Accepted: 8 May 2024

DOI: 10.1111/ocr.12811

Orthodontics & Craniofacial Research WILEY

RESEARCH ARTICLE

Evaluation of different machine learning algorithms for extraction decision in orthodontic treatment

Begüm Köktürk[1] | Hande Pamukçu[1] | Ömer Gözüaçık[2]

[1]Department of Orthodontics, Faculty of Dentistry, Başkent University, Ankara, Turkey

[2]Independent Researcher, Ankara, Turkey

Correspondence
Hande Pamukçu, Department of Orthodontics, Faculty of Dentistry, Başkent University, Yukarıbahçelievler Mah. 82. Sokak No: 26, 06490 Bahçelievler, Ankara, Turkey.
Email: handeorkun@yahoo.com

Funding information
Baskent Üniversitesi

Abstract

Introduction: The extraction decision significantly affects the treatment process and outcome. Therefore, it is crucial to make this decision with a more objective and standardized method. The objectives of this study were (1) to identify the best-performing model among seven machine learning (ML) models, which will standardize the extraction decision and serve as a guide for inexperienced clinicians, and (2) to determine the important variables for the extraction decision.

Methods: This study included 1000 patients who received orthodontic treatment with or without extraction (500 extraction and 500 non-extraction). The success criteria of the study were the decisions made by the four experienced orthodontists. Seven ML models were trained using 36 variables; including demographic information, cephalometric and model measurements. First, the extraction decision was performed, and then the extraction type was identified. Accuracy and area under the curve (AUC) of the receiver operating characteristics (ROC) curve were used to measure the success of ML models.

Results: The Stacking Classifier model, which consists of Gradient Boosted Trees, Support Vector Machine, and Random Forest models, showed the highest performance in extraction decision with 91.2% AUC. The most important features determining extraction decision were maxillary and mandibular arch length discrepancy, Wits Appraisal, and ANS-Me length. Likewise, the Stacking Classifier showed the highest performance with 76.3% accuracy in extraction type decisions. The most important variables for the extraction type decision were mandibular arch length discrepancy, Class I molar relationship, cephalometric overbite, Wits Appraisal, and L1-NB distance.

Conclusion: The Stacking Classifier model exhibited the best performance for the extraction decision. While ML models showed a high performance in extraction decision, they could not able to achieve the same level of performance in extraction type decision.

KEYWORDS
ensemble models, extraction decision, machine learning, treatment plan

1 | INTRODUCTION

One of the most important dilemmas encountered in orthodontic treatment is the decision of extraction. This decision is primarily based on factors such as the severity of the malocclusion, the patient's soft tissue, the treatment philosophy, and the orthodontic education of the clinician.[1–4] The popularity of orthodontic treatments with extraction has decreased over the years.[5] This decline can be attributed to various alternative treatments as maxillary expansion, interproximal reduction, usage of skeletal anchorage, and effective distalization systems.[1,6] Nevertheless, extraction still has a significant effect on solving severe crowding, eliminating dental or dentoalveolar protrusions, reducing the vertical dimension, increasing treatment stability, closing open-bites, reducing arch widths and orthodontic decompensation before orthognathic surgery.[7–9] Furthermore, extractions in orthodontic treatments can have a positive effect on soft tissue parameters.[4]

Nowadays, one of the most popular fields of technology is artificial intelligence (AI). Machine learning (ML), a subset of AI, involves using specific algorithms with a computer to model the properties of data for inference or other tasks. ML involves training a computer to predict new data by learning from existing data and generalizing patterns effectively. A key consideration in ML is the sample size of the data, as ML algorithms tend to perform better with larger datasets. Another important concern is overfitting, which occurs when a model trained on specific data is unable to generalize the problem and predict unseen data.[10]

AI has brought a new perspective to dentistry.[11,12] In orthodontics, AI has been utilized for tasks such as landmark detection in lateral cephalograms, determining the maturation of cervical vertebrae, assessing the need for orthodontic treatment, classifying skeletal structures, and determining the requirement of orthognathic surgery.[13–15] Several AI studies have focused on the extraction decision in orthodontic treatment. The majority of the studies in this field have used Artificial Neural Networks (ANN).[16–19] ANN is a ML component that mimics the behaviour of biological neurons and their interactions to make decisions based on provided data. Besides this, some recent studies have explored on different techniques such as Random Forest[20,21] and Automated Machine Learning (AutoML) systems.[22] Random Forest is a ML model that consists of multiple decision trees working together as an ensemble.[23] AutoML automatically selects and optimizes ML models for optimal performance, accessible to non-experts,[24] whereas there are important ML models for classification tasks that have not been used in past studies. Different validation strategies for overfitting were applied in the previous studies. Nevertheless, the results of these studies are difficult to compare and generalize due to low sample sizes, imbalanced label distributions, and potentially overfitted models. A better validation scheme and standardization are needed in extraction decisions that enable comparisons between different studies.

Different clinicians could make multiple treatment plans for a single orthodontic case.[25] However, it is important to enhance the objectivity and standardization of extraction decisions using evidence-based methods. The aims of this study were (1) to evaluate the performance of seven different machine learning models in extraction decision for clinical support and identify the model with the highest performance, and (2) to determine the key variables significantly influencing extraction decision. This will enable inexperienced practitioners to provide support during treatment planning. The null hypothesis of this study posits that there is no difference in extraction decision-making among the seven ML models.

2 | MATERIALS AND METHODS

This study was reviewed and approved by the Başkent University Institutional Review Board (Project number: D-KA 22/20) and supported by the Başkent University Research Fund.

2.1 | Data collection

Başkent University is a multicentered university hospital that provides medical/dental care to patients from different provinces of Türkiye. The orthodontic treatment records of patients treated at Başkent University, Faculty of Dentistry, Department of Orthodontics between the years 2012 and 2022 were retrospectively evaluated. All individuals included in the study belonged to the Caucasian ethnic group. Written informed consent was obtained from all patients at the beginning of their treatment, as per the standard procedure of the university. The inclusion criteria for this study were patients who had undergone one of four types of orthodontic treatment plans: non-extraction, extraction of maxillary and mandibular first premolars (44-44), extraction of maxillary first premolars and mandibular second premolars (44-55), and extraction of maxillary first premolars (44-00). The patients selected for this study were in the permanent dentition stage. Patients with caries in premolars, any missing records, a history of previous orthodontic treatment, craniofacial syndromes, maxillofacial deformities, orthognathic surgery, malformed or impacted or missing teeth (excluding third molars), and those who had undergone phase 1 orthodontic treatment (growth modification, treated with functional appliance or headgear) were excluded from the study. The data set consists of 500 patients with extraction and 500 patients with non-extraction who met the specified inclusion criteria. Among the patients with extractions, 268 were in the 44-00 group, 172 were in the 44-44 group and 60 were in the 44-55 group. The treatment mechanics of the non-extraction group were as follows: intermaxillary Class 2 elastics (34.8%), upper molar distalization (19.3%), IPR (29.8%), maxillary expansion (11.5%), TADs (8.5%) and intermaxillary Class 3 elastics (6.4%).

Four orthodontists with at least 10 years of academic and clinical orthodontic experience from Orthodontic Department of Başkent University had planned treatments and made the extraction decisions. The success criteria of this study was that the decisions of the ML models were consistent with the decisions of experienced orthodontists. A total of 36 variables, consisting of demographic information, cephalometric and model measurements were chosen to train the ML models. Twenty-nine measurements were chosen for cephalometric evaluation (Table 1). Gender, initial age, and lip posture data were obtained from the digital anamnesis forms in the university's patient archive. Angle's molar classification was determined based on the patient's initial models and intraoral photographs of the patients. Additionally, maxillary-mandibular arch length discrepancy and the curve of Spee were measured with a digital calliper.

2.2 | Reliability analysis

All data were analysed and evaluated by the same investigator (BK). Pre-treatment lateral cephalometric radiographs were measured with the Dolphin Imaging® program (Vers 11.95, Patterson Dental). The intra-examiner reliability was tested by remeasuring 20% of the radiographs 2 weeks after the first measurement. The intra-examiner reliability level was assessed using intra-class correlation coefficients (ICCs).

2.3 | Machine learning models

The extraction decision of ML models was analysed in a two-step approach. First, extraction decisions were identified by using a binary classifier. Extraction cases were labelled as 1, and non-extraction cases were labelled as 0. Then, the extraction types were identified by a multiclass classifier. The outputs of this classifier are the probabilities for each extraction type. The extraction type with maximum probability is chosen as the final decision (Figure 1).

ML models were trained with a total of 36 variables (Table 2). Seven different types of ML models were applied: Logistic Regression, Support Vector Machines, Random Forests, Gradient Boosted Trees, Multi-layer Perceptron (ANN), Voting Classifier, and Stacking Classifier. These models were implemented with Python 3.9.12 (Python Software Foundation) using the ML packages: Scikit-Learn module[26] and LightGBM.[27] The models had been made accessible as open-source on: https://github.com/ortho-research/extraction-decision.

In order to explain the models and identify how the variables were affecting the inference, SHAP (Shapley Additive explanation)[28] package was used.

2.4 | Experimental set-up

While analysing the performance of each algorithm, five-fold cross-validation was used. The data was split into five equal-sized folds (Figure 2). In each iteration, four of these folds were used as training data (800 patients) and one-fold was used as test data (200 patients). The ML models were trained with the four-folds, and their performance on the test data was recorded. This process was repeated five times, with each fold serving as the test data once. The average performance across all iterations was calculated. This process was repeated 10 times, shuffling the data and re-dividing data into folds. As a final metric, the average of the repeated cross-validation results was calculated. Different regularization techniques specified for each algorithm were applied to prevent overfitting. Additionally, each algorithm was tuned with different hyper-parameters, maximizing the repeated cross-validation metric on test data, and the best hyper-parameter selections for each algorithm were found.

2.5 | ML models' performance evaluation

ML models were evaluated by comparing their predictions to the actual treatment plans determined by the orthodontists for each sample. As a metric of success, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, accuracy, precision and recall were measured. The ROC curve is a graphical representation of a model's classification performance at different thresholds, and the AUC is calculated by measuring the area under this curve. A higher AUC value indicates a more successful ML model, with a value closer to 1 signifying superior performance. The equations and explanations for accuracy, precision, and recall metrics were provided in Figure S1. Three of the algorithms that achieve the highest performance were selected and combined using ensemble models known as voting and stacking. The percentage of patients correctly assigned to the appropriate treatment plan, along with 95% confidence intervals (CIs), was calculated.

MLxtend library was used for statistical analysis. Cochran's Q test was utilized to assess the performance of the classifiers and Pair-wise McNemar tests were conducted to examine specific differences at the $P < .05$ level of significance (Table S1).

3 | RESULTS

3.1 | Reliability analysis

ICC values were ranged from 0.91 to 0.98, representing excellent reliability.

3.2 | Data analysis

The mean age of all patients was 14.98 ± 4.87. For the non-extraction group, the mean age was 14.95 ± 5.53, while for the extraction group was 15.00 ± 4.12; and there was no statistically significant difference between the two groups ($P \geq .05$). Among the extraction cases, Class

TABLE 1 Variable definitions.

Variable	Description of feature
Age	Patient's age at the beginning of treatment
Sex	Patient's biological sex
Lip posture	Competent—Upper and lower lip separated by less than 3 mm in relaxed position Incompetent—Upper and lower lip separated by 3 mm or more in relaxed position
Angle's molar classification	Class I relation: The mesiobuccal cusp of upper first permanent molar occludes with mesiobuccal groove of the lower first permanent molar. Class II relation: According to class I malocclusion, the lower molar is distal to the upper molar • Division 1: The overjet is increased with upright or proclined upper incisors • Division 2: The upper incisors are retroclined, with a normal or occasionally increased overjet Class III relation: According to class I malocclusion, the lower molar is mesial to the upper molar
Maxillary arch length discrepancy (mm)	Amount of maxillary arch crowding
Mandibular arch length discrepancy (mm)	Amount of mandibular crowding
Curve of Spee (mm)	The perpendicular distance between the deepest cusp tip and a flat plane on the mandibular occlusal surface. The measurement was made on the right and left side, then the mean value was determined
SNA (°)	The angle between Sella, Nasion and A point
SNB (°)	The angle between Sella, Nasion and B point
ANB (°)	The angle between A point, Nasion and B point
Wits appraisal (mm)	The distance between the perpendiculars of A point and B point on the occlusal plane
GoGnSN (°)	The angle between Gonion-Gnathion plane and Sella-Nasion plane
FMA (°)	The angle between Frankfort horizontal plane and mandibular plane
Sum of the angles (°)	Sum of the saddle angle, articular angle and gonial angle
ANS-Me (mm)	The distance between the ANS point and Me point
U1-NA (mm)	The distance between the tip of the upper incisor and the plane Nasion to A point
U1-NA (°)	The angle between the long axis of the upper incisor and the plane Nasion to A point
U1-FH (°)	The angle between the long axis of the upper incisor and Frankfort horizontal plane
U1-PP (°)	The angle between the long axis of the upper incisor and palatal plane
L1-NB (mm)	The distance between the tip of the lower incisor and the plane Nasion to B point
L1-NB (°)	The angle between the long axis of the lower incisor and the plane Nasion to B point
IMPA (°)	The angle between the long axis of the lower incisor and mandibular plane
L1-A-Pg (mm)	The distance between the tip of the lower incisor and the plane A point to Pg
Interincisal angle (°)	The angle between the long axis of the upper and lower incisor
Cephalometric overjet (mm)	The distance between the tip of the upper and lower incisor measured throughout the occlusal plane
Cephalometric overbite (mm)	The distance between the tip of the upper and lower incisor measured perpendicular to the occlusal plane
Upper lip to E line (mm)	Most prominent distance from the upper lip to the E-line
Lower lip to E line (mm)	Most prominent distance from the lower lip to the E-line
Nasolabial angle (°)	The angle between columella tangent and upper lip tangent
Facial convexity angle (°)	The angle between the plane soft tissue glabella to subnasale and the plane subnasale to soft tissue pogonion
Z angle (°)	The angle between Frankfort horizontal plane and the plane most anterior point of the most protrusive lip to soft tissue pogonion
Maxillary sulcus (°)	The angle between subnasale, soft tissue A and upper lip point
Mandibular sulcus (°)	The angle between lower lip, soft tissue Pogonion, soft tissue B point
Upper lip thickness (mm)	The horizontal distance between the upper lip most labial and maxillary incisor most labial
Lower lip thickness (mm)	The horizontal distance between the lower lip most labial and maxillary incisor tip
Upper incisor exposure (mm)	The distance between the tip of the upper incisor and upper stomion along the vertical plane

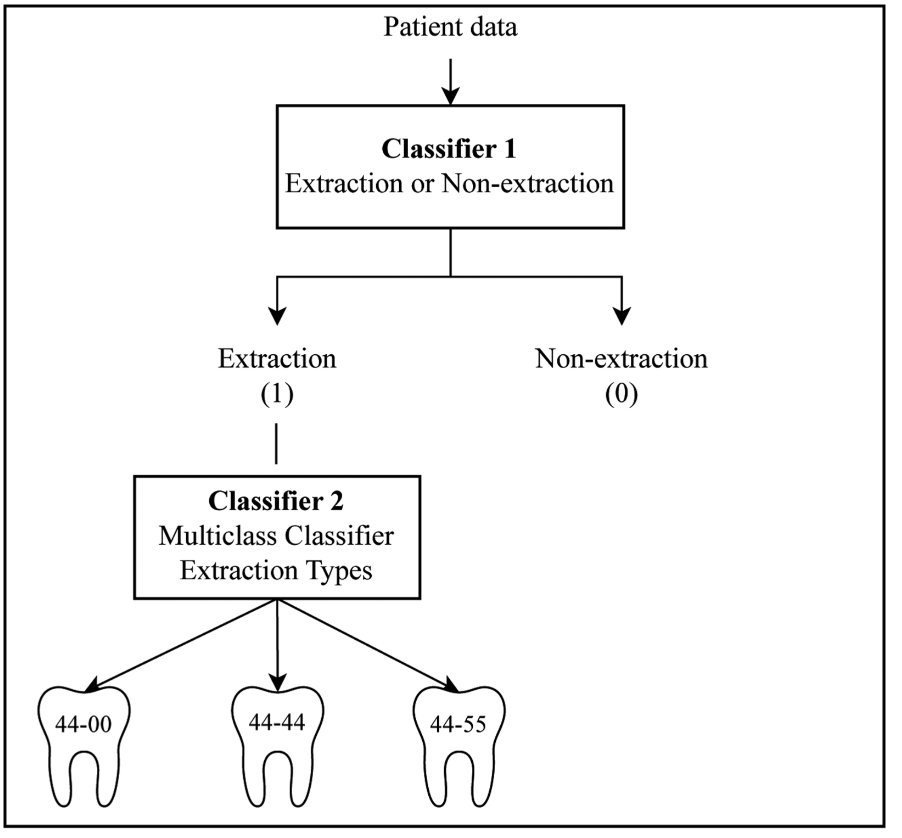

FIGURE 1 Two-step data processing flow.

II division 1 was the most common malocclusion with 61.2%. On the other hand, Class I malocclusion was the most common among the non-extraction cases with 54.4% (Table 2). The mean values of the variables for the extraction and non-extraction groups were given in Table 2.

3.3 | ML models' performance results

The null hypothesis was rejected because there were differences between the seven ML models, and the model with the highest performance was the Stacking Classifier for both decisions. Although the Stacking Classifier was the best model, the ML models' success was found to be close to each other. Accuracy, AUC, precision and recall of the ML models were given in Table 3. The performances of the five models, excluding the ensemble models, were given below in order of success based on AUC values.

Models' performance for the extraction/non-extraction decision as follows:

Gradient Boosted Trees—83.3% accuracy (95% CI of 82.8%–83.8%) and 0.899 AUC (95% CI of 0.897–0.902).
Support Vector Machine—82.6% accuracy (95% CI of 82.4%–82.8%) and 0.898 AUC (95% CI of 0.897–0.900).
Random Forest—81.7% accuracy (95% CI of 81.3%–82.1%) and 0.894 AUC (95% CI of 0.892–0.896).
Logistic Regression—82.4% accuracy (95% CI of 82.1%–82.7%) and 0.888 AUC (95% CI of 0.889–0.897).
Multi-Layer Perceptron—81.5% accuracy (95% CI of 81%–82%) and 0.881 AUC (95% CI of 0.875–0.887).

Models' performance for the extraction type decision as follows:

Gradient Boosted Trees—75.9% accuracy (95% CI of 75.1%–76.7%).
Support Vector Machine—75.3% accuracy (95% CI of 75.1%–75.5%).
Random Forest—75.1% accuracy (95% CI of 74.6%–75.6%).
Logistic Regression—75.7% accuracy (95% CI of 75.3%–76.1%).
Multi-Layer Perceptron—75% accuracy (95% CI of 74.3%–75.7%).

The performances of the ensemble models were presented below in order of success based on AUC values.

Stacking Classifier model consists of three models with the highest performance: Random Forest, Support Vector Machine, and Gradient Boosted Trees. This model showed 84.1% accuracy (95% CI of 83.8%–84.4%) and 0.912 AUC (95% CI of 0.908–0.916) for the extraction/non-extraction decision; 76.3% accuracy (95% CI of 76.0%–76.6%) for the extraction type decision. The ROC curve graph of the Stacking Classifier is given in Figure S2.

Voting Classifier model consists of three models with the highest performance: Random Forest, Support Vector Machine, and Gradient Boosted Trees. This model showed 83% accuracy (95% CI of 82.8%–83.2%) and 0.905 AUC (95% CI of 0.903–0.907) for the extraction/non-extraction decision; 76.1% accuracy (95% CI of 75.7%–76.5%) for the extraction type decision.

TABLE 2 Demographics, clinical characteristics, model and cephalometric measurements of patients.

Variables	Total	Treatment type: Extraction	Treatment type: Non-extraction	*P*-value
Mean age ± SD (years)	14.98 ± 4.87	15.00 ± 4.12	14.95 ± 5.53	.888
Sex (*n*) (%)				
Female	618	314 (62.8%)	304 (60.8%)	.515
Male	382	186 (37.2%)	196 (39.2%)	
Angle's molar relationship (*n*) (%)				
Class I	439	167 (33.4%)	272 (54.4%)	<.001*
Class II Division 1	500	306 (61.2%)	194 (38.8%)	
Class II Division 2	47	23 (4.6%)	24 (4.8%)	
Class III	14	4 (0.8%)	10 (2%)	
Molar key (*n*) (%)				
Class III key	14	4	10	
Super Class I key	101	15	86	
Class I key	338	132	206	<.001*
End-on key	211	98	113	
Class II key	336	231	105	
Lip posture (*n*) (%)				
Compotent	947	457 (91.4%)	490 (98%)	<.001*
Incompotent	53	43 (8.6%)	10 (2%)	
Model measurements (mean ± SD)				
Maxillary arch length discrepancy (mm)	−3.91 ± 3.69	−5.8 ± 3.48	−2.02 ± 2.81	<.001*
Mandibular arch length discrepancy (mm)	−2.45 ± 2.63	−3.5 ± 2.80	−1.4 ± 1.95	<.001*
Curve of Spee (mm)	2.68 ± 0.50	2.72 ± 0.52	2.64 ± 0.47	.012*
Cephalometric measurements (mean ± SD)				
SNA (°)	80.43 ± 3.79	80.31 ± 3.91	80.55 ± 3.67	.312
SNB (°)	76.59 ± 3.77	75.97 ± 3.59	77.22 ± 3.84	<.001*
ANB (°)	3.84 ± 2.60	4.34 ± 2.57	3.33 ± 2.54	<.001*
Wits Appraisal (mm)	1.41 ± 3.93	2.31 ± 4.05	0.51 ± 3.59	<.001*
GoGnSN (°)	33.88 ± 5.89	34.87 ± 5.97	32.88 ± 5.64	<.001*
FMA (°)	27.56 ± 5.51	28.55 ± 5.70	26.57 ± 5.14	<.001*
Sum of the angles (°)	396.84 ± 5.83	397.8 ± 5.94	395.88 ± 5.57	<.001*
ANS-Me (mm)	62.90 ± 6.50	63.88 ± 5.92	61.91 ± 6.89	<.001*
U1-NA (mm)	4.46 ± 2.95	4.83 ± 3.07	4.09 ± 2.78	<.001*
U1-NA (°)	22.74 ± 8.80	23.69 ± 9.17	21.80 ± 8.32	.001*
U1-FH (°)	112.45 ± 8.92	113.25 ± 9.27	111.65 ± 8.49	.005*
U1-PP (°)	111.56 ± 8.51	112.21 ± 8.92	110.92 ± 8.03	.016*
L1-NB (mm)	5.23 ± 2.51	5.70 ± 2.55	4.76 ± 2.38	<.001*
L1-NB (°)	26.06 ± 6.83	26.87 ± 6.93	25.25 ± 6.65	<.001*
IMPA (°)	92.62 ± 7.71	93.10 ± 7.86	92.15 ± 7.54	.053
L1-A-Pg (mm)	2.23 ± 2.57	2.43 ± 2.66	2.03 ± 2.46	.015*
Interincisal angle (°)	127.36 ± 12.2	125.1 ± 12.34	129.62 ± 11.64	<.001*
Cephalometric overjet (mm)	4.5 ± 2.59	5.03 ± 2.77	3.98 ± 2.28	<.001*
Cephalometric overbite (mm)	2.46 ± 2.46	2.21 ± 2.55	2.71 ± 2.35	.001*
Upper lip to E line (mm)	−2.53 ± 2.45	−2.20 ± 2.46	−2.86 ± 2.40	<.001*
Lower lip to E line (mm)	−0.87 ± 2.64	−0.43 ± 2.69	−1.31 ± 2.51	<.001*

TABLE 2 (Continued)

Variables	Total	Treatment type: Extraction	Treatment type: Non-extraction	*P*-value
Nasolabial angle (°)	109.76 ± 10.38	109.4 ± 10.64	110.12 ± 10.11	.273
Facial convexity angle (°)	17.03 ± 6.27	18.15 ± 5.92	15.92 ± 6.43	<.001*
Z angle (°)	72.31 ± 7	70.88 ± 6.62	73.75 ± 7.08	<.001*
Maxillary sulcus depth (mm)	151.58 ± 12.18	152.83 ± 11.98	150.32 ± 12.26	.001*
Mandibulary sulcus depth (mm)	128.70 ± 14.76	128.58 ± 15.30	128.81 ± 14.21	.812
Upper lip thickness (mm)	11.69 ± 2.20	11.39 ± 2.15	11.98 ± 2.21	<.001*
Lower lip thickness (mm)	10.71 ± 2.43	10.40 ± 2.41	11.02 ± 2.42	<.001*
Upper incisor exposure (mm)	2.90 ± 1.92	2.86 ± 2	2.94 ± 1.85	.513

Note: *T*-test to compare age, model and cephalometric measurements between extraction and non-extraction groups. Chi-square test to compare sex, Angle's molar relationship and lip posture between extraction and non-extraction groups.

Abbreviation: SD, standard deviation.

*Statistically significant differences ($P \leq .05$).

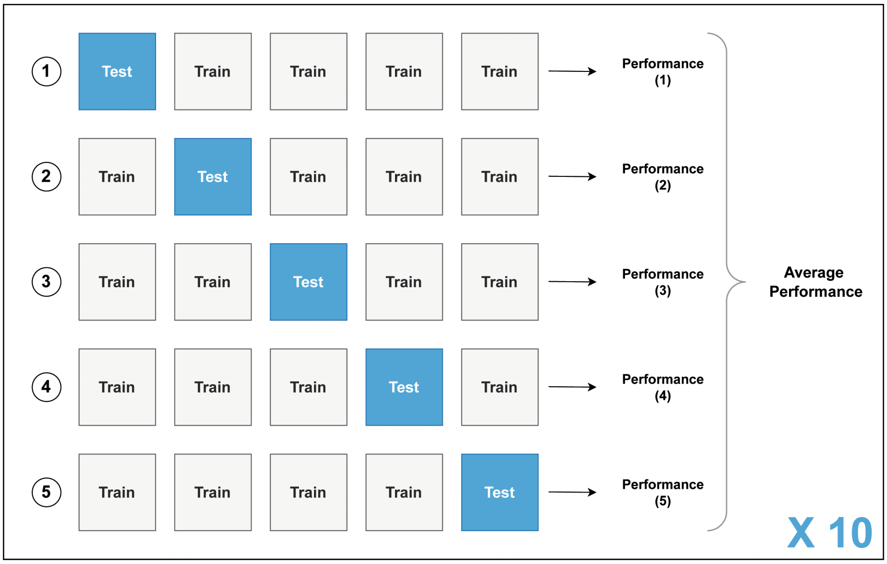

FIGURE 2 The cross-validation method used in this study.

The performance assessment of the seven ML models is given in Table S1.

3.4 | Important variables of the best-performing model

The most important variables affecting the extraction/non-extraction decisions (classifier 1) were shown in Figure 3: maxillary (30.7%) and mandibular (10.56%) arch length discrepancies, Wits Appraisal (6.73%) and ANS-Me length (6.65%). There was a negative correlation between the probability of extraction and maxillary/mandibular arch length discrepancy. On the other hand, there is a positive correlation between Wits Appraisal and ANS-Me length.

The most important variables affecting the extraction type decision were mandibular arch length discrepancy (14.56%), Class I molar relationship (12.2%), cephalometric overbite (7.69%), Wits Appraisal (7.25%) and L1-NB distance (7.19%).

For the 44-00 and 44-44 groups, Class I molar relationship and mandibular arch length discrepancy were found the most important variables. Patients with molar relationships other than Class I and patients with high values for mandibular arch length discrepancies were more likely to be in the 44-00 group, whereas the opposite was true for the 44-44 group (Figure S3 and S4). The important variables for the 44-55 group are given in Figure S5.

4 | DISCUSSION

ML is widely used in both dentistry and continuously advancing, especially in the fields of diagnosis and treatment planning. In this study, the success of seven ML models in extraction decisions was

TABLE 3 Performance values of machine learning models with 95% confidence interval.

Model	Multi-layer perceptron	Random forest	Logistic regression	Support vector machine	Gradient boosted trees	Voting classifier	Stacking classifier
Extraction/non extraction (classifier 1)							
Accuracy	81.5% (81%–82%)	81.7% (81.3%–82.1%)	82.4% (82.1%–82.7%)	82.6% (82.4%–82.8%)	83.3% (82.8%–83.8%)	83% (82.8%–83.2%)	84.1% (83.8%–84.4%)
AUC	0.881 (0.875–0.887)	0.894 (0.892–0.896)	0.888 (0.889–0.897)	0.898 (0.897–0.900)	0.899 (0.897–0.902)	0.905 (0.903–0.907)	0.912 (0.908–0.916)
Precision	80.9% (80.2%–81.6%)	80.5% (80.2%–80.8%)	81.5% (81.1%–81.9%)	81.3% (81.2%–81.4%)	82.8% (82.2%–83.4%)	82.6% (82.4%–82.8%)	83.2% (82.8%–83.6%)
Recall	82.6% (80%–83.2%)	83.8% (83.4%–84.2%)	83.4% (82.9%–83.9%)	83.7% (83.5%–83.9%)	84.1% (83.7%–84.5%)	83.8% (83.7%–83.9%)	84.4% (84.1%–84.7%)
Extraction type (classifier 2)							
	Multi-layer perceptron	Random forest	Logistic regression	Support vector machine	Gradient boosted trees	Voting classifier	Stacking classifier
Accuracy	75% (74.3%–75.7%)	75.1% (74.6%–75.6%)	75.7% (75.3%–76.1%)	75.3% (75.1%–75.5%)	75.9% (75.1%–76.7%)	76.1% (75.7%–76.5%)	76.3% (76.0%–76.6%)

Abbreviation: AUC, area under the ROC curve.

compared with decisions made by orthodontists with a minimum of 10 years of academic and clinical experience. The objective was to identify the most effective ML method that could provide support to clinicians with less clinical experience. To achieve this, seven different ML models were trained using data from 1000 patients, incorporating 36 parameters. The Stacking Classifier model, which combines the three most successful models' outputs, achieved the highest accuracy of 84.1% and the highest AUC of 0.912 for the extraction/non-extraction decision. The three best-performing models are Gradient Boosted Trees, Support Vector Machines, and Random Forest. The highest accuracy values in extraction decision were obtained 94% with the Multilayer Perceptron model in Li et al.[18] study. Also, Del Real et al.[22] found 93.9% accuracy with Automated AI and Jung et al.[17] found 93% accuracy with Back Propagation. While evaluating these results with high accuracy, it should be taken into account that Jung et al.[17] and Li et al.[18] did not apply any cross-validation techniques. In the literature, two studies measured AUC with the Multilayer Perceptron model as 98% and 82%, respectively.[18,21] However, the possibility of overfitting should be considered in these studies. On the other hand, Support Vector Machine was found to be the most successful model with 92.5% AUC in the study of Mason et al.,[29] which was in agreement with the present study, being one of the best-performing models.

The highest accuracy in the extraction type decision was obtained with 76.3% with the Stacking Classifier model. There were only two studies that include the type of extraction in their research. Jung et al.[17] found 84% accuracy, and Li et al.[18] found 84.2% accuracy for this decision. It is hard to compare studies with different demographic groups and extraction type distributions. Labels for the extraction type were also different in the study of Li et al.; the 44-00 extraction group was not included, whereas this extraction type was dominant in the present study. However, the rate of 76.3% obtained from the present study was not deemed sufficiently high to contribute to the extraction decisions.

Previous studies had used Artificial Neural Networks,[16–18,20,21,29] Random Forest,[20,21,29] Logistic Regression,[20,29] and Support Vector Machine[29] for extraction decision. Gradient Boosted Trees, a state-of-the-art algorithm used for classification tasks, was included in this study.[30] Ensemble models are popular in the area of ML, providing better performances than single models. Each ML model has its own strengths and weaknesses. By appropriately merging their outputs, their strengths can be combined and their weaknesses can be eliminated. Therefore, two ensemble models using voting and stacking methods were added in this study. Voting classifier uses multiple models to make predictions, and chooses the most popular class. Stacking classifier aggregates predictions of multiple models, using a meta-algorithm to make the final prediction.

This is the first study in the orthodontic literature to train ML models using a dataset of 1000 patients. Etemad et al.[21] evaluated the highest number of patients in the literature, with a total of 838 cases. It is important to note that having a larger dataset is crucial for achieving improved and more generalizable performance in ML models. However, even 1000 patients with 36 parameters were not

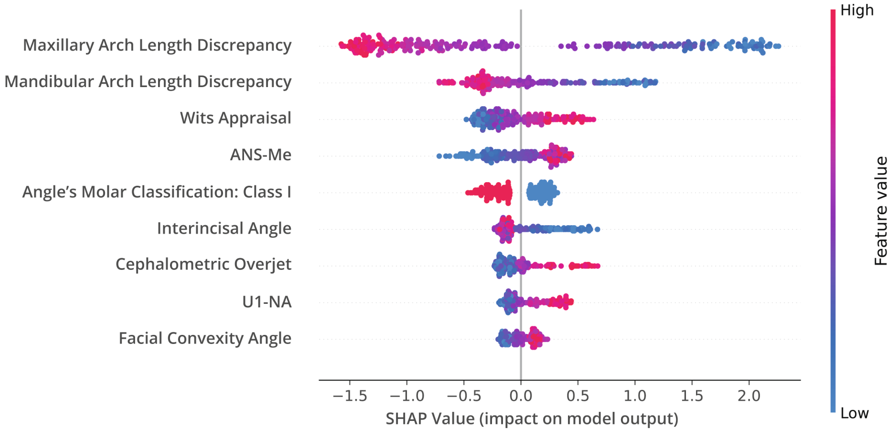

FIGURE 3 SHAP summary plot of the top nine variables of the Stacking Classifier model in extraction decision. Positive (extraction) and negative (non-extraction) values represent the classes for the extraction decision. Higher SHAP values indicate a higher impact on the model decision. For each variable, a dot was created indicating the patient's value in that variable and the colour of this point is determined by the value of the variable. Red represents higher values of that variable and blue represents lower values. The definitions of the variables were given in Table 2.

a large sample for ML studies. Another problem is the imbalanced sample distribution which has negative consequences. The distribution of extraction and non-extraction groups was generally around 40%–60% to 24.8%–75.2% in the previous studies.[16–18,21,22] In this study, while the sample size was kept as high as possible, the ratio of extraction and non-extraction was kept equal (%50). However, due to the limited number of patients in each extraction group, this balance could not be achieved for the extraction types.

One of the major challenges in ML is overfitting. This phenomenon occurs when the model becomes unnecessarily complex, essentially memorizing the training data and failing to generalize to solve the problem. In such conditions, metrics on the training sets are high, but they are significantly lower on the unseen test data. In order to prevent overfitting, samples can be divided as train and test sets, regularization techniques specific to each ML model can be used, and cross-validation can be applied. In this study, all data were used both during train and test phases (Figure 2). In traditional approaches where data are divided into train and test sets directly, the output scores cannot be generalized as they may be dependent on a specific split, or the order of the data. When the previous studies were investigated, three studies used cross-validation methods.[21,22,29] The 5-fold cross-validation method was used and repeated 10 times to establish a stronger validation scheme in this research.

The accuracy and AUC were used as success metrics. AUC is a metric that identifies how two classes are separable and gives a better understanding of the performance of the model when data is imbalanced.[31] In a ML problem where 90% of the data is positive, and 10% is negative, accuracy can be a misleading metric as a model always classifies positive would reach 90% accuracy. AUC was utilized as a success metric in three previous studies.[18,21,29] Mason et al.[29] calculated the CIs of the four ML models and found larger intervals for each model compared to this study. The reason for the significantly smaller CIs in current study is the larger dataset, using 5-fold cross-validation repeated 10 times, and tuning of each model.

In this study, the decisions of four orthodontists who have been working together at the same university for at least 10 years have been utilized. Even though these four orthodontists have received the same training and have been working together for years, they may have different tendencies in treatment decisions. In the literature, there is no consensus among ML studies on whether the extraction decision is made by a single expert, by a committee, or by multiple experts. Jung et al.[17] and Li et al.[18] achieved a more stable model by, respectively, using one and two experts, leading to high performance. Unfortunately, the generalizability of models trained with a single perspective is questionable. Leavitt et al.[32] extensively discussed this fact in their study, in which extraction decisions were made by 30 experts. Treatment bias can be prevented in models trained with various perspectives and treatment philosophies. However, the inclusion of too many experts in treatment planning can also reduce consistency.

Parameters that affect the extraction decision of clinicians are one of the important topics for orthodontics. Numerous studies have demonstrated that one of the most important parameters is crowding.[1,33–36] However, Evrard et al.[34] found that the soft tissue profile has a greater impact on the decision to extract teeth in their survey study. This indicates that orthodontists are now placing more emphasis on facial aesthetics. Iard et al. revealed that the initial lip protrusion is a helpful parameter for the extraction decision in borderline cases.[37] However, the extent of soft tissue changes caused by tooth and alveolar process movement is still uncertain.[38,39] Another parameter for the extraction decision is the vertical pattern of the patient. Guo et al.[33] found that the growth pattern was the most important parameter for the extraction decision. While extraction treatments tend to be performed in hyper-divergent patients, non-extraction treatments are mostly performed in mesio-divergent patients. However, some studies have argued that treatment with or without extraction does not cause a significant vertical change.[8,40] Previous ML studies concluded that the extraction decision of ML models is mostly based on maxillary/mandibular arch length discrepancy,[18,21,22,29] U1-NA(mm),[18,29] L1-NB (mm),[21,29] molar relationship,[22] incompetent lips and IMPA.[16] Current study revealed that the arch length discrepancy is the most important variable in the extraction decision for ML models, which was in agreement with the findings of previous studies.[18,21,22] This indicates that although tooth extraction decisions are made for various conditions, the primary reason is still the lack of space. Unlike previous research, this study found Wits Appraisal and ANS-Me length as other important parameters for the extraction decision. These parameters also point to the sagittal and vertical problems of the patients.

4.1 | Clinical significance

Based on these results, the Stacking Classifier, which consists of Gradient Boosted Trees, Support Vector Machine, and Random Forest models, appears to be the best ML model. As algorithms continue to be developed, these models could greatly contribute to the planning phase of orthodontic treatment. It is important to note that no ML model can replace the judgement of an experienced orthodontist.

4.2 | Limitations

One of the limitations was that some parameters, including the patient's cooperation, desire for the treatment plan, presence of dental asymmetry, maturation stage, gingival biotype, photographic soft tissue analysis (gummy smile, etc.), which could potentially influence the extraction decision were not included in this study. The quality assessment of treatment outcomes could not be conducted as some of the patients' treatments are still ongoing. In future studies, including only patients who have successfully completed their treatment can increase the quality of the study. Another limitation was that a sufficient number of patients could only be obtained just for some extraction groups since the most common types of extractions were 44-00, 44-44, and 44-55 in Başkent University's Orthodontic clinic. Other combinations of premolar extractions, molar, and lower incisor extractions could also be included in future studies. Due to the retrospective nature of this study, the inter-examiner and intra-examiner reliability of orthodontists' treatment decisions could not be analysed. Future research involving multiple orthodontists with different perspectives and backgrounds contributes to the generalization of the study. Finally, further investigations can enhance the reliability of the study by testing it with an external test set.

5 | CONCLUSIONS

1. This study compared seven machine learning models for extraction decision and extraction type. The highest performing model was the Stacking Classifier with 84.1% accuracy and 0.912 AUC value.
2. While ML models showed a high performance in extraction decisions, they could not show this performance in extraction type decisions.
3. The variables that play an important role in the decision of extraction for ML models were maxillary/mandibular arch length discrepancy, Wits Appraisal, and ANS-Me length. On the other hand, the most important variables in the extraction type decision were mandibular arch length discrepancy, Class I molar relationship, cephalometric overbite, Wits Appraisal, and L1-NB distance.
4. Future studies should focus not only on the development of algorithms, but also on matching clinical scenarios in order to evaluate the algorithms and support decision-making in extraction.

AUTHOR CONTRIBUTIONS

B.K. contributed to the concept of this study, determined the methodology, explored the sources, and wrote the original text. H.P. contributed to the concept design of this study, determined the methodology, reviewed and edited the original text, and supervised. Ö.G. determined the methodology, implemented the software, performed the data curation, and wrote the original text. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

This study was supported by Başkent University Research Fund. All authors have nothing to disclose.

CONFLICT OF INTEREST STATEMENT

The authors have nothing to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICAL APPROVAL

Written informed consents from the patients had been collected at the beginning of the treatment as a standard procedure of the Başkent University's orthodontic clinic.

ORCID

Hande Pamukçu https://orcid.org/0000-0003-4242-5114

REFERENCES

1. Jackson TH, Guez C, Lin FC, Proffit WR, Ko CC. Extraction frequencies at a university orthodontic clinic in the 21st century: demographic and diagnostic factors affecting the likelihood of extraction. *Am J Orthod Dentofacial Orthop*. 2017;151(3):456-462. doi:10.1016/j.ajodo.2016.08.021
2. Saghafi N, Heaton LJ, Bayirli B, Turpin DL, Khosravi R, Bollen AM. Influence of clinicians' experience and gender on extraction

decision in orthodontics. *Angle Orthod*. 2017;87(5):641-650. doi:10.2319/020117-80.1
3. Al-Ani MH, Mageet AO. Extraction planning in orthodontics. *J Contemp Dent Pract*. 2018;19(5):619-623. doi:10.5005/jp-journals-10024-2307
4. Konstantonis D, Vasileiou D, Papageorgiou SN, Eliades T. Soft tissue changes following extraction vs. nonextraction orthodontic fixed appliance treatment: a systematic review and meta-analysis. *Eur J Oral Sci*. 2018;126(3):167-179. doi:10.1111/eos.12409
5. Peck S. Extractions, retention and stability: the search for orthodontic truth. *Eur J Orthod*. 2017;39(2):109-115. doi:10.1093/ejo/cjx004
6. Baek ES, Hwang S, Kim KH, Chung CJ. Total intrusion and distalization of the maxillary arch to improve smile esthetics. *Korean J Orthod*. 2017;47(1):59-73.
7. Germec-Cakan D, Taner TU, Akan S. Arch-width and perimeter changes in patients with borderline class i malocclusion treated with extractions or without extractions with air-rotor stripping. *Am J Orthod Dentofacial Orthop*. 2010;137(6):734.e1-734.e7. doi:10.1016/j.ajodo.2009.12.023
8. Kirschneck C, Proff P, Reicheneder C, Lippold C. Short-term effects of systematic premolar extraction on lip profile, vertical dimension and cephalometric parameters in borderline patients for extraction therapy—a retrospective cohort study. *Clin Oral Investig*. 2016;20(4):865-874. doi:10.1007/s00784-015-1574-5
9. Herzog C, Konstantonis D, Konstantoni N, Eliades T. Arch-width changes in extraction vs nonextraction treatments in matched class I borderline malocclusions. *Am J Orthod Dentofacial Orthop*. 2017;151(4):735-743. doi:10.1016/j.ajodo.2016.10.021
10. Domingos P. A few useful things to know about machine learning. *Commun ACM*. 2012;55(10):78-87. https://dl.acm.org/citation.cfm?id=2347755
11. Tandon D, Rajawat J, Banerjee M. Present and future of artificial intelligence in dentistry. *J Oral Biol Craniofacial Res*. 2020;10(4):391-396. doi:10.1016/j.jobcr.2020.07.015
12. Shan T, Tay FR, Gu L. Application of artificial intelligence in dentistry. *J Dent Res*. 2021;100(3):232-244. doi:10.1177/0022034520969115
13. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: a scoping review. *Am J Orthod Dentofacial Orthop*. 2021;160(2):170-192. doi:10.1016/j.ajodo.2021.02.013
14. Bichu YM, Hansa I, Bichu AY, Premjani P, Flores-Mir C, Vaid NR. Applications of artificial intelligence and machine learning in orthodontics: a scoping review. *Prog Orthod*. 2021;22(1):1-11.
15. Duran GS, Gökmen Ş, Topsakal KG, Görgülü S. Evaluation of the accuracy of fully automatic cephalometric analysis software with artificial intelligence algorithm. *Orthod Craniofac Res*. 2023;26:481-490.
16. Xie X, Wang L, Wang A. Artificial neural network modeling for deciding if extractions are necessary prior to orthodontic treatment. *Angle Orthod*. 2010;80(2):262-266. doi:10.2319/111608-588.1
17. Jung SK, Kim TW. New approach for the diagnosis of extractions with neural network machine learning. *Am J Orthod Dentofacial Orthop*. 2016;149(1):127-133. doi:10.1016/j.ajodo.2015.07.030
18. Li P, Kong D, Tang T, et al. Orthodontic treatment planning based on artificial neural networks. *Sci Rep*. 2019;9(1):2037. doi:10.1038/s41598-018-38439-w
19. Takada K. Artificial intelligence expert systems with neural network machine learning may assist decision-making for extractions in orthodontic treatment planning. *J Evid Based Dent Pract*. 2016;16(3):190-192. doi:10.1016/j.jebdp.2016.07.002
20. Suhail Y, Upadhyay M, Chhibber A. Machine learning for the diagnosis of orthodontic extractions: a computational analysis using ensemble learning. *Bioengineering*. 2020;7(2):55. doi:10.3390/bioengineering7020055
21. Etemad L, Wu TH, Heiner P, et al. Machine learning from clinical data sets of a contemporary decision for orthodontic tooth extraction. *Orthod Craniofac Res*. 2021;24:193-200. doi:10.1111/ocr.12502
22. Del Real A, Del Real O, Sardina S, Oyonarte R. Use of automated artificial intelligence to predict the need for orthodontic extractions. *Korean J Orthod*. 2022;52(2):102-111. doi:10.4041/kjod.2022.52.2.102
23. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18-22.
24. Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA: automatic model selection and hyperparameter optimization in WEKA. *Automated Machine Learning: Methods, Systems, Challenges*. Cham, Switzerland AG: Springer Nature; 2019:81-95.
25. Lee R, MacFarlane T, O'Brien K. Consistency of orthodontic treatment planning decisions. *Clin Orthod Res*. 1999;2(2):79-84. doi:10.1111/ocr.1999.2.2.79
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
27. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-3154.
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4768-4777.
29. Mason T, Kelly KM, Eckert G, Dean JA, Dundar MM, Turkkahraman H. A machine learning model for orthodontic extraction/non-extraction decision in a racially and ethnically diverse patient population. *Int Orthod*. 2023;21(3):100759. doi:10.1016/j.ortho.2023.100759
30. Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. Supervised machine learning algorithms: classification and comparison. *Int J Comput Trends Technol*. 2017;48(3):128-138.
31. Wardhani NWS, Rochayani MY, Iriany A, Sulistyono AD, Lestantyo P. Cross-validation metrics for evaluating classification performance on imbalanced data. *2019 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*. 2019;IEEE:14-18. doi:10.1109/IC3INA48034.2019.8949568
32. Leavitt L, Volovic J, Steinhauer L, et al. Can we predict orthodontic extraction patterns by using machine learning? *Orthod Craniofac Res*. 2023;26:552-559.
33. Guo Y, Han X, Xu H, Ai D, Zeng H, Bai D. Morphological characteristics influencing the orthodontic extraction strategies for Angle's class II division 1 malocclusions. *Prog Orthod*. 2014;15:1-7. doi:10.1186/s40510-014-0044-y
34. Evrard AS, Tepedino M, Cattaneo PM, Cornelis MA. Which factors influence orthodontists in their decision to extract? A questionnaire survey. *J Clin Exp Dent*. 2019;11(5):e432-e438. doi:10.4317/jced.55709
35. Bishara SE, Cummins DM, Jakobsen JR. The morphologic basis for the extraction decision in class II, division 1 malocclusions: a comparative study. *Am J Orthod Dentofacial Orthop*. 1995;107(2):129-135.
36. Konstantonis D, Anthopoulou C, Makou M. Extraction decision and identification of treatment predictors in class I malocclusions. *Prog Orthod*. 2013;14:1-8. doi:10.1186/2196-1042-14-47
37. Iared W, da Silva EMK, Iared W, Macedo CR. Esthetic perception of changes in facial profile resulting from orthodontic treatment with extraction of premolars. *J Am Dent Assoc*. 2017;148(1):9-16. doi:10.1016/j.adaj.2016.09.004
38. Hodges A, Rossouw PE, Campbell PM, Boley JC, Alexander RA, Buschang PH. Prediction of lip response to four first premolar extractions in white female adolescents and adults. *Angle Orthod*. 2009;79(3):413-421. doi:10.2319/050208-247.1
39. Erdinc AE, Nanda RS, Dandajena TC. Profile changes of patients treated with and without premolar extractions. *Am J Orthod Dentofacial Orthop*. 2007;132(3):324-331. doi:10.1016/j.ajodo.2005.08.045

40. Kocadereli İ. The effect of first premolar extraction on vertical dimension. *Am J Orthod Dentofacial Orthop*. 1999;116(1):41-45.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Köktürk B, Pamukçu H, Gözüaçık Ö. Evaluation of different machine learning algorithms for extraction decision in orthodontic treatment. *Orthod Craniofac Res*. 2024;27(Suppl. 2):13-24. doi:10.1111/ocr.12811