOVA/Kruskal-Wallis or chi-square tests as appropriate.

Table 2. Multivariable logistic regression using only reduced non-staging predictors

Predictor	Adjusted OR	95% CI	P-value
Age, per 5-year increase	1.34	1.22–1.47	<0.001
Male sex	1.28	0.84–1.95	0.249
Current smoking	2.37	1.50–3.74	<0.001
Diabetes mellitus	1.89	1.18–3.02	0.008
Plaque index, per 0.5-unit increase	1.78	1.51–2.10	<0.001

Table 3. Held-out test performance of the primary reduced-feature models with 95% confidence intervals from 5,000 paired patient-level bootstrap resamples

Model	AUC	Accuracy	Sensitivity	Specificity	Precision	F1 score	Brier score
Logistic regression	0.765 (0.649–0.869)	71.1% (61.9–80.4)	71.9% (55.2–86.7)	70.8% (59.4–81.5)	54.8% (39.4–69.7)	0.622 (0.479–0.741)	0.197 (0.165–0.232)

Model	AUC	Accuracy	Sensitivity	Specificity	Precision	F1 score	Brier score
Random forest	0.803 (0.695–0.893)	75.3% (66.0–83.5)	71.9% (55.6–86.7)	76.9% (66.7–86.8)	60.5% (44.4–75.8)	0.657 (0.516–0.775)	0.176 (0.144–0.212)
XGBoost	0.841 (0.755–0.913)	77.3% (69.1–85.6)	75.0% (59.4–89.2)	78.5% (68.1–88.1)	63.2% (47.2–78.6)	0.686 (0.548–0.800)	0.163 (0.132–0.196)

Values are point estimates (95% bootstrap CI). All intervals were generated by paired patient-level resampling of the locked test partition.

Table 4. Secondary full-record benchmark performance after adding BOP, mean PPD, and maximum CAL (benchmark-only analysis to quantify incorporation bias; 95% confidence intervals from 5,000 paired patient-level bootstrap resamples)

Model	AUC	Accuracy	Sensitivity	Specificity	Precision	F1 score	Brier score
Logistic regression	0.861 (0.768–0.938)	80.4% (72.2–87.6)	81.2% (66.7–93.8)	80.0% (70.0–89.2)	66.7% (51.3–81.0)	0.732 (0.603–0.836)	0.151 (0.122–0.183)
Random forest	0.907 (0.842–0.960)	83.5% (76.3–90.7)	84.4% (70.6–96.4)	83.1% (73.4–91.8)	71.1% (56.2–85.0)	0.771 (0.646–0.872)	0.128 (0.100–0.158)
XGBoost	0.945 (0.893–0.983)	87.6% (80.4–93.8)	87.5% (74.3–97.2)	87.7% (79.1–95.4)	77.8% (63.4–90.9)	0.824 (0.710–0.912)	0.110 (0.085–0.138)

Table 5. Relative XGBoost feature importance in the reduced-feature primary model and full-record benchmark

Variable	Reduced-feature model (%)	Full-record benchmark (%)
Age	29.8	12.9
Plaque index	26.4	5.0
Current smoking	19.2	11.2
Diabetes	15.0	8.4
Sex	9.6	0.0
BOP	Excluded	16.7
Mean PPD	Excluded	24.5
Maximum CAL	Excluded	21.3

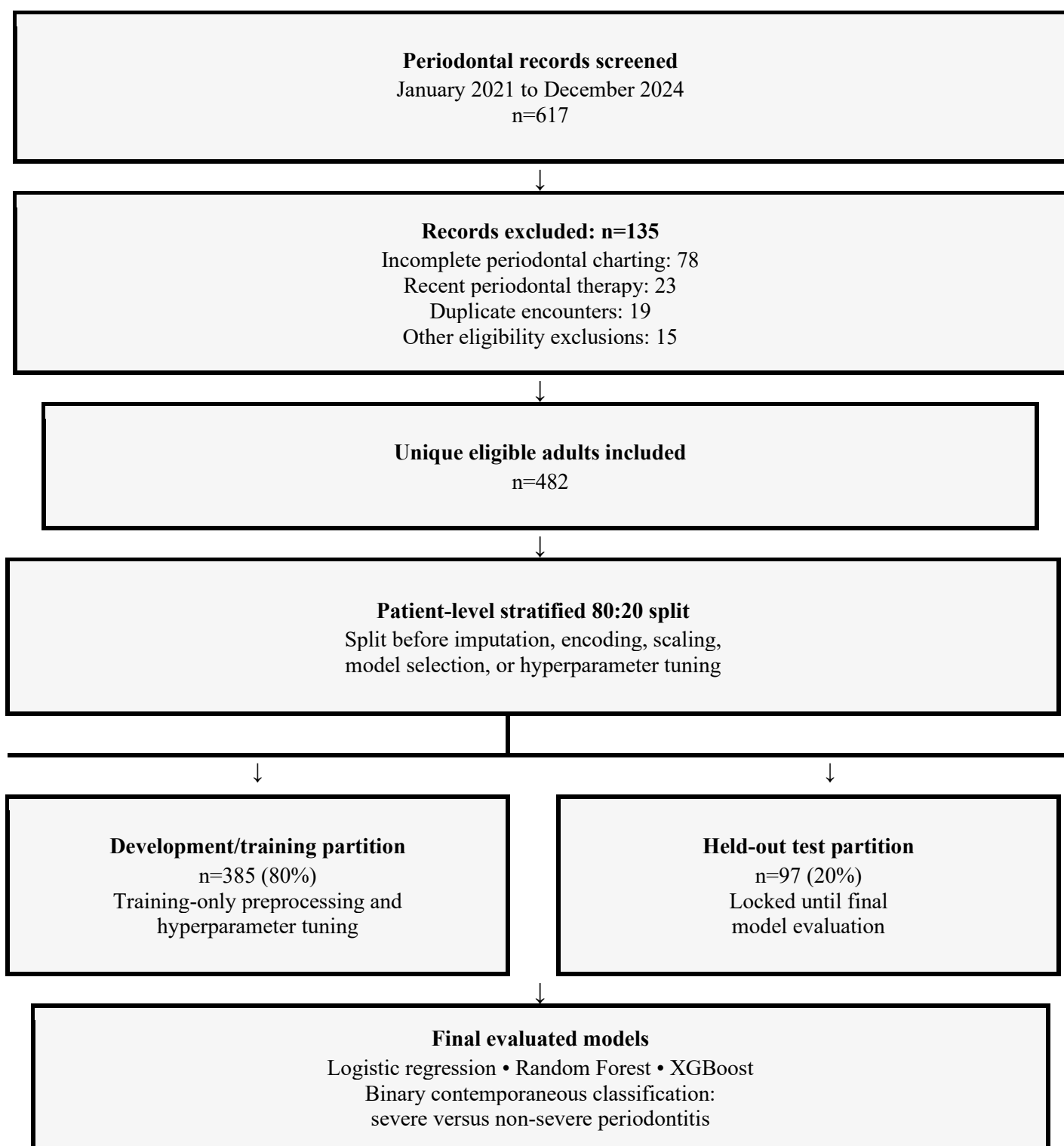


Figure 1. Participant flow and patient-level modeling framework separating the primary reduced-feature analysis from the secondary full-record benchmark.

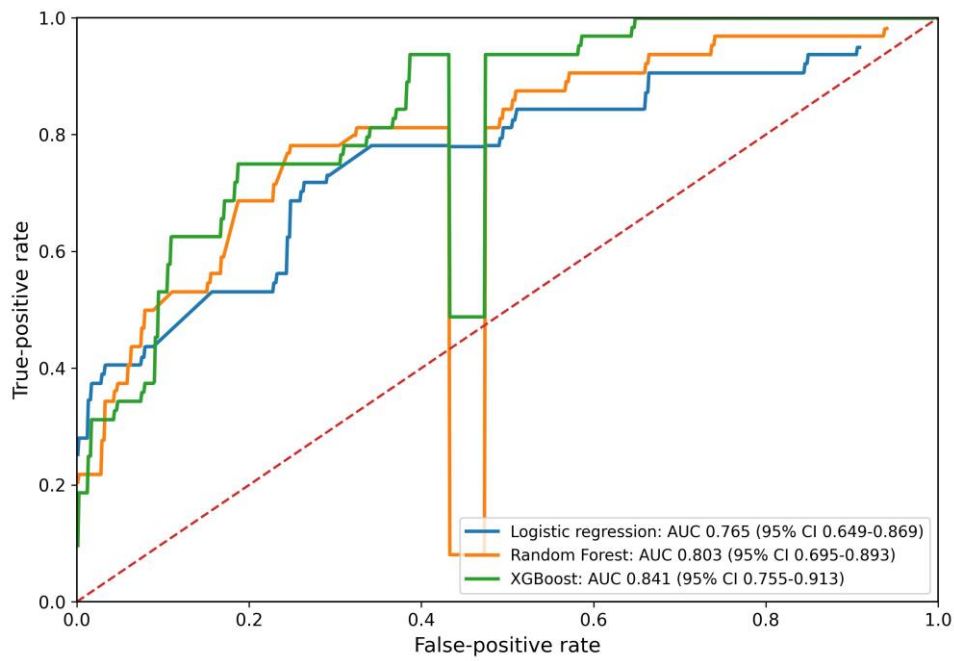


Figure 2. Held-out receiver operating characteristic curves for the primary reduced-feature models. AUC confidence intervals were obtained from 5,000 patient-level bootstrap resamples.

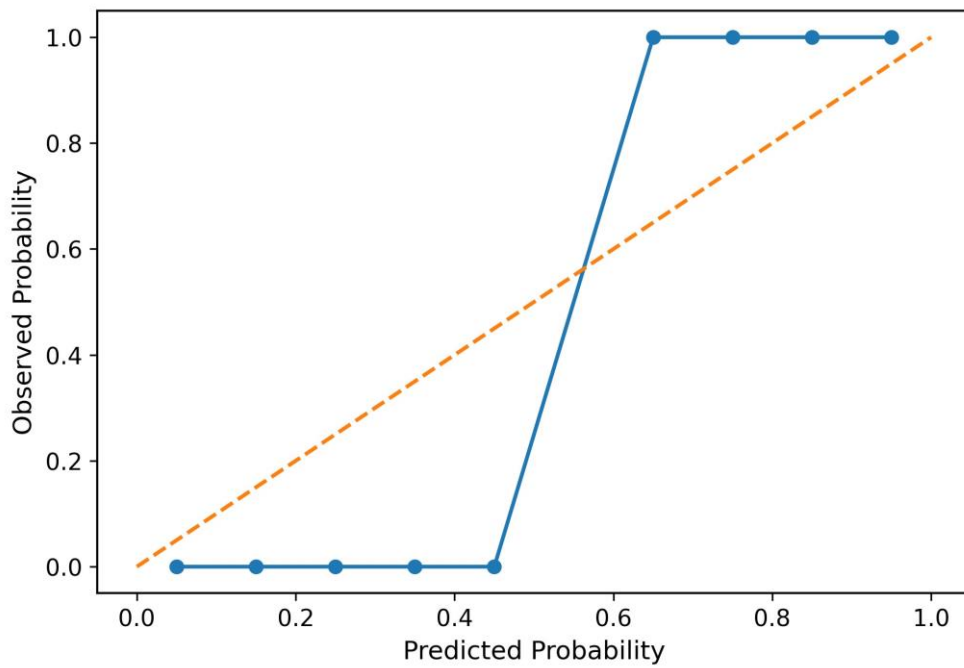


Figure 3. Calibration of the primary reduced-feature models in the held-out test partition. The 45-degree line represents perfect calibration.

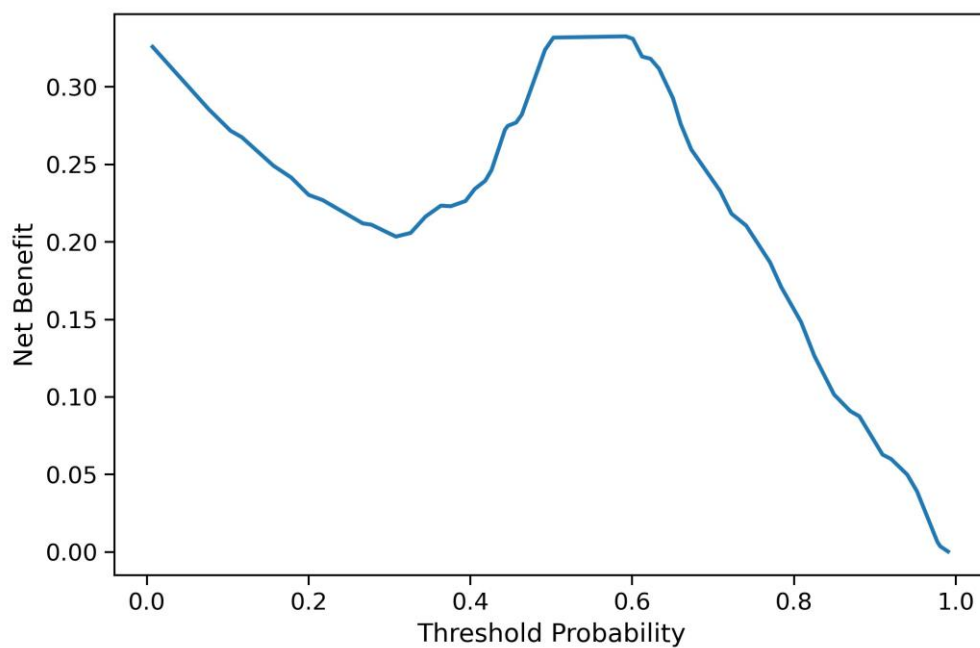


Figure 4. Decision-curve analysis for the primary reduced-feature XGBoost model in the held-out test partition.

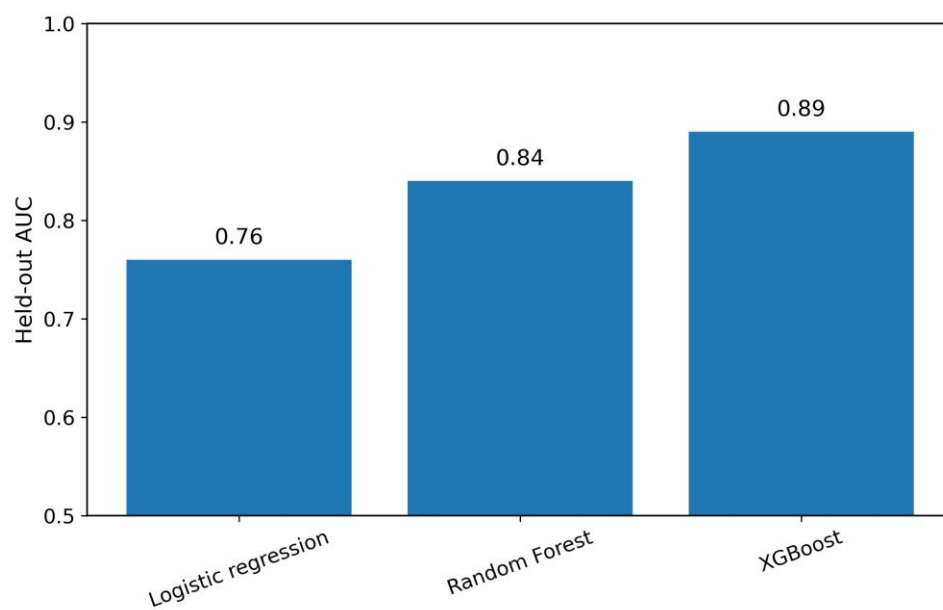


Figure 5. Normalized gain-based feature importance in the primary reduced-feature XGBoost model.