oral

MDPI

Article

Utilization of Machine Learning Methods for Predicting Orthodontic Treatment Length

Mohammed H. Elnagar[1,2], Allen Y. Pan[3], Aryo Handono[4], Flavio Sanchez[1], Sameh Talaat[4,5], Christoph Bourauel[5], Ahmed Kaboudan[6], and Budi Kusnoto[1]

[1] Department of Orthodontics, College of Dentistry, University of Illinois Chicago, Chicago, IL, USA.
[2] Department of Orthodontics,Tanta University, Tanta, Egypt.
[3] Midwestern University, Downers Grove, IL, USA. Master student information management system, Harvard University, Cambridge, MA, USA.
[4] Master student Telkom University, Bandung, West Java, Indonesia.
[4] Department of Orthodontics, Faculty of Dental Medicine, Future University in Egypt, Cairo, Egypt.
[5] Department of Oral Technology, University of Bonn, Bonn, Germany.
[6] DigiBrain4 Inc., Chicago, IL, USA.

* Correspondence: Mohammed H. Elnagar DDS, MSc, PhD, Department of Orthodontics (M/C 841), College of Dentistry, University of Illinois Chicago, 801 S. Paulina Street, RM 131, Chicago, IL 60612-7211; E-mail: melnagar@uic.edu

ABSTRACT

Treatment duration is one of the most important factors that patients consider when deciding whether to have orthodontic treatment or not. This study aimed to build and compare Machine Learning (ML) models for prediction of orthodontic treatment length and to identify factors affecting the duration of orthodontic treatment using the ML approach. Records of 518 patients who successfully finished orthodontic treatment were used in this study. 70% of the patient data was used for training ML models, and 30% of data was used for testing these models. We applied and compared nine machine-learning algorithms: Simple Linear Regression, Modified Simple Linear Regression, Polynomial Linear Regression, K Nearest Neighbor, Simple Decision Tree, Bagging Regressor, Random Forest, Gradient Boosting Regression, and AdaBoost Regression. We then calculated the importance of patient data features for the ML models with the highest performance. The best overall performance was obtained through Bagging Regressor and AdaBoost Regression ML methods. The most important features in predicting treatment length were age, crowding, artificial intelligence case difficulty score, overjet, and overbite. Without patient information, several ML algorithms showed comparable performance for predicting treatment length. Bagging and AdaBoost showed the best performance when patient information, including age, malocclusion, and crowding, was provided.

Keywords: Artificial Intelligence; Machine Learning; Orthodontic Treatment Length

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Oral* **2022**, *2*, Firstpage–Lastpage. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Treatment duration is one of the most important factors that patients consider when deciding whether to have orthodontic treatment [1]. An exact and accurate prediction of the duration of the total orthodontic treatment might motivate patients or prepare them for what to expect (Mavreas and Athanasiou, 2008) [2]. Additionally, a reliable idea of the treatment duration helps the orthodontist to better plan the overall treatment and the sequence of appointments (Fink and Smith, 1992; Mavreas and Athanasiou, 2008) [1,2]. Earlier studies reported that orthodontic treatment employing fixed appliances typically lasts 14 to 33 months (Kafle et al., 2019; Tsichlaki et al., 2016) [3,4] with a mean around 22 to 24 months, depending on the discrepancy being treated (Aljehani and Baeshen, 2018; Simister, 2007) [5].

Factors influencing the duration of orthodontic treatment are manifold. They range from general malocclusion, anatomic/biologic factors (like bone morphology, patient age, and disease), the type of treatment (extraction versus non-extraction), to the planned treatment technique (Bhikoo et al., 2018) [6]. A further aspect might be patient cooperation, which is closely related to socio-economic factors and education (Kafle et al., 2019; Mavreas and Athanasiou, 2008; Tsichlaki et al., 2016) [2–4].

As teeth have to be moved through the bone, one decisive factor influencing the speed of orthodontic tooth movement and thus treatment duration is bone metabolism, i.e. the ability of bone to remodel as a result of the applied force systems (Abbing et al., 2020) [7]. Bone metabolism depends, in part, on age, the bony structure itself, and/or systemic disease (Abbing et al., 2020; Kaur and El-Bialy, 2020; Landin-Ramos, 2020) [7–9]. One could approach the prediction of treatment duration via bone morphology. Here, the bone structure and density, the thickness of the cortical bone, and the structure of the spongious bone would have to be analyzed in detail. An approach using fractal analysis of panoramic x-ray images has recently been presented (Cesur et al., 2020) [10], while more classical approaches use indices of severity, such as The American Board of Orthodontics Discrepancy Index (ABO-DI), to give an answer to patients' frequent question, "When do I get my braces off?" (Aljehani and Baeshen, 2018) [5].

Artificial intelligence (AI) is bringing a paradigm shift to healthcare, powered by the increasing availability of healthcare data and the rapid progress of analytics techniques [11]. Machine learning (ML) is a subset of AI techniques, used to determine complex models and extract knowledge. In clinical practice, ML predictive models can assist the clinician in decision-making regarding individual patient care [12,13].

To our knowledge, ML has not been used to predict orthodontic treatment length. Therefore, our study aimed to build and compare ML models to predict orthodontic treatment length and to identify factors affecting the duration of orthodontic treatment using a ML approach.

2. Materials and Methods

We retrospectively evaluated the records of 631 patients who completed orthodontic treatment at All Care Orthodontics, Chicago, IL. Ethical approval (IRP Number 20193360) for this study was obtained from the research ethics committee of WIRB-Copernicus. All experiments were done in accordance with approved guidelines.

The inclusion criteria were as follows: Patients who had 1) comprehensive orthodontic treatment; 2) successfully finished their orthodontic treatment without disruption during the treatment period; 3) a complete set of standard orthodontic records pre-treatment and at de-bond appointment; 4) treatment by a board-certified orthodontist. The exclusion criteria were patients who had: 1) limited orthodontic treatment; 2) phase 1 orthodontic treatment; 3) treatment disrupted and, consequently, increased treatment length; 4) more than four failed appointments; 5) treatment under Medicaid coverage; and 6) craniofacial syndromes. A total of 518 patients met the inclusion and exclusion criteria, and their records were used in this study.

The following parameters were collected for each patient: 1) Gender, race, and age when treatment started; 2) commute distance to the orthodontic office in miles; 3) overjet, overbite, maxillary and mandibular arch crowding calculated in mm; 4) malocclusion classification (I, II and III); 5) actual treatment length, in months, starting from bonding to debonding appointment; 6) estimated treatment length determined by an orthodontist; 7) treatment difficulty estimated by artificial intelligence (AI Score: 1, Easy to 5, Very Difficult) using a deep learning model, previously published by Talaat et al. 2021[13].

Implementation of Machine Learning Models

A total of nine machine learning algorithms were tested. These included: 1) Simple Linear Regression (Baseline Model); 2) Modified Simple Linear Regression; 3) Polynomial

Linear Regression; 4) K Nearest Neighbor (KNN); 5) Simple Decision Tree; 6) Bagging Regressor; 7) Random Forest; 8) Gradient Boosting Regression; and 9) AdaBoost Regression[14].

The cases corresponding to each of the possible outcomes were divided into two groups: 70% of cases were used for ML training and the remaining 30% for ML testing. The same training and testing sets were used with every model to ensure fair comparison. After each model was trained and optimized using 70% of the patient sample, the remaining 30% of cases served as the testing dataset to evaluate the model's predictive ability. We compared all models using three indicators: mean squared error (MSE) on training data, MSE on testing data, and coefficient of determination (R2) of the model on the entire dataset. Ideally, the testing MSE should be as low as possible. A training MSE that is much lower than testing MSE usually indicates model overfitting on the training dataset. In addition, a higher R2 score is desirable, representing the proportion of the variance for the dependent variable (actual treatment time) that's explained by independent variables in a regression model. Furthermore, we analyzed residual values according to the statistical best practices and generated feature Importance and Permutation Importance for each model.

3. Results

This study used data from 518 patients, 281 females and 237 males. The mean patient age was 17.49 +/- 8.15 years, and the mean patient treatment time was 26.10 +/- 8.15 months; the mean crowding was 3.18 +/- 3.64 mm for the maxillary arch and 2.79 +/- 3.56 mm for the mandibular arch (negative crowding represents spacing); Class I malocclusion was present in 299 cases, Class II in 145, and Class III in 74. The mean treatment difficulty estimated by AI Score = 2.53 +/- 0.81. The mean patient commute distance to the orthodontic office was = 3.44 +/- 4.979 miles (Table 1) (Figures 1 and 2).

Table 1. Description of Patient Demographic Data

N = 518	Actual treatment time (months)	Overjet (mm)	Overbite (mm)	Maxillary crowding (mm)	Mandibular crowding (mm)	AI score	Patient age (years)	Distance to treatment office (miles)
Mean	26.101	2.49	2.844	3.178	2.792	2.527	17.49	3.445
STD	8.146	2.699	1.752	3.644	3.56	0.808	8.15	4.979
Min	2.6	-12	-6	-10	-16	2	8.48	1.06
25%	20	1	2	1	1	2	12.59	1.06
50%	25.6	2	3	3	3	2	14.23	1.41
75%	31.575	4	4	5	5	3	19.977	3.77
Max	47.8	14	8	18	15	5	62.12	34.11

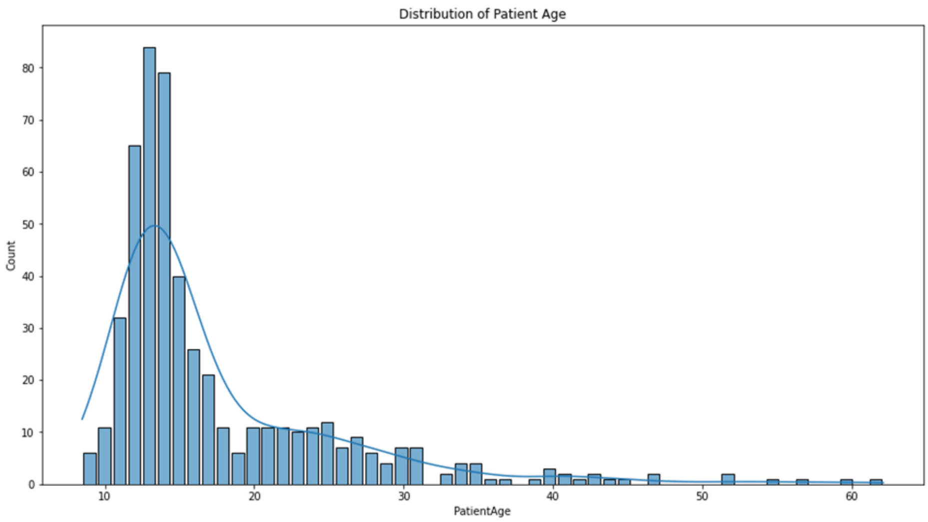

Figure 1. Histogram of patient age distribution.

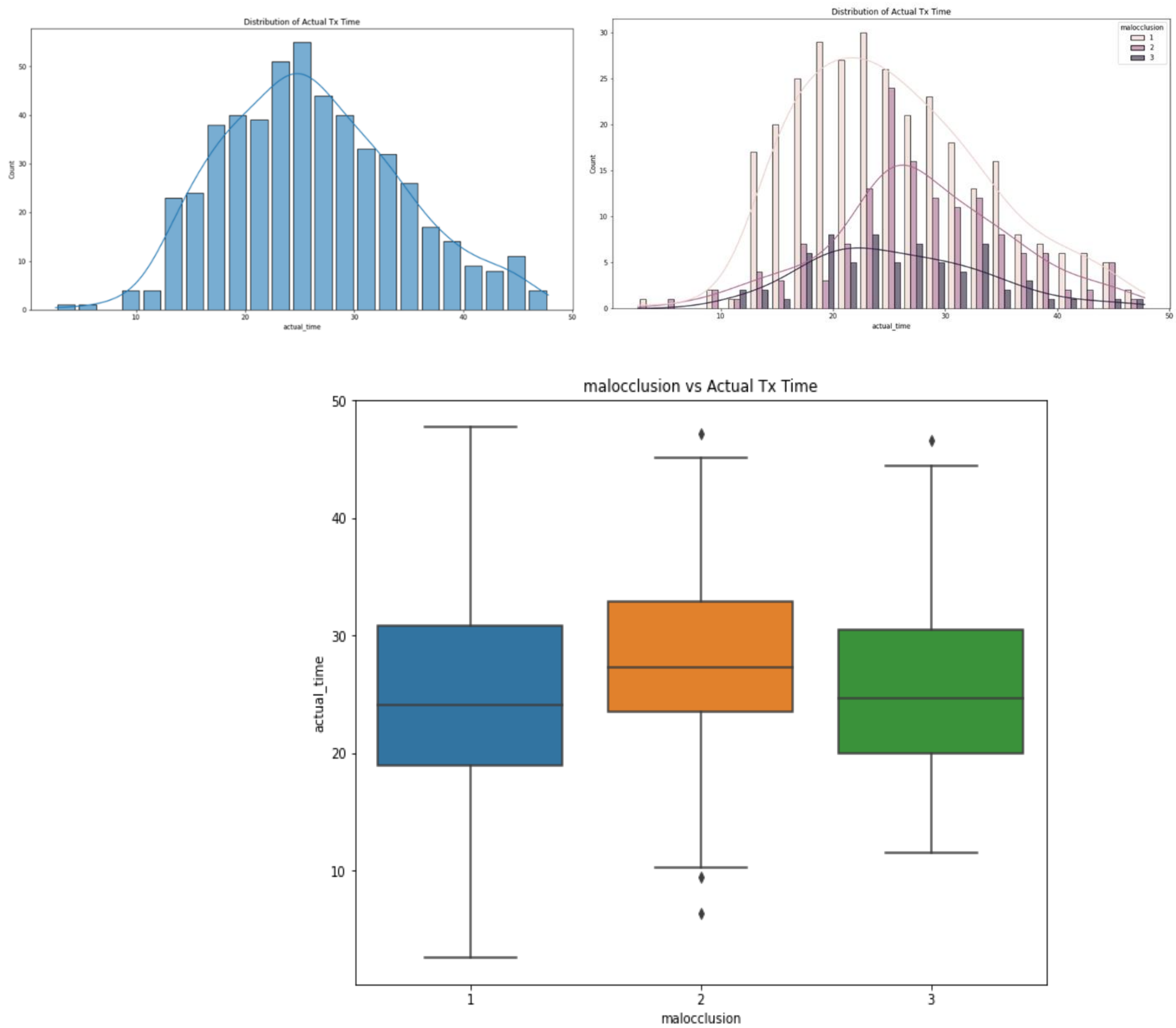

Figure 2. Histograms showing (a) actual treatment time distribution and (b) actual treatment time based on malocclusion Class. (c) Boxplot showing malocclusion versus actual treatment time.

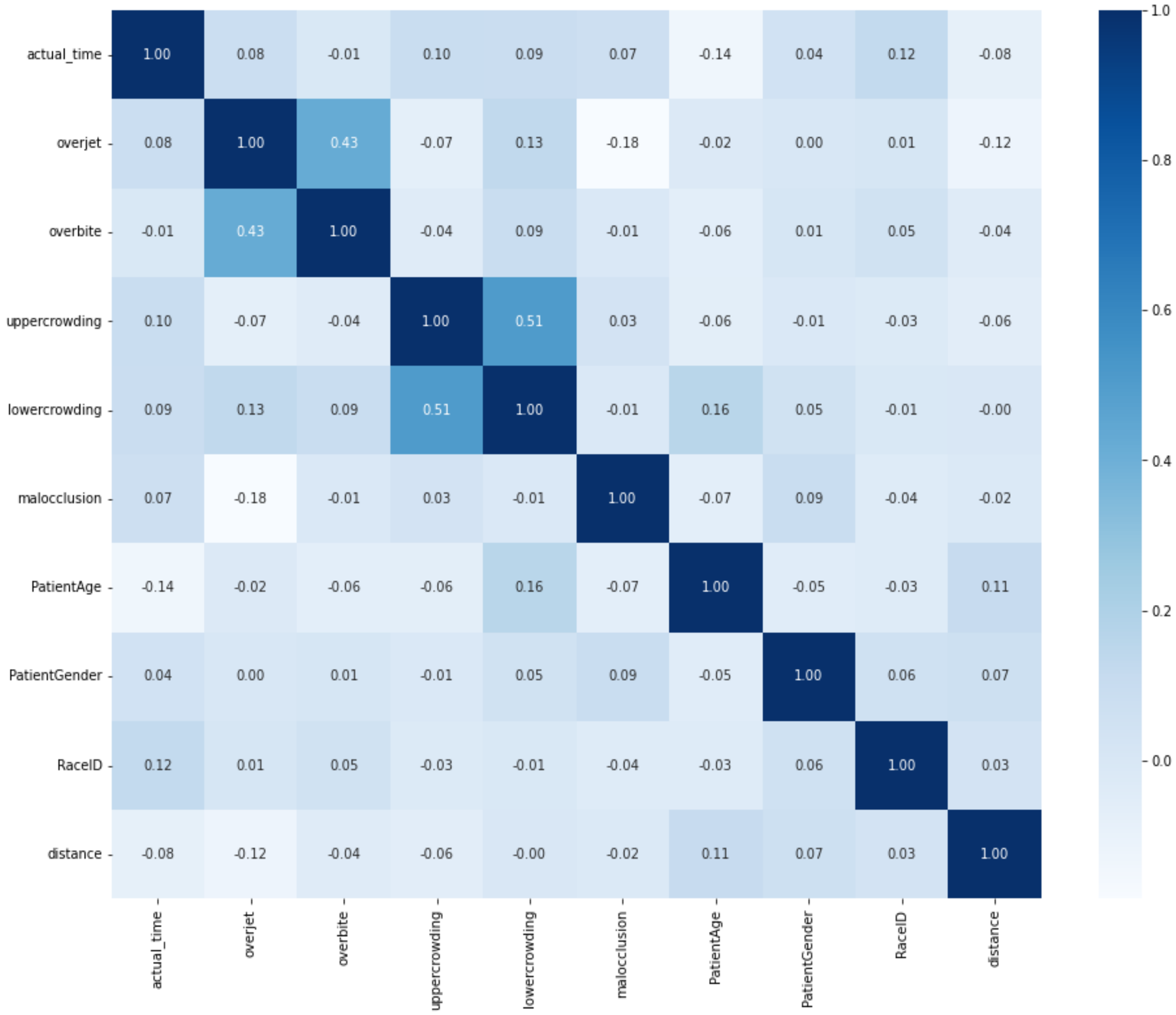

Figure 3. Heat map showing the correlation between variables.

The correlation between variables shown in Figure 3 reveals that the overbite and overjet values were highly correlated (0.43). In addition, both maxillary and mandibular crowding values are highly correlated (0.51). All other pairs did not show significant correlations.

Different ML models behave differently when processing the inputs. Accordingly, the performance of these models also varies. For the ML algorithms evaluated, the following was observed: Bagging and AdaBoost are the best models, with much lower MSE values for both training and testing datasets and a higher R2 score to explain the variances (Table 2) (Figure 4 and 5) .

Table 2. Performance Comparison of the ML Models

ML Model	Training MSE	Testing MSE	R2 Score
Simple Linear Regression	59.65	66.76	0.067
Modified Simple Linear Model	58.85	65.21	0.082
Polynomial Linear Regression	48.85	79.25	0.124
KNN (best k = 9)	79.40	81.25	-0.266
Decision Tree (w/AI score)	51.99	71.97	0.124

Decision Tree (w/o AI score)	55.65	58.20	0.148
Bagging (w/AI score)	40.86	60.95	0.308
Bagging (w/o AI score)	43.08	55.31	0.276
Random Forest (w/AI score)	47.02	58.65	0.237
Random Forest (w/o AI score)	50.29	54.32	0.222
Gradient Boosting (w/AI score)	59.85	54.08	0.122
Gradient Boosting (w/o AI score)	61.76	54.80	0.100
AdaBoost (w/AI score)	38.55	58.10	0.329
AdaBoost (w/o AI score)	42.38	55.08	0.302

MSE, mean squared error; R2, coefficient of determination

The charts shown in the following figures identify the importance of each indicator in the ML models through Feature Importance and Permutation Importance. The R2 scores of between 0.27 and 0.33 are significantly larger than the chance level, making it possible to subtract individual feature importance and permutation importance to probe which features are most predictive.

Feature importance, as the name suggests, shows the importance of each feature variable in the model. For a complex such as Bagging, Random Forest, or AdaBoost, feature importance is the average of all sub-models. Permutation importance measures the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target. Therefore, this drop in the model score indicates how much the model depends on the feature.

With or without AI scores, the feature importance shows that patient age, maxillary crowding, and mandibular crowding are the three most predictive components in the Bagging model (Figure 6). Overjet, overbite, and race identification also have quite significant feature importance.

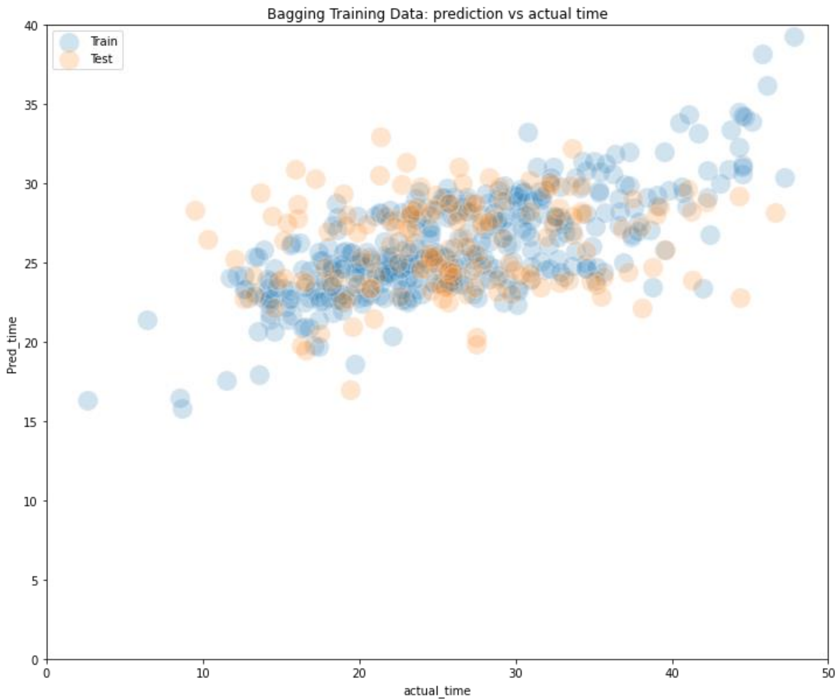

Figure 4. scatterplots comparing actual treatment time vs predicted treatment time for bagging model.

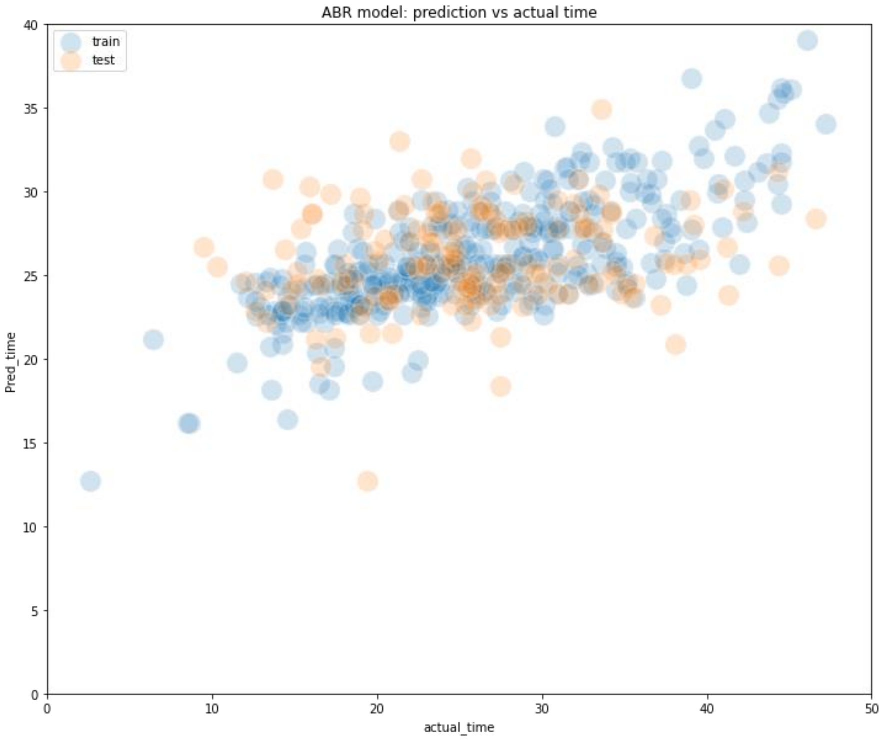

Figure 5. scatterplots comparing actual treatment time vs predicted treatment time for AdaBoost model.

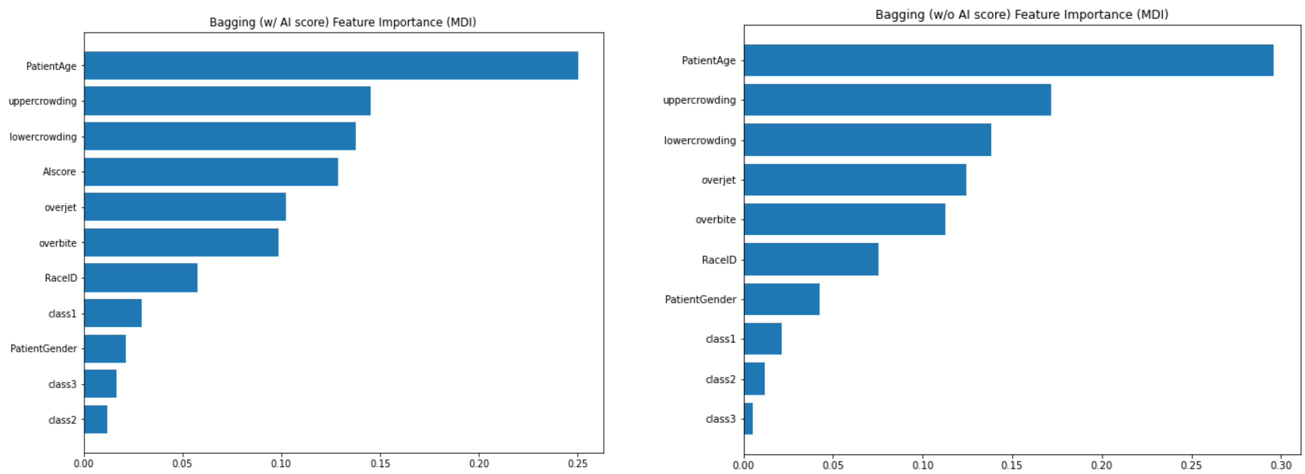

Figure 6. Feature importance for the Bagging model.

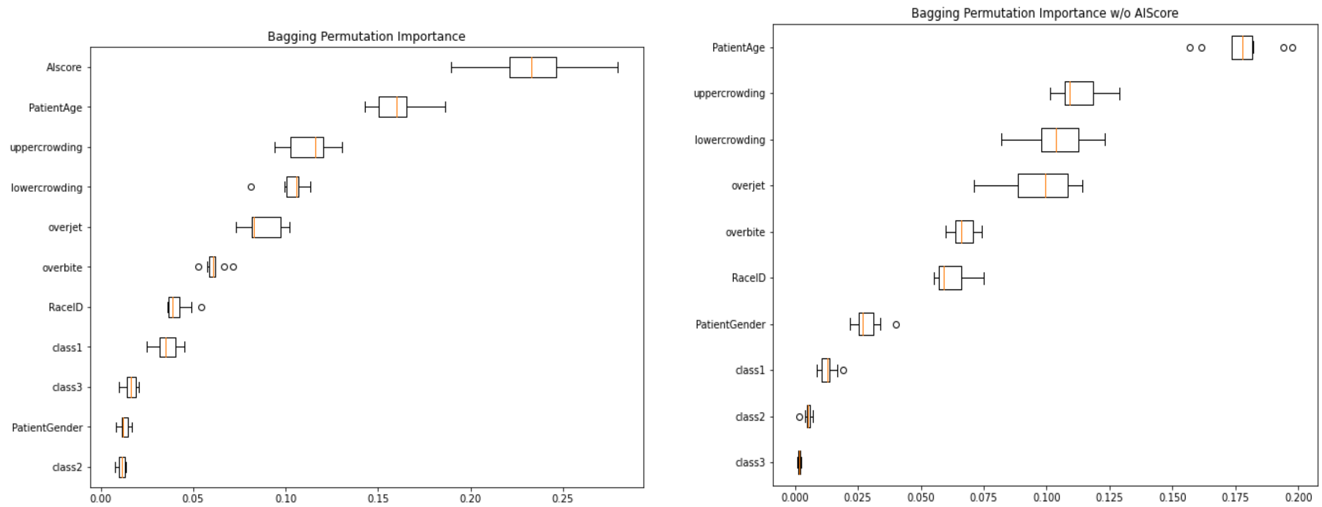

Figure 7. Permutation importance for the Bagging model

We can see that patient age, maxillary crowding, and mandibular crowding are also the top predictive variables measured by permutation importance in the Bagging model (Figure 7). In addition, figure 5 shows that the AI score played an important role in the model including the AI score as a predictive variable.

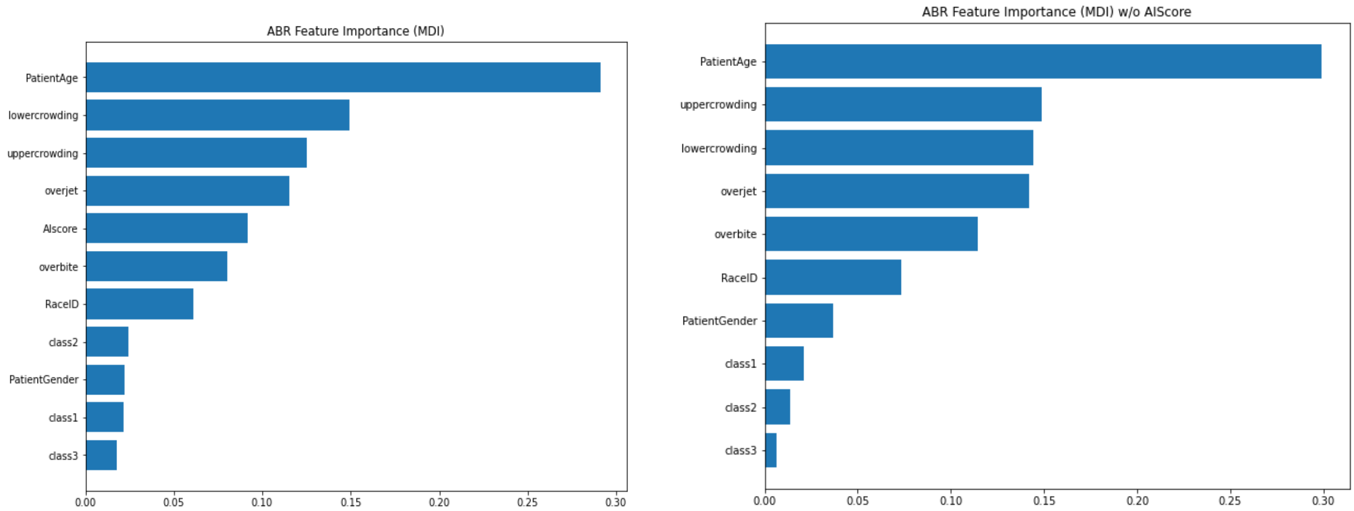

Figure 8. Feature importance in AdaBoost model.

With results very similar to the Bagging model, the feature importance of the AdaBoost model (with or without AI score; Figure 8) shows that patient age, maxillary crowding, and mandibular crowding are the three most predictive components. Overjet, overbite and race identification also have significant feature importance.

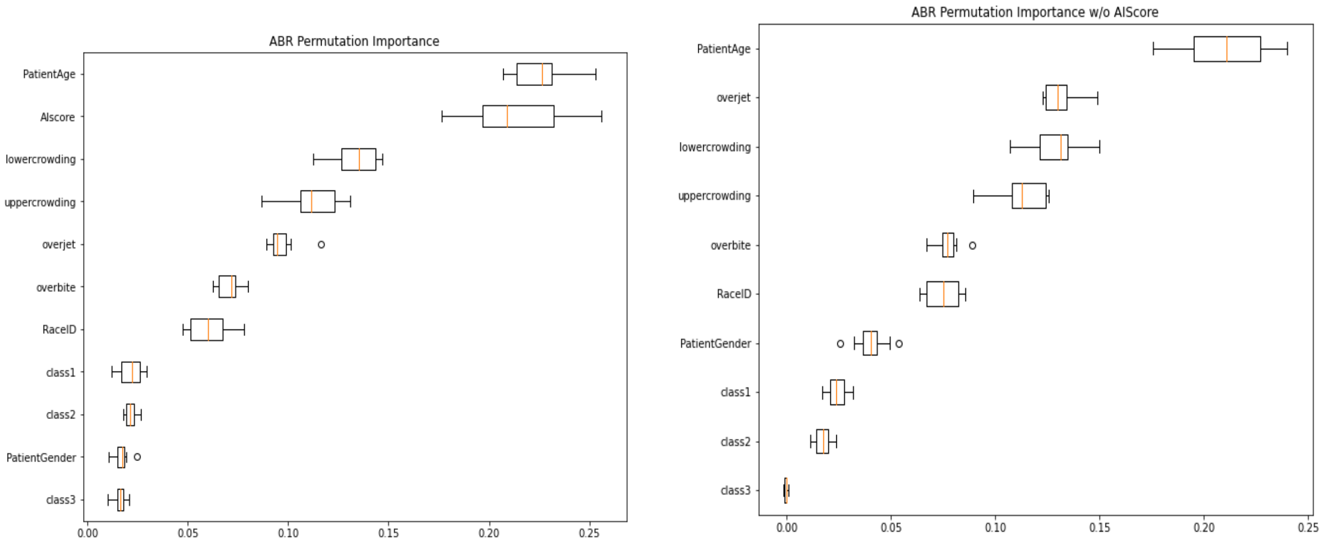

Figure 9. Permutation importance for AdaBoost model.

The permutation importance results of AdaBoost (Figure 9) show results similar to those of the Bagging models, with patient age, maxillary and mandibular crowding being more significant than other variables. In the AdaBoost model without AI score, overjet stood out as the second most important variable.

4. Discussion

The ML models built in this study were used to predict the orthodontic treatment length based on multiple factors, including patient demographics, types of malocclusions, and measures of malocclusion severity such as crowding, overjet and AI score for treatment difficulty. When we evaluated the performance of different ML models, we found that the Bagging and AdaBoost models had better performance than the other ML models tested. Bagging, or Bootstrap Aggregating, is based on the decision tree model. It generates multiple samples of training data via bootstrapping, training a deeper decision tree on each sample of training data, then outputs the averaged results of all models, i.e., aggregating. Compared to regular decision tree models, bagging enjoys the benefits of high expressiveness and low variances. AdaBoost is a complex boosting decision tree regression model that uses multiple subsequent trees of residuals to build a combined, e.g., boosting. AdaBoost assigns larger weights to outliers in each iteration of the boosting model building. This makes AdaBoost especially efficient compared to other boosting methods [15]. We tested the performance of the ML models with and without the AI score [13] Adding the AI score improves the ML models' performances, especially evident with the Bagging and AdaBoost models. The AI score is based on malocclusion detection and assessment by AI from clinical images, including crowding, spacing, deep bite, open bite, and crossbite [13]. AI score is a novel method for assessing the case difficulty, confirming that the more difficult the case, the longer the treatment duration.

We assessed the feature importance for the ML predictive models; patient age, maxillary crowding, and mandibular crowding were the top features. Patient age could be a contributing factor due to the biological differences between adolescents and adults. Vayda et al. in 1995 reported significant differences in treatment length between adults and adolescents [16]. Other studies reported no significant differences in treatment length between adults and adolescents [17]. Additional parameters contribute to treatment length prediction by ML. For example, crowding, overjet, overbite, and AI score are all measures for the severity of the malocclusion; previous studies found that quantitative malocclusion indices, such as peer assessment rating (PAR) and the objective grading system (OGS), correlated with treatment length [3]. Other factors were found to have less contribution, such as gender, race, and malocclusion classification into Class I, II, and III;

this aligns with previous findings [1,3,7]. Unexplored factors may also contribute to treatment length, including orthodontic technique employed, the operator skill and experience, and patient compliance. However, the impact of these factors is unknown and needs to be examined.

The scope of this study was to build a predictive model that can be used at initial patient screening or consultation. Other parameters can be used for fine-tuning of the ML models in the future. Furthermore, individual/subjective issues create more variations than the quantifiable factors presented in the study. However, we can perform additional studies to correlate those numeric variables to better understanding of impact on treatment length. A clinical application of the ML predictive models presented in this study could be a software or a mobile application with a graphical user interface (GUI) that can be used during the orthodontic screening or consultation to provide helpful information for both the patient and the orthodontist (Figure 10). Furthermore, these ML models could be integrated with orthodontic software currently available.

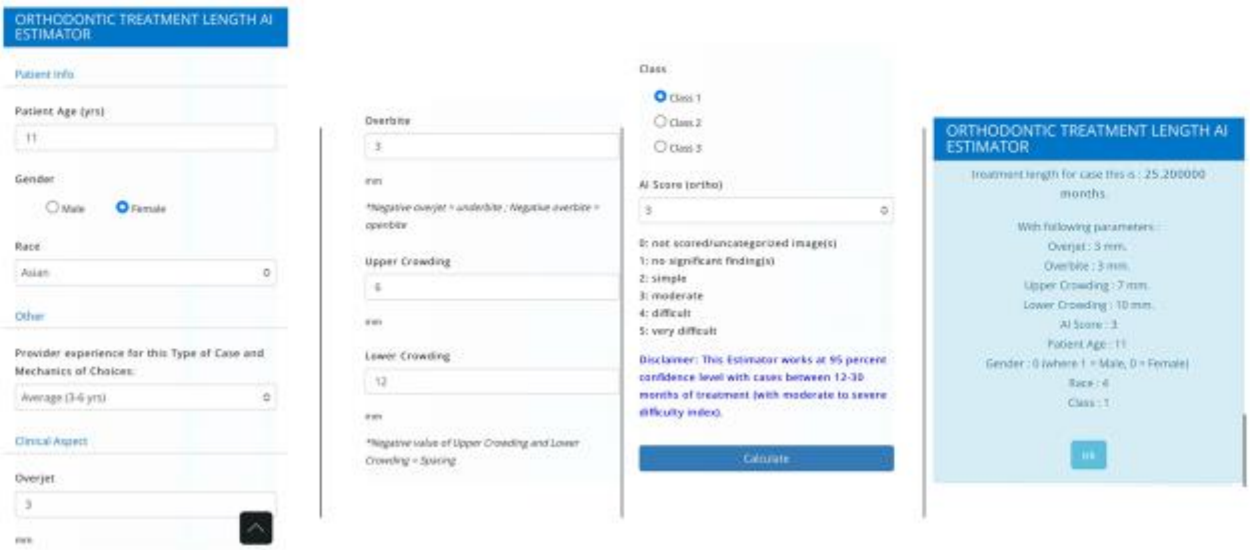

Figure 10. Graphical user interface for mobile application for treatment length prediction.

5. Conclusion

We achieved our objective of developing predictive models-based ML methods. Bagging and AdaBoost ML methods provided good predictability for orthodontic treatment length when patient information, such as age, malocclusion, and crowding, was provided. Furthermore, the study demonstrated the relative importance of each factor. Additional studies should be done on large, diverse datasets to include more variables and improve the performance of ML models for understanding orthodontic treatment length.

REFERENCES

1. Fink DF, Smith RJ. The duration of orthodontic treatment. *Am J Orthod Dentofacial Orthop*. 1992;102(1):45-51. doi:10.1016/0889-5406(92)70013-Z
2. Mavreas D, Athanasiou AE. Factors affecting the duration of orthodontic treatment: a systematic review. *Eur J Orthod*. 2008;30(4):386-395. doi:10.1093/EJO/CJN018
3. Tsichlaki A, Chin SY, Pandis N, Fleming PS. How long does treatment with fixed orthodontic appliances last? A systematic review. *Am J Orthod Dentofacial Orthop*. 2016;149(3):308-318. doi:10.1016/J.AJODO.2015.09.020
4. Kafle D, Mishra RK, Mahto RK, Luintel S, Shrestha S, Sangroula S. Comparison of Orthodontic Treatment Duration Among

Extraction Versus Non Extraction Therapies. *Orthod J Nepal*. 2019;9(2):57-60. doi:10.3126/OJN.V9I2.28416

5. Aljehani D, Baeshen HA. Effectiveness of the American Board of Orthodontics Discrepancy Index in predicting Treatment Time. *J Contemp Dent Pract*. 2018;19(6):647-650.
6. Bhikoo C, Xu J, Sun H, Jin C, Jiang H, Hu R. Factors affecting treatment duration of labial inversely impacted maxillary central incisors. *Am J Orthod Dentofacial Orthop*. 2018;153(5):708-715. doi:10.1016/J.AJODO.2017.09.017
7. Abbing A, Koretsi V, Eliades T, Papageorgiou SN. Duration of orthodontic treatment with fixed appliances in adolescents and adults: a systematic review with meta-analysis. *Prog Orthod*. 2020;21(1). doi:10.1186/S40510-020-00334-4
8. Kaur H, El-Bialy T. Shortening of Overall Orthodontic Treatment Duration with Low-Intensity Pulsed Ultrasound (LIPUS). *J Clin Med*. 2020;9(5). doi:10.3390/JCM9051303
9. Landin-Ramos M, Yadav S, Gandhi V, Upadhyay M, Tadinada A. Is there a relationship between mandibular cortical bone thickness and orthodontic treatment time? *Angle Orthod*. 2020;90(6):794-800. doi:10.2319/012220-42.1
10. Cesur E, Bayrak S, Kursun-Çakmak EŞ, Arslan C, Köklü A, Orhan K. Evaluating the effects of functional orthodontic treatment on mandibular osseous structure using fractal dimension analysis of dental panoramic radiographs. *Angle Orthod*. 2020;90(6):783-793. doi:10.2319/012020-39.1
11. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243. doi:10.1136/svn-2017-000101
12. Venugopalan SR, Elnagar MH, Karhade DS, Allareddy V. Assessment of Outcomes by Using Machine Learning. In: Ko C-C, Shen D, Wang L, eds. *Machine Learning in Dentistry*. Springer International Publishing; 2021:131-143. doi:10.1007/978-3-030-71881-7_11
13. Talaat S, Kaboudan A, Talaat W, et al. The validity of an artificial intelligence application for assessment of orthodontic treatment need from clinical images. *Semin Orthod*. 2021;27(2):164-171. doi:10.1053/j.sodo.2021.05.012
14. Evaluate and select a machine learning algorithm - IBM Garage Practices. Accessed May 30, 2022. https://www.ibm.com/garage/method/practices/reason/evaluate-and-select-machine-learning-algorithm/
15. Borstelmann SM. Machine Learning Principles for Radiology Investigators. *Acad Radiol*. 2020;27(1):13-25. doi:10.1016/J.ACRA.2019.07.030
16. Vayda D , Korsch S , W eyant R , V ig K V. Does patient age affect duration or success of orthodontic treatment? *J Dent Res*. Published online 1995:74 : 163.
17. Dyer GS, Harris EF, Vaden JL. Age effects on orthodontic treatment: adolescents contrasted with adults. *Am J Orthod Dentofacial Orthop*. 1991;100(6):523-530. doi:10.1016/0889-5406(91)70092-B