Prediction of Root Canal Treatment Using Machine Learning

Matej Jelenc
Jožef Stefan Institute
Ljubljana, Slovenia
jelenc11matej@gmail.com

Miljana Shulajkovska
Jožef Stefan Institute
Ljubljana, Slovenia
miljana.sulajkovska@ijs.si

Rok Jurič
Odontos, Private Endodontic Practice
Ljubljana, Slovenia
rok.juric@odontos.si

Anton Gradšek
Jožef Stefan Institute
Ljubljana, Slovenia
anton.gradisek@ijs.si

Abstract

Root canal treatment is a medical procedure aimed at preventing or treating apical periodontitis, which is an inflammation around the apex of a tooth root. In this study, we analyzed a dataset collected by an experienced practitioner over the course of several years, and developed a forecasting model, based on the XGBoost algorithm, to predict the outcome of the treatment. The trained models achieved a mean area under the receiver-operating-characteristic curve (AUROC) of 0.92 and average precision (AP) of 0.77. We discuss the importance of individual features in view of expert dental knowledge. To assist the practitioner in daily practice, we developed a web-based application to provide an assessment of treatment outcomes.

Keywords

root canal treatment outcome, feature importance, gradient boosting machines

1 Introduction

Apical periodontitis is an inflammation of tissues around the apex of a tooth. It is a major health burden in the general population, with 6% of all teeth showing signs of this condition. Root canal treatment (RCT) is aimed to either prevent the onset of apical periodontitis or to help the tissue to heal if it is already present [13]. Predicting treatment outcomes in RTC is of high interest both to the patients and the dentists, as well as to the insurance companies, as information about the likelihood of successful treatment can lead to better allocation of resources and avoid potentially more invasive procedures, such as tooth removal and its replacement with an implant.

Machine learning has previously been used to study some aspects of the root canal treatment, including association between patient-, tooth- and treatment-level factors and root canal treatment failure [10], predicting root fracture after root canal treatment and crown installation [6], and non-surgical root canal treatment prognosis [2]. In this study, we analyze the data collected by Jurič et al. [13]. This dataset is of special interest since it relies on the RCT patient data obtained by a single experienced practitioner (ensuring a high level of consistency in the treatment approach), as opposed to studies where numerous dentists were treating patients and different choices between them could have resulted in a less representative dataset. The aim of the study was to develop and evaluate an algorithm that predicts the outcome of the RCT, as well as to analyze how robust the algorithm is and which features influence the outcome the most. This study goes hand-in-hand with the study by Jurič et al. [13] where the analysis was conducted solely using statistical methods.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.skui.1849

2 Related Work

To our knowledge, utilization of machine learning in endodontics is still relatively unresearched, specifically when predicting treatment outcome only using tabular data. Among the related papers, [10] employs XGBoost to explore the association between patient-, tooth- and treatment-level factors and root canal treatment failure, while [2] used Random Forests (RF), K Nearest Neighbours (KNNs), Logistic Regression (LR) and Naive-Bayes (NB) to predict the outcome of non-surgical root canal treatments, similarly to this study. Paper [8] explores the prediction of treatment longevity using Support Vector Machines (SVMs), LR and NB, while [14] investigates the relation between root canal morphology and root canal treatment using both statistical and machine learning methods, specifically, using RF, SVMs and Gradient Boosting Machines (GBMs). Moreover, papers [19, 18] investigate the prediction of case difficulty and prognosis of endodontic microsurgery, while [6, 9] explore the prediction of root fracture and postoperative pain after root canal treatment. Additionally, multiple papers have been found to investigate root canal treatment outcome or related factors using deep learning (DL) on X-ray images, specifically panoramic or periapical radiographs, such as [3, 22, 11, 1, 5].

3 Data

The dataset analyzed in this study contains treatment details, clinical and radiographic data regarding primary or secondary root canal treatment of mature permanent teeth collected and curated in [13]. Three different types of outcome were determined - clinical, radiographic, and combined, for which both a strict (no clinical or radiographical sign of disease) and loose (only negligible sign of disease) assessment criteria were used. In this paper, only strict assessments were considered and used as prediction targets. All assessments were binary, with 1 representing successfull and 0

representing unsuccessfull treatment outcome. The dataset was fairly imbalanced, with 88% of all cases representing successfull radiographic outcome, 92% successfull clinical outcome and 83% successfull combined outcome. The study cohort consisted of 740 patients and 1264 teeth, resulting in 3153 root canal treatment cases and 84 features in total. The majority of features represented either categorical or binary values, such as variables representing gender, tooth type, root canal etc., while variables such as age and working length were treated as continous.

4 Methods

This section outlines the methods used for ranking feature importance and finally training baseline models that can be used as a tool for prediction of root canal treatment outcome.

4.1 Data Preprocessing

First, data regarding second visits was removed, to ensure consistency among cases. Next, features directly dependent or derived from a specific feature were excluded from the dataset to minimize the dimensionality of the data, as well as any post-operative factors that were directly used to determine the treatment outcome. The dataset was further reduced by removing redundant features, which can only have one value or their value is missing for more than 50% of all cases. Similarly, cases for which more than 50% of features are missing were excluded, resulting in 3153 cases and 84 features in total. Lastly, the dataset was preprocessed using label encoding and evenly split into training (80%) and testing (20%) sets. Furthermore, the training set was split into training (80%) and validation (20%) sets when ranking feature importance, to avoid overfitting.

4.2 Model Architecture

For the underlying model, gradient boosting machines were used, specifically the XGBoost algorithm [7], as it remains widely regarded as the state-of-the-art and preferred choice for tabular data tasks, over the more and more popular deep learning algorithms, as shown in [4, 12, 20]. Furthermore, algorithms based on transparent methods, such as decision trees, are strongly preferred for applications in medicine when compared to the "black box" approaches typically associated with deep learning.

4.3 Metrics

Due to the dataset's high imbalance between negative (87%) and positive (13%) cases, standard classification metrics such as accuracy or area under the receiver-operating-characteristic curve (AUROC) can be highly misleading, therefore average precision (AP) was chosen as the key metric for estimating model's performance and ability to produce quality predictions, specifically using the formula:

$$AP = \sum_{i=2}^{n} (R_i - R_{i-1}) \cdot P_i$$

where R_i and P_i are recall and precision at the i-th threshold when testing on n samples [17], while AUROC was only used to provide additional insight when interpreting results.

4.4 Grid Search

To obtain reasonable starting training hyperparameters and a baseline model that utilizes all available information, we performed cross-validated grid-search over a simple manually defined parameter grid, using the scikit-learn library [17].

4.5 Correlation Clustering

When a subset of features in a dataset is highly correlated, standard methods such as feature permutation importance or performing an ablation study often produce inaccurate results, since the model can highly depend only on a specific feature and discard correlated features. Similarly, methods such as SHapley Additive exPlanations (SHAP) [16] or XGBoost's built-in feature importances only account for the contribution of a specific feature to the model's prediction, which can again be misleadingly low due to the feature's correlation to another.

To address this problem, clustering was performed based on the correlation between features. Let $X \in \mathbb{R}^{m \times n}$ represent the dataset with m cases and n features. By calculating the Spearman rank correlation coefficient [15, 17, 23] on X, a symmetric feature correlation matrix $C \in \mathbb{R}^{n \times n}$ was obtained and transformed into a distance matrix $D \in \mathbb{R}^{n \times n}$. To group correlated features, hierarchical clustering using Ward's method [17, 21] was performed on D to obtain a hierarchical clustering tree, which was then flattened into discrete clusters containing features with high absolute correlation.

4.6 Ranking Feature Importance

To determine the significance of a specific feature f, a separate XGBoost model M_f was trained and evaluated on a reduced dataset X_f to obtain baseline results. Next, permutation testing was conducted by permuting the feature f in the testing set and calculating the drop in performance of M_f compared to the baseline results. Each feature was tested 20 times. Lastly, a mean drop and p-value were calculated on the observed performance drops by performing a t-test, where a high mean drop represented high feature importance and a low p-value represented a low chance that the observed drop in performance was caused by an outside factor and not by the random distribution of f in the test set. To ensure that the feature's importance estimation was not corrupted by any correlated features and at the same time account for the feature's possible non-linear connections with other features, while also minimizing the computational cost as much as possible, the reduced dataset X_f was determined as follows.

First, using the model trained on all features (see 4.4), SHAP values [17, 16] were calculated to determine the most contributing feature inside of each cluster. Let $F = \{f_1, \ldots, f_n\}$ represent the set of all features and $S : F \to \mathbb{R}^m$ the transformation that returns SHAP values for a specific feature. The most contributing feature inside of the i-th correlation cluster $C_i = \{f_j \mid j \in I_i\}$ was calculated by taking the feature with the highest mean absolute SHAP value i.e. such $f^* \in C_i$ that $\forall j \in I_i : \overline{|S(f_j)|} \leq \overline{|S(f^*)|}$.

The reduced dataset $X^* \in \mathbb{R}^{m \times r}$, containing only representative features, was then transformed into X_f for a feature $f \in C_i$ by replacing f_i^* by f in X^*. Such approach allows eliminating features highly correlated to f and reduces computational cost by only utilizing the most contributing feature within each cluster, while

still accounting for any non-linear connections between f and features in other clusters. The procedure is visualized in Figure 1.

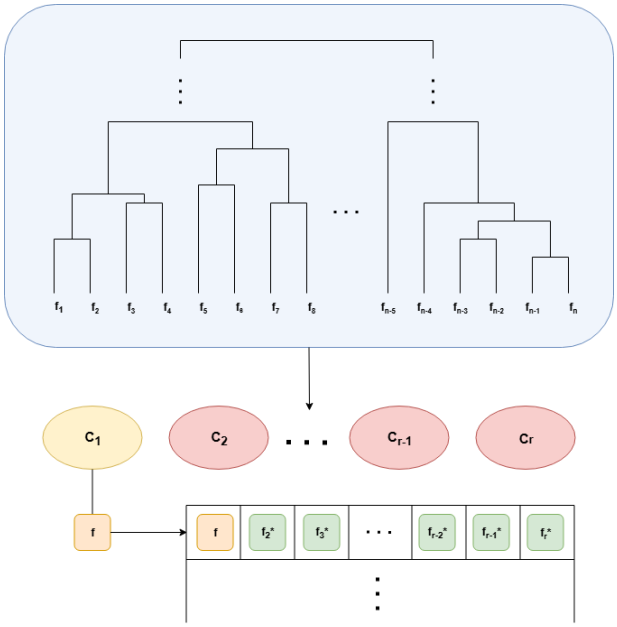


Figure 1: The hierarchical correlation tree is first flattened into clusters $C_1, \ldots, C_r$, for which representative features $f_1^*, \ldots, f_r^*$ define the base dataset X^*, from which we get X_f for $f \in C_i$ by replacing f_i^* by f.

4.7 Evaluation

After obtaining feature importances, features with p-value < 0.05 were deemed as significant. Next, a model using starting parameters found in 4.4 was trained on features belonging in the k-th percentile in terms of feature importance, for k in 1%, 5%, 10%, 25%, 50%, 75%, and 100% (the latter corresponding to all significant features).

5 Results

Figures 2 show the comparison of performances in terms of AP of models trained on different percentiles. The highest performance was achieved when utilizing the entire preprocessed dataset consisting of 84 distinct features in total, achieving AUROC of 0.90 and AP of 0.70 when predicting radiographic outcome, AUROC of 0.94 and AP of 0.86 when predicting clinical outcome and finally AUROC of 0.91 and AP of 0.77 when predicting combined outcome. Out of the 84 chosen features, our method deemed 39 of them significant for radiographic assessment, 54 significant for clinical assessment, and 65 for combined assessment, which produced AUROC of 0.88, 0.85, 0.87 and AP of 0.66, 0.75 and 0.70 respectively.

6 Discussion and Conclusion

Achieving high performance, our paper shows promise in using machine learning techniques for predicting the outcome of endodontic treatments. Moreover, we developed a web application, which allows predicting the outcome of root canal treatments using the models trained on different subsets of data, serving as a tool to assist in assessing the quality and success of a treatment, as well as to give insight for possible further patient care.

Furthermore, all the statistically significant factors found in the original study [13], are found as significant by our method as well. Specifically, "lesion diameter" was found to be the most relevant factor, with "root PAI" and "canal code" being in the top 5%, "tooth type" ("tooth group" and "canal number") in the top 10%, "type of sealer" and "quality of coronal restoration" in the top 25%, "tenderness to periapical palpation" and "quality of root filling" in the top 50% and lastly "injury history" in the top 100% of all significant features. Here, we exclude factors such as "number of visits" and "number of canals per root", since they were not used in this study. Moreover, among the most important factors that this study found and were not accounted for or found as insignificant in [13], are "age" as the second most important factor, "cumulative time" being in the top 5% and "alergic disorders", "working length", "treatment type", "obturation", "PD local", "vertical percussion", "fistulation" and "pain bite" being in the top 25%. Such results suggest that machine learning techniques can perhaps be a better or alternative approach for ranking feature significance in comparison with standard statistical methods such as logistic regression models, since they better account for possible non-linear relationships between different factors and the treatment outcome.

To further refine our approach of selecting significant features, we plan to test different p-values, as the models trained on only significant features achieved a lower performance than the models trained on the entire dataset, with a 5% drop in AUROC and a 7% drop in AP on average, suggesting that there are features which our method deemed insignificant despite enhancing the models' ability to learn and produce accurate results. Future work will also involve analysis of third-party datasets to investigate whether the results obtained in this study are generalizable and to what degree the data collected by a single experienced practitioner is different to a dataset that is typically collected over a course of several years by a number of dentists-in-training. Additionally, we wish to incorporate various explainability techniques, to better justify the models' predictions, in turn giving a deeper insight into how specific factors affect the outcome of root canal treatments as well as better assist a doctor in understanding and interpreting the predicted outcome.

References

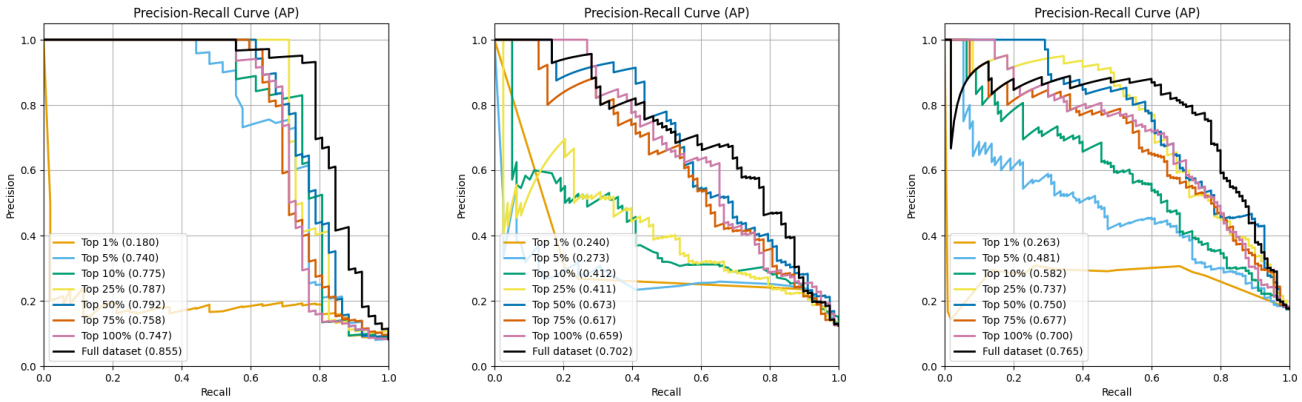
[1] Muhammed Ayhan, İsmail Kayadibi, and Berkehan Aykanat. 2025. Rcfla-yolo: a deep learning-driven framework for the automated assessment of root canal filling quality in periapical radiographs. *BMC Medical Education*, 25, 1, 894. DOI: 10.1186/s12909-025-07483-2.

[2] Catalina Bennasar, Irene García, Yolanda Gonzalez-Cid, Francesc Pérez, and Juan Jiménez. 2023. Second opinion for non-surgical root canal treatment prognosis using machine learning models. *Diagnostics*, 13, 17, 2742. DOI: 10.3390/diagnostics13172742.

[3] Catalina Bennasar, Antonio Nadal-Martínez, Sebastiana Arroyo, Yolanda Gonzalez-Cid, Ángel Arturo López-González, and Pedro Juan Tárraga. 2025. Integrating machine learning and deep learning for predicting non-surgical root canal treatment outcomes using two-dimensional periapical radiographs. *Diagnostics*, 15, 8, 1009. DOI: 10.3390/diagnostics15081009.

[4] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2024. Deep neural networks and tabular data: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35, 6, 7499–7519. DOI: 10.1109/TNNLS.2022.3229161.

[5] Berrin Çelik, Mehmet Zahid Genç, and Mahmut Emin Çelik. 2025. Evaluation of root canal filling length on periapical radiograph using artificial intelligence. *Oral Radiology*, 41, 1, 102–110. DOI: 10.1007/s11282-024-00781-3.



(a) Strict clinical assessment **(b) Strict radiographic assessment** **(c) Strict combined assessment**

Figure 2: Average precision (AP) achieved by XGBoost when predicting strict clinical, radiographic and combined assessment, utilizing different subsets of features - all features, all significant features, top 75%, top 50%, top 25%, top 10%, top 5% and top 1% significant features.

[6] Wan-Ting Chang, Hsun-Yu Huang, Tzer-Min Lee, Tsen-Yu Sung, Chun-Hung Yang, and Yung-Ming Kuo. 2024. Predicting root fracture after root canal treatment and crown installation using deep learning. *Journal of Dental Sciences*, 19, 1, 587–593. DOI: 10.1016/j.jds.2023.10.019.

[7] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). ACM, (Aug. 2016), 785–794. DOI: 10.1145/2939672.2939785.

[8] Pragati Choudhari, Anand Singh Rajawat, and S. B. Goyal. 2023. Longevity recommendation for root canal treatment using machine learning. *Engineering Proceedings*, 59, 1, 193. DOI: 10.3390/engproc2023059193.

[9] Xin Gao, Xing Xin, Zhi Li, and Wei Zhang. 2021. Predicting postoperative pain following root canal treatment by using artificial neural network evaluation. *Scientific Reports*, 11, 1, 17243. DOI: 10.1038/s41598-021-96777-8.

[10] Chantal S. Herbst, Falk Schwendicke, Joachim Krois, and Sascha R. Herbst. 2022. Association between patient-, tooth- and treatment-level factors and root canal treatment failure: a retrospective longitudinal and machine learning study. *Journal of Dentistry*, 117, 103937. DOI: 10.1016/j.jdent.2021.103937.

[11] Sascha Rudolf Herbst, Vinay Pitchika, Joachim Krois, Aleksander Krasowski, and Falk Schwendicke. 2023. Machine learning to predict apical lesions: a cross-sectional and model development study. *Journal of Clinical Medicine*, 12, 17, 5464. DOI: 10.3390/jcm12175464.

[12] Yejin Hwang and Jongwoo Song. 2023. Recent deep learning methods for tabular data. *Communications for Statistical Applications and Methods*, 30, 2, (Mar. 2023), 215–226. DOI: 10.29220/CSAM.2023.30.2.215.

[13] Rok Jurič, G. Vidmar, R. Blagus, and Janja Jan. 2024. Factors associated with the outcome of root canal treatment—a cohort study conducted in a private practice. *International Endodontic Journal*, 57, 4, 377–393. DOI: 10.1111/iej.14022.

[14] Mohmed Isaqali Karobari, Vishnu Priya Veeraraghavan, P. J. Nagarathna, Sudhir Rama Varma, Jayaraj Kodangattil Narayanan, and Santosh R. Patil. 2025. Predictive analysis of root canal morphology in relation to root canal treatment failures: a retrospective study. *Frontiers in Dental Medicine*, 6. DOI: 10.3389/fdmed.2025.1540038.

[15] Maurice G. Kendall and Alan Stuart. 1973. *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. (1st ed.). See Section 31.18. Charles Griffin, London, UK. ISBN: 978-0852640111.

[16] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. (2017). https://arxiv.org/abs/1705.07874 arXiv: 1705.07874 [cs.AI].

[17] F. Pedregosa et al. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

[18] Yang Qu, Zhenzhe Lin, Zhaojing Yang, Haotian Lin, Xiangya Huang, and Lisha Gu. 2022. Machine learning models for prognosis prediction in endodontic microsurgery. *Journal of Dentistry*, 118, 103947. DOI: 10.1016/j.jdent.2022.103947.

[19] Yang Qu, Yiting Wen, Ming Chen, Kailing Guo, Xiangya Huang, and Lisha Gu. 2023. Predicting case difficulty in endodontic microsurgery using machine learning algorithms. *Journal of Dentistry*, 133, 104522. DOI: 10.1016/j.jdent.2023.104522.

[20] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: deep learning is not all you need. *Information Fusion*, 81, 84–90. DOI: 10.1016/j.inffus.2021.11.011.

[21] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 301, 236–244.

[22] Weiwei Wu, Surong Chen, Pan Chen, Min Chen, Yan Yang, Yuan Gao, Jingyu Hu, and Jingzhi Ma. 2024. Identification of root canal morphology in fused-rooted mandibular second molars from x-ray images based on deep learning. *Journal of Endodontics*, 50, 9, 1289–1297.e1. DOI: 10.1016/j.joen.2024.05.014.

[23] Daniel Zwillinger and Stephen Kokoska. 2000. *CRC Standard Probability and Statistics Tables and Formulae*. (1st ed.). Section 14.7. Chapman & Hall/CRC, Boca Raton, FL. ISBN: 978-0-8493-0026-4.