DOI: 10.4274/TurkJOrthod.2026.2025.190

TURKISH JOURNAL OF
ORTHODONTICS

Original Article

Assessing the Diagnostic Performance of ChatGPT-5.0 versus Machine Learning in Orthodontics: A Comparative Analysis for Extraction Treatment Planning

Artun Yangın, Hasan Camcı, Mehmet Soybelli

Afyonkarahisar Health Sciences University Faculty of Dentistry, Department of Orthodontics, Afyonkarahisar, Türkiye

Cite this article as: Yangın A, Camcı H, Soybelli M. Assessing the diagnostic performance of ChatGPT-5.0 versus machine learning in orthodontics: a comparative analysis for extraction treatment planning. *Turk J Orthod.* [Epub Ahead of Print]

Main Points

- ChatGPT-5.0 achieved 75.77% accuracy in extraction treatment decisions.
- XGBoost achieved the highest overall accuracy (78.08%).
- ChatGPT-5.0 showed the highest sensitivity for extraction cases (76.68%).
- ChatGPT-5.0 outperformed random forest, support vector machine, logistic regression, and multi-layer perceptron.

ABSTRACT

Objective: To make accurate orthodontic extraction decisions, various clinical and cephalometric variables must be evaluated. This study aims to evaluate ChatGPT-5.0's performance in distinguishing orthodontic extraction decisions and to compare it with five supervised machine learning (ML) algorithms.

Methods: Of 550 retrospectively evaluated orthodontic records, 30 were reserved for calibration, leaving 520 for the main analysis. The reference standard was the consensus treatment decision of three expert orthodontists with more than 5 years of clinical experience. Overall, 23 variables were analyzed, including 13 clinical parameters, 7 cephalometric measurements, and photographs. ChatGPT-5.0's performance was evaluated using a 5-fold cross-validation design. It was compared with XGBoost, random forest, support vector machine (SVM), logistic regression, and multi-layer perceptron (MLP). Performance metrics included accuracy, sensitivity, specificity, precision, F1-score, and balanced accuracy, with 95% confidence intervals calculated. Statistical analyses utilized Cochran's Q test and the McNemar test with Holm-Bonferroni correction.

Results: Of the 520 main cases, 223 (42.88%) were extraction treatments and 297 (57.12%) were non-extraction treatments. XGBoost achieved the highest accuracy (78.08%), followed closely by ChatGPT-5.0 (75.77%). The overall performance difference among models was significant ($p \leq 0.001$). In pairwise comparisons, ChatGPT's accuracy was significantly higher than those of random forest, SVM, logistic regression, and MLP, but was found to be similar to XGBoost. ChatGPT-5.0 showed the highest sensitivity (76.68%), whereas XGBoost showed the highest specificity (82.15%).

Conclusion: ChatGPT-5.0 demonstrated performance comparable to, and in some cases superior to, traditional ML models for orthodontic extraction decisions. While XGBoost yielded the highest overall classification accuracy, ChatGPT-5.0's high sensitivity in detecting extraction cases was noteworthy.

Keywords: Artificial intelligence, ChatGPT, machine learning, orthodontic treatment planning, tooth extraction

Corresponding author: Res. Assist. Artun Yangın, **e-mail:** yangnartun@gmail.com **ORCID:** orcid.org/0009-0006-8229-6109
Received: January 05, 2026 **Accepted:** June 17, 2026 **Epub:** August 07, 2026

INTRODUCTION

Along with technological advancements, the disciplines of dentistry and orthodontics have undergone a comprehensive digital transformation.[1] Digital planning and workflows have become an integral part of routine practice.[2] In recent years, rapid progress in artificial intelligence (AI) has further accelerated this transformation.[3] Artificial intelligence applications in numerous areas such as cephalometric analysis,[4] patient cooperation prediction,[5] evaluation of cervical vertebral maturation stage,[6] unerupted tooth size prediction,[7] and treatment planning[8] have been widely documented in the literature. In this process, natural language processing systems that were initially text-based have evolved into more advanced, multimodal systems that include image-analysis capabilities, and have revealed remarkable potential in visually oriented areas such as the evaluation of orthodontic photographs and treatment planning.[9]

The use of large language models (LLMs) in orthodontics and dentistry has increased gradually in recent years.[1,10-12] However, there are significant differences, in terms of performance, consistency, and output structure, among various model families (such as ChatGPT, Gemini, DeepSeek, and Copilot) and among their versions.[13] In the literature, LLM applications are generally categorized into two main frameworks: text-based prompts and image-based or multimodal prompts.[12] While text-based applications mostly focus on patient communication, treatment planning, and clinical literature queries, performance on these tasks is evaluated using accuracy, readability, F1 score, and concordance metrics.[10] On the other hand, image-based applications requiring the interpretation of panoramic, cephalometric, or hand-wrist radiographs assess the visual processing capacity of multimodal LLMs and are generally evaluated using stricter quantitative metrics, such as mean absolute error, root mean square error, precision, and sensitivity.[14] Although previous studies have demonstrated that LLMs can demonstrate high reliability in text-based tasks, they also indicate that performance in image-based analyses varies significantly depending on the model and task, and in most cases such performance fails to reach the level of accuracy expected for clinical diagnostic use.[15-17]

In parallel with these developments, the systems used in orthodontics for diagnosis and determination of developmental stages are rapidly evolving from traditional deep learning (DL) approaches toward multimodal architectures.[18,19] Particularly for skeletal age determination, new-generation models that combine chaotic functional-connectivity matrices with metaheuristic methods, such as Puma Optimization, to analyze the nonlinear dynamics of cervical vertebrae have been reported to offer high accuracy and computational efficiency.[20] Similarly, task-specific DL architectures capable of pixel-level image recognition and feature extraction yield highly robust results in tasks such as cephalometric landmarking and radiological age determination.[21,22] Even when LLMs are guided by atlas-supported or example-enriched prompt strategies, they often lag behind task-specific machine learning (ML)/DL models in distinguishing fine visual details. Therefore, in light of current knowledge, LLMs are positioned as auxiliary clinical decision support tools rather than independent diagnostic systems.[22,23]

Orthodontic treatment planning is inherently a multidimensional process that requires the simultaneous evaluation of numerous variables, such as the patient's age, malocclusion type, cephalometric measurements, and dental relationships.[24-26] The decision between extraction and non-extraction treatment, one of the most critical stages of this process, can be particularly challenging for less-experienced clinicians, as it directly affects treatment mechanics, facial esthetics, and long-term stability.[25,27] For this reason, tools capable of processing multiple clinical inputs in a structured manner and providing support to the decision-making process are highly valuable in clinical practice.[28,29]

Although ML algorithms, which produce predictive results by identifying patterns in data, are increasingly used in orthodontic planning, studies directly comparing the performance of LLM-based approaches, particularly in multi-factor clinical decision tasks, with traditional ML/DL methods remain limited. Existing comparisons often remain superficial; the differences in data structures, training paradigms, evaluation strategies, and interpretability between these groups of models are not addressed in sufficient detail. This study aims to fill this gap by comparing ChatGPT's performance in distinguishing orthodontic extraction and non-extraction cases with that of traditional ML models on the same dataset.

The aim of the study was to compare the performance of ChatGPT with traditional ML models in predicting orthodontic extraction/non-extraction treatment decisions and to evaluate how the decision-making structures of these two approaches differ in terms of interpretability. A novel aspect of the study is that, on the one hand, it compares ChatGPT and traditional ML models on the same sample, and on the other hand, it demonstrates that the explanatory responses obtained from LLMs are not directly, methodologically, equivalent to quantitative feature importance metrics in ML models. The study presents a framework that jointly evaluates the potential and limitations of LLM use in orthodontics with respect to both performance and explainability. The null hypothesis (H_0) was that the LLM-based approach will not differ from ML algorithms in terms of its ability to distinguish cases; while the alternative hypothesis (H_1) predicts that there will be a statistically significant difference between the classification performances of these two approaches.

METHODS

This study was approved by the Institutional Afyonkarahisar Health Sciences University Non-Intervention Scientific Research Ethics Committee (approval no: 2024/8, date: 04.10.2024).

The principles of the Declaration of Helsinki were adhered to during the study process; written and verbal informed consent was obtained from all patients and their legal guardians.

Orthodontic records of patients who had completed treatment at our institution were retrospectively reviewed. Cases meeting the predefined inclusion and exclusion criteria were included in the study. The original plans of these patients had been determined and applied in accordance with a joint clinical decision by three expert orthodontists, each with at least five years of clinical experience, independent of the current study. Therefore, the treatment decisions made and applied in clinical practice were accepted as the reference standard in the study. To further validate the reference standard, the patients' post-treatment intraoral and extraoral records were also re-evaluated by three independent expert orthodontists. The evaluators were blinded to each other's evaluations, and the treatment outcomes were examined to determine whether they met the American Board of Orthodontics (ABO) criteria. Cases not meeting the ABO criteria were excluded from the study. Following this screening process, 550 cases fulfilled the study criteria. Of these, 520 cases comprised the primary analysis dataset, whereas the remaining 30 cases were used for prompt calibration.

The inclusion criteria for the study were: (1) fully erupted permanent dentition, (2) absence of dentofacial anomalies, (3) no indication for functional treatment, and (4) specific extraction patterns. In terms of treatment protocol, the study included non-extraction cases and cases in which the following premolars were extracted: upper and lower first premolars; upper first and lower second premolars; or only upper first premolars.

Exclusion criteria included impacted or supernumerary teeth, congenital tooth agenesis, retained primary teeth, diagnostic photographs of inadequate quality, changes in the treatment plan during treatment, and post-treatment records that failed to satisfy the ABO criteria.

The collected dataset consisted of intraoral and extraoral photographs, demographic information, diagnostic records, model analyses, and cephalometric measurements (Table 1). A total of 23 variables were used: 13 clinical parameters, 7 cephalometric measurements, 2 photograph

Table 1. Clinical, cephalometric, photographic, and demographic variables included in the analysis

Values	Parameter 1	Parameter 2	Parameter 3
Clinical values			
Age			
Gender	Female (0)	Male (1)	
Jaw position	Retrognathic (0)	Orthognathic (1)	Prognathic (2)
Molar relationship	Class I (0)	Class II (1)	Class III (2)
Overbite (mm)			
Overjet (mm)			
Vertical values	Decreased (0)	Normal (1)	Increased (2)
Upper midline	X mm to the right in the upper (+X)	X mm to the left in the upper (-X)	
Lower midline	X mm to the right in the lower (+X)	X mm to the left in the lower (-X)	
Upper crowding			
Lower crowding			
Upper diastema			
Lower diastema			
Cephalometric values			
SNA (°)			
SNB (°)			
ANB (°)			
FMA (°)			
U1-SN (°)			
IMPA (°)			
Nasolabial angle (°)			
Photos			
Intraoral			
Extraoral			
Treatment decision	Non-extraction (0)	Extraction (1)	

SNA, sella-nasion-A point angle; SNB, sella-nasion-B point angle; ANB, A point-nasion-B point angle; FMA, Frankfort-mandibular plane angle; U1-SN, upper incisor to sella-nasion plane angle; IMPA, incisor mandibular plane angle.

types, and 1 treatment decision. In the primary analysis, the treatment decision was operationalized as a binary outcome variable. Cases without tooth extraction were classified as "non-extraction", and treatment plans involving extraction of upper and/or lower premolars were classified as "extraction". Among the 520 cases in the main analysis dataset, 223 (42.88%) were in the extraction group and 297 (57.12%) were in the non-extraction treatment group (Table 2). Although the class distribution was not perfectly balanced, it was not significantly imbalanced; therefore, in addition to accuracy, balanced accuracy, sensitivity, specificity, precision, and F1-score were used in the performance evaluation to better reflect the potential impact of class imbalance.

All intraoral and extraoral photographs were obtained using a Canon EOS 60D digital camera (Tokyo, Japan), a Sigma 105-mm f/2.8 macro lens (Kawasaki, Japan), and a Godox ring flash (Shenzhen, China). Intraoral photographs (PNG format) were taken from five standard angles (frontal, right and left lateral, upper occlusal, and lower occlusal); extraoral photographs (PNG format) consisted of two images (frontal at rest and profile).

Evaluation of ChatGPT

All analyses were performed by a single researcher using a MacBook Air (MacBook Air, Apple Inc., Cupertino, California; Apple M1 chip, 8-core CPU, 8 GB RAM, 256 GB SSD). The analyses were conducted using ChatGPT-5.0 (OpenAI, San Francisco, CA, USA; October 2024 version), accessed via the web-based interface. To minimize information transfer between sessions, five separate sessions were opened, and the "memory" feature was disabled. Thus, each session was conducted independently, limiting contextual effects that could arise from previous conversations.

Thirty of the 550 cases in the initial dataset were set aside for prompt calibration purposes, and the remaining 520 cases formed the main analysis dataset. The 30 cases reserved for prompt calibration were not included in the final performance evaluation. The dataset split in this way and the schematic of the analysis workflow are shown in Figure 1.

The 520 cases constituting the main analysis dataset were divided into five equal subsets (fold 1 to fold 5), each consisting of 104 cases (Figure 1). A 5-fold cross-validation design was used.[20,30] In each iteration, one subset was used as the test set, while the remaining four subsets were presented to ChatGPT in the same session to provide contextual guidance. Accordingly, in each iteration, 416 cases were used for contextual prompting, and 104 cases were used for testing purposes. At the end of five iterations, the out-of-fold predictions from all folds were combined, and the final statistical analysis was performed on them. This process is summarized in Figure 1. In this study, the parameters of the ChatGPT-5.0 model were not updated, and no retraining or fine-tuning was applied.[31,32]

A standardized prompt template was created to test ChatGPT. To ensure standardization, all photographs were resized to 224×224 pixels. The final prompt used in the study was determined as follows:

"You are an orthodontic assistant. Based on the clinical data, cephalometric measurements, and photographs I provide, evaluate the patient's treatment plan. Decide whether the treatment should be non-extraction or extraction. Do not make any additional comments-only state the treatment decision".

Iterative prompt calibration (IPC) was performed to determine the final prompt. Seven prompts were tested in separate sessions using the 30 cases within a calibration subset which was not included in the overall performance analysis. The initial prompt was: "Based on this data, create the patient's treatment plan; decide whether it will be extraction or non-extraction". The final prompt presented above was selected after the calibration process. Figure 1 schematically shows the prompt calibration stage and its separation from the analysis dataset; it also presents the accuracy rates of the developed prompts. The 30 calibration cases used for prompt calibration were not included in the 5-fold cross-validation process. Therefore, the final performance metrics were calculated solely over the out-of-fold predictions obtained from the main analysis dataset.

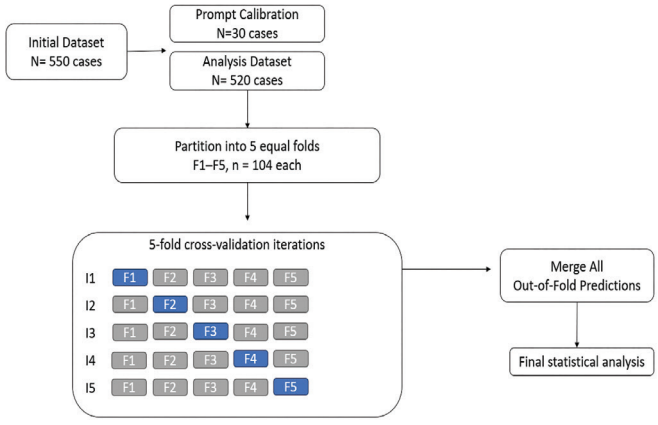


Figure 1. Flow diagram of the study.

Table 2. Demographic characteristics and treatment distribution of the study population

		Extraction type		
		Extraction	Non-extraction	Total
Sex	Female	151 (67.71%)	193 (64.98%)	344 (66.15%)
	Male	72 (32.29%)	104 (35.02%)	176 (33.85%)
Age		14.97	14.46	14.68
Total		**223 (42.88%)**	**297 (57.12%)**	**520**

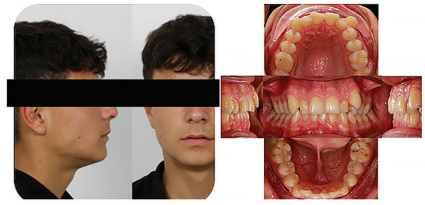
To qualitatively illustrate ChatGPT's decision process, a chain-of-thought (CoT) approach was additionally applied to a single sample case. However, this was not designed as a systematic analysis component and was not included in the overall performance analysis. The prompt given to ChatGPT and the sample reasoning output obtained are summarized in Table 3. The responses provided by ChatGPT for each case were recorded by the primary researcher and independently reviewed and verified by three expert orthodontists. A confusion matrix was created by comparing the actual results with the ChatGPT predictions. Furthermore, final performance metrics were calculated by combining the out-of-fold predictions obtained from each cross-validation iteration (Figure 2).

Evaluation of Machine Learning Algorithms

The dataset includes photographs, clinical data, cephalometric measurements, and treatment decisions of 520 patients (Table 1). Although ML algorithms can process numerical data, they lack the ability to directly analyze images. Therefore, a convolutional neural network (CNN) was used to convert visual data into a numerical format. To digitize visual data, the VGG16 CNN with pre-trained weights on the ImageNet dataset was used. A transfer learning approach was adopted in the training of the model. In this context, the final dense layers of the model responsible for classification were removed, and only the lower layers responsible for feature extraction were preserved. These lower layers were frozen and served as fixed feature extractors,

Table 3. Sample ChatGPT-5.0 output generated using a chain-of-thought (CoT)-optimized prompt

Prompt



You are my orthodontic assistant. Based on the clinical data, cephalometric measurements and photos I provide, evaluate the patient's treatment plan. Decide whether the treatment should be non-extraction or extraction. Do not make any other comments—only state the decision. Then, extract the chain of thought for this process.

CLINICAL VALUES

- Age: 17 years 8 months
- Gender: Male (1)
- Jaw position: Orthognathic (0)
- Molar relationship: Class II (1)
- Overbite (mm): 3
- Overjet (mm): 3.5
- Vertical values: Normal (1)
- Upper Midline: 3 mm to the left (–3)
- Lower Midline: Centered (0)
- Upper Crowding (mm): 10
- Lower Crowding (mm): 4.5

CEPHALOMETRIC VALUES

- SNA (°): 78.9
- SNB (°): 75.7
- ANB (°): 3.1
- FMA (°): 22.1
- U1-SN (°): 97.9
- IMPA (°): 94.7
- Nasolabial Angle (°): 114.5

Response

1. Crowding severity: The patient presents with significant crowding-10 mm in the upper arch and 4.5 mm in the lower-which cannot be resolved with non-extraction mechanics alone.
2. Skeletal Relationship: ANB is 3.1°, suggesting mild skeletal class II, but the jaw position is orthognathic, indicating that the discrepancy can be managed dentally.
3. Incisor inclinations: U1-SN is 97.9° and IMPA is 94.7°, both close to normal, indicating that there is no significant dental compensation; thus, space for alignment and retraction needs to be created.
4. Nasolabial angle: Measured at 114.5°, suggesting a favorable soft tissue profile that can tolerate retraction without causing aesthetic compromise.
5. Midline discrepancy: A 3 mm upper midline shift to the left is noted; bilateral extractions can aid in symmetry correction.
6. Overall alignment strategy: Extraction of all first premolars (14, 24, 34, 44) will provide adequate space for resolving crowding and allow correction of the molar and canine relationships, midline deviation, and incisor inclinations while maintaining soft tissue harmony.

producing 512-dimensional abstract feature vectors from the resized photographs (224×224 pixels). These 512-dimensional vectors obtained from each of the 7 photographs (5 intraoral, 2 extraoral) were concatenated, creating a single combined visual feature vector of 3584 dimensions (7×512) for each patient. The digitized image data were then combined with the patients' clinical and cephalometric data into a single CSV file, creating a numerical dataset suitable for testing ML algorithms. These numerical data were used to predict the binary outcome: non-extraction versus extraction treatment.

The ML algorithms used in this study are classified into different types according to their structural characteristics. As shown in Table 4, multi-layer perceptron (MLP) is a neural-network-based algorithm that processes data through forward and backward propagation. Random Forest is a tree-based method that performs classification by majority voting among numerous decision trees. XGBoost is also tree-based, but relies on a boosting approach and tries to correct previous errors by sequentially training decision trees. Support vector machines (SVM) and logistic regression are mathematical model-based methods. While SVM aims to determine the hyperplane that best separates the data, logistic regression calculates the probability of data points belonging to a specific class. Analyses were conducted on the Jupyter Notebook platform (Python 3.8.5, Anaconda, macOS). While the algorithms were trained with data obtained from 416 patients, 104 patients were reserved for testing. A strict evaluation strategy was followed to minimize the risk of overfitting during model training and to increase model generalizability. To prevent the algorithms from memorizing the training data, 5-fold cross-validation was applied to the dataset of 520 cases. In each iteration, 80% of the data (n=416) were used for training, while the remaining 20% (n=104) that the model had never seen before were reserved as an independent test set. Furthermore, the hyperparameter optimization process (GridSearchCV) of the models was performed solely on the training sets (n=416). The test sets (n=104) were completely isolated from the optimization process, strictly preventing data leakage. Confusion matrices were generated using the combined classification results obtained for each case during the cross-validation process.

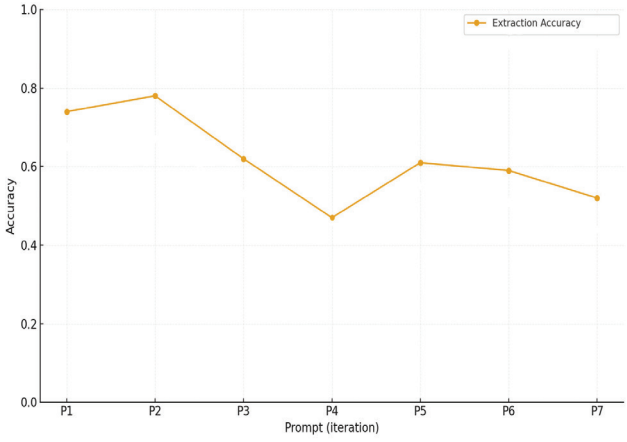


Figure 2. Iterative prompt calibration process and accuracy rates of 7 alternative prompts.

Statistical Analysis

All statistical analyses were performed using IBM SPSS Statistics for Windows (version 26.0; IBM Corp., Armonk, NY, USA). The performance of ChatGPT and ML algorithms in making extraction decisions was evaluated using accuracy, precision, sensitivity (recall), F1-score, balanced accuracy, and specificity (Table 5, Figure 3). 95% confidence intervals (CIs) for these metrics were obtained using the Wilson method.

Since the same cases were classified by all models, inter-model comparisons were made taking the paired data structure into account. Cochran's Q test was used to compare the overall accuracy of the six models. For pairwise comparisons between ChatGPT and each ML model, the McNemar test, which is appropriate for paired binary outcomes, was applied. A Holm-Bonferroni correction was performed to control type I error inflation due to multiple pairwise comparisons. Additionally, to strengthen clinical interpretation, differences between ChatGPT and other models for sensitivity among positive (extraction) cases and specificity among negative (non-extraction) cases were re-evaluated using the McNemar test, and the Holm-Bonferroni correction was applied. In all tests, the statistical significance level was accepted as two-tailed $p \leq 0.05$.

RESULTS

Data obtained from 520 patients (mean age: 14.68 years; standard deviation =2.75) who were included in the evaluation were analyzed (Table 2). Model performance is presented in Figure 4. Among the evaluated models, the highest accuracy was obtained by the XGBoost model (0.7808; 95% CI: 0.7432, 0.8142).

Table 4. The algorithms employed in the current study

Algorithm	Algorithm type	Working mechanism
MLP	Neural network-based	Processes data using artificial neurons; learns through forward and backward propagation.
Random forest	Tree-based	Builds multiple decision trees and classifies through majority voting.
XGBoost	Boosting (tree-based)	Trains decision trees sequentially, correcting previous errors during learning.
SVM	Mathematical model	Finds the hyperplane that best separates the data.
Logistic regression	Mathematical model	Calculates the probability of data points belonging to a class, separates them using the sigmoid function.

MLP, multi-layer perceptron; SVM, support vector machine.

Table 5. Metrics measuring model performance and their specific use or application

Metric	What it measures	When to use it
Accuracy	The overall correctness of the model's predictions (ratio of correct predictions to total predictions).	When the dataset is balanced and all classes are equally important.
Precision	The proportion of true positives out of all predicted positives.	When false positives (e.g., false alarms) need to be minimized.
Sensitivity (recall)	The proportion of true positives out of all actual positives.	When false negatives (e.g., missed detections) need to be minimized.
F1 score	The harmonic mean of precision and recall, balancing the two metrics.	When both precision and recall are equally important, especially for imbalanced datasets.
Balanced accuracy	The average of sensitivity and specificity, useful for imbalanced datasets.	When the dataset is imbalanced and overall performance across classes is needed.
Specificity	The proportion of true negatives out of all actual negatives.	When false positives need to be minimized, especially in cases where detecting negatives correctly is crucial.

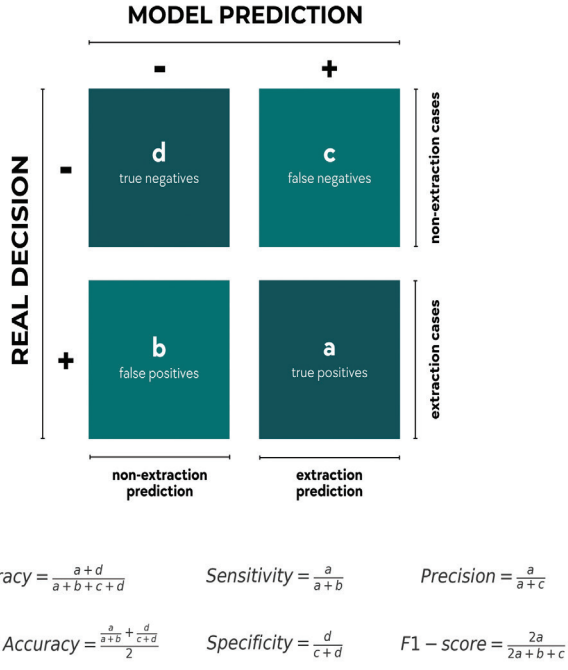


Figure 3. Confusion matrix structure and calculation of metrics for measuring model performance (positive values represent aggregated cases, while negative values represent non-aggregated cases).

ChatGPT ranked second with an accuracy of 0.7577 (95% CI: 0.7191, 0.7925) and outperformed random forest, SVM, logistic regression, and MLP. The lowest accuracy was observed for the MLP model (0.6288).

In terms of sensitivity, the highest value was determined for ChatGPT (0.7668; 95% CI: 0.7071, 0.8175). On the other hand, XGBoost showed the highest performance in terms of specificity, precision, F1-score, and balanced accuracy. When these findings are evaluated together, ChatGPT is seen to have a higher tendency to capture extraction cases, whereas XGBoost exhibits a stronger overall classification balance.

When the correct/incorrect classification patterns of the six models on the same cases were compared with Cochran's Q test, an overall statistically significant difference was found among the models ($Q=47.432$; $p\leq0.001$) (Table 6). This result indicates that the evaluated models do not have equivalent performance.

Pairwise comparisons between ChatGPT and other models were performed using the McNemar test, and the Holm-Bonferroni correction was applied for multiple comparisons. Accordingly, ChatGPT showed a significantly higher accuracy compared to random forest (adjusted $p=0.017$), SVM (adjusted $p=0.010$), logistic regression (adjusted $p\leq0.001$), and MLP (adjusted $p<0.001$) models. In contrast, no significant difference in accuracy was detected between ChatGPT and XGBoost (adjusted $p=0.393$) (Table 6).

In the analyses performed on extraction cases, the sensitivity of ChatGPT was found to be significantly higher than random forest, SVM, logistic regression, and MLP models (all adjusted p-values ≤0.001). No significant difference in sensitivity was detected between XGBoost and ChatGPT (adjusted $p=0.368$; Table 6). This finding supports the conclusion that ChatGPT shows a stronger tendency, especially in detecting extraction cases.

In the specificity analyses performed within non-extraction cases, although some differences were observed in crude comparisons, no statistically significant difference remained in terms of specificity between ChatGPT and other models after Holm-Bonferroni correction (all adjusted $p\geq0.05$) (Table 6). This result indicates that, although XGBoost and some classical ML models have numerically higher specificity values, this superiority was not statistically confirmed after multiple comparisons.

Confusion Matrices

The confusion matrices presented in Figure 5 show the combined classification results obtained for all 520 cases. Examination of the confusion matrices from classification results for 520 cases showed that distributions closest to the ideal pattern were obtained for the XGBoost and ChatGPT models (Figure 5). All models classified non-extraction cases more accurately than they classified extraction cases. The model with the highest number of true positive classifications was ChatGPT (n=171), followed by XGBoost (n=162).

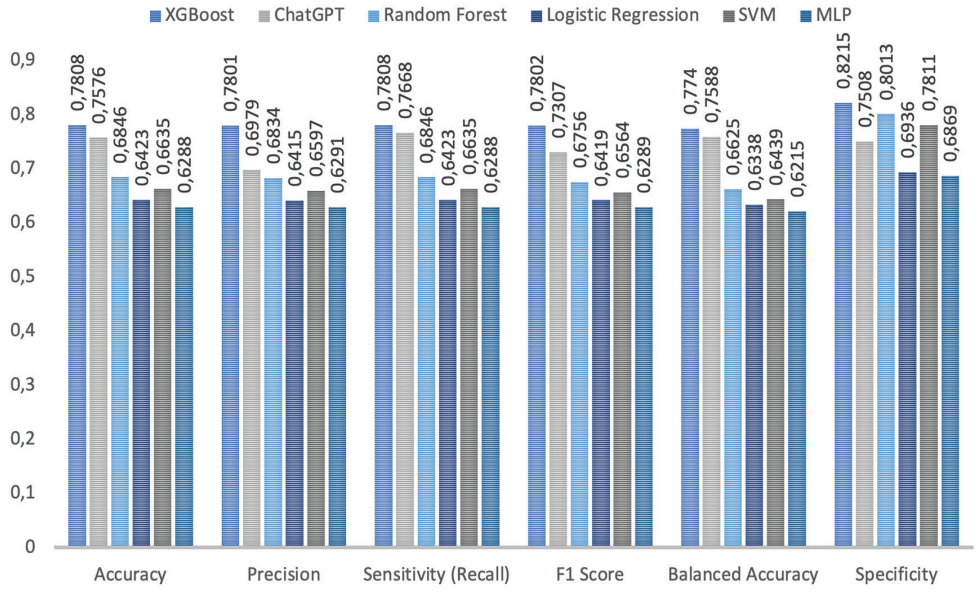


Figure 4. Comparison of metrics used to evaluate model performance in treatment decisions, including extraction and non-extraction outcomes. SVM, support vector machine; MLP, multi-layer perceptron.

Table 6. Pairwise comparison of ChatGPT and machine learning models

Pairwise comparison of ChatGPT and machine learning models in terms of accuracy					
Comparison	**ChatGPT accuracy**	**Comparator model accuracy**	**Unadjusted p-value**	**Holm-adjusted p-value**	**Interpretation**
Cochran's Q test (6 models)	-	-	<0.001	-	A statistically significant overall difference was observed among the models
ChatGPT vs. XGBoost	0.7577	0.7808	0.393	0.393	Not statistically significant
ChatGPT vs. random forest	0.7577	0.6846	0.009	0.017	Statistically significant
ChatGPT vs. SVM	0.7577	0.6731	0.003	0.010	Statistically significant
ChatGPT vs. logistic regression	0.7577	0.6500	<0.001	<0.001	Statistically significant
ChatGPT vs. MLP	0.7577	0.6288	<0.001	<0.001	Statistically significant
Pairwise comparisons of ChatGPT and machine learning models in terms of sensitivity and specificity					
Metric	**Comparison**	**ChatGPT**	**Other model**	**Raw p**	**Holm-adjusted p**
Sensitivity	ChatGPT vs. XGBoost	0.7668	0.7265	0.368	0.368
Sensitivity	ChatGPT vs. random forest	0.7668	0.5067	<0.001	<0.001
Sensitivity	ChatGPT vs. SVM	0.7668	0.5067	<0.001	<0.001
Sensitivity	ChatGPT vs. logistic regression	0.7668	0.5740	<0.001	<0.001
Sensitivity	ChatGPT vs. MLP	0.7668	0.6009	<0.001	<0.001
Specificity	ChatGPT vs. XGBoost	0.7508	0.8215	0.032	0.128
Specificity	ChatGPT vs. random forest	0.7508	0.8182	0.043	0.128
Specificity	ChatGPT vs. SVM	0.7508	0.7980	0.198	0.396
Specificity	ChatGPT vs. logistic regression	0.7508	0.7071	0.250	0.396
Specificity	ChatGPT vs. MLP	0.7508	0.6498	0.012	0.061

p<0.05 was considered statistically significant.
SVM, support vector machine; MLP, multi-layer perceptron.

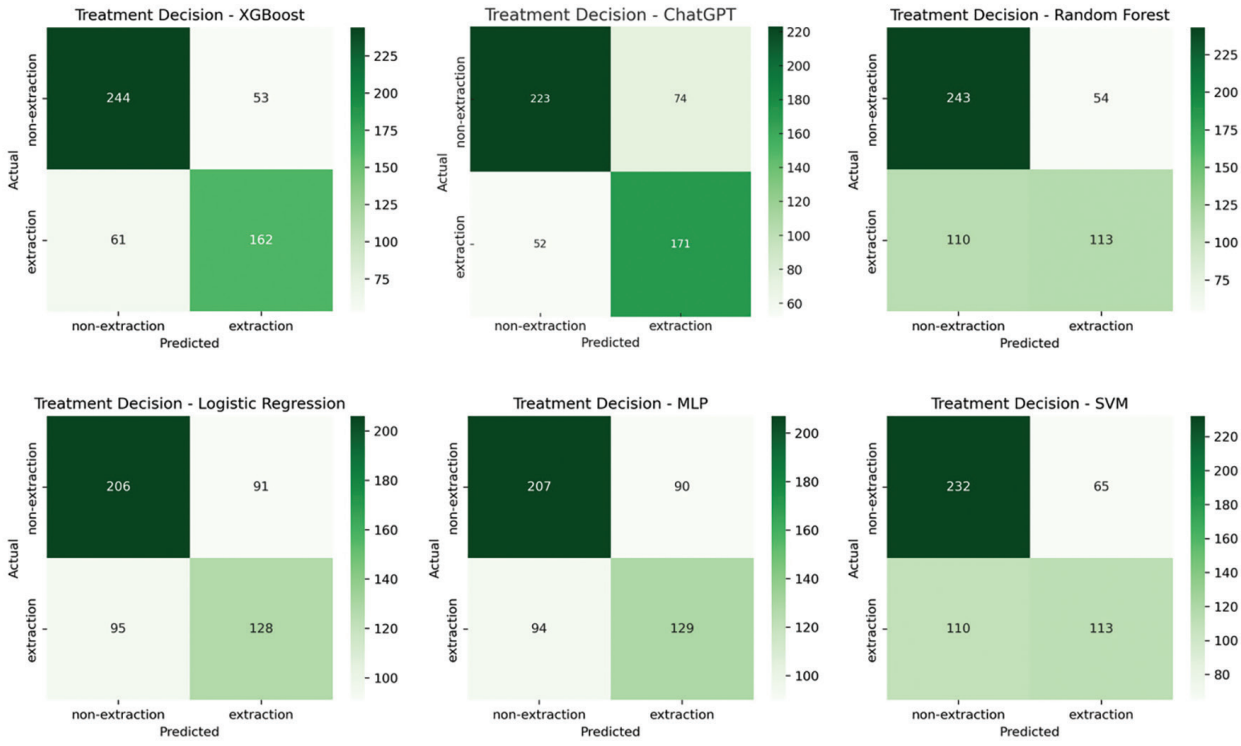


Figure 5. Visualization of model performance in the treatment decision presented as a confusion matrix. MLP, multi-layer perceptron; SVM, support vector machine.

This finding indicates that, although ChatGPT lagged behind XGBoost in overall performance metrics, it exhibited higher sensitivity in identifying extraction cases. In contrast, the number of false-negative cases was highest in the random forest and SVM models; in each model, 110 cases that were actual extractions were classified as non-extraction. In terms of the number of false positive cases, that is, cases that were actually non-extraction but were classified as extraction, the lowest values were observed for the XGBoost (n=53) and random forest (n=54) models.

Feature Importance

Feature importance was evaluated using clinical and cephalometric variables (Figure 6). Across XGBoost, random forest, and logistic regression models, variable contributions were examined using quantitative feature-importance metrics. For ChatGPT, a quantitative feature importance calculation in the classical sense was not performed. Instead, the prominent variables in the explanatory responses accompanying the model's classification decision were evaluated qualitatively. Support vector machine and MLP models were not included in the relevant analysis because they did not provide directly interpretable variable-contribution outputs with the method used in this study.

In ML models, maxillary crowding was identified as one of the most influential variables and was frequently emphasized in ChatGPT's explanatory responses (Figure 6). ChatGPT referred to the nasolabial angle more prominently in its decision justifications, whereas the relative contribution of

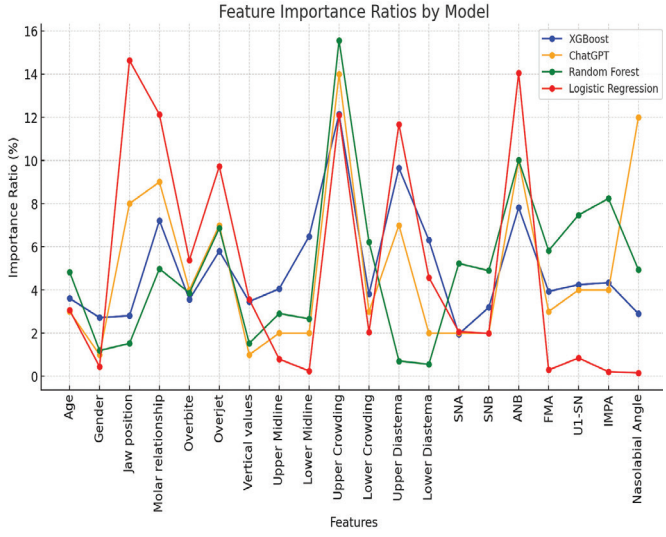


Figure 6. Feature importance percentages for the artificial intelligence models.

this parameter was lower in the XGBoost model. The logistic regression model assigned higher weight to jaw position and ANB angle than other ML models. While the contribution of the diastema variable remained low in the random forest model, it was more pronounced in the XGBoost and logistic regression models. In ChatGPT's explanations, maxillary arch diastemas were identified as particularly noteworthy parameters in the decision-making process.

DISCUSSION

The use of AI in orthodontics has expanded rapidly in recent years. In particular, the accuracy of different algorithms for diagnosis, treatment planning, and cephalometric analysis has been investigated.[11,21,33] While early studies mostly focused on traditional ML models trained for specific tasks,[33] the potential of LLMs such as ChatGPT, Microsoft Copilot, and Gemini in clinical decision processes has recently started to draw attention.[10,13] In this context, the present study aimed to directly compare the performance of ChatGPT-5.0 in distinguishing between extraction and non-extraction treatment decisions with that of traditional ML algorithms. To the best of our knowledge, this study is the first to perform the mentioned comparison on the same sample.

Traditional ML algorithms are designed and optimized for specific tasks.[34] In contrast, LLMs like ChatGPT-5.0 cannot be retrained by the user. Their performance can be guided only by prompt engineering.[31] For this reason, the two approaches handle the clinical problem in fundamentally different ways. General-purpose models such as ChatGPT-5.0, thanks to their multimodal architectures, can simultaneously evaluate different data types, including age, cephalometric measurements, crowding severity, and extraoral photographs, both visually and textually within a clinical context.[35] However, since they are not specialized for learning from numerical data, traditional algorithms are expected to perform better on pure classification tasks. In the study, the parameters of the ChatGPT-5.0 model were not updated, and no retraining or fine-tuning was applied.[31,32] Therefore, the approach used should not be considered model training in the classical sense. Instead, the study should be defined as a large-language-model evaluation design supported by a 5-fold cross-validation performed under many-shot in-context guidance using numerous examples, conducted after prompt calibration.[36] Consequently, this method is not part of the standard zero-shot or few-shot evaluation paradigm but rather of a contextually guided, cross-validated evaluation approach.

The findings of our study support this framework. While XGBoost achieved the highest performance with 78.08% accuracy, ChatGPT-5.0 ranked second with 75.77% accuracy. The performance ranking being XGBoost ≥ ChatGPT > Random Forest > SVM > Logistic Regression > MLP is consistent with the large-scale study conducted by Prasad et al.[28] and the models reaching the highest accuracy rates being XGBoost, Random Forest, and Decision Tree-based algorithms. Similarly, in the study of Köktürk et al.,[37] the most successful predictions in the orthodontic extraction decision were obtained with Gradient Boosted Trees (the algorithm family forming the basis of XGBoost), SVM, and random forest models. Considering the complex structure of clinical, cephalometric, and demographic variables in our dataset, it is an expected result and is consistent with the literature that algorithms that optimize decision trees, such as XGBoost, show superior performance compared with neural networks, such as the traditional (MLP; 62.88% accuracy). However, ChatGPT-5.0 demonstrated performance comparable to traditional models. In particular, the model, which exhibits higher sensitivity than many algorithms, suggests that it has a strong tendency to detect extraction cases. That ChatGPT-5.0 obtained its highest score in sensitivity (0.7668) and its lowest score in precision (0.6979) indicates that the model is more aggressive in detecting extraction cases but has a higher margin of error. Although ChatGPT-5.0 appears promising as a tool to support clinical decision-making, ask-specific models trained with expert consensus are more reliable for precise treatment decisions.

The current findings should be interpreted in the context of previous studies on the use of AI in orthodontics. For example, Kunz et al. [38] reported an accuracy of over 90% using CNNs in cephalometric landmark detection. Similarly, studies evaluating extraction/non-extraction treatment planning with CNN architectures such as ResNet and VGG have reported success rates varying between 70% and 90%.[33] When evaluated from this perspective, the performance demonstrated by ChatGPT-5.0 in this study falls within the reported success range of traditional algorithms and is at an acceptable level.

That orthodontic planning is inherently subjective raises the question of how the "ground truth" will be defined in such studies. While approaches based on a single expert opinion were used by Jung and Kim,[27] designs that brought together the opinions of multiple experts were reported by Prasad et al.[28] and Mason et al.[39] While the single expert approach increases the risk of individual bias, the inclusion of multiple experts can decrease data consistency. The consensus method used in this study by three expert orthodontists offered a more balanced solution between the two extreme approaches.

The data distribution in our study appears consistent with clinical reality. Of the 520 patients, 223 (42.88%) were in the extraction treatment group and 297 (57.12%) were in the non-extraction treatment group. Although examples balancing the case distribution as 50% to 50% have been reported by Köktürk et al.[37] and Mason et al.,[39] there are also studies preserving natural clinical distributions where non-extraction cases reach 75%. The distribution in our study reflects the slight class imbalance encountered in clinical practice. Indeed, Etemad et al.[40] emphasized that this imbalance is not a data flaw, but a natural consequence of the conservative approach adopted by clinicians in borderline cases.

When the performance profiles of the models were examined, the vast majority achieved their highest scores for specificity. This finding is consistent with the low sensitivity (0.29, 0.53) and high specificity (0.94, 0.96) results reported by Etemad et al.[40] In other words, the models appear to be more successful in identifying non-extraction cases.

With respect to the variables affecting the extraction decision, our findings are consistent with the literature.[37,39,40] Mason et al.[39] reported that crowding, facial profile, and lower incisor

inclination are among the most influential parameters in the extraction/non-extraction treatment decision. In our study, maxillary crowding also emerged as one of the most important variables in traditional ML models in which quantitative feature importance was calculated. This result is consistent with the literature, which indicates that the discrepancy between available and required arch lengths is one of the main determinants in extraction decision-making. In contrast, the greater prominence of variables, such as the nasolabial angle and upper-arch diastemas, in ChatGPT-5.0's explanatory responses suggests that the model processes clinical information on a different representational plane.

The difference in explainability between traditional ML models and LLMs should not be ignored when interpreting these findings. In models such as XGBoost, feature importance is a quantitative metric that indicates each feature's contribution to model predictions. Conversely, LLMs, such as ChatGPT, lack a similarly standardized method for calculating statistical feature importance. By their nature, these models are black-box systems and their decision mechanisms are not fully transparent.[41] In the literature, to partially mitigate this opacity, prompting strategies such as CoT are used, in which the model is asked to generate step-by-step explanations.[42,43] However, these explanations are not quantitative evidence that directly reflects the internal classification mechanism of the model; rather, they are textual justifications generated post-decision. Therefore, it is not correct to evaluate the deterministic feature importance values obtained from ML models and ChatGPT's more frequent references to certain parameters on the same methodological grounds. ChatGPT's frequent reference to certain parameters should be interpreted not as a finding equivalent to quantitative feature importance but as an post-hoc explanatory output generated after the decision. For this reason, the importance of correct prompt selection to ensure that generated responses meet the same standard has been reported in the literature, and the IPC method has been proposed.[44,45] Iterative prompt calibration can guide models that tend to generate open-ended, variable responses to produce more structured and consistent outputs.[15,43,45] This approach is particularly important for reducing the risk of hallucination.[46] However, it is necessary to conduct calibration processes on small test subsets that are independent of the main dataset to prevent data leakage and to maintain scientific validity.[45] In our study, an additional dataset of 30 patients was used for IPC, and the most appropriate prompt was obtained using the aforementioned method.

The findings of the present study provide insights into how LLMs such as ChatGPT-5.0 and ML algorithms such as XGBoost can be integrated into routine orthodontic decision-making. These models can be particularly useful in two clinical scenarios. First, these algorithms can be used as a second opinion and verification tool for orthodontic students and physicians with limited clinical experience. In situations where complex, multifactorial data (age, degree of crowding, cephalometric angles) need to be evaluated simultaneously, AI systems can provide a standardized decision-making process by systematically filtering clinical parameters that clinicians might otherwise overlook. Second, especially in borderline cases when the physician is undecided between extraction and non-extraction treatment, the predictions offered by these models can serve as a clinical prompt, triggering are-evaluation of treatment risks. In our study, ChatGPT-5.0's relatively higher sensitivity in detecting extraction cases suggests that the model might have a more aggressive tendency when making such risky decisions and might help physicians critically question their own decisions. Consequently, the aim of these systems is not to replace the physician but to create an evidence-based, reproducible hybrid decision-support system that minimizes clinical oversights as patient loads increase.

Study Limitations

Certain limitations should be considered when evaluating the findings obtained from this study. Since this is, to the best of our knowledge, one of the first studies to evaluate ChatGPT-5.0's performance in orthodontic extraction versus non-extraction treatment decisions, the direct comparability of the findings with previous literature is limited. Data obtained from a single center may limit the generalizability of the treatment protocols. Although the local, isolated system structure used in the study was chosen for data security, it does not completely eliminate the risk of unauthorized access or other systemic vulnerabilities. For ChatGPT-5.0, intra-model consistency between sessions or across time was not directly evaluated. To reduce variability, all evaluations were conducted on the same day by a single researcher using a standardized prompt structure; a new session was started after each case. However, this approach is a methodological standardization measure, not a direct consistency analysis, and it does not completely eliminate the risk of user-dependent bias. Furthermore, ChatGPT-5.0's image-processing capacity is still limited compared to that of traditional ML algorithms, and the performance of NLP-based models is sensitive to prompt quality; therefore, the obtained results reflect ChatGPT-5.0's capabilities at the time the research was conducted. Despite this, by comparing ChatGPT-5.0 with traditional ML models using the same sample, this study makes an original contribution to the existing literature. Additionally, it provides a methodological framework by demonstrating that ChatGPT-5.0's explanatory outputs are not directly equivalent to quantitative feature importance metrics in traditional models. Future studies should focus on improving LLMs with multimodal training approaches, evaluating clinician-LLM hybrid decision support systems, and testing these models with longitudinal clinical validation designs. In this respect, the study constitutes a starting point for future research involving misclassification patterns, subgroup performance, and more detailed explanatory analyses.

CONCLUSION

This study compared the performance of ChatGPT-5.0's with that of traditional ML algorithms in distinguishing between orthodontic extraction and non-extraction treatment decisions. The findings showed a statistically significant difference in performance when all models were evaluated together. Therefore, the null hypothesis (H0), which predicted no difference between the LLM-based approach and ML algorithms in classification performance, was rejected, and the alternative hypothesis (H1), which predicted a statistically significant difference between the two approaches, was accepted.

However, the fact that ChatGPT-5.0 did not show a statistically significant difference in accuracy compared to XGBoost in pairwise comparisons, while it exhibited significantly higher performance than random forest, SVM, logistic regression, and MLP models, indicates that this result is not consistent across models. Although ChatGPT-5.0 performed better than or differently from some traditional ML models, it yielded results similar to those of the highest-performing model, XGBoost.

These findings indicate that ChatGPT-5.0 has notable potential as a clinical decision support tool in orthodontics; however, some task-oriented optimized ML algorithms can maintain their superiority, especially in terms of overall classification balance and specificity.

Ethics

Ethics Committee Approval: This study was approved by the Institutional Afyonkarahisar Health Sciences University Non-Intervention Scientific Research Ethics Committee (approval no: 2024/8, date: 04.10.2024).

Informed Consent: Written and verbal informed consent was obtained from all patients and their legal guardians.

Footnotes

Author Contributions: Concept - A.Y.; Design - A.Y., H.C.; Data Collection and/or Processing - M.S.; Analysis and/or Interpretation - H.C., M.S.; Literature Search - A.Y., M.S., H.C.; Writing - A.Y., H.C., M.S.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Camcı H, Salmanpour F. Comparing the esthetic impact of virtual mandibular advancement, bichectomy, jawline, and their combination. *Am J Orthod Dentofacial Orthop.* 2023;163(6):756-765. [CrossRef]
2. Camcı H, Salmanpour F. Impact of intraoral scanning conditions on the accuracy virtual aligners (VA). *Australasian Orthodontic Journal.* 2022;38(1):102-110. [CrossRef]
3. Salmanpour F, Camcı H. Prediction of patient cooperation before orthodontic treatment: handwriting and artificial intelligence. *J World Fed Orthod.* 2024;13(6):303-309. [CrossRef]
4. Ahmed N, Abbasi MS, Zuberi F, et al. Artificial intelligence techniques: analysis, application, and outcome in dentistry-a systematic review. *Biomed Res Int.* 2021;2021:9751564. [CrossRef]
5. Salmanpour F, Camci H. Artificial intelligence for predicting orthodontic patient cooperation: voice records versus frontal photographs. *APOS Trends Orthod.* 2024;14:255-263. [CrossRef]
6. Radwan MT, Sin Ç, Akkaya N, Vahdettin L. Artificial intelligence-based algorithm for cervical vertebrae maturation stage assessment. *Orthod Craniofac Res.* 2023;26(3):349-355. [CrossRef]
7. Camcı H, Salmanpour F. Estimating the size of unerupted teeth: moyers vs deep learning. *Am J Orthod Dentofacial Orthop.* 2022;161(3):451-456. [CrossRef]
8. Fawaz P, Sayegh PE, Vannet BV. What is the current state of artificial intelligence applications in dentistry and orthodontics? *J Stomatol Oral Maxillofac Surg.* 2023;124(5):101524. [CrossRef]
9. Dipalma G, Inchingolo AD, Inchingolo AM, et al. Artificial intelligence and its clinical applications in orthodontics: a systematic review. *Diagnostics (Basel).* 2023;13(24):3677. [CrossRef]
10. Salmanpour F, Camcı H, Geniş Ö. Comparative analysis of AI chatbot (ChatGPT-4.0 and microsoft copilot) and expert responses to common orthodontic questions: patient and orthodontist evaluations. *BMC Oral Health.* 2025;25(1):896. [CrossRef]
11. Hakami Z, Saheb SAK, Bawazeer OA. Orthodontic knowledge assessment: a comparison of five AI chatbots. *Saudi Dent J.* 2026;38(3):20. [CrossRef]
12. Zheng J, Ding X, Pu JJ, et al. Unlocking the potentials of large language models in orthodontics: a scoping review. *Bioengineering.* 2024;11(11):1145. [CrossRef]
13. Çelik İH, Camcı H, Salmanpour F. Bridging the information gap in pediatric dentistry: a comparison of ChatGPT-4o, Google Gemini Advanced, and expert responses based on evaluations by parents and pediatric dentists. Journal of Clinical Pediatric Dentistry. 2026. 50(1);147-155. [CrossRef]
14. Ozkan E, Koyun M. Atlas-assisted bone age estimation from hand-wrist radiographs using multimodal large language models: a comparative study. *Diagnostics.* 2026;16(3):487. [CrossRef]
15. Akpınar M, Salmanpour F. Chat generative pretrained transformer-4.0's accuracy in assessing cervical vertebrae and hand-wrist maturation stages: a retrospective study. *Am J Orthod Dentofacial Orthop.* 2025;168(6):753-763. [CrossRef]
16. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. *J Med Internet Res.* 2023;25:e51580. [CrossRef]
17. Morishita M, Fukuda H, Muraoka K, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: a challenge explored. *J Dent Sci.* 2024;19(3):1595-1600. [CrossRef]
18. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature.* 2023;616(7956):259-265. [CrossRef]
19. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172-180. Erratum in: *Nature.* 2023;620(7973):E19. [CrossRef]
20. Cicek O, Özçelik YB, Altan A. A new approach based on metaheuristic optimization using chaotic functional connectivity matrices and fractal dimension analysis for AI-driven detection of orthodontic growth and development stage. *Fractal and Fractional.* 2025;9(3):148. [CrossRef]
21. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric analysis in orthodontics using artificial intelligence-a comprehensive review. *Biomed Res Int.* 2022;2022:1880113. [CrossRef]

22. Yıldırım A, Cicek O, Genç YS. Can AI-based chatgpt models accurately analyze hand-wrist radiographs? A comparative study. *Diagnostics (Basel)*. 2025;15(12):1513. [CrossRef]
23. Farzanegan P, Zarabadi M, Najary S, Tahmasbi MA, Behnaz M. The role of artificial intelligence in orthodontics for determining skeletal age based on cervical vertebra maturation degree: a comprehensive review. *Health Sci Rep*. 2025;8(11):e71487. [CrossRef]
24. Suhail Y, Upadhyay M, Chhibber A, Kshitiz. Machine learning for the diagnosis of orthodontic extractions: a computational analysis using ensemble learning. *Bioengineering (Basel)*. 2020;7(2):55. [CrossRef]
25. Choi HI, Jung SK, Baek SH, et al. Artificial intelligent model with neural network machine learning for the diagnosis of orthognathic surgery. *J Craniofac Surg*. 2019;30(7):1986-1989. Erratum in: J Craniofac Surg. 2020;31(4):1156. [CrossRef]
26. Ganguly R, Suri L, Patel F. A literature review of t extraction decision and outcomes in orthodontic treatment. *J Mass Dent Soc*. 2016;65(2):28-31. [CrossRef]
27. 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. [CrossRef]
28. Prasad J, Mallikarjunaiah DR, Shetty A, Gandedkar N, Chikkamuniswamy AB, Shivashankar PC. Machine learning predictive model as clinical decision support system in orthodontic treatment planning. *Dent J (Basel)*. 2022;11(1):1. [CrossRef]
29. Kılınç DD, Mansız D. Examination of the reliability and readability of chatbot generative pretrained transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am J Orthod Dentofacial Orthop*. 2024;165(5):546-555. [CrossRef]
30. Kök H, Acilar AM, İzgi MS. Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. *Prog Orthod*. 2019;20(1):41. [CrossRef]
31. Abuabara A, do Nascimento TVPM, Trentini SM, et al. Evaluating the accuracy of generative artificial intelligence models in dental age estimation based on the Demirjian's method. *Front Dent Med*. 2025;6:1634006. [CrossRef]
32. Çelik İH, Salmanpour F. Can ChatGPT-5 estimate dental age? A comparative study with Demirjian and Willems methods in paediatric dentistry. *BMC Oral Health*. 2026;26(1):849. [CrossRef]
33. 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.e4. [CrossRef]
34. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - a systematic review. *J Dent Sci*. 2021;16(1):482-492. [CrossRef]
35. OpenAI. Karşınızda GPT-5. [CrossRef]
36. Vassis S, Powell H, Petersen E, et al. Large-language models in orthodontics: assessing reliability and validity of ChatGPT in pretreatment patient education. *Cureus*. 2024;16(8):e68085. [CrossRef]
37. 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(Suppl 2):13-24. [CrossRef]
38. Kunz F, Stellzig-Eisenhauer A, Zeman F, et al. Artificial intelligence in orthodontics. *J Orofac Orthop*. 2010;81:52-68. [CrossRef]
39. 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. [CrossRef]
40. Etemad LE, Heiner JP, Amin AA, et al. Effectiveness of machine learning in predicting orthodontic tooth extractions: a multi-institutional study. *Bioengineering (Basel)*. 2024;11(9):888. [CrossRef]
41. Nordblom NF, Büttner M, Schwendicke F. Artificial intelligence in orthodontics: critical review. *J Dent Res*. 2024;103(6):577-584. [CrossRef]
42. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst*. 2022;35. [CrossRef]
43. Salmanpour F, Akpınar M. Performance of chat generative pretrained transformer-4.0 in determining labiolingual localization of maxillary impacted canine and presence of resorption in incisors through panoramic radiographs: a retrospective study. *American Journal of Orthodontics and Dentofacial Orthopedics*. 2025;168(2):220-231. [CrossRef]
44. Jha S, Jha SK, Lincoln P, Bastian ND, Velasquez A, Neema S. dehallucinating large language models using formal methods guided iterative prompting. Proceedings - 2023 IEEE International Conference on Assured Autonomy ICAA. 2023:149-152. [CrossRef]
45. Wang J, Sun Y, Liang Y, Li X, Gong B. Iteratively calibrating prompts for unsupervised diverse opinion summarization. *Frontiers in Artificial Intelligence and Applications*. 2024;392:3939-3946. [CrossRef]
46. Zhao Z, Wang B, Ouyang L, Dong X, Wang J, He C. Beyond hallucinations: enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv*. 2023. [CrossRef]