# Predicting the Severity of Postoperative Symptoms Following Mandibular Third Molar Extractions Using Machine Learning Techniques

*Ann. Ital. Chir.*, 2026 97, 1: 162–174
https://doi.org/10.62713/aic.4090

Qianqian Hou[1], Huan Ge[1], Jiayue Xiang[1], Yiming Gao[1,2]

[1]Department of Stomatology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 200025 Shanghai, China
[2]College of Stomatology, Shanghai Jiao Tong University, National Center for Stomatology, National Clinical Research Center for Oral Diseases, Shanghai Key Laboratory of Stomatology, 200011 Shanghai, China

AIM: This study aims to develop and externally validate machine-learning models that effectively predict the risk and severity of postoperative symptoms one week following mandibular third molar extractions.
METHODS: This retrospective cohort study included 321 patients (18–35 years old) who underwent lower third-molar surgery. Demographics, Pell–Gregory vertical (PGV) and Pell-Gregory level (PGL) classifications, surgical variables, and day-7 pain visual analogue scale (VAS) were recorded for all participants. The data were randomly divided into training (70%) and validation (30%) datasets. Five machine-learning algorithms—Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), and Neural Network (NNET)—were developed using nested cross-validation. Model performance was assessed through area under the receiver operating characteristic (AUROC) values, Brier scores, and calibration slopes, with a nomogram constructed from the best-performing model.
RESULTS: GBM achieved the highest discrimination on the validation dataset with an AUROC of 0.687 (95% CI, 0.624–0.744), followed by the Neural Network (AUROC = 0.677). The GBM model yielded a calibration slope of 0.98 and a Brier score of 0.225, indicating excellent predictive accuracy. However, the top six predictors were found to be operative time, mouth opening, PGV, PGL, smoking, and preoperative symptoms. The GBM model, which underlies the nomogram, achieved an area under the curve (AUC) value of 0.666, indicating its discrimination capability. Additionally, the calibration curve confirmed the model's accuracy, and the decision curve analysis (DCA) suggested that the nomogram provides clinically promising potential for effective risk stratification.
CONCLUSIONS: A GBM-based nomogram provides moderate yet clinically useful discrimination for healthy adults aged 18–35 years at risk for severe early symptoms after third-molar extraction. However, this approach requires external validation in older or medically complex patients before it is recommended for clinical predictions.

**Keywords:** third molar; machine learning; postoperative symptom severity; surgery

## Introduction

The surgical extraction of mandibular third molars (M3Ms), often indicated for clinical conditions such as impaction and recurrent infections, is a fundamental aspect of oral and maxillofacial surgery [1,2]. Although routinely performed, this procedure is commonly associated with postoperative complications, including pain, swelling, trismus, and alveolar osteitis, that can significantly impact a patient's quality of life [3,4]. Accurately predicting the severity of these symptoms could enhance postoperative management and overall patient outcomes.

*Submitted: 31 March 2025 Revised: 30 June 2025 Accepted: 15 July 2025 Published: 10 January 2026*
*Correspondence to: Yiming Gao, Department of Stomatology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, 200025 Shanghai, China; College of Stomatology, Shanghai Jiao Tong University, National Center for Stomatology, National Clinical Research Center for Oral Diseases, Shanghai Key Laboratory of Stomatology, 200011 Shanghai, China (e-mail: gaoyiming0808@163.com).*
*Editor: Guido Gabriele*

Recent advances in machine learning (ML) have demonstrated significant potential in predicting postoperative outcomes across various medical disciplines [5]. ML algorithms have been effectively employed to develop predictive models that anticipate postoperative complications [5,6]. Deep learning methods have been utilized to assess surgical difficulty in extracting impacted M3Ms [7]. These algorithms offer a promising framework for crafting predictive models that enable personalized treatment plans and enhance clinical outcomes. Such models can significantly inform clinical decision-making, optimize resource allocation, and ultimately result in better patient management and overall outcomes [8]. Moreover, the establishment of ML-based predictive nomograms offers an innovative approach in medicine [9], providing intuitive and individualized risk assessments through visually accessible tools.

Advancements in precision medicine have highlighted the critical role of individualized prevention and management strategies in reducing surgical trauma and minimizing postoperative complications [10]. Traditionally, the assessment of risk factors associated with postoperative complications

following M3M extraction has been guided by clinical experience and observational studies, emphasizing radiological, anatomical, and intraoperative factors [11,12]. However, these traditional approaches often lack the precision and objectivity that ML models can offer. Although some research has incorporated these factors to address extraction difficulty, there remains a need to compare various ML algorithms in terms of predictive accuracy. By integrating ML techniques with radiographic assessments and clinical data, more accurate models can be developed to predict the severity of postoperative symptom severity following M3Ms extractions, thereby improving treatment planning and patient outcomes.

This study aims to leverage the potential of ML to develop a robust predictive model for evaluating the severity of postoperative symptoms following M3M extractions. By comparing the performance of various ML algorithms, we seek to identify the most effective approach and create a clinically applicable, user-friendly nomogram. The findings seek to contribute to the advancement of personalized oral healthcare by facilitating more informed surgical planning and improving patient care.

## Methods

### *Study Design and Population*

This study employed a retrospective cohort design to develop and evaluate an ML-based predictive model for assessing the risk of postoperative symptom severity (PSS) following M3Ms extractions. The cohort comprised 321 patients who underwent M3M extractions at the Department of Stomatology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, China, between 1 June 2022 and 30 April 2023. The study protocol was rigorously reviewed and approved by the Institutional Ethics Committee of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine (Approval No: ruijin-eth-2023-268), ensuring compliance with ethical guidelines and the protection of patient confidentiality. Informed consent was obtained from all participants before their inclusion in the study.

### *Inclusion and Exclusion Criteria*

The inclusion criteria for patient recruitment were as follows: patients aged between 18 and 35 years, scheduled for elective extraction of one or both mandibular third molars under local anaesthesia, classified as American Society of Anesthesiologists (ASA) physical status I or II, and those capable of providing written informed consent and completing all follow-up assessments.

Exclusion criteria during patient selection were set as follows:

Presence of systemic disorder increasing the ASA status to III or higher, including but not limited to:

- Cardiovascular disease requiring monitoring, implanted cardiac devices, or anti-arrhythmic therapy.
- Uncontrolled endocrine disease (e.g., HbA1c $>8\%$ or insulin-dependent diabetes).
- Coagulopathy or ongoing anticoagulant treatment that cannot be interrupted for $\geq 6$ hours.
- Chronic renal, hepatic, or respiratory failure.
- Immunodeficiency or ongoing chemotherapy.
- Pregnancy or lactation.
- Diagnosis of acute pericoronitis or other active oral infection.
- Use of analgesics, anti-inflammatory drugs, or antibiotics within 48 hours pre-surgery.
- History of mandibular trauma or previous jaw surgery.
- Known allergy to local anaesthetics.

All 321 participants included in the final model satisfied these criteria, and none required intraoperative cardiac monitoring.

### *Data Collection and Variables*

Demographic, anatomical, radiographic, and operative data were collected for all participants: Demographic variables included gender, age, body mass index (BMI), and smoking status. Tooth-specific variables, including root morphology, root number, and root curve, were also recorded. Anatomical variables included mouth opening, measured as the interincisal distance (in mm) during surgical positioning [13].

Radiographic variables were subcategorized into tooth position and tooth-specific features. Tooth position was determined using Winter's classification (vertical, mesioangular, horizontal, or distoangular) and the Pell-Gregory classification system [14]. Root number was defined as single, double, and triple roots.

Operative variables included the medical history of pericoronitis, operative time, type of surgical procedure, flap design, and surgeon's experience level. The presence or absence of preoperative pericoronitis symptoms was recorded as part of medical history. Procedure type was categorized as elevator/forceps alone, bone removal and/or tooth sectioning, or combined bone removal and tooth sectioning. The flap design was categorized as none, relaxing, or triangular. Surgical experience or expertise was defined the number of years since the completion of residency and categorized as 0–5 years, 5–10 years, or $>10$ years [13]. The observed operative time was defined as the interval between the initial incision and the placement of the final suture [15]. Cone-beam computed tomography (CBCT)-derived metrics, such as inferior alveolar nerve (IAN)-root distance and crown-integrity grading were not analyzed, as CBCT imaging was not routinely conducted for all patients.

### *Surgical Procedure*

Third molar extractions were performed using a standardized approach established at Ruijin Hospital. Prior to surgery, patients underwent a 1-minute rinse with 0.2% chlorhexidine mouthwash, followed by local disinfection of the oral cavity with iodine solution. However, postop-

erative management included a 2-day course of oral antibiotics, comprising cefprozil 0.5 g (Yinlishu® Cefprozil Dispersible Tablets, 0.25 g per tablet; National Drug Approval No. H20052514; Guangzhou Baiyunshan Pharmaceutical Group Co., Ltd., Guangzhou, China) in combination with metronidazole 1.2 g (Wuyao® metronidazole tablets, 0.2 g per tablet; National Drug Approval No. H42021947; Farmito (China) Co., Ltd., Wuhan, China). Sutures were usually removed 7 days after surgery [16]. A preoperative assessment of panoramic radiographs was performed by an experienced surgeon blinded to the clinical outcomes.

*Outcome Assessment*

Among the available evaluation instruments, the postoperative symptom severity (PoSSe) questionnaire is specifically designed for mandibular third-molar surgery and shows greater sensitivity to acute postoperative changes compared to general tools such as the medical outcomes study (MOS) item short from health survey, the short form-36 health survey (SF-36) or oral health impact profile-14 (OHIP-14) [17–19]. Therefore, PoSSe was adopted as the primary outcome measure in the present study. The PoSSe approach was administered on postoperative day 7 when the sutures were removed. Patients were divided into high-risk and low-risk groups based on median total PoSSe scores [17,20]. Patients with PoSSe scores strictly above the median were classified as the high-risk group (value = 1), while those with scores equal to or below the median were classified as the low-risk group (value = 0). A higher PoSSe score indicates more severe postoperative symptoms.

*Machine Learning Model Development*

To develop and validate the predictive models, the dataset was randomly partitioned into a training set (n = 224, 70%) and a validation set (n = 97, 30%). Five machine learning algorithms were utilized to develop models predicting postoperative symptom severity risk (PSSR): Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), and Neural Network (NNET). The performance of the models was evaluated using the area under the receiver operating characteristic (AUROC) curve, along with sensitivity, specificity, and accuracy. These assessment indicators were selected a priori based on documented performance in similar medical-predictive tasks and their diverse bias-variance profiles. Deep-learning and distance-based classifiers were excluded due to the high risk of over-fitting and poor calibration associated with small, heterogeneous datasets.

*Statistical Analyses*

All statistical analyses were performed using R software, version 4.2.1 (R Foundation for Statistical Computing, Vienna, Austria). The ML models were developed using the R packages "caret", "e1071", "randomforest", "net", "gbm", part", "GLM", and "pROC", while the "rms" packages were used for constructing the nomogram. Model performance was assessed based on sensitivity, specificity, and accuracy, F1 score, and Brier score. Calibration plots and decision curve analysis (DCA) were generated with 1000-sample bootstrapping using the rms and rmda R packages. All statistical tests were two-sided, and a $p$-value of $<0.05$ was considered statistically significant.

Categorical variables were compared with the Pearson $\chi^2$ test. For $2 \times 2$ contingency tables where all expected frequencies were $>5$, Pearson's $\chi^2$ test was applied without Yates' continuity correction. If any expected frequency was $\leq 5$, either Yates' correction or Fisher's exact test was used, as appropriate. Normality of continuous variables was assessed employing the Shapiro–Wilk test ($\alpha = 0.05$). Skewed variables were expressed as median with interquartile range (IQR) and compared using the Mann–Whitney U test.

For ML model development, the dataset was randomly stratified into two groups: the training (70%) and validation (30%) sets. This random splitting was repeated until an equal distribution of patient characteristics was achieved across both sets. Consequently, five ML algorithms were developed: GBM, XGBoost, RF, DT, and NNET. During the training process, tuning was considered for ML-based models to minimize overfitting and model hyperparameters were optimized through 5-fold cross-validation.

Then, each model was trained to predict the risk of PSSR using the training data and performance indicators, including AUROC, sensitivity, specificity, and overall accuracy, were all calculated in the validation set. Models with AUR values closer to 1 were considered better in classification performance. Afterwards, based on the best-performing model, a nomogram was created to provide individualized risk predictions for those undergoing M3M extractions, thereby facilitating decision-making and improving postoperative management.

## Results

*Assessment of Baseline Characteristics Across the Study Cohort*

The final study cohort comprised 321 participants with a mean age of 23.4 $\pm$ 4.2 years (range: 18–35); no patients older than 35 years were included. The study cohort included 64.5% female participants, resulting in a male-to-female ratio of about 1:1.82 (Table 1). Severe postoperative symptoms were observed in 155 patients, representing 48.3% of the study population. Among these patients, the average age was 22.15 $\pm$ 10.49 years, with an age range of 18 to 35 years (Table 1). The mean duration of the surgical procedure was 16.04 minutes, with a median of 14.83 minutes. The study cohort predominantly consisted of healthy young adults (median age: 23 years), with a balanced distribution of sex and no significant inter-group demographic differences. However, several clinical and surgical variables (e.g., smoking status, impaction classification, flap design) differed significantly between the two groups (Ta-

**Table 1. Baseline characteristics of all the patients undergoing third molar surgery.**

| Variable | PoSSe = 0 | PoSSe = 1 | $\chi^2/Z$ | *p*-value |
|---|---|---|---|---|
| Sample (n) | 166 | 155 | NA | NA |
| Gender (male, %) | 67 (40.0%) | 47 (30.0%) | 3.52 | 0.060 |
| Smoke (yes, %) | 35 (21%) | 53 (34%) | 6.92 | 0.008 |
| Age (≤25 years, %) | 74 (45%) | 57 (37%) | 2.02 | 0.155 |
| BMI (%), kg/m$^2$ | | | 1.87 | 0.394 |
| <18.5 | 43 (26%) | 50 (32%) | | |
| 18.5–24.9 | 60 (36%) | 55 (35%) | | |
| ≥25 | 63 (38%) | 50 (32%) | | |
| Pell-Gregory vertical, No. (%) | | | 7.07 | 0.029 |
| I | 123 (74%) | 118 (76%) | | |
| II | 35 (21%) | 20 (13%) | | |
| III | 8 (4.8%) | 17 (11%) | | |
| Pell-Gregory horizontal, No. (%) | | | 10.04 | 0.007 |
| A | 53 (32%) | 29 (19%) | | |
| B | 91 (55%) | 90 (58%) | | |
| C | 22 (13%) | 36 (23%) | | |
| Winter's, No. (%) | | | 2.99 | 0.224 |
| Mesioangular | 48 (29%) | 32 (21%) | | |
| Horizontal | 75 (45%) | 80 (52%) | | |
| Vertical | 43 (26%) | 43 (28%) | | |
| Root curvature (n, %) | | | 0.22 | 0.634 |
| One root | 128 (77%) | 116 (75%) | | |
| More than one root | 38 (23%) | 39 (25%) | | |
| Root morphology (n, %) | | | 0.70 | 0.706 |
| Conical | 113 (68%) | 104 (67%) | | |
| Spherical | 10 (6.0%) | 13 (8.4%) | | |
| Bifurcation | 43 (26%) | 38 (25%) | | |
| Preoperative symptoms (n, %) | | | 6.26 | 0.012 |
| No | 61 (37%) | 37 (24%) | | |
| Yes | 105 (63%) | 118 (76%) | | |
| Flap design (n, %) | | | 7.03 | 0.030 |
| No flap | 21 (13%) | 14 (9.0%) | | |
| Relaxing incision | 46 (28%) | 27 (17%) | | |
| Triangular flap | 99 (60%) | 114 (74%) | | |
| Procedure type (n, %) | | | 8.88 | 0.012 |
| Elevator/forceps alone | 52 (31%) | 28 (18%) | | |
| Bone removal/tooth sectioning | 57 (34%) | 54 (35%) | | |
| Bone removal + tooth sectioning | 57 (34%) | 73 (47%) | | |
| Surgical experience (n, %) | | | 0.68 | 0.713 |
| >10 years | 73 (44%) | 73 (47%) | | |
| 5–10 years | 63 (38%) | 52 (34%) | | |
| <5 years | 30 (18%) | 30 (19%) | | |
| Number of roots, (n, %) | | | 2.12 | 0.347 |
| Single root | 42 (25%) | 45 (29%) | | |
| Double roots | 117 (70%) | 99 (64%) | | |
| ≥three roots | 7 (4.2%) | 11 (7.1%) | | |
| Mouth opening (cm), median (IQR) | 4.29 (0.49) | 4.14 (0.50) | 2.63 | 0.009 |
| Operation time (minutes), median (IQR) | 13.34 (13.09) | 16.90 (14.04) | –3.57 | <0.001 |

Abbreviations: NA, not applicable; BMI, body mass index; PoSSe, post-operative symptom severity; IQR, interquartile range.

*Categorical variables: For 2 × 2 contingency tables with all expected frequencies >5, Pearson's $\chi^2$ test was applied without Yates' continuity correction. When any expected frequency was ≤5, Yates' correction or Fisher's exact test was used, as appropriate. Continuous variables: Student's *t*-test if Shapiro–Wilk $p \geq 0.05$; otherwise, Mann–Whitney U.

Note: Patients with PoSSe scores strictly greater than the total median were classified as the high-risk group (value = 1); those with scores equal to or below the median were classified as the low-risk group (value = 0). This single threshold was applied consistently to both the training and validation sets. *p*-value was derived from the bivariate association analyses between each of the study variables and PoSSe.

**Table 2. Baseline characteristics of the patients in the training data set after third molar surgery.**

| Variable | PoSSe = 0 | PoSSe = 1 | $\chi^2/Z$ | *p*-value |
|---|---|---|---|---|
| Sample (n) | 112 | 112 | NA | NA |
| Gender (male, %) | 41 (37%) | 33 (29%) | 1.29 | 0.250 |
| Smoke (yes, %) | 22 (20%) | 38 (34%) | 5.82 | 0.015 |
| Age (≤25 years, %) | 56 (50%) | 44 (39%) | 2.60 | 0.107 |
| BMI (%), kg/m$^2$ | | | 0.31 | 0.856 |
| <18.5 | 33 (29%) | 35 (31%) | | |
| 18.5–24.9 | 42 (38%) | 38 (34%) | | |
| ≥25 | 37 (33%) | 39 (35%) | | |
| Pell-Gregory vertical, No. (%) | | | 4.59 | 0.101 |
| I | 84 (75%) | 89 (79%) | | |
| II | 21 (19%) | 11 (9.8%) | | |
| III | 7 (6.3%) | 12 (11%) | | |
| Pell-Gregory horizontal, No. (%) | | | 5.68 | 0.058 |
| A | 34 (30%) | 22 (20%) | | |
| B | 64 (57%) | 65 (58%) | | |
| C | 14 (13%) | 25 (22%) | | |
| Winter's, No. (%) | | | 1.37 | 0.505 |
| Mesioangular | 30 (27%) | 23 (21%) | | |
| Horizontal | 52 (46%) | 59 (53%) | | |
| Vertical | 30 (27%) | 30 (27%) | | |
| Root curvature (n, %) | | | 0.37 | 0.541 |
| One root | 85 (76%) | 81 (72%) | | |
| More than one root | 27 (24%) | 31 (28%) | | |
| Root morphology (n, %) | | | 3.50 | 0.174 |
| Conical | 80 (71%) | 75 (67%) | | |
| Spherical | 4 (3.6%) | 11 (9.8%) | | |
| Bifurcation | 28 (25%) | 26 (23%) | | |
| Preoperative symptoms (n, %) | | | 5.10 | 0.024 |
| No | 46 (41%) | 30 (27%) | | |
| Yes | 66 (59%) | 82 (73%) | | |
| Flap design (n, %) | | | 8.24 | 0.016 |
| No flap | 12 (11%) | 10 (8.9%) | | |
| Relaxing incision | 34 (30%) | 17 (15%) | | |
| Triangular flap | 66 (59%) | 85 (76%) | | |
| Procedure type (n, %) | | | 6.63 | 0.036 |
| Elevator/forceps alone | 36 (32%) | 20 (18%) | | |
| Bone removal/tooth sectioning | 39 (35%) | 42 (38%) | | |
| Bone removal + tooth sectioning | 37 (33%) | 50 (45%) | | |
| Surgical experience (n, %) | | | 0.28 | 0.870 |
| >10 years | 51 (46%) | 50 (45%) | | |
| 5–10 years | 42 (38%) | 40 (36%) | | |
| <5 years | 19 (17%) | 22 (20%) | | |
| Number of roots, (n, %) | | | 0.76 | 0.683 |
| Single root | 33 (29%) | 31 (28%) | | |
| Double roots | 74 (66%) | 73 (65%) | | |
| ≥three roots | 5 (4.5%) | 8 (7.1%) | | |
| Mouth Opening (cm), median (IQR) | 4.28 (0.49) | 4.15 (0.48) | 1.94 | 0.053 |
| Operation time (minutes), median (IQR) | 12.58 (11.56) | 16.54 (13.42) | –3.68 | <0.001 |

*Categorical variables: For 2 × 2 contingency tables with all expected frequencies >5, Pearson's $\chi^2$ test was applied without Yates' continuity correction. When any expected frequency was ≤5, Yates' correction or Fisher's exact test was used, as appropriate. Continuous variables: Student's *t*-test if Shapiro–Wilk $p \geq 0.05$; otherwise, Mann–Whitney U.

Note: Patients with PoSSe scores strictly greater than the total median were classified as the high-risk group (value = 1); those with scores equal to or below the median were classified as the low-risk group (value = 0). This single threshold was applied consistently to both the training and validation sets.

*p*-value was derived from the bivariate association analyses between each of the study variables and PoSSe.

ble 1) and were therefore included in the multivariable predictive models described below.

*Machine Learning Model Performance*

A total of 224 patients were included in the training set to establish the nomogram-based predictive model, while

**Table 3. Baseline characteristics of the patients in the test data set undergoing third molar surgery.**

| Variable | PoSSe = 0 | PoSSe = 1 | $\chi^2/Z$ | *p*-value |
|---|---|---|---|---|
| Sample (n) | 54 | 43 | NA | NA |
| Gender (male, %) | 26 (48%) | 14 (33%) | 2.40 | 0.121 |
| Smoke (yes, %) | 13 (24%) | 15 (35%) | 1.36 | 0.243 |
| Age (≤25 years, %) | 18 (33%) | 13 (30%) | 0.10 | 0.744 |
| BMI (%), kg/m$^2$ | | | 5.94 | 0.051 |
| <18.5 | 10 (19%) | 15 (35%) | | |
| 18.5–24.9 | 18 (33%) | 17 (40%) | | |
| ≥25 | 26 (48%) | 11 (26%) | | |
| Pell-Gregory vertical, No. (%) | | | 4.03 | 0.133 |
| I | 19 (35%) | 7 (16%) | | |
| II | 27 (50%) | 25 (58%) | | |
| III | 8 (15%) | 11 (26%) | | |
| Pell-Gregory horizontal, No. (%) | | | 4.90 | 0.086 |
| A | 39 (72%) | 29 (67%) | | |
| B | 14 (26%) | 9 (21%) | | |
| C | 1 (1.9%) | 5 (12%) | | |
| Winter's, No. (%) | | | 1.87 | 0.393 |
| Mesioangular | 18 (33%) | 9 (21%) | | |
| Horizontal | 23 (43%) | 21 (49%) | | |
| Vertical | 13 (24%) | 13 (30%) | | |
| Root curvature (n, %) | | | 0.04 | 0.827 |
| One root | 43 (80%) | 35 (81%) | | |
| More than one root | 11 (20%) | 8 (19%) | | |
| Root morphology (n, %) | | | 1.36 | 0.506 |
| Conical | 33 (61%) | 29 (67%) | | |
| Spherical | 6 (11%) | 2 (4.7%) | | |
| Bifurcation | 15 (28%) | 12 (28%) | | |
| Preoperative symptoms (n, %) | | | 1.80 | 0.179 |
| No | 15 (28%) | 7 (16%) | | |
| Yes | 39 (72%) | 36 (84%) | | |
| Flap design (n, %) | | | 1.13 | 0.568 |
| No flap | 9 (17%) | 4 (9.3%) | | |
| Relaxing incision | 12 (22%) | 10 (23%) | | |
| Triangular flap | 33 (61%) | 29 (67%) | | |
| Procedure type (n, %) | | | 2.87 | 0.239 |
| Elevator/forceps alone | 16 (30%) | 8 (19%) | | |
| Bone removal/tooth sectioning | 18 (33%) | 12 (28%) | | |
| Bone removal + tooth sectioning | 20 (37%) | 23 (53%) | | |
| Surgical experience (n, %) | | | 1.73 | 0.422 |
| >10 years | 22 (41%) | 23 (53%) | | |
| 5–10 years | 21 (39%) | 12 (28%) | | |
| <5 years | 11 (20%) | 8 (19%) | | |
| Number of roots, (n, %) | | | 4.28 | 0.117 |
| Single root | 9 (17%) | 14 (33%) | | |
| Double roots | 43 (80%) | 26 (60%) | | |
| ≥three roots | 2 (3.7%) | 3 (7.0%) | | |
| Mouth Opening (cm), median (IQR) | 4.31 (0.47) | 4.12 (0.56) | 1.77 | 0.080 |
| Operation time (minutes), median (IQR) | 15.93 (14.09) | 17.68 (15.08) | –1.12 | 0.265 |

Data is presented as median (IQR) or numbers, with percentages in parentheses.

*Categorical variables: For 2 × 2 contingency tables with all expected frequencies >5, Pearson's $\chi^2$ test was applied without Yates' continuity correction. When any expected frequency was ≤5, Yates' correction or Fisher's exact test was used, as appropriate. Continuous variables: Student's *t*-test if Shapiro–Wilk $p \geq 0.05$; otherwise, Mann–Whitney U.

Note: Patients with PoSSe scores strictly greater than the total median were classified as the high-risk group (value = 1); those with scores equal to or below the median were classified as the low-risk group (value = 0). This single threshold was applied consistently to both the training and validation sets.

*p*-value was derived from the bivariate association analyses between each of the study variables and PoSSe.

97 patients were included in the validation set to evaluate model performance (Table 2). Among the training dataset, 44.6% of patients were under 25 years, 33.0% were male, 73.2% were nonsmokers, and 14.3% were classified as Pell-

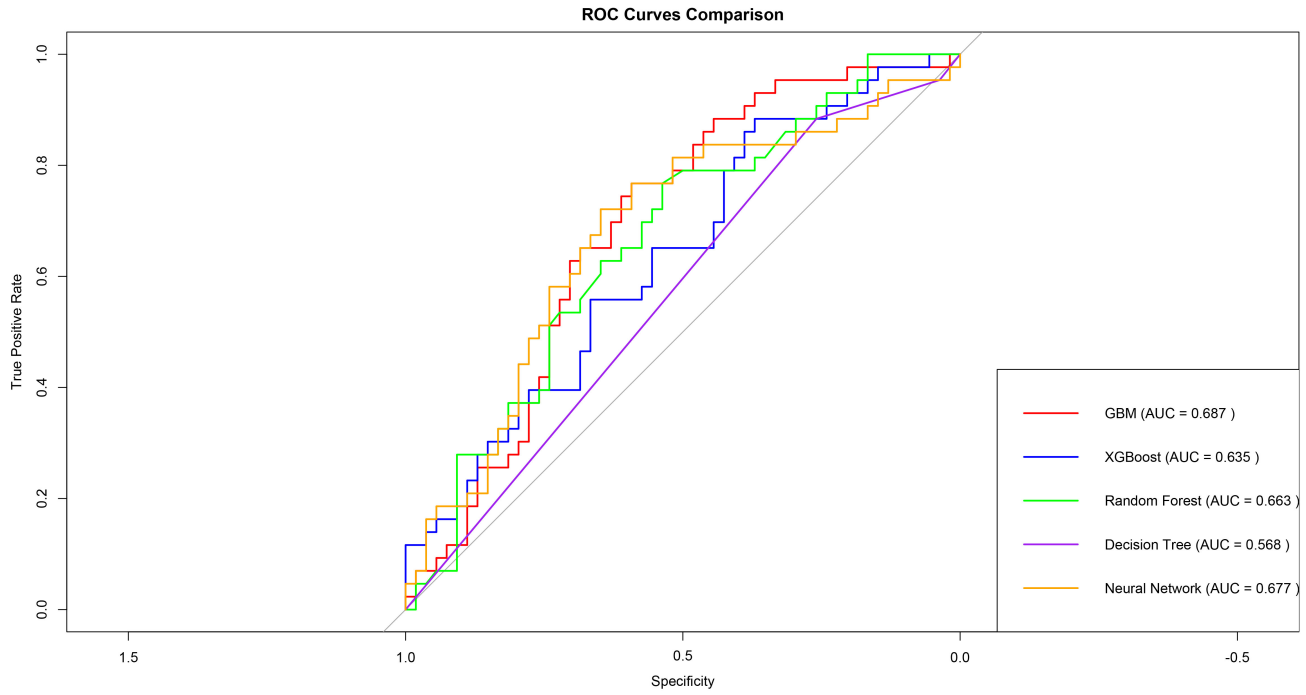


**Fig. 1. Receiver operating characteristic (ROC) curve for five machine learning algorithms.** GBM, Gradient Boosting Machine; XGBoost, Extreme gradient boosting; AUC, area under the curve.

**Table 4. Comparison of predictive performance across the five types of machine learning algorithms in the validation sets.**

| Methods | AUC (95% CI) | Sensitivity | Specificity | Accuracy | F1 score | Brier |
|---|---|---|---|---|---|---|
| GBM | 0.687 (0.624–0.744) | 0.698 | 0.630 | 0.600 | 0.645 | 0.225 |
| XGBoost | 0.635 (0.594–0.696) | 0.581 | 0.556 | 0.510 | 0.543 | 0.248 |
| Random Forest | 0.663 (0.617–0.716) | 0.605 | 0.648 | 0.578 | 0.591 | 0.232 |
| Decision Tree | 0.568 (0.524–0.674) | 0.884 | 0.259 | 0.487 | 0.628 | 0.252 |
| Neural Network | 0.677 (0.614–0.726) | 0.837 | 0.463 | 0.554 | 0.667 | 0.243 |

**Table 5. Predictive performance of all five machine-learning algorithms after using the six SHAP-derived predictors in internal cross-validation (validation dataset).**

| Methods | AUC (95% CI) | Sensitivity | Specificity | Accuracy | F1 score | Brier |
|---|---|---|---|---|---|---|
| GBM | 0.666 (0.602–0.728) | 0.689 | 0.618 | 0.650 | 0.634 | 0.228 |
| XGBoost | 0.622 (0.580–0.682) | 0.560 | 0.548 | 0.553 | 0.527 | 0.251 |
| Random Forest | 0.645 (0.599–0.700) | 0.595 | 0.630 | 0.614 | 0.578 | 0.236 |
| Decision Tree | 0.550 (0.508–0.654) | 0.870 | 0.245 | 0.522 | 0.617 | 0.255 |
| Neural Network | 0.660 (0.600–0.714) | 0.820 | 0.452 | 0.615 | 0.653 | 0.246 |

Notes: Cut-offs were chosen by maximizing Youden's index; 1000 sample bootstraps were used for the CIs.

Gregory ramus class II (Table 2). The incidence of serious postoperative symptoms in the training dataset was 50.0% and 44.3% in the testing dataset. The baseline characteristics of the patients in the testing dataset undergoing third molar extraction are detailed in Table 3.

*Performance of Machine Learning Algorithms*

The predictive performance of five machine learning algorithms was evaluated for determining PSSR. The GBM algorithm demonstrated superior performance, achieving the highest AUROC curve value of 0.687 (Fig. 1).

The GBM outperformed the other four algorithms when all variables were incorporated (Fig. 1, Table 4). Even after restricting the models to the six predictors selected through SHapley Additive exPlanations (SHAP) (Fig. 2), GBM continued to exhibit the highest discriminative performance, with an area under the curve (AUC) of 0.666 (Table 5).

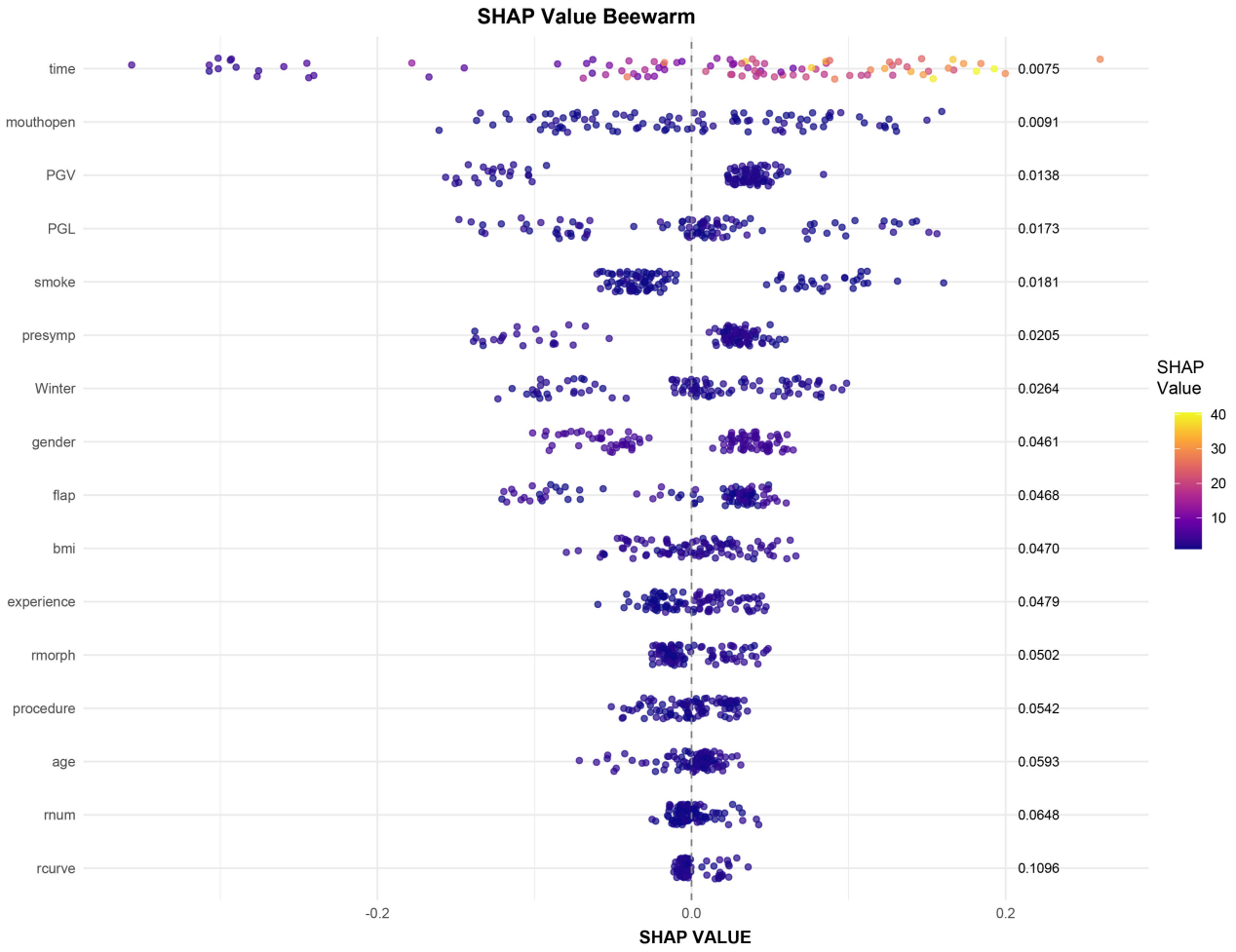


**Fig. 2. SHapley Additive exPlanations (SHAP) values in a beeswarm plot for GBM.** SHAP values in a beeswarm plot summarize the contribution of each feature to the model's predictions. Each point on the plot represents the SHAP value of a single observation for a particular feature. The x-axis represents the SHAP value, which quantifies the impact of the feature on the model output. A positive SHAP value indicates that the feature value for that observation increased the prediction, while a negative SHAP value indicates that it decreased the prediction. The features are ranked in descending order of their overall importance, determined by the average absolute SHAP value across all observations (represented by the values on the y-axis to the right of each feature's beeswarm). PGV, Pell-Gregory vertical; PGL, Pell-Gregory level.

*SHAP Value of Variables in GBM Machine Learning Algorithms*

Fig. 2 displays SHAP values in a beeswarm plot, illustrating the impact of each feature on the output of the GBM model predicting PSSR. Among the assessed variables, 'operation time' exhibited the most significant influence on the model's prediction, with longer duration consistently associated with higher predicted risk scores. Across all ensemble models, the most influential predictors were impaction depth, operative time, and mouth opening. In contrast, features such as 'number of root', 'age', and 'root of curve' showed comparatively lower impact, as indicated by their smaller average absolute SHAP values and distributions more tightly clustered near zero (Fig. 2). Based on these findings, six key predictors were selected for constructing the nomogram (Fig. 3).

All models showed acceptable calibration, with the GBM model exhibiting the most favorable outcomes. GBM yielded calibration slopes close to ideal (0.98) and showed the lowest Brier score (0.225), suggesting strong agreement between predicted and observed values. Comparative performance metrics for the five ML algorithm models on the validation set are detailed in Table 4 and Fig. 1. Among them, the GBM model demonstrated the highest performance in predicting PSSR, yielding a sensitivity of 0.698, specificity of 0.630, F1 score of 0.645, Brier score of 0.225, and overall accuracy of 0.600. Based on these findings, the GBM model was selected as the final predictive model (Table 4). Model calibration was further quantified using the Brier score, where a lower value indicates improved prediction reliability (Fig. 4). Additionally, the predictive performance of all five machine-learning algorithms, using the six SHAP-derived predictors in internal cross-validation (test dataset), is shown in Table 5.

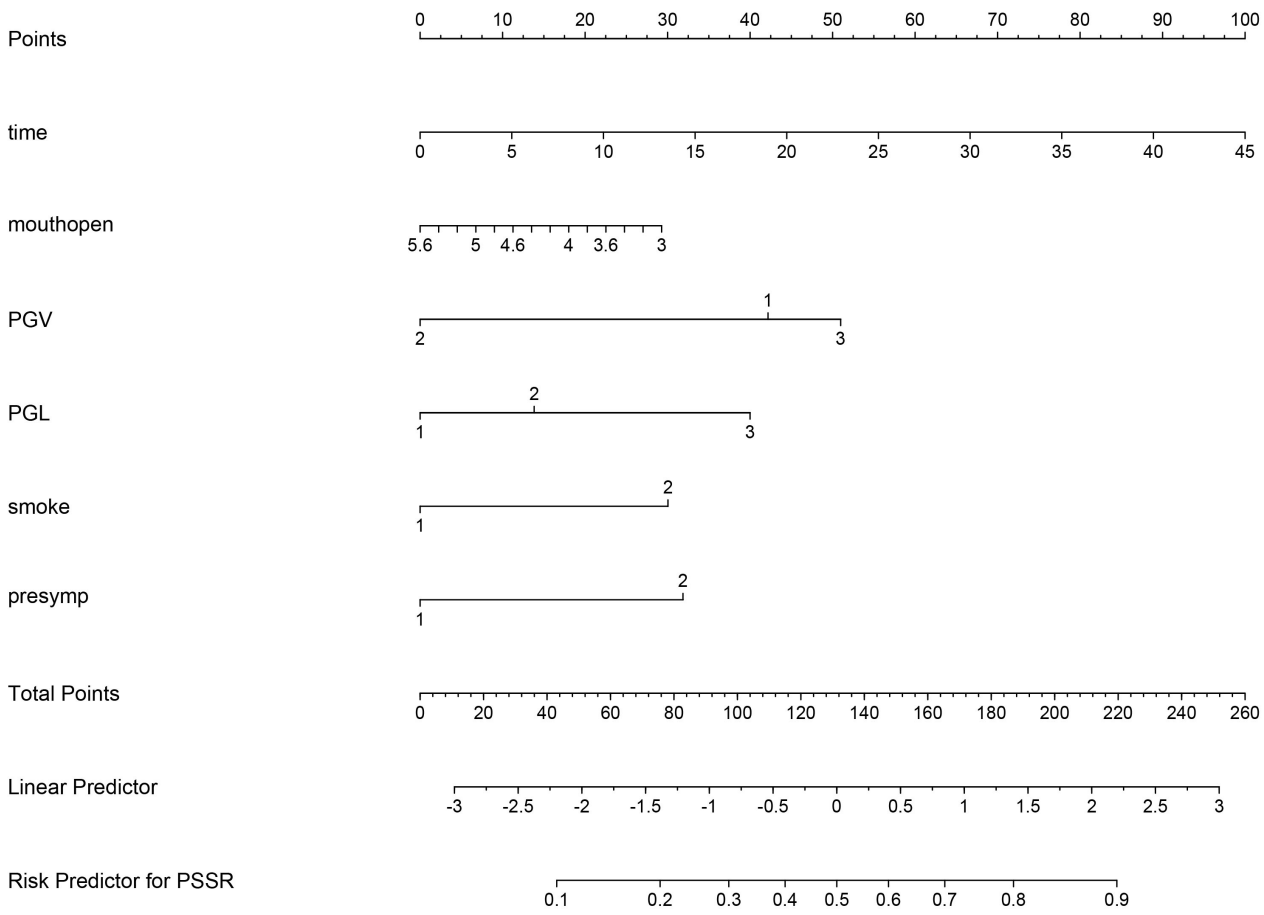


**Fig. 3. Nomogram for predicting postoperative symptom severity following mandibular third molar extractions using GBM.** PSSR, postoperative symptom severity risk.

*Development of a Predictive Nomogram*

Utilizing the GBM model, a predictive nomogram was constructed to estimate the individual probability of experiencing severe postoperative symptoms following M3M surgery. The nomogram integrates key clinical variables, including operative time, mouth opening, Pell-Gregory classification, and preoperative symptomatology, offering a personalized risk assessment tool (Fig. 3). The nomogram was based on a compact GBM model containing six predictors that together accounted for over 80% of the total SHAP importance. The GBM model, which underlies the nomogram, achieved an AUC value of 0.666, indicating its discrimination capability. The nomogram demonstrated excellent calibration (Fig. 4) and yielded favorable net benefit across clinically relevant threshold probabilities, as illustrated by decision-curve analysis (Fig. 5), underscoring its practical utility in guiding clinical decision-making.

How to use the nomogram guide:

(1) Identify the patient's category or numerical value on each predictor axis.

(2) Draw a vertical line upward to intersect the "Points" scale.

(3) Sum the individual points to obtain the "Total Points".

(4) Draw a vertical line downward from the "Total Points" to assess the predicted probability of a high PoSSe score (>60).

## Discussion

In this study, a machine learning-based analysis of patient data was conducted to develop a personalized and accurate system for predicting postoperative symptoms in patients undergoing M3Ms. A comparison of ML algorithms identified that the GBM model demonstrates the highest predictive performance. To support clinical application of the model, a nomogram was subsequently established based on GBM to estimate the individual probability of PSSR.

Mobilio *et al.* [21] found that surgical duration significantly influences acute postoperative symptoms after lower third molar extraction. Similarly, Farhadi *et al.* [22] demonstrated that the difficulty index is useful in predicting the probability of postoperative infection after impacted mandibular third molar surgery. Furthermore, Luo *et al.* [23] highlighted that factors such as preoperative panoramic radiographs, computed tomography (CT) imaging, patient age, the experience of the surgeon, and postoperative bleeding can predict postoperative complications in M3M extractions. Moreover, Kocyigit *et al.* [24] developed an artificial intelligence-based system to estimate postoperative

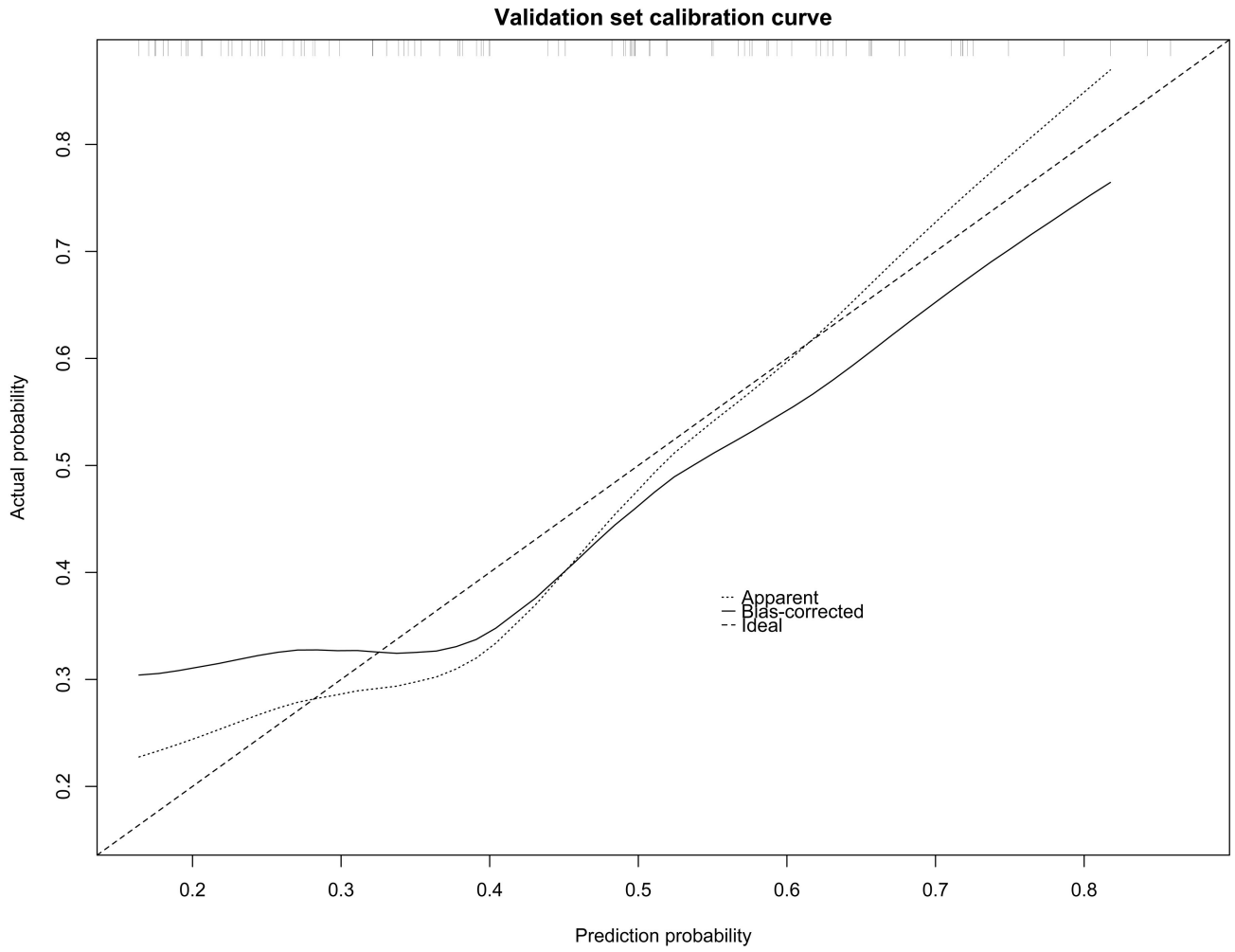


**Fig. 4. Calibration plot for predicting postoperative symptom severity following mandibular third molar extractions using GBM.** A bootstrap-bias–corrected calibration plot is now presented. Intercept = –0.01 (ideal = 0); slope = 0.96 (ideal = 1). Hosmer–Lemeshow $\chi^2 = 6.21$, $p = 0.41$; Brier score = 0.210.

discomfort after impacted third molar surgery, indicating the potential of machine learning algorithms in predicting postoperative outcomes. Overall, these findings provide a comprehensive understanding of the multifaceted nature of postoperative symptoms and support the integration of machine learning algorithms for predicting the postoperative severity of symptoms.

The SHAP value analysis provided valuable insights into the factors affecting the model's predictions. Notably, operation time emerged as the most prominent variable, with longer duration consistently associated with an increased predicted risk of PSSR, aligning with findings from a study by Qiao *et al*. [18]. Other variables, albeit less consistent, including 'Mouth Opening', 'Pell-Gregory vertical classification', and 'Pell-Gregory horizontal classification' also contributed substantially to the model's predictions. However, the variability in their impact, as reflected by the wider distribution of their SHAP values, suggests potential interactions with other features that necessitate further investigation.

In contrast, features such as age and number of roots showed a relatively smaller overall impact on the model predictions. While these features may still play a crucial role in model accuracy, their individual contributions were less pronounced in this dataset. This does not alleviate their clinical significance; rather, it suggests that their influence on PSSR may be primarily mediated through interactions with other variables or may become more apparent within specific subgroups of patients. Therefore, further studies are warranted to elucidate these interactions and optimize their role in predictive models.

Machine learning algorithms have been increasingly applied in dentistry to predict various postoperative outcomes [25,26]. In our study, five different machine learning methods, such as GBM, XGBoost, RF, DT, and NNET, were evaluated to assess their predictive performance. GBM demonstrated the highest predictive accuracy, effectively handling complex data relationships. XGBoost offers improved computational efficiency and built-in regularization, which helps mitigate overfitting but requires careful hyperparameter tuning and may be sensitive to noisy

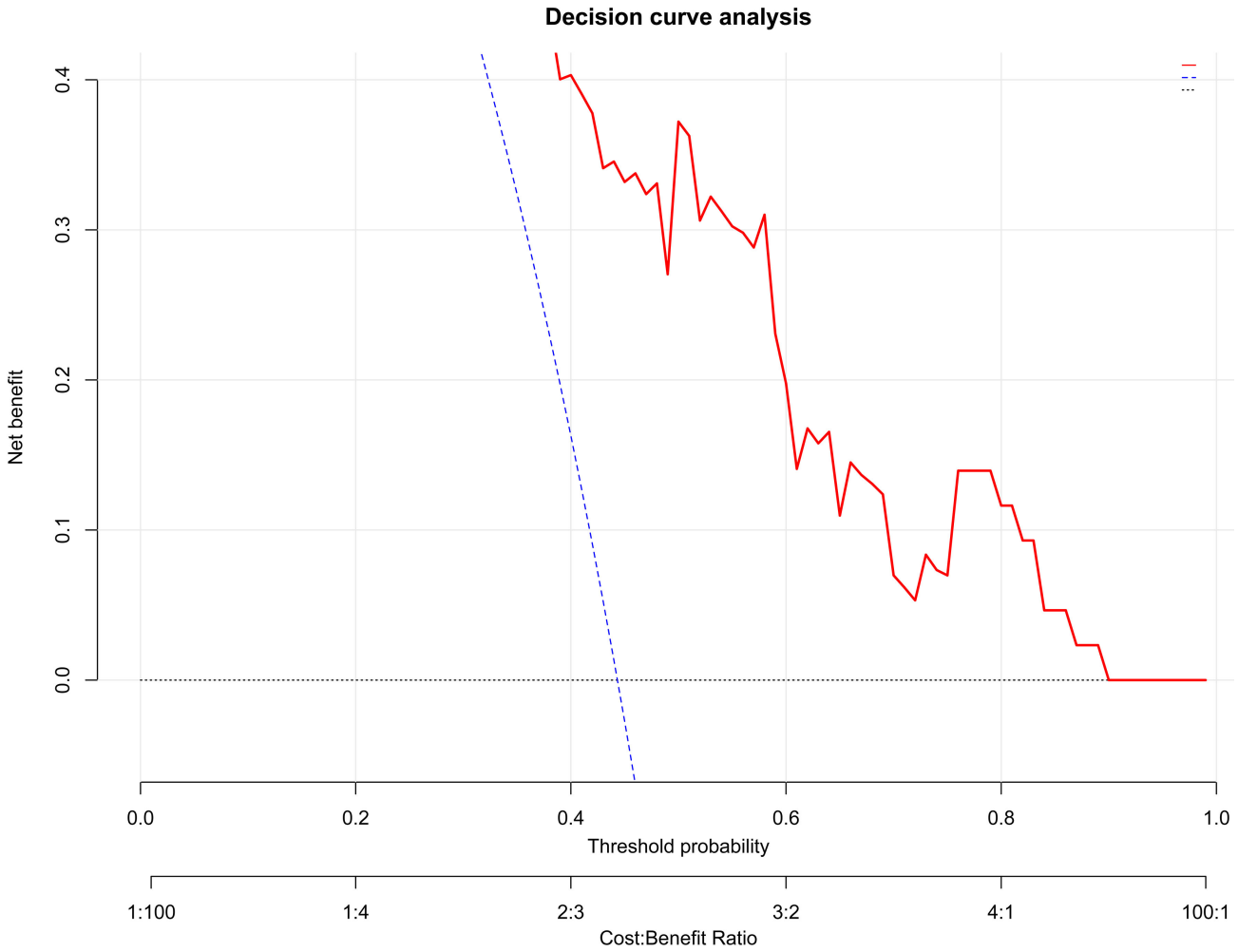


**Fig. 5. Decision curve analysis (DCA) plot for predicting postoperative symptom severity following mandibular third molar extractions using GBM**. The red curve represents the predictive model, the blue dashed line indicates the scenario of treatment for all patients, and the black dashed line reflects the scenario of no treatment. The model demonstrates good performance when the net benefit exceeds 0.4. The nomogram provides a higher net benefit than either the "treat-all" or "treat-none" strategies throughout threshold probabilities ranging from 0.40–0.90.

data [27]. RF is robust to overfitting and can handle large datasets; however, it may perform poorly with noisy data and can be slow to train on very large datasets [28]. DT models are simple to understand and can handle both numerical and categorical data, but they are prone to overfitting, especially with deep trees, and can be sensitive to small variations in the data. NNETs can capture complex, non-linear relationships and learn effectively from large datasets; however, they usually require substantial training data and are particularly prone to overfitting when applied to smaller datasets. Notably, the predictive model established in this study was based on a cohort of healthy young adults (ASA I–II, 18–35 years); therefore, its applicability to paediatric, middle-aged, or medically complex patients remains uncertain. Additional external validation in broader populations is required before the model can be adopted for widespread clinical use.

Despite several promising outcomes, our study has some limitations. First, all data were sourced exclusively from a single institution, Ruijin Hospital, potentially limiting the generalizability of our findings to other populations or healthcare settings, where surgical practices and patient demographics may differ. Furthermore, the exclusion of variables such as IAN-root distance and crown integrity, due to the lack of routine CBCT imaging, may have reduced the model's predictive accuracy. Subsequent studies incorporating routine CBCT imaging data are planned to evaluate the incremental value of these variables. Second, applying multiple complex algorithms to a relatively small dataset introduces a risk of overfitting, potentially resulting in models that perform well on training data but poorly on external or unseen datasets. Third, SHAP analysis provides valuable insights into overall variable importance; it does not fully capture individual-level predictions or complex feature interactions. Finally, the model should not be applied to patients with significant systemic disease (ASA $\geq$III) or those requiring cardiac monitoring, as such individuals were excluded from training datasets and may exhibit dif-

ferent postoperative risk profiles. Thus, further validation in a larger and more diverse population is warranted before clinical implementation.

## Conclusions

Comparison across distinct algorithmic families demonstrated that gradient-boosting ensembles offer a meaningful but still moderate improvement in predictive performance over the four-alternative machine learning models (XGBoost, Random Forest, Decision Tree and Neural Network). These observations reinforce existing evidence that tree-based boosting approaches are particularly well-suited for clinical prediction tasks. The resulting compact GBM-based nomogram incorporating these variables provides patient-specific risk estimates at the point of care, assisting clinicians in patient counselling, analgesic planning, and scheduling postoperative follow-up.

## Availability of Data and Materials

The data analyzed are available from the corresponding author upon reasonable request.

## Author Contributions

QQH conducted the research and wrote the paper. HG analyzed the data. JYX investigated and analyzed data. YMG conceived and designed the study and conducted experiments. All authors have been involved in revising it critically for important intellectual content. All authors gave final approval of the version to be published. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

## Ethics Approval and Consent to Participate

This study was approved by the Institutional Ethics Committee of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine (Approval No. ruijin-eth-2023-268). All procedures adhered to the Declaration of Helsinki, and written informed consent was obtained from every participant before their enrolment in the study.

## Acknowledgment

The authors thank the clinical nursing team for their assistance with patient recruitment and data collection, and a colleague for helpful statistical discussions.

## Funding

This research received no external funding.

## Conflict of Interest

The authors declare no conflict of interest.

## References

[1] Chen YW, Chi LY, Lee OKS. Revisit incidence of complications after impacted mandibular third molar extraction: A nationwide population-based cohort study. PloS One. 2021; 16: e0246625. https://doi.org/10.1371/journal.pone.0246625.

[2] Bouloux GF, Steed MB, Perciaccante VJ. Complications of third molar surgery. Oral and Maxillofacial Surgery Clinics. 2007; 19: 117–128. https://doi.org/10.1016/j.coms.2006.11.013.

[3] Gojayeva G, Tekin G, Saruhan Kose N, Dereci O, Kosar YC, Caliskan G. Evaluation of complications and quality of life of patient after surgical extraction of mandibular impacted third molar teeth. BMC Oral Health. 2024; 24: 131. https://doi.org/10.1186/s12903-024-03877-8.

[4] Colorado-Bonnin M, Valmaseda-Castellón E, Berini-Aytés L, Gay-Escoda C. Quality of life following lower third molar removal. International Journal of Oral and Maxillofacial Surgery. 2006; 35: 343–347. https://doi.org/10.1016/j.ijom.2005.08.008.

[5] Ladinez MJV, Figueroa CBD, Guartatanga PGP, Veloz BAA, Arias CAL, Dominguez YF. Efficacy of machine learning algorithms versus conventional assessment techniques in predicting postoperative complications in general surgery: a comprehensive literature review. Ibero-American Journal of Health Science Research. 2024; 4: 89–96. https://doi.org/10.56183/iberojhr.v4is.627.

[6] Simonini A, Murugan J, Vittori A, Pallotto R, Bignami EG, Calevo MG, *et al*. Data-driven Machine Learning Models for Risk Stratification and Prediction of Emergence Delirium in Pediatric Patients Underwent Tonsillectomy/Adenotonsillectomy. Annali Italiani di Chirurgia. 2024; 95: 944–955. https://doi.org/10.62713/aic.3485.

[7] Trachoo V, Taetragool U, Pianchoopat P, Sukitporn-Udom C, Morakrant N, Warin K. Deep Learning for Predicting the Difficulty Level of Removing the Impacted Mandibular Third Molar. International Dental Journal. 2025; 75: 144–150. https://doi.org/10.1016/j.identj.2024.06.021.

[8] Nwaimo CS, Adegbola AE, Adegbola MD. Transforming healthcare with data analytics: Predictive models for patient outcomes. GSC Biological and Pharmaceutical Sciences. 2024; 27: 025–035.

[9] Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. The Lancet. Oncology. 2015; 16: e173–80. https://doi.org/10.1016/S1470-2045(14)71116-7.

[10] Kassab GS, An G, Sander EA, Miga MI, Guccione JM, Ji S, *et al*. Augmenting Surgery via Multi-scale Modeling and Translational Systems Biology in the Era of Precision Medicine: A Multidisciplinary Perspective. Annals of Biomedical Engineering. 2016; 44: 2611–2625. https://doi.org/10.1007/s10439-016-1596-4.

[11] Ali D. Risk factors of complications subsequent third molar extractions: A prospective cohort study: Risk factors of complications subsequent third molar extractions. Brazilian Dental Science. 2021; 24. https://doi.org/10.14295/bds.2021.v24i4.2759.

[12] Stacchi C, Daugela P, Berton F, Lombardi T, Andriulionis T, Perinetti G, *et al*. A classification for assessing surgical difficulty in the extraction of mandibular impacted third molars: Description and clinical validation. Quintessence International. 2018; 49: 745–753. https://doi.org/10.3290/j.qi.a40778.

[13] Qiao F, He B, Zhang J, Sun J, Dong R, Zhang X. Establishment and validation of a predictive nomogram for extended operation time following mandibular third molar removal. Clinical Oral Investigations. 2021; 25: 1915–1923. https://doi.org/10.1007/s00784-020-03499-8.

[14] Susarla SM, Dodson TB. Predicting third molar surgery operative time: a validated model. Journal of Oral and Maxillofacial Surgery: Official Journal of the American Association of Oral and Maxillofacial Surgeons. 2013; 71: 5–13. https://doi.org/10.1016/j.joms.2012.08.004.

[15] Qiao F, Li L, Zhang J, Dong R, Sun J. Operation time is independent associated with serious postoperative symptom in patients with mandibular third molar removal. Annals of Palliative Medicine. 2021; 10: 4080–4089. https://doi.org/10.21037/apm-20-2340.

[16] Ventä I. Current care guidelines for third molar teeth. Journal of Oral and Maxillofacial Surgery: Official Journal of the American Association of Oral and Maxillofacial Surgeons. 2015; 73: 804–805. https://doi.org/10.1016/j.joms.2014.12.039.

[17] Ruta DA, Bissias E, Ogston S, Ogden GR. Assessing health outcomes after extraction of third molars: the postoperative symptom severity (PoSSe) scale. The British Journal of Oral & Maxillofacial Surgery. 2000; 38: 480–487. https://doi.org/10.1054/bjom.2000.0339.

[18] Qiao F, Huang X, Li B, Dong R, Huang X, Sun J. A Validated Model to Predict Postoperative Symptom Severity After Mandibular Third Molar Removal. Journal of Oral and Maxillofacial Surgery: Official Journal of the American Association of Oral and Maxillofacial Surgeons. 2020; 78: 893–901. https://doi.org/10.1016/j.joms.2020.02.007.

[19] Maferano EF, Filho EL, Silva PG, Granville-Garcia AF, Firmino RT, Perazzo MD, *et al*. Evaluating quality of life in third molar surgery: a scoping review of the postoperative symptom severity (PoSSe) scale. Medicina Oral, Patologia Oral Y Cirugia Bucal. 2025; 30: e232–e239. https://doi.org/10.4317/medoral.26839.

[20] Duarte-Rodrigues L, Miranda EFP, Souza TO, de Paiva HN, Falci SGM, Galvão EL. Third molar removal and its impact on quality of life: systematic review and meta-analysis. Quality of Life Research: an International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation. 2018; 27: 2477–2489. https://doi.org/10.1007/s11136-018-1889-1.

[21] Mobilio N, Vecchiatini R, Vasquez M, Calura G, Catapano S. Effect of flap design and duration of surgery on acute postoperative symptoms and signs after extraction of lower third molars: A randomized prospective study. Journal of Dental Research, Dental Clinics, Dental Prospects. 2017; 11: 156–160. https://doi.org/10.15171/joddd.2017.028.

[22] Farhadi F, Emamverdizadeh P, Hadilou M, Jalali P. Evaluation of Infection and Effective Factors in Impacted Mandibular Third Molar Surgeries: A Cross-Sectional Study. International Journal of Dentistry. 2022; 2022: 8934184. https://doi.org/10.1155/2022/8934184.

[23] Luo Q, Diao W, Luo L, Zhang Y. Comparisons of the Computed Tomographic Scan and Panoramic Radiography Before Mandibular Third Molar Extraction Surgery. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research. 2018; 24: 3340–3347. https://doi.org/10.12659/MSM.907913.

[24] Kocyigit S, Özgönenel O, Baş B, Ozden B, Hosgor H, Kaya OA. Development of an artificial intelligence system to estimate postoperative discomfort after impacted third molar surgery. Selcuk Dental Journal. 2020; 7: 148–154. https://doi.org/10.15311/selcukdentj.535365.

[25] Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, *et al*. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. Journal of Dentistry. 2021; 107: 103610. https://doi.org/10.1016/j.jdent.2021.103610.

[26] Arsiwala-Scheppach LT, Chaurasia A, Müller A, Krois J, Schwendicke F. Machine Learning in Dentistry: A Scoping Review. Journal of Clinical Medicine. 2023; 12: 937. https://doi.org/10.3390/jcm12030937.

[27] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794). 2016.

[28] Speiser JL, Miller ME, Tooze J, Ip E. A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. Expert Systems with Applications. 2019; 134: 93–101. https://doi.org/10.1016/j.eswa.2019.05.028.