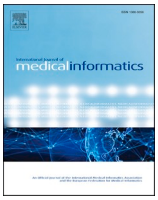
Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Machine learning in orthodontics: Transforming Invisalign treatment planning through precision, interpretability, and ethical practices

Sanisa Trakulmututa [a,b], Khin Than Win [c,*]

[a] University of Wollongong, Faculty of Engineering and Information Science, Master of Health Informatics, Orthodontist and Invisalign Provider, Bangkok, Thailand
[b] NSW 2522, Australia
[c] School of Computing and Information Technology, Faculty of Engineering and Information Science, University of Wollongong, NSW 2522, Australia

ARTICLE INFO

Keywords:
Invisalign
Machine Learning
Predictive Analytics
Orthodontics
Ethical AI
SHAP Analysis
Treatment Planning
Sensitivity
Specificity

ABSTRACT

Background: The integration of artificial intelligence (AI) in healthcare, particularly in orthodontics, is evolving rapidly. This study leverages a unique dataset from Thai patients undergoing Invisalign treatment to explore the synergy between AI and clinical orthodontics.
Objective: This research aims to augment the predictability and personalization of Invisalign treatment plans via advanced machine learning (ML) models. The focus is on enhancing clinical decision-making by predicting treatment outcomes, identifying key influencing factors, and improving the models' interpretability and explainability within a framework of ethical AI.
Methods: We analyzed 657 de-identified patient records from five dental clinics in Thailand. ML techniques, including Decision Trees (DT), Random Forest (RF), XGBoost, and Artificial Neural Networks (ANNs) were employed. Emphasis was placed on model transparency using SHapley Additive exPlanations (SHAP), integrating clinical expertise with predictive analytics to deepen understanding of treatment dynamics.
Results: XGBoost outperformed other models in predicting Invisalign outcomes, achieving an accuracy of 93.94%, sensitivity of 97.12%, specificity of 90.00%, and F1-score of 94.39%. SHAP analysis enhanced interpretability, offering detailed insights into how clinical and demographic features influence predictions.
Conclusions: This research advances the precision of orthodontic treatment predictions significantly and pioneers the ethical application of AI in orthodontic care. By improving model transparency and accountability, the study cultivates trust among stakeholders and enhances the overall effectiveness and satisfaction associated with treatment. This work sets a new benchmark for data-driven, patient-centric orthodontic care using a patient dataset from Thailand.

1. Introduction

The advent of artificial intelligence (AI) in healthcare has emerged in a transformative era, particularly within the domain of orthodontics, where precision and patient-specific treatment are significant. Invisalign, a system of clear orthodontic devices, has become increasingly popular due to its aesthetic appeal and comfort. However, the efficacy of Invisalign treatment plans encounters a significant challenge in orthodontics due to the highly diverse nature of dental anatomy and patient compliance requirements [1,2]. Unlike traditional braces, Invisalign aligners rely on precise, incremental adjustments to move teeth as in a staging plan, which must be planned based on detailed predictive modelling of patient-specific orthodontic conditions. The complexity is compounded by factors such as the biological response of teeth to aligning forces, properties of aligner materials, slipping motions, and the variability in patient adherence to wearing schedules [2]. These uncertainties make the accurate prediction of treatment outcomes challenging yet crucial for ensuring effective and efficient treatment, necessitating the development of advanced predictive tools that enhance treatment planning and outcome predictability [3]. Inaccuracies in predicting treatment trajectories can result in extended treatment durations, increased costs, and, in some cases, compromised treatment effectiveness. Financially, these inaccuracies can escalate treatment costs significantly, considering that orthodontic treatments might need to be extended or adjusted unexpectedly [4]. Health-wise, prolonged treatment can lead to patient discomfort, increased risk of complications

* Corresponding author.
E-mail addresses: st100@uowmail.edu.au (S. Trakulmututa), win@uow.edu.au (K.T. Win).

https://doi.org/10.1016/j.ijmedinf.2025.106139
Received 23 May 2024; Received in revised form 26 September 2025; Accepted 6 October 2025
Available online 10 October 2025

such as tooth decay or gum disease due to extended wear of aligners, and lower patient satisfaction [5].

Current orthodontic treatment planning with Invisalign involves predicting tooth movements that must occur in precise increments, planned from the outset of treatment. This process is complicated by several factors. Such variables introduce uncertainties that can significantly impact treatment effectiveness. Traditional methods often rely on generalized predictions that may not account adequately for individual variability, leading to less than optimal treatment outcomes and frequent need for mid-course adjustments. As digital orthodontic treatments evolve, the potential for machine learning (ML) to provide crucial insights and enhance clinical decision-making has become evident [6]. A review from 2011 to 2020 highlights a surge in studies applying AI and ML in Orthodontics [7]. Orthodontic studies employing AI have predominantly concentrated on automating the detection of anatomical landmarks to enhance diagnostic accuracy and streamline treatment planning [8]. These applications significantly contribute to precision aligning treatments with individual growth patterns and expected treatment outcomes. However, much focus has remained on diagnostic support and procedural planning.

Previous studies employed algorithms, including Artificial Neural Networks (ANNs) and regression analysis, in the Invisalign treatment planning [9]. The current research efforts and limitations in ML models applied to dental care, and an overview of clinical problems, have been presented in the Appendix. The summary table in the study, as shown in Appendix J, provides a valuable overview of existing machine learning (ML) applications in dental informatics, demonstrating how various algorithms have been employed for different clinical problems. While ML has been successfully applied to tasks such as cephalometric landmark detection, mandibular morphology prediction, and periodontal disease classification, its application in orthodontics, particularly in treatment planning, remains limited. The studies listed in the table predominantly focus on diagnostic and classification tasks rather than predictive modeling for treatment success, highlighting a significant gap in the field.

One of the key challenges noted in previous ML applications in orthodontics is the lack of interpretability. Many of the models referenced in the table rely on complex neural networks, such as convolutional neural networks (CNNs) [10] and artificial neural networks (ANNs) [11], which often function as “black boxes.” This opacity makes it difficult for clinicians to understand how the models arrive at their predictions, which is a critical barrier to adoption in clinical settings [12]. The absence of explainable AI techniques in these studies limits their practical applicability, as orthodontic treatment decisions require transparency and justifiable reasoning, as shown in Appendix J.

This study addresses these limitations by integrating SHapley Additive exPlanations (SHAP) into the ML framework for Invisalign treatment planning. By employing SHAP, the study enhances model interpretability, allowing clinicians to visualize the impact of individual predictors on treatment outcomes. The inclusion of SHAP distinguishes this research from previous studies as it provides a novel approach to bridging the gap between ML precision and clinical explainability.

By utilizing a dataset composed of real patient records from Thailand, this research not only adds a geographical specificity to its findings but also enhances the relevance and applicability of its results to similar demographics.

1.1. Evolution of machine learning techniques and orthodontics

In orthodontic research, commonly used AI algorithms, including ANNs, convolutional neural networks (CNNs), support vector machines (SVM), and regression analysis, have markedly influenced the field [7]. These models have shown varied adoption and effectiveness, with certain nuances in their application specifically to orthodontics and, more critically, to Invisalign treatment [13].

The Random Forest (RF) model, an ensemble technique building on the Decision Tree (DT) ’s foundation, has gained traction due to its superior accuracy and robustness against overfitting [14]. As demonstrated by Malaga and Alvaro (2023), they showcased RF’s efficacy in segmenting Invisalign treatment durations into clinically relevant timeframes, considering variables like patient age and specific dental movement [9]. This model’s capacity to integrate a vast array of inputs, including 3D dental movements, offers a significant improvement over simpler models [9,15].

ANNs, for instance, excel in modeling complex interactions within treatment data but require substantial expertise to deploy effectively, which limits their broader use [16]. In contrast, XGBoost’s application, while still nascent, promises substantial benefits due to its scalability and high performance on structured data [17]. Studies such as those by Wang et al. (2022) and Lu Xing et al. (2023) highlight XGBoost’s potential in orthodontic predictive modeling, demonstrating its ability to enhance treatment outcome predictions through detailed feature analysis. [18,19]. This evolution reflects a broader trend of leveraging sophisticated data analysis techniques to improve the precision of Invisalign treatment outcomes, supporting the integration of AI in ways that are both scientifically rigorous and aligned with clinical needs. The ongoing challenge lies in balancing the sophistication of these models with the need for interpretability and ethical AI practices.

1.2. Explainable artificial intelligence (XAI) and orthodontics

Explainable AI provides an explanation of internal functions and presents the model that is understandable by humans or decision makers. It aims to ensure that users can comprehend, trust, and effectively oversee the new wave of AI systems [20]. The integration of ML in Invisalign treatment planning has marked a significant advancement in orthodontics, providing sophisticated predictive tools that refine treatment outcomes. Nonetheless, the efficacy of these ML models extends beyond mere predictive accuracy; it also critically considers their interpretability and explainability [21]. These aspects are fundamental in ensuring that ML-driven decisions are transparent, comprehensible, and aligned with the intricate demands of orthodontic care. Interpretability, as outlined by Kazimierczak et al. (2024), involves the extent to which the rationale behind a model’s decision can be understood by a human [22]. In the orthodontic context, where treatment decisions have profound implications on patient health and satisfaction, the ability to illustrate the algorithmic recommendations ensures that clinicians can integrate machine-suggested pathways with established clinical expertise and ethical standards, promoting informed decision-making [23].

Explainability, a central pillar of XAI, concerns the degree to which the operations of ML models are transparent, elucidating how data inputs are processed into outputs [24]. In the specialized field of orthodontics, where treatment is highly tailored to individual needs, grasping the underpinnings of ML predictions is crucial. This clarity builds trust and supports informed consent. The application of SHAP (SHapley Additive exPlanations), investigated by Sheu et al. (2022), exemplifies how individual features affect model predictions, thereby enhancing transparency [25]. These comprehensive explanations enable clinicians to effectively articulate the logic behind algorithm-based recommendations to patients.

Recent studies, by Lee et al. (2024) and Gomez-Rios et al. (2023), highlight a growing inclination toward adopting more explainable and interpretable ML methodologies in orthodontics [26,27]. This trend is propelled by the need to align AI-generated treatment strategies with detailed clinical insights and patient-specific factors, ensuring that AI deployments are not merely data-centric but also adhere to ethical and clinical standards. The integration of these evaluative and explanatory frameworks, characteristic of XAI, is imperative for the progression of ML in orthodontics. It ensures that the incorporation of these technologies into clinical practice is scientifically solid and practically advantageous.

The objective of this work is twofold: firstly, to develop and validate

ML models that can predict Invisalign treatment outcomes with high accuracy, and secondly, to ensure these models adhere to the principles of ethical AI by being transparent and understandable to clinicians. This dual focus aims to bridge the gap between technological innovation and practical orthodontic applications, ensuring that the benefits of AI are realized in a manner that supports clinical needs and upholds ethical standards.

The motivation behind selecting specific machine learning models for this study stems from their proven capabilities in handling complex, multidimensional data typical in orthodontic treatment scenarios. Machine learning models like Decision Trees (DT), Random Forest (RF), XGBoost, and Artificial Neural Networks (ANNs) were chosen for their robustness in extracting patterns and making predictions from complex datasets. These models are particularly adept at managing the non-linear relationships and high dimensionality found in orthodontic data, which includes variables ranging from patient anatomical features to behavioral factors influencing treatment compliance.

DT and RF are favored for their interpretability and ease of use, providing clear insights into decision-making processes by illustrating the paths taken to reach a prediction. XGBoost offers exceptional performance and efficiency, handling large datasets with speed while providing configurable options to balance bias and variance effectively. ANNs are selected for their ability to model intricate relationships through their layered structure, making them suitable for capturing the subtle nuances in treatment response, which are often missed by more straightforward models.

SHapley Additive exPlanations SHAP analysis provides the explanability of the blackbox nature of neural networks by providing explanations of features and their importance. The fundamental value of SHAP analysis lies in its ability to deconstruct the output of complex machine learning models into understandable contributions of each feature [28]. This is especially crucial in orthodontics, where clinicians must make precise and personalized decisions based on a variety of factors. The application of SHAP analysis in orthodontics goes beyond just enhancing understanding. It directly contributes to building trust and facilitating communication between clinicians and patients. By making AI's decision-making process transparent, SHAP ensures that the predictions generated by algorithms, such as those predicting tooth extraction in orthodontic planning [29]. Instead, they are seen as credible and scientifically substantiated. This aspect of SHAP is particularly significant in an era where patient involvement in treatment decisions is increasingly encouraged. Patients are more likely to feel confident about the course of their treatment when they can understand the reasoning behind the proposed plans. Ultimately, SHAP not only bolsters the precision of orthodontic treatments but also enhances patient satisfaction and adherence by clarifying how personalized treatments are devised and optimized based on individual data [30].

2. Methods

Our methodological framework is designed to leverage the full potential of ML technologies to refine and enhance the predictive accuracy of Invisalign treatment planning, as shown in Appendix B. This study adopts Design Science Research (DSR) as the overarching research methodology, which provides the structure for problem identification, artifact (model) development, evaluation, and communication. Within this methodological cycle, the Cross-Industry Standard Process for Data Mining (CRISP-DM) is employed as the structured process model to guide the technical implementation of machine learning workflows. DSR ensures methodological rigor and clinical relevance, while CRISP-DM operationalizes the technical steps of model development.

2.1. Data context and integration

The Cross-Industry Standard Process for Data Mining (CRISP-DM) approach established a process model that offers a structured framework for executing data mining projects. This framework is designed to be independent of the industry sector and the technology employed. [31]. This study employs a robust methodology rooted in the principles of CRISP-DM to construct predictive models for Invisalign treatment outcomes, as shown in Table 1 and Appendix A. A comprehensive dataset was compiled from 657 anonymized orthodontic treatment records from five diverse dental clinics across Thailand.

2.2. Ethics approval and data protection

This study received approval from the University of Wollongong Health and Medical Human Research Ethics Committee (Approval #2023/328), ensuring that all research activities conform to the highest ethical standards.

2.3. Data attributes

The dataset includes extensive attributes ranging from basic demographic details to complex treatment-specific information, as shown in Table 2, such as the type and duration of Invisalign treatments. Each record is categorized into outcomes labeled as "Success" or "Failure," with failure indicating the need for additional aligners beyond the plan. The distribution of the target variable in our dataset reveals a near-balanced split with 52.2 % of cases classified as successful Invisalign treatments and 47.8 % as failures.

2.4. Data preprocessing and handling of missing values

Our research utilized machine learning techniques to develop predictive models for orthodontic treatment outcomes, as Appendix C shows. We chose Python for its robust ecosystem of data science

Table 1
Machine learning development and CRISP-DM.

	Phase	Description	Focus	Outcome
1	Business Understanding	Identified the need to enhance Invisalign treatment plan predictability through consultations and literature review.	Ethical AI	Improved patient outcomes with ethical AI
2	Data Understanding	Analyzed 657 de-identified orthodontic records to scope key trends and related factors influencing treatment outcomes.	Key Trends / Factors.	Insights into treatment outcomes
3	Data Preparation	Cleaned and preprocessed data using techniques like median imputation and one-hot encoding to ensure suitability for machine learning.	Data Suitability	Structured and clean dataset ready for modeling
4	Modeling	Developed and tested DT, RF, XGBoost, and ANNs, focusing on models with high interpretability and ethical AI alignment.	High Interpretability / Ethical AI	Robust, interpretable models aligned with ethical standards
5	Evaluation	Assessed model performance using metrics like accuracy and SHAP analysis to ensure transparent and explainable insights.	Transparent / Explainable Insights	Validated models with transparent insights
6	Deployment	Integrated models into a clinical decision support system (CDSS) prototype and obtained feedback from domain experts, emphasizing explainability and visualization.	Explainability / Visualization	Feedback from experts, enhancing user trust and model transparency without real clinical trials

libraries, as shown in Appendix H. The data, derived from de-identified records from private dental clinics across Thailand, underwent rigorous preprocessing to ensure its readiness for effective analysis and modelling. An essential part of this preprocessing was the transformation of categorical variables into numerical formats through one-hot encoding. This encoding method is crucial as machine learning algorithms require numerical input to perform optimally, ensuring both enhanced interpretability and model accuracy [32,33].

In our dataset, missing data was a notable challenge, particularly in the variables "Treatment Time" and "Number of Aligners." To mitigate the potential biases and preserve the integrity of our dataset, we employed median imputation. This method was selected due to its resilience against outliers, ensuring that the extreme values did not skew the results [34]. This approach is particularly suitable for our clinical dataset, where the distribution of data can be asymmetrical and outliers are common [35].

For the variables in question, missing values constituted approximately 1.52 % for "Treatment Time" and 0.76 % for "Number of Aligners." Handling these missing entries effectively was crucial since both attributes are significant predictors of Invisalign treatment success. The use of median imputation involved replacing missing values with the median of existing values for each attribute, thus maintaining the central tendency of the data distribution without the influence of outliers.

This method not only helped preserve the full dataset for analysis but also ensured the robustness of our statistical findings by maintaining the sample size and preventing the biases that could arise from listwise deletion, as shown in Appendix D and E. Listwise deletion could potentially lead to a reduction in sample representativeness, especially if the missingness is systematically related to other key variables [36]. By opting for median imputation, we upheld the accuracy and generalizability of our study's outcomes, adhering to recommended practices in clinical and epidemiological research for handling missing data [37].

Incorporating this method aligns with our commitment to upholding the highest standards of data integrity and reliability in our predictive modeling, essential for ensuring that our findings are both scientifically valid and practically applicable in clinical settings. This step is part of our broader data preprocessing efforts, which are critical in laying a solid foundation for the successful application of machine learning techniques in our research.

Table 2
A summary of data attributes and related information.

	Attributes	Data type	Description
1	ID	Integer	An anonymized identifier assigned to each patient record
2	Year	Integer	The year in which the patient's treatment commenced.
3	Dentist	Nominal	Classification of the dentist's expertise, whether a general dentist or an orthodontist.
4	Age	Integer	The age of the patient at the start of treatment.
5	Gender	Nominal	The gender of the patient.
6	Submission processes	Nominal	The method used to obtain the dental impressions or scans for the aligners.
7	Classification	Nominal	The classification of dental malocclusion includes Class I, Class II-1, Class II-2, and Class III.
8	Submission option	Nominal	The specific format in which the Invisalign production was ordered, including options like Full, Lite, I-7, or Teen.
9	Treatment time	Real	Recorded in months, this is the duration of the initial set of treatments as planned.
10	No. of aligners (Max)	Integer	The total number of upper aligners used in the first phase of treatment.
11	No. of aligners (Mand)	Integer	The total number of lower aligners used in the first phase of treatment.
12	Wearing time	Real	The duration, typically in days, for which each pair of aligners was worn.
13	Result	Nominal	Evaluation of how effectively the treatment planning software predicted the treatment course and outcomes

Accuracy and integrity of data are critical in clinical research focused on predicting orthodontic treatment outcomes. To address missing data, we utilize median imputation, recognized for its robustness and appropriateness for our numerical datasets. This technique replaces missing entries with the median value of the available data for each attribute, thereby preserving the central tendency and avoiding distortion by outliers, as validated by Schafer and Graham (2002) [38].

In addition, Patient ID was treated strictly as an identifier and excluded from all model training and SHAP interpretability analyses to prevent data leakage and distortion of feature importance. Only clinically relevant variables (e.g., treatment time, number of aligners, submission options, age, and wear time) were included in the modeling pipeline.

2.5. Machine learning algorithm selection

The selection of ML algorithms is tailored to meet the specific requirements of predictive modeling in orthodontics. DT is chosen for its simplicity and ease of interpretation, making it highly suitable for clinical environments where decisions must be transparent and easily understandable. To address the limitations of DT, particularly overfitting, we integrate RF, an ensemble method that effectively enhances predictive accuracy and handles complex, high-dimensional datasets. ANN is deployed to capture intricate non-linear relationships within large datasets, offering deep insights into complex treatment dynamics. Additionally, XGBoost is selected for its high performance, especially in managing overfitting and excelling in structured data analysis. These models are developed and assessed using Python.

2.5.1. Decision Tree DT

Decision Trees (DTs) have been selected for their transparency and simplicity, offering clear, visual representations of decision-making processes through binary rules [39]. This quality renders them particularly valuable in clinical environments where outcomes must be easily interpretable by healthcare professionals, and communication of how decisions are derived, which is crucial for clinical decision support systems.

The development of the Decision Tree (DT) model is correspondingly aligned with the objectives of predictability and interpretability, crucial in medical contexts. The Decision Tree classifier was chosen for its simplicity and clarity, making it ideal for clinical environments where decisions need to be both interpretable and justifiable. The initial data preparation involved converting certain columns to numeric formats and handling missing values, followed by encoding categorical variables. This preparation is vital as it ensures the data is apt for machine learning applications, directly impacting the effectiveness of the predictive models.

The Decision Tree was trained using the Python Scikit-learn library's GridSearchCV, optimizing hyperparameters such as "max_depth", "min_samples_split", and "min_samples_leaf". The optimal "max_depth" determined was 20, to balance the model's complexity and generalizability, thus capturing essential data patterns effectively. The model underwent critical evaluation using metrics like accuracy, sensitivity (recall), specificity, and F1-score on 30 % of the test data, ensuring robustness and reliability.

Visualization techniques such as the "plot_tree" function simplified the understanding of the model's decision-making process, while the graphical representation of feature importance helped communicate key predictive variables. These methods enhance the model's interpretability, supporting its practical application in clinical settings where clear explanations are necessary. Overall, the project's approach ensures that the Decision Tree model is not only effective and efficient but also transparent and comprehensible for end-users in real-world medical scenarios.

2.5.2. Random Forest RF

Random Forest (RF), an ensemble of decision trees, is utilized for its robustness in handling large datasets and its capability to enhance prediction accuracy while mitigating the risk of overfitting, which is a common challenge with single decision trees [40]. By aggregating the predictions of multiple decision trees, RF provides more stable and reliable predictive performance, making it well-suited for complex medical predictive tasks where high reliability is crucial.

The RF model, renowned for its robustness against overfitting through its ensemble approach of multiple decision trees, was meticulously chosen to handle the complexities inherent in medical data effectively.

The data preparation phase involved the conversion of select columns to numeric formats, handling of missing values, and encoding of categorical variables to prepare the dataset for advanced machine learning operations. This thorough preparation ensures the data's compatibility with the sophisticated requirements of the RF model. Following this, the model was trained using the GridSearchCV function from the Scikit-learn library, optimizing crucial hyperparameters such as "n_estimators", "max_depth", "min_samples_split", and "min_samples_leaf".

To address any potential imbalance and enhance model performance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. The technique helps to balance the class distribution without introducing duplicate records, maintaining the diversity of the data. This mitigates the risk of the model being biased toward the majority class, improving its ability to correctly classify both successful and failed treatments. This approach ensures that the predictive framework remains robust and reliable, facilitating its application in a clinical setting. The training process, particularly the use of SMOTE for balancing the dataset, highlights the project's alignment with DSR principles, emphasizing systematic parameter optimization to enhance model performance.

The RF model's effectiveness was evaluated through various metrics such as accuracy, sensitivity (recall), specificity, and F1-score, showcasing its robustness and reliability for clinical deployment. Enhanced interpretability was achieved through the visualization of feature importance and simplified decision tree diagrams, making the model's decisions accessible and understandable. These visualizations not only clarify the model's internal mechanics but also facilitate its practical application in clinical settings where clear, interpretable results are paramount. The project's approach, by aligning with established data science practices and focusing on the model's practical application in real-world scenarios, ensures that the RF model is not only technically sound but also practically relevant in medical contexts.

2.5.3. Artificial Neural networks ANNs

Artificial Neural Networks (ANNs), known for their proficiency in modelling complex nonlinear relationships, are employed to manage the intricate and multifaceted nature of medical data [41]. ANNs' ability to learn from vast amounts of data and capture subtle patterns makes them indispensable for tasks where traditional statistical methods might falter. Their deep learning capabilities enable the handling of unstructured data types common in medical settings, such as free-text clinical notes or imaging [42].

The development of an Artificial Neural Network (ANN) model is designed to meet the accuracy and interpretability needed in medical settings. The model development starts with thorough data preprocessing, including numerical conversion of variables, missing value imputation, and categorical feature encoding, optimizing the dataset for neural network training. The ANN is then trained on a normalized dataset, with architecture settings such as hidden layer sizes and iteration numbers finely tuned to effectively capture data patterns.

Post-training, the model's effectiveness is evaluated using metrics such as accuracy, sensitivity (recall), specificity, and F1-score, supplemented by the generation of a Receiver Operating Characteristic (ROC) curve to assess its discriminatory ability. The ROC curve and the area under the curve (AUC) provide deeper insights into the model's performance across various thresholds, reflecting its capability to distinguish between classes effectively. Moreover, the integration of SHAP (SHapley Additive exPlanations) further enhances the model's transparency, illustrating the impact of each feature on the predictive outcomes. This detailed evaluation ensures that the ANN not only meets the high standards required for medical diagnostics but is also interpretable, making it a valuable tool for clinical decision-making.

2.5.4. Xgboost

XGBoost has been integrated into the framework for its exceptional performance in scenarios involving imbalanced datasets, which are typical in medical diagnostics, where certain conditions or outcomes are rare [43]. XGBoost's gradient boosting framework optimizes predictive accuracy by sequentially correcting the mistakes of prior trees, focusing on difficult-to-predict instances, and effectively improving the model's performance over iterations [44].

The development process began with thorough data preparation, including the conversion of certain columns to numeric formats, addressing missing values, and encoding categorical variables. This preparation was critical to ensure the data's suitability for the sophisticated algorithms used in XGBoost. Following data preparation, the XGBoost model was trained, with hyperparameters such as tree depth and learning rate optimized using GridSearchCV. This optimization helped in fine-tuning the model to achieve the best balance between bias and variance, enhancing its predictive accuracy and generalizability across different medical scenarios.

The model's performance was evaluated using a variety of metrics, including accuracy, sensitivity (recall), specificity, and F1-score, reflecting its effectiveness in a real-world clinical setting. Additionally, the integration of SHAP (SHapley Additive exPlanations) provided insights into the contribution of each feature to the prediction outcomes, enhancing the model's transparency and interpretability. This is crucial in medical settings where understanding the rationale behind diagnostic predictions is as important as the accuracy of the predictions themselves.

These algorithms were systematically chosen to ensure a comprehensive evaluation of the dataset, encompassing various machine learning paradigms to address different aspects of the predictive modeling task. Each model underwent training and validation processes, adhering to a 70/30 training/testing data split, reflecting the CRISP-DM's emphasis on critical data evaluation. Performance metrics such as accuracy, precision, recall, and F1-score were calculated to assess and compare the effectiveness of each algorithm, providing insights into their respective strengths and suitability for the specific challenges posed by the medical dataset. This multi-model approach not only facilitates a thorough understanding of each algorithm's potential but also aids in determining the most appropriate model or ensemble of models for deployment in clinical settings, ensuring the reliability and validity of the predictions in practical applications.

2.6. Performance evaluation

The models underwent testing and validation to assess their predictive accuracy and robustness. A 70/30 split was used for training and testing datasets. In this study, several classification metrics were employed to evaluate model performance, including accuracy, sensitivity (recall), specificity, precision (positive predictive value), and F1-score. To ensure clarity across clinical and technical audiences, it is important to note that **recall is mathematically equivalent to sensitivity**, representing the model's ability to correctly identify true positive outcomes—specifically, cases in which Invisalign treatment is predicted to be successful.

While **precision** indicates the proportion of correctly predicted successful outcomes among all positive predictions, **specificity** captures the ability of the model to correctly classify unsuccessful treatment cases (true negatives). Including both sensitivity and specificity ensures a more **clinically balanced evaluation**, where minimizing false positives and false negatives is crucial for optimizing treatment decisions and resource allocation.

These metrics were selected not only for their technical relevance but also for their alignment with clinical outcome assessment. For consistency, these terms have been harmonized across Tables, Figures, and the Discussion section to support both data science rigor and clinical interpretability.

2.7. Mitigating overfitting in predictive modeling

In response to the inherent risks of overfitting associated with complex models like XGBoost and ANN, particularly when dealing with smaller datasets, our study has deployed several advanced techniques to enhance the generalizability of our predictive models. The dataset, encompassing records from 657 patients across five dental clinics in Thailand, presents unique challenges due to its limited size and the high dimensionality of the models employed.

2.7.1. Cross-Validation methods

To ensure robustness and reliability, k-fold cross-validation was extensively applied. This technique not only tests the models' stability and performance across different data subsets but also ensures that they are capturing underlying patterns rather than merely memorizing the data. Such validation is crucial for affirming the consistency and accuracy of the model predictions across various clinical settings.

2.7.2. Regularization strategies

In the training phases of our ANN models, dropout regularization played a pivotal role. By randomly disabling neurons during the learning process, this technique prevents the model from becoming overly dependent on specific features, thus promoting a more generalized learning process. This approach is essential for developing models that perform reliably on new, unseen datasets.

2.7.3. Hyperparameter optimization

Through rigorous and methodical hyperparameter tuning, including both grid and random search techniques, we finely adjusted the parameters of our models, particularly XGBoost. Critical parameters like the maximum depth of the trees and the minimum child weight were optimized to strike an ideal balance between model complexity and predictive power, enhancing the models' generalizability.

2.7.4. Model Simplification

To further reduce the risk of overfitting, we simplified our models by pruning unnecessary inputs and reducing the complexity of network architectures. This strategy focuses the learning process on the most impactful features, which is vital for maintaining predictive accuracy without overfitting noise in the data.

2.7.5. Early stopping mechanism

We integrated early stopping in our training protocols to curtail the learning process as soon as the validation performance begins to deteriorate. This technique is instrumental in preventing the models from overfitting by monitoring their performance on a held-out validation set and stopping training before deterioration sets in.

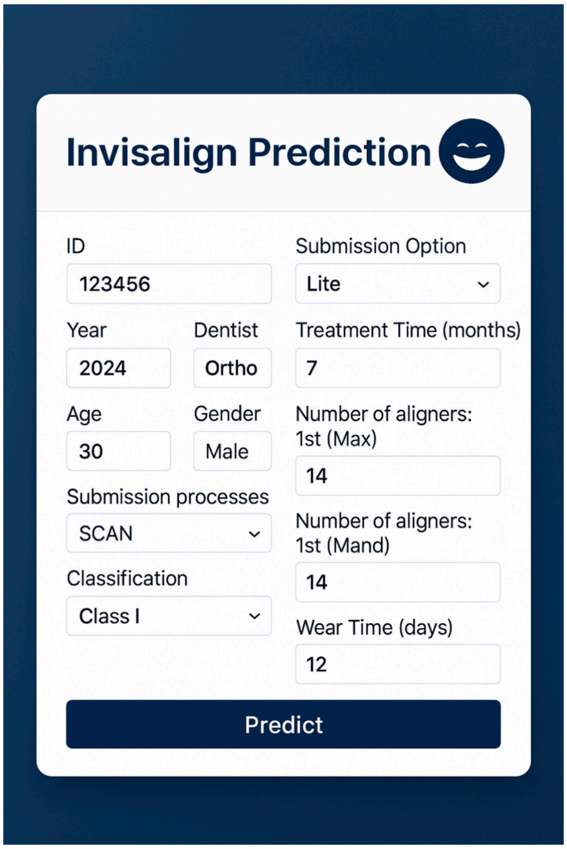
2.8. Ethical and practical implementation

Our research introduces a novel Clinical Decision Support Model (CDSM) specifically engineered for enhancing Invisalign treatments through machine learning techniques. The proposed study ensures the ethical AI perspectives as it identifies the value perspectives for conducting the study, adopting the CRISPDM and DSR as shown in Appendices F and G. The data attribute selection and modelling is conducted by the domain expert who is trained in orthodontics and health informatics, thus understanding the clinical importance of the data and the model. Thus, the decision model developed in this study applies the domain knowledge, appropriate attribute selection and different model development, ensuring transparency of the model, interpretability and support in decision making for clinicians.

Documentation of AI model predictions is crucial for transparency and accountability in clinical settings. The study describes how the model was developed, the data that was trained on, the statistical accuracy of its predictions, and its limitations. To promote usability and interpretability, a clinician-facing interface (Figs. 1-3) was designed, allowing input of treatment parameters and displaying real-time predictions (success or failure). This interactive dashboard supports transparent documentation and explainable decision-making, enabling clinicians to interpret predictions without relying solely on opaque algorithmic outputs.

2.9. Model-specific explanations and interpretability analysis

Recognizing the critical need for interpretability and explainability in clinical applications of ML, we implement SHAP analysis alongside other techniques like tree-based models and Receiver Operating Characteristic (ROC) curves. SHAP enhances the transparency of complex



Invisalign Prediction

ID: 123456

Submission Option: Lite

Year: 2024

Dentist: Ortho

Treatment Time (months): 7

Age: 30

Gender: Male

Number of aligners: 1st (Max): 14

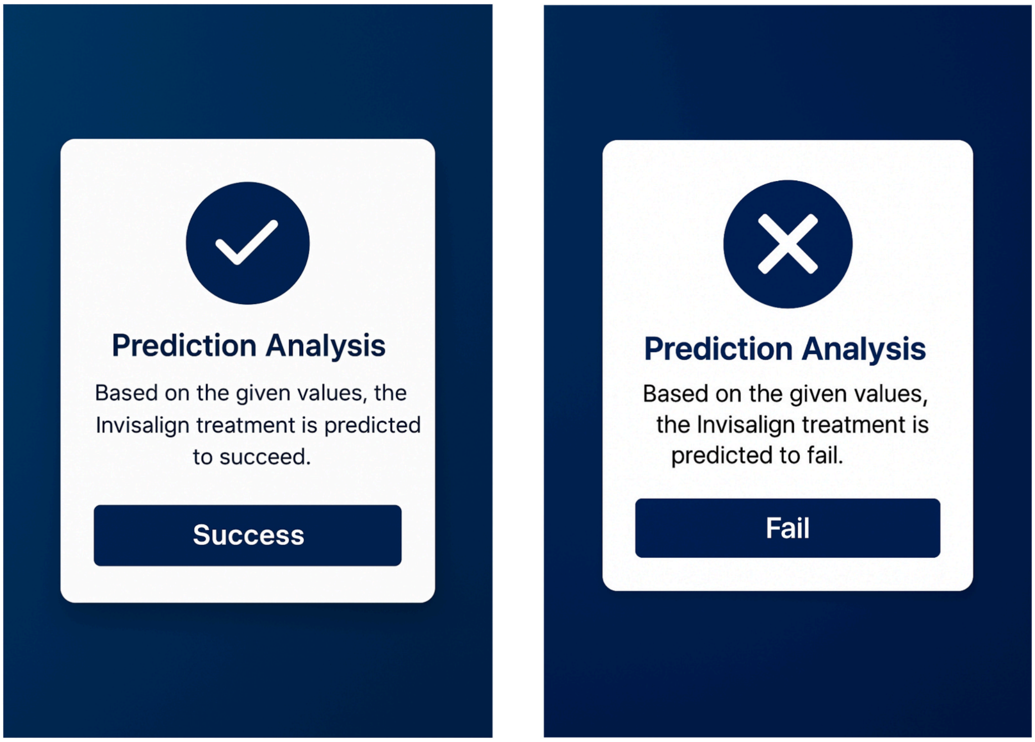
Submission processes: SCAN

Number of aligners: 1st (Mand): 14

Classification: Class I

Wear Time (days): 12

Fig. 1. Filled in page.



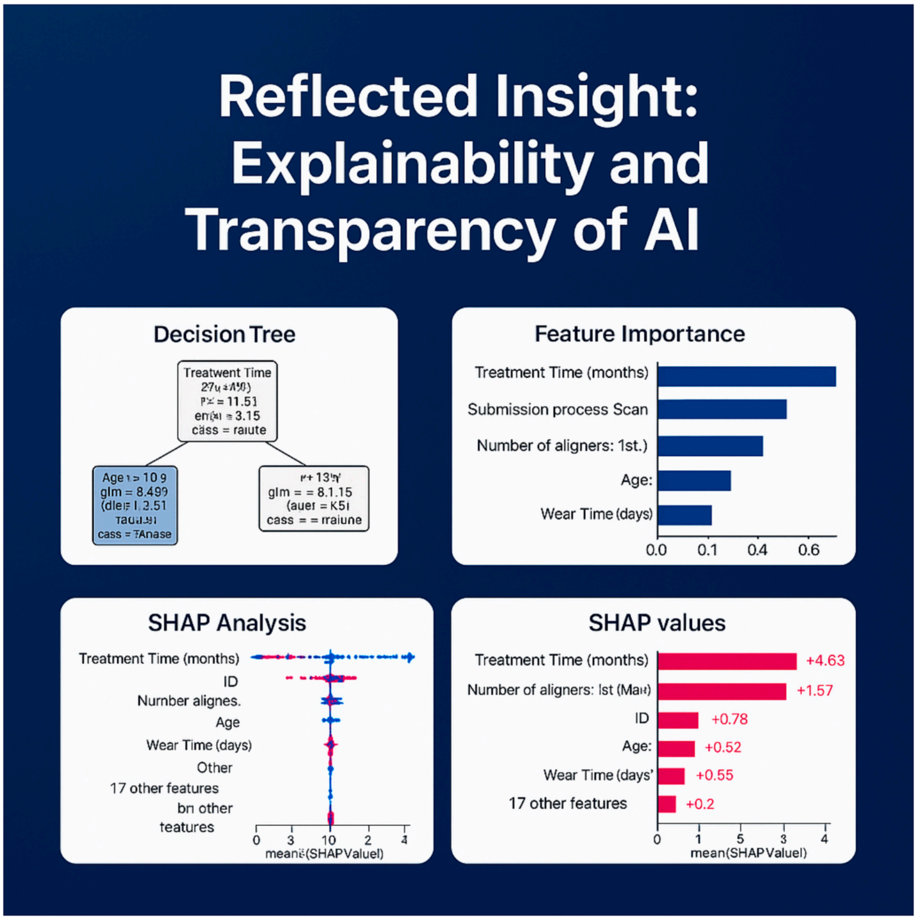
Prediction Analysis

Based on the given values, the Invisalign treatment is predicted to succeed.

Prediction Analysis

Based on the given values, the Invisalign treatment is predicted to fail.

Fig. 2. Result page – Success (Left) / Fail (Right).



Reflected Insight: Explainability and Transparency of AI

Fig. 3. Inside page.

models such as XGBoost and ANNs by depicting how individual features impact predictions, offering a detailed view of feature importance. Tree-based models further aid in understanding the decision-making process by visually representing decision paths, while ROC curves evaluate the sensitivity and specificity of models, as shown in Appendix H and I.

By merging technological innovation with orthodontic expertise, our framework sets a new standard for personalized orthodontic care, aligning with the current trends toward precision medicine in orthodontics. This strategic approach not only optimizes Invisalign treatment outcomes but also fosters a deeper integration of data-driven methodologies in clinical practice, paving the way for enhanced patient-centric care in orthodontics. This will not only enhance treatment accuracy but also improve patient satisfaction by tailoring interventions to individual needs.

3. Results

3.1. Performance parameters and model evaluation

Our investigation into DT, RF, ANN, and XGBoost models has significantly enhanced the precision of predicting Invisalign treatment outcomes. The XGBoost model, in particular, demonstrated superior performance, achieving an accuracy of 93.94 %, sensitivity (recall) of 97.12 %, specificity of 87.88 %, and a positive predictive value (precision) of 91.82 %. These metrics, alongside a detailed analysis of the confusion matrix (Fig. 4), validate the model's capacity to accurately classify both successful and unsuccessful treatment outcomes. The confusion matrix further highlights the model's ability to minimize false positives and false negatives—an essential aspect in clinical settings where incorrect predictions can lead to suboptimal treatment planning or patient dissatisfaction.

3.2. Cross-Validation and final model assessment

3.2.1. Decision Tree model Validation: K-fold Cross-Validation and holdout comparison

To validate the generalizability and stability of the proposed Decision Tree (DT) model, a 5-fold stratified cross-validation was conducted. This evaluation framework allowed the dataset to be partitioned into five equal subsets while preserving class distribution. The model was

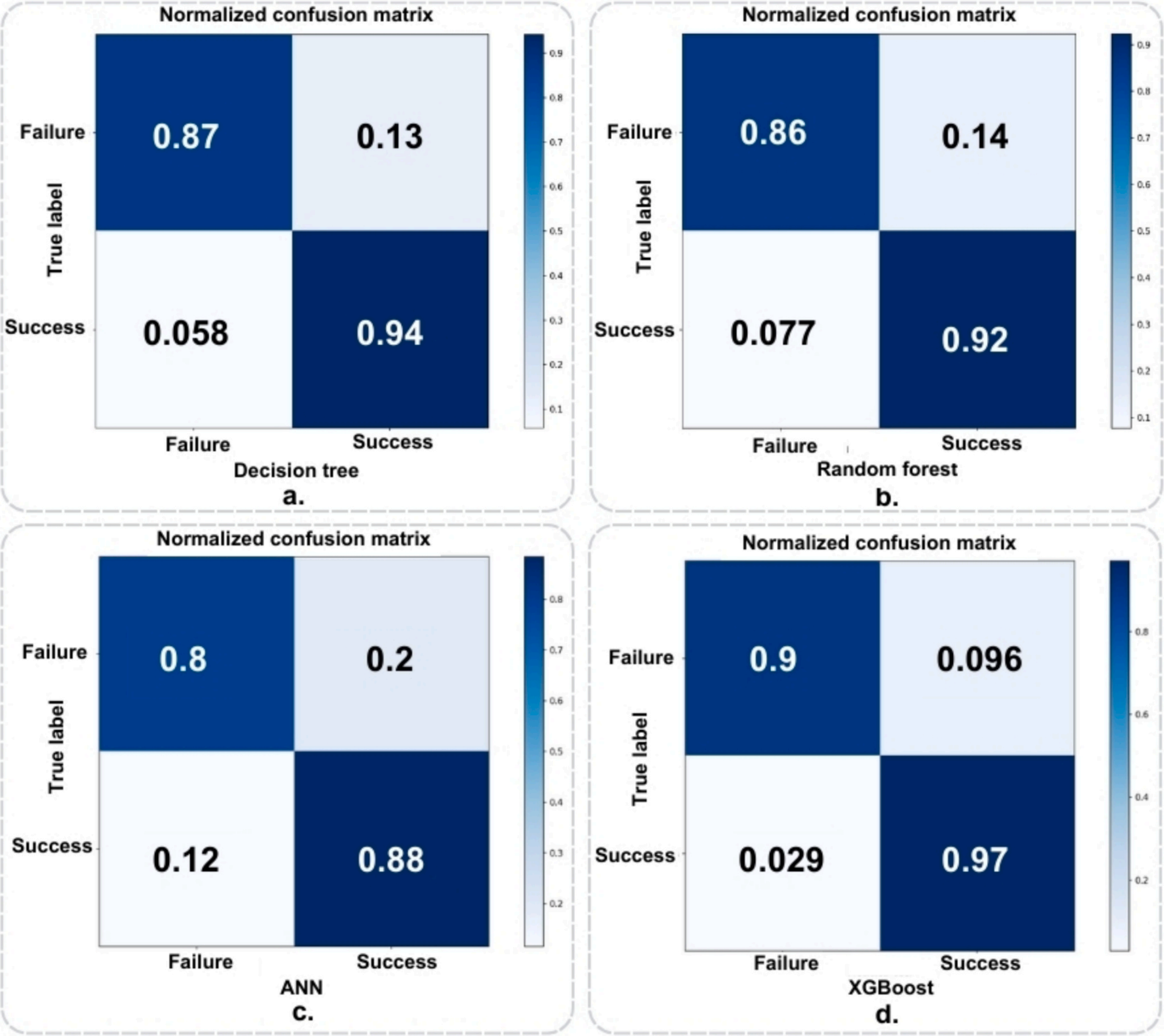


Fig. 4. Confusion matrix of ML development (a. DT, b. RF, c. ANN, and d. XGBoost).

trained on four subsets and validated on the remaining fold, rotating iteratively across all folds. The results demonstrated consistent performance, with an average accuracy of 87.67 % (±2.11), sensitivity of 87.18 % (±2.09), specificity of 86.91 % (±2.42), positive predictive value (PPV; formerly precision) of 89.03 % (±2.38), and an F1-score of 88.08 % (±1.96), as visualized in the corresponding box plot (Fig. 5). These metrics indicate that the model maintained balanced sensitivity, specificity, and PPV throughout multiple data splits, suggesting a reliable learning process.

Furthermore, a comparative bar chart (Fig. 6) illustrates the alignment between cross-validation results and the final evaluation metrics derived from a separate holdout test set. The final DT model achieved superior performance with an accuracy of 90.91 %, sensitivity of 94.23 %, specificity of 81.82 %, PPV of 89.09 %, and an F1-score of 91.59 %. The marginal performance gains observed in the test set, particularly in sensitivity and F1-score, highlight the model's strong predictive capability on unseen data while minimizing overfitting. This dual evaluation approach strengthens the credibility of the proposed model and reinforces its potential utility in supporting clinical decision-making for orthodontic treatment planning.

3.2.2. Random Forest model Validation: K-fold Cross-Validation and holdout comparison

To rigorously evaluate the generalizability of the developed Random Forest (RF) model for orthodontic treatment outcome prediction, a 5-fold stratified cross-validation strategy was implemented in addition to the conventional holdout test set evaluation. The cross-validation results yielded a strong average performance across all folds, with an accuracy of 89.95 % (±1.78), sensitivity of 89.79 % (±1.62), specificity of 86.98 % (±2.10), positive predictive value (PPV; formerly precision) of 91.02 % (±3.56), and an F1-score of 90.35 % (±1.54). These figures are illustrated in Fig. 7 using box plots to visualize metric dispersion, and the comparative bar chart in Fig. 8 highlights the consistency between the final model's test performance and cross-validation averages.

The final RF model, tested on the holdout dataset, demonstrated even slightly superior results—an accuracy of 90.40 %, sensitivity of 93.27 %, specificity of 86.0 %, PPV of 88.99 %, and an F1-score of 91.08 %. These results affirm the model's robustness and stability across unseen data, with only minor variance from the cross-validation means. The slight fluctuations are within acceptable bounds and do not indicate overfitting. This dual-evaluation framework enhances the methodological integrity and reinforces the clinical applicability of the RF model in supporting data-driven orthodontic treatment decisions.

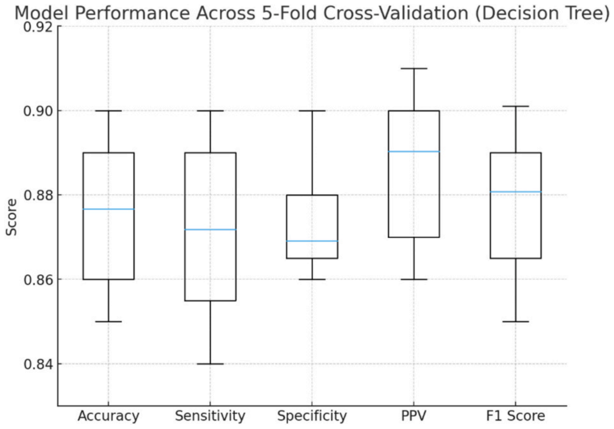


Fig. 5. Performance distribution of the Decision Tree model across 5-fold cross-validation (showing accuracy, sensitivity, specificity, positive predictive value (PPV), and F1-score. The balanced spread of sensitivity and PPV across folds indicates robust and consistent performance.).

3.2.3. ANN model Validation: K-fold Cross-Validation and holdout comparison

To evaluate the generalizability and consistency of the ANN-based predictive model, a 5-fold stratified cross-validation was employed in parallel with the conventional holdout test set approach. The results of the cross-validation process yielded stable and commendable performance metrics with an average accuracy of 82.30 % (±4.20), sensitivity of 88.90 % (±3.60), specificity of 82.50 % (±4.00), positive predictive value (PPV; formerly precision) of 80.10 % (±3.90), and an F1-score of 84.40 % (±4.20), as shown in Fig. 8. These scores suggest strong sensitivity with moderate PPV and balanced specificity across the validation folds.

When benchmarked against the final test set evaluation, the ANN model achieved an accuracy of 84.34 %, sensitivity of 88.46 %, specificity of 81.00 %, PPV of 82.88 %, and an F1-score of 85.58 %, indicating consistent and slightly improved results compared to the cross-validation averages. The accompanying bar chart (Fig. 10) provides a visual comparison between the average K-Fold scores (with standard deviation) and final test metrics, highlighting the ANN model's capacity to maintain reliable predictive performance across unseen subsets. This dual-validation strategy confirms the model's applicability to broader clinical contexts with reduced risk of overfitting.

3.2.4. Xgboost model Validation: K-fold Cross-Validation and holdout comparison

To ensure the robustness and generalizability of the XGBoost classifier, a 3-fold stratified cross-validation was conducted in parallel with the holdout test set evaluation. The model demonstrated stable and high average performance across the folds, yielding an accuracy of **90.41 % (±1.63)**, sensitivity of **90.10 % (±2.47)**, specificity of **87.50 % (±2.20)**, positive predictive value (PPV; formerly precision) of **91.44 % (±1.57)**, and an F1-score of **90.74 % (±1.59)**. These consistent results suggest that the model generalizes well to unseen subsets of data, as shown in Fig. 11.

Moreover, the final evaluation on the test set resulted in an even higher accuracy of **94.44 %**, sensitivity of **98.08 %**, specificity of **83.33 %**, PPV of **91.89 %**, and an F1-score of **94.88 %**. The small deviation between the cross-validated and holdout metrics confirms the model's strong predictive capacity while minimizing concerns of overfitting, as shown in Fig. 12. The combination of ensemble learning strength and regularization inherent to XGBoost appears well-suited for modeling complex patterns in orthodontic treatment success prediction.

3.3. Explainable Artificial intelligence (XAI) in orthodontic predictive modeling

3.3.1. Tree-Based model explainability

Tree-based models like RF and DT inherently offer transparent decision paths, making them especially valuable in clinical settings for their explainability, as shown in Fig. 13. This feature enables orthodontists to comprehend the reasoning behind each model prediction, fostering a deeper trust in AI-driven decisions. Feature importance visualizations from these models emphasize 'Treatment Time' as the most critical factor, underscoring its impact on treatment success.

3.3.2. Sensitivity evaluation using ROC curves

The sensitivity and specificity of our models were assessed using Receiver Operating Characteristic (ROC) curves, particularly for the ANN models, as shown in Fig. 14. These curves provide essential insights into the models' diagnostic accuracy, ensuring their appropriateness for clinical use and upholding the standards of ethical AI.

3.3.3. Xgboost predictive performance visualization

The XGBoost distribution graph presented in this study offers a detailed look at the predictive accuracy and variability of the model, as shown in Fig. 15. It showcases the mean predicted values along with

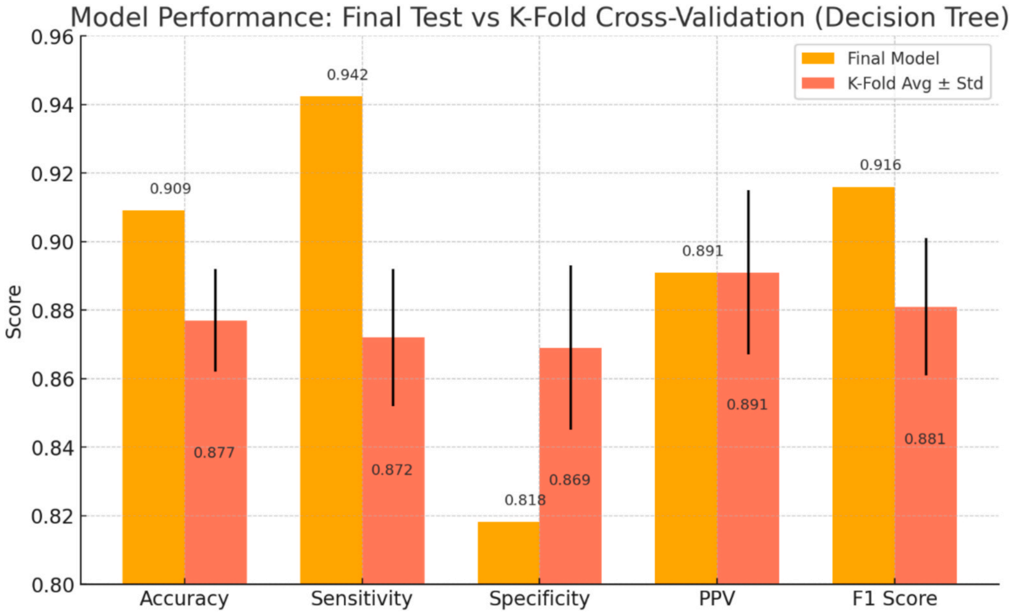


Fig. 6. Comparative evaluation of the Decision Tree model: final holdout test set vs. 5-fold cross-validation averages (with standard deviations). Metrics include accuracy, sensitivity, specificity, PPV, and F1-score. The final model demonstrates slightly higher sensitivity and F1-score than the cross-validation average, confirming strong predictive performance on unseen data while maintaining clinical interpretability.

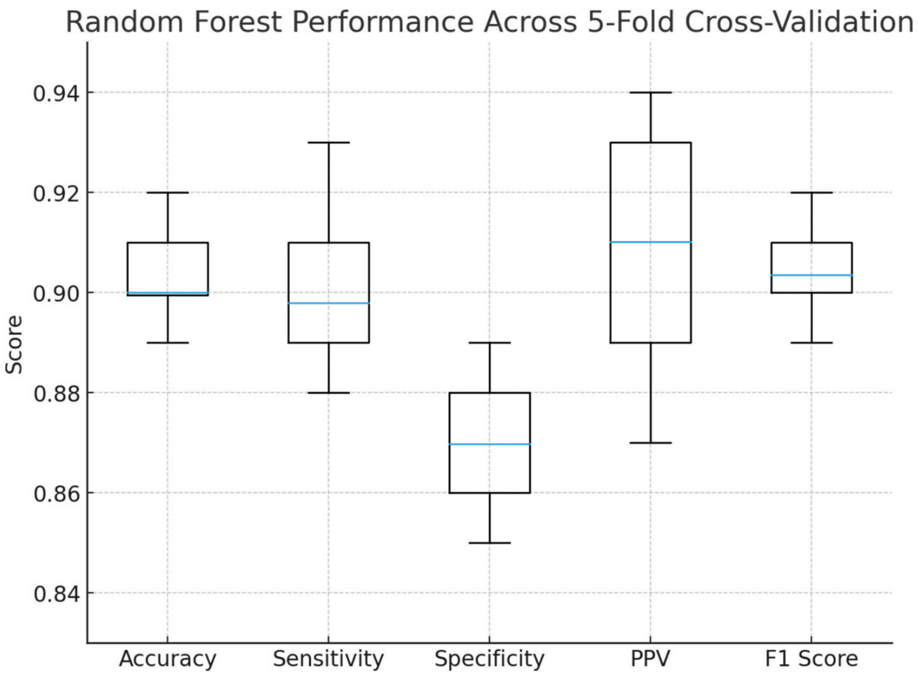


Fig. 7. Performance distribution of the Random Forest model across 5-fold cross-validation. (showing accuracy, sensitivity, specificity, positive predictive value (PPV), and F1-score. The narrow dispersion across folds highlights the stability and consistency of the model's performance.).

their standard deviations, crucial for evaluating the model's consistency and reliability. This visualization serves as a practical tool for clinicians, providing them with a clear understanding of the model's performance across diverse scenarios. It reinforces the model's robustness by transparently illustrating prediction ranges.

3.3.4. SHAP analysis for advanced interpretability

Further enhancing our model's transparency, SHAP analysis was applied to the XGBoost and ANN models, as shown in Fig. 16. This analysis detailed the influence of variables like 'Treatment Time' and 'Number of Aligners' on the predictive outcomes. Such detailed insights are imperative for ethical AI practices, allowing clinicians to make informed, patient-specific decisions. The feature importance charts from SHAP analysis further illustrate the significant roles of these variables, facilitating a tailored approach to treatment planning.

To further promote transparency, clinical applicability, and ethical use of artificial intelligence in orthodontic decision-making, SHAP (SHapley Additive exPlanations) analysis was conducted on both the

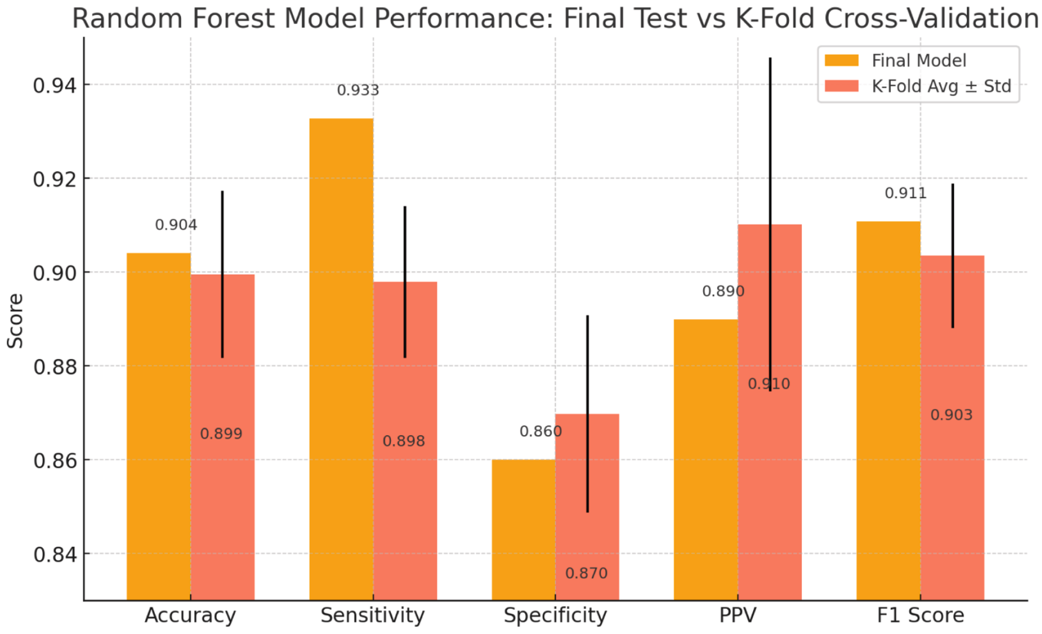


Fig. 8. Comparative evaluation of the Random Forest model: final holdout test set vs. 5-fold cross-validation averages (with standard deviations). Metrics include accuracy, sensitivity, specificity, PPV, and F1-score. The final model demonstrates slightly higher sensitivity and F1-score compared to cross-validation averages, supporting its robustness and clinical applicability for orthodontic treatment outcome prediction.

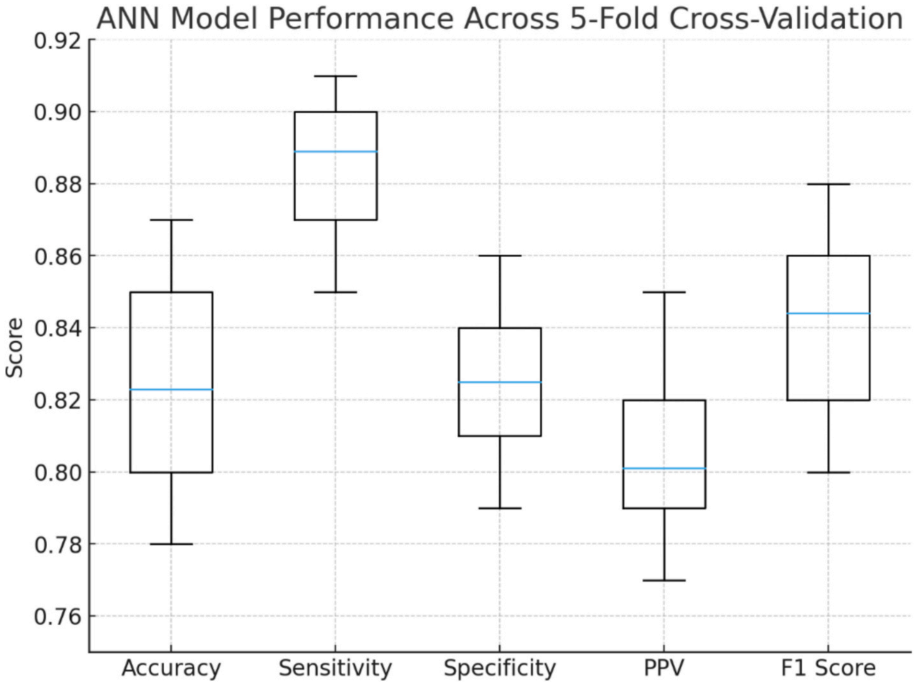


Fig. 9. Performance distribution of the ANN model across 5-fold cross-validation (showing accuracy, sensitivity, specificity, positive predictive value (PPV), and F1-score. The spread across folds highlights consistent performance with higher sensitivity compared to PPV, indicating strong detection ability but moderate precision.).

XGBoost and ANN models. This interpretability framework provides a rigorous, individualized understanding of how each feature contributes to the model's predictions, reinforcing the credibility and explainability of our approach. The outputs are summarized in Fig. 16, which comprises four panels (A–D), each offering a distinct perspective on model interpretability. These analyses align with the *Evaluation* and *Reflection* phases of the Design Science Research (DSR) methodology and directly support the integration of the models into clinical workflows.

Panel A: Global Feature Summary (ANN – SHAP Summary Plot).

This panel ranks features by their global influence across all predictions, highlighting *treatment time (months)* and *submission options* as dominant predictors in the ANN model. These findings are consistent with clinical expectations, as treatment duration and submission types significantly shape aligner protocols. In the context of DSR, this aligns

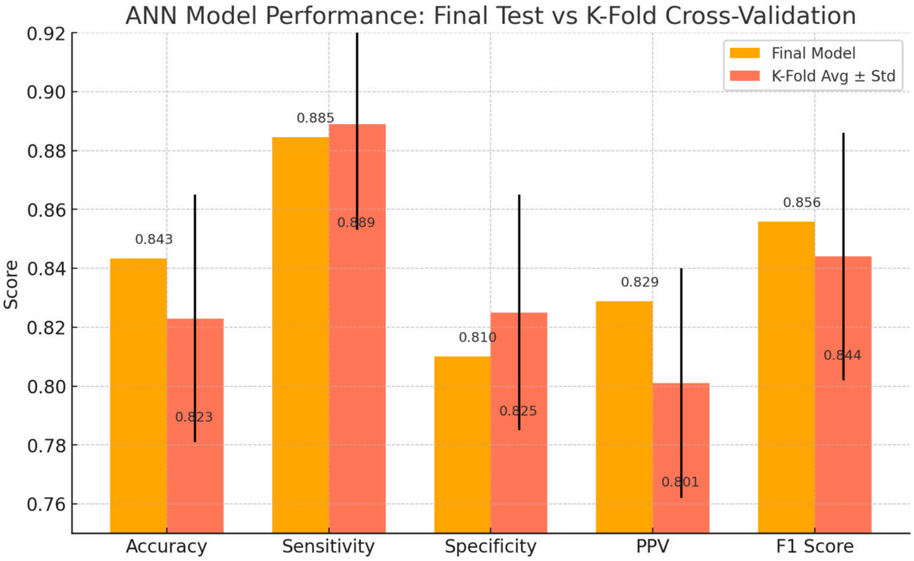


Fig. 10. Comparative evaluation of the ANN model: final holdout test set vs. 5-fold cross-validation averages (with standard deviations). Metrics include accuracy, sensitivity, specificity, PPV, and F1-score. The final model demonstrates slightly higher accuracy and PPV compared to cross-validation averages, confirming consistent predictive performance across unseen data.

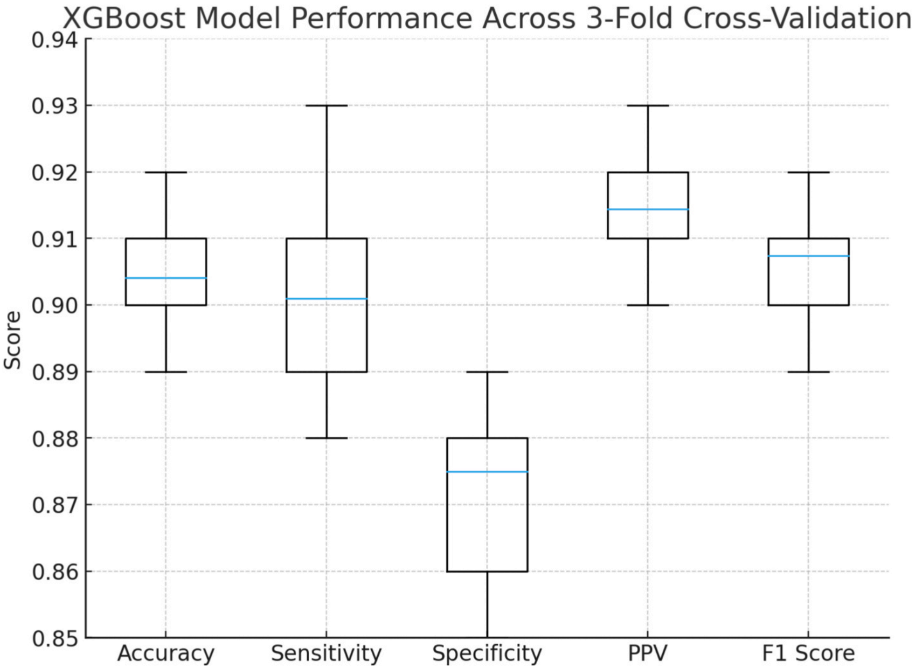


Fig. 11. Performance distribution of the XGBoost model across 3-fold cross-validation (showing accuracy, sensitivity, specificity, positive predictive value (PPV), and F1-score. The results demonstrate consistently high sensitivity and PPV, with balanced specificity, indicating stable performance across folds.).

with the *Design and Development* phase by operationalizing theoretically grounded constructs into the model.

Panel B: Global Feature Contribution (XGBoost – SHAP Summary Plot).

This panel presents a feature-level breakdown of how individual variable values (e.g., higher age or wear time) influence prediction outcomes in the XGBoost model. This visualization reflects the model's adaptability to patient-specific profiles and supports clinical interpretation of individual variability. Within the DSR cycle, this corresponds to the *Demonstration* phase, providing evidence of how the model behaves under real-world-like conditions.

Panel C: Mean SHAP Values (XGBoost – Bar Chart Summary).

Panel C distills the global impact of features into average SHAP values, facilitating a more accessible interpretation of variable importance. Notably, *treatment time* and *number of aligners* emerge as the most influential features. This serves as an evidence-based prioritization guide for clinical decisions and aligns with the *Evaluation* phase of DSR by translating model behavior into actionable insights.

Panel D: Local Explanation (ANN – SHAP Force Plot).

This panel illustrates the SHAP explanation for a single patient case, showing how specific features increase or decrease the prediction relative to the model's base value. Such local interpretability is crucial in

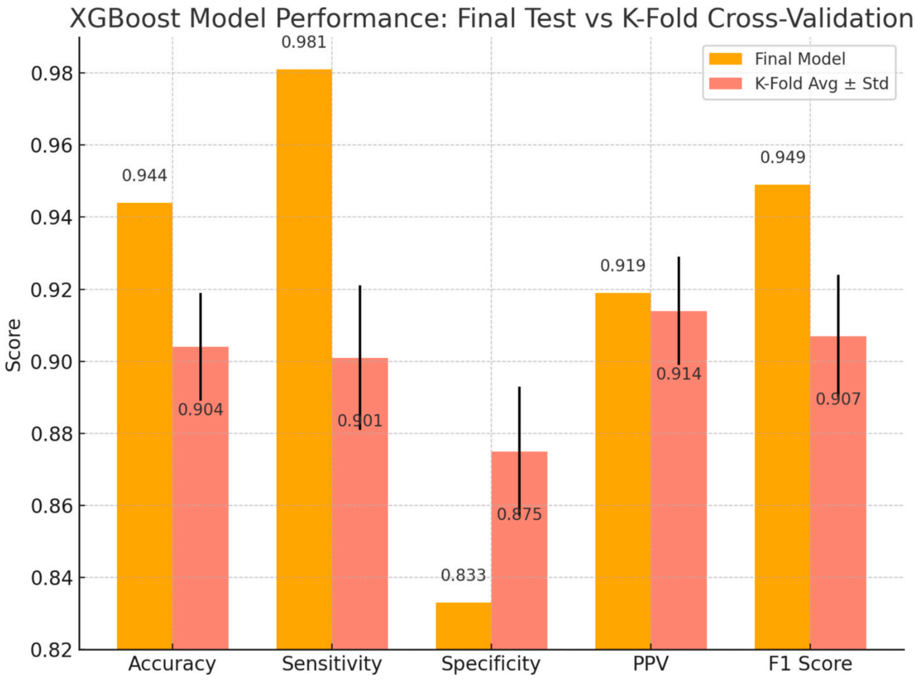


Fig. 12. Comparative evaluation of the XGBoost model: final holdout test set vs. 3-fold cross-validation averages (with standard deviations). Metrics include accuracy, sensitivity, specificity, PPV, and F1-score. The final model exhibits higher sensitivity and F1-score compared to cross-validation means, confirming the strong predictive ability of the XGBoost classifier on unseen data.

clinical practice, enabling case-level justifications that support shared decision-making and enhance transparency in patient communication. This corresponds to the *Reflection* and *Communication* phases in DSR, closing the loop between model output and clinical utility.

Integration into Clinical Workflow.

Taken together, the SHAP analyses presented in Panels A–D demonstrate not only technical robustness but also the potential for meaningful integration into clinical practice. Global patterns (Panels A and B) offer insights that can inform clinical guidelines and system-wide policies, while local explanations (Panel D) support individualized treatment planning and informed consent. These interpretable outputs can be readily embedded into clinical dashboards or electronic health records (EHRs), facilitating seamless adoption of predictive models within routine orthodontic care.

4. Discussion

The integration of ML models such as DT, RF, ANN, and XGBoost into Invisalign treatment planning marks a seminal shift in orthodontics, propelling the field toward a data-driven and ethically-informed clinical decision-making paradigm, as shown in Table 4. This transformation is underpinned by our pioneering use of XAI principles, which ensure that treatments are not only tailored to the unique temporal dynamics of each patient's needs but are also transparent and understandable to practitioners.

Sensitivity (true positive rate) measures the ability of the model to correctly identify positive cases, which is critical for clinical scenarios like identifying severe orthodontic issues. A focus on sensitivity ensures that critical cases are not overlooked, aligning the model's predictions with the primary goal of clinical interventions.

While sensitivity is crucial, it cannot stand alone in evaluating model performance. The F1 score, which combines precision and recall (sensitivity) into a harmonic mean as in Table 3, provides a balanced view of the model's ability to manage both false positives and false negatives. This metric is particularly relevant when the cost of misclassifications differs significantly between classes. High sensitivity indicates the model's ability to identify cases requiring attention, reducing the risk of missing critical interventions and ensuring appropriate treatment. In contrast, high specificity minimizes unnecessary treatments by accurately ruling out cases that do not require intervention, enhancing clinical efficiency. In current practice, sensitivity-focused tools are often used for preliminary screening to avoid missed cases, while specificity-focused tools validate findings in confirmatory stages.

Our study is the first to deploy SHAP within the orthodontic field, enhancing the interpretability and explainability of complex ML models as shown in Table 5. This application is a crucial step towards clarifying the 'black box' nature of algorithms like XGBoost and ANN, which, while powerful, have traditionally posed challenges in terms of transparency. By illustrating how variables such as treatment time and the number of aligners impact treatment outcomes, SHAP analysis fosters a deeper understanding and trust in these models, ensuring that algorithmic decisions can be ethically justified and aligned with individual treatment needs.Table 6.

To further ensure robustness and transparency, only clinically meaningful variables were included in both model training and SHAP analyses. Non-clinical identifiers such as Patient ID were strictly excluded, thereby preventing potential data leakage and ensuring that interpretability results genuinely reflect clinical factors such as treatment time, number of aligners, submission options, and wear time. This safeguards both the methodological integrity of our study and the ethical standards of clinical applicability.

The distribution visualization of the XGBoost model, illustrating the predicted mean and standard deviation for each test data point, provides clinicians with essential insights into the model's reliability and interpretability. The mean values represent the central predictions, while the standard deviation captures the variability or uncertainty in these

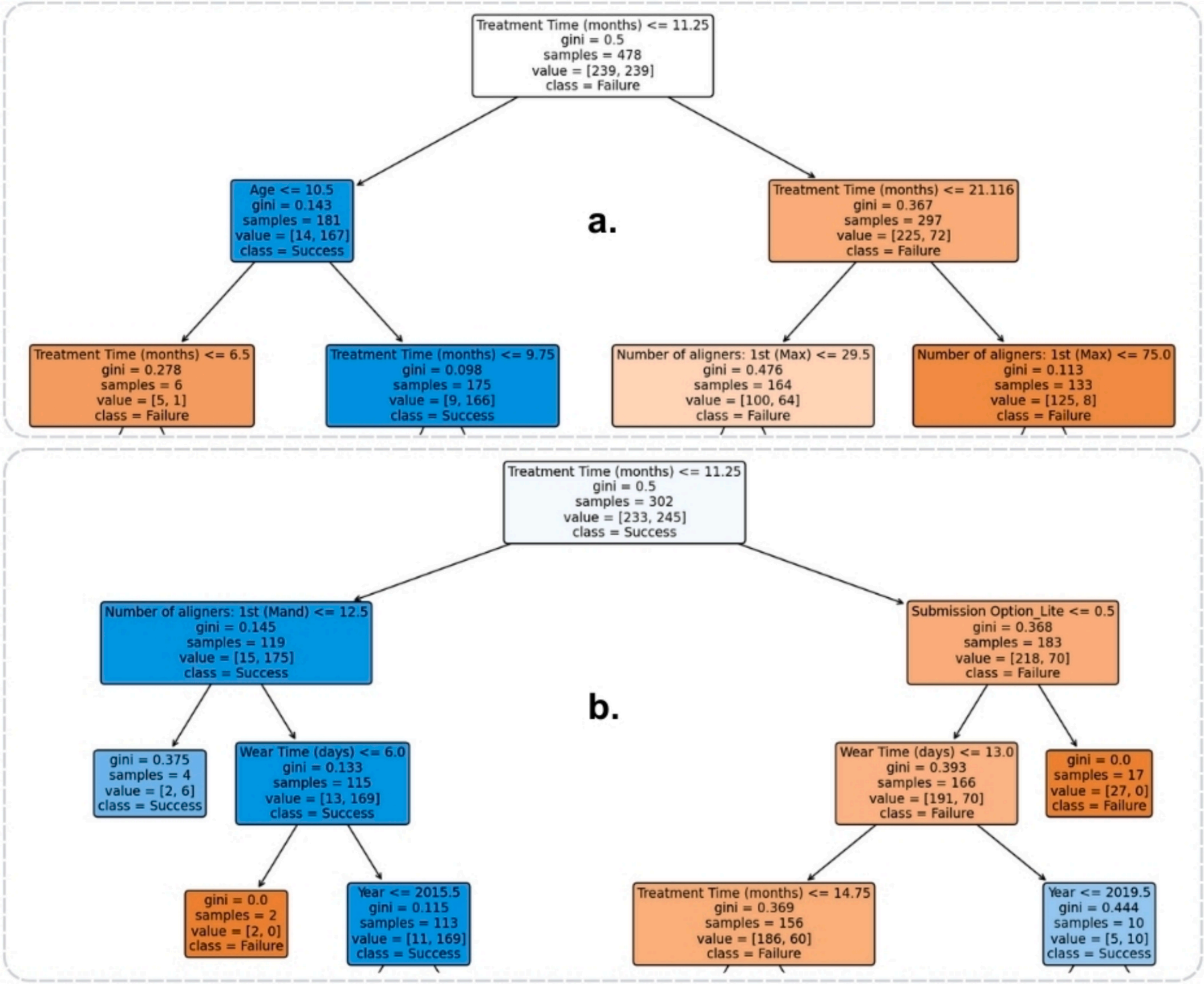


Fig. 13. Tree-based visualization (a. DT and b. RF).

predictions. Clinicians can use this information to assess the confidence of the model's outputs, ensuring predictions are robust and consistent. Data points with high variability, indicated by larger standard deviations, suggest areas of greater uncertainty. This prompts clinicians to exercise caution when making decisions for these cases and consider integrating additional clinical data or expertise.

Moreover, the visualization offers a broader evaluation of predictive trends. The color-coded representation of mean values allows clinicians to easily identify patterns in high and low prediction probabilities across the dataset. This assists in determining whether the model aligns with clinical expectations and effectively captures the complexity of orthodontic treatment scenarios. By linking these insights to individual patient cases, the visualization supports clinicians in assessing the practical utility of the model, facilitating evidence-based decision-making. Overall, this figure acts as a critical tool for bridging the gap between machine learning predictions and clinical applicability, enabling clinicians to make more informed and confident treatment decisions.

Furthermore, the inclusion of real patient data from Thailand into our models is a novel approach that enhances the cultural and regional relevance of our findings. This integration not only improves the generalizability of our treatment predictions but also ensures that our predictive models are robust, reflecting a broad spectrum of patient demographics and clinical scenarios. This aspect of our research is critical for the design of personalized treatment plans that are informed by an authentic understanding of diverse patient needs.

Treatment success is prone to be higher among younger patients ($\leq$ 10.5 years) with shorter treatment times ($\leq$ 6.5 months) and those whose treatment duration does not exceed 11.25 months. Specifically, success is significantly associated with the number of upper aligners (U $\leq$ 29.5), whereas failure is more likely with upper aligners exceeding 75, underscoring the critical need for patient adherence and compliance. For lower aligners (L $\leq$ 12.5) and wear times of $\leq$ 6.0 days, the success rate is notably higher, highlighting the importance of balancing functional movement and patient compliance.

The SHAP analysis reveals positive impacts of submission options, particularly the Teen submission option, likely due to higher biological

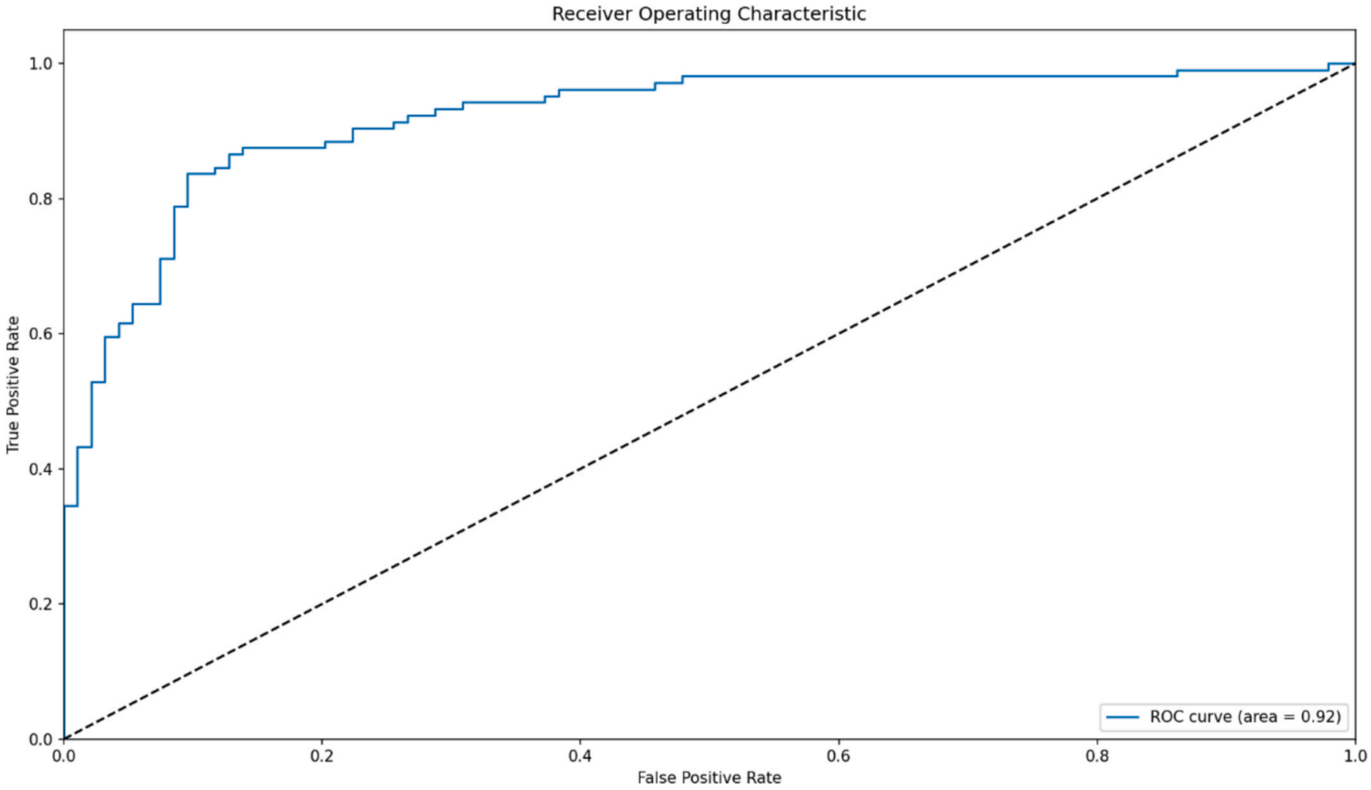


Fig. 14. ROC curve of ANN.

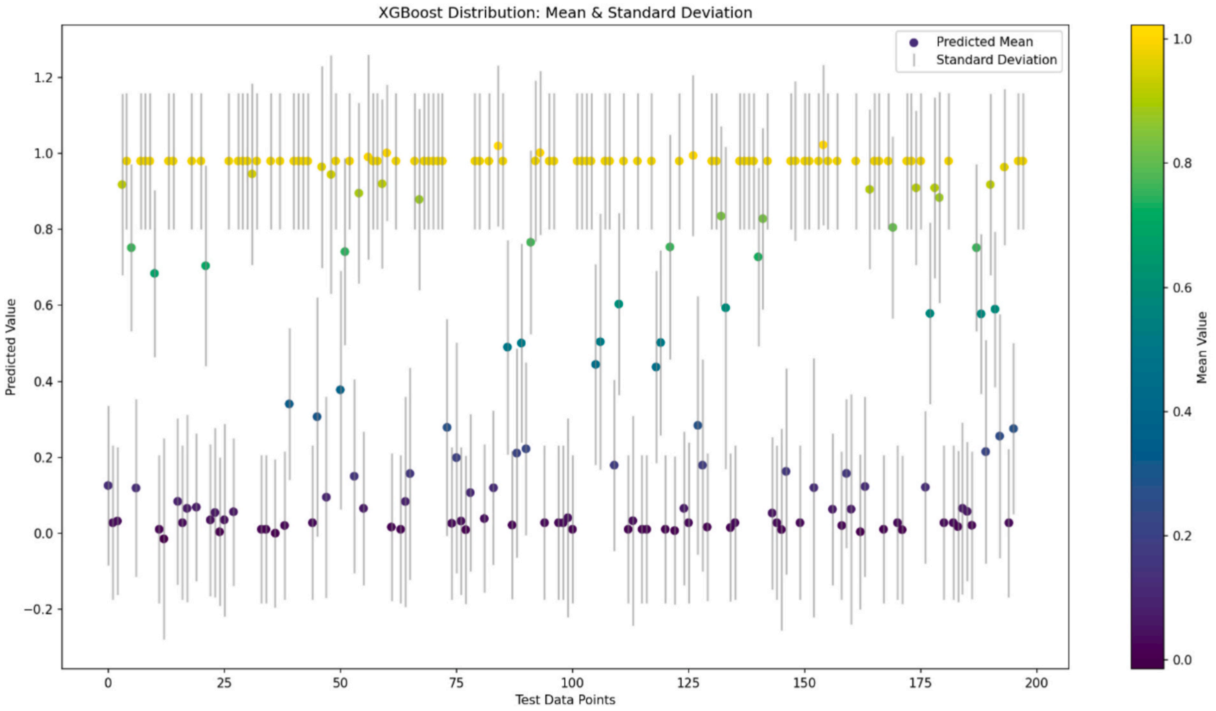


Fig. 15. The distribution visualization of XGBoost (mean and standard deviation).

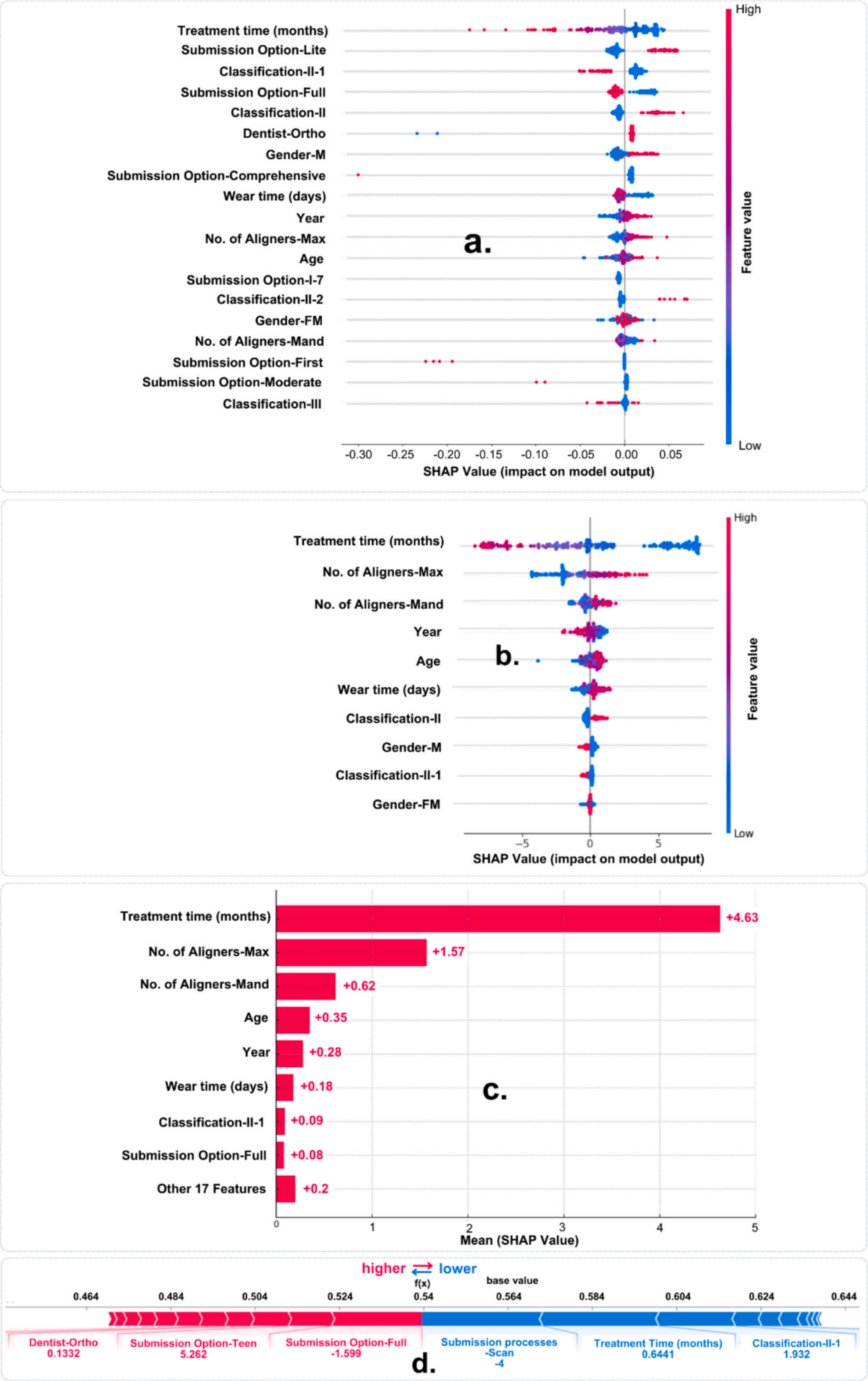


Fig. 16. SHAP analysis (a. SHAP value analysis of ANN, b. SHAP value analysis of XGBoost, c. SHAP relative importance Bar chart of XGBoost, and d. SHAP force analysis of ANN.

Table 3
An Overview of Performance Metrics.

	ML	Accuracy	Sensitivity	Specificity	F1-Score
1	DT	0.9091	0.9423	0.87	0.9159
2	RF	0.8939	0.9231	0.86	0.9014
3	ANN	0.8434	0.8846	0.8	0.8558
4	XGBoost	0.9434	0.9712	0.9036	0.9439

responsiveness, as corroborated by various studies emphasizing the accelerated tooth movement in younger patients. Conversely, the scan submission process negatively impacts outcomes, potentially due to training and quality control issues, as well as the complexities involved in correcting Class II-1 protrusion cases that might require tooth extraction or specialized techniques. The SHAP force analysis also indicates that comprehensive cases with extended treatment times and full

Table 4
Confusion matrix analysis.

Model	TP	TN	FP	FN	Insight and interpretation
DT	0.94	0.87	0.13	0.058	• High accuracy in predicting treatment success (94 %) and failure (87 %). • Low false positive and negative rates indicate reliable performance. • The model's simplicity aids interpretability but may overfit training data, affecting generalization in real-world scenarios.
RF	0.92	0.86	0.14	0.077	• High accuracy but slightly lower than DT • Balances bias-variance trade-off better due to ensemble learning. • The complexity of multiple trees can reduce interpretability, necessitating feature importance analysis for clarity.
ANN	0.88	0.80	0.20	0.12	• Good accuracy but outperformed by tree-based models. • Black-box nature poses challenges for interpretability, requiring techniques like SHAP for better insights. • Performance could be enhanced with more data and fine-tuning hyperparameters.
XGBoost	0.97	0.90	0.096	0.029	• Highest true positive and true negative rates. • Strong predictive power and lower error rates highlight robustness. • Gradient boosting algorithm's ability to handle various data types and imbalances makes it ideal, but interpretability must be addressed through SHAP analysis.

Table 5
Tree-based and SHAP analysis insight.

Model	Key features (identified)	Significant nodes / SHAP features	Clinical interpretation and proposed assumption
DT	• Treatment time • Age • No. of aligners	• Treatment Time $\leq$ 11.25 months • Age $\leq$ 10.5 years • Treatment Time $\leq$ 6.5 months • No. of Aligners (U) $\leq$ 29.5 • No. of Aligners (U) $>$ 75.0	• Success (prone to) : Patent adherence and compliance issues o Age $\leq$ 10.5 years ■ Shorter treatment times (6.5 years) o Treatment time $\leq$ 11.25 months o No. of aligners ■ U $\leq$ 29.5 • Fail (prone to) o No. of aligners ■ U $>$ 75.0
RF	• Treatment time • No. of aligners • Wear time	• Treatment Time $\leq$ 11.25 months • No. of Aligners (L) $\leq$ 12.5 • Wear Time (days) $\leq$ 6.0	: Balancing of functional movement issues and patient compliance • Success (prone to) o No. of aligners ■ L $\leq$ 12.5 o Wear time $\leq$ 6.0
ANN	• Submission Option • Submission processes • Classification • Treatment time	• Submission Option-Teen, Lite, Full • Submission processes-SCAN • Classification-II-1 • The higher treatment time	(SHAP-Force) : Biological responsiveness (SHAP) : Lesser tooth movement and special intervention or technique (SHAP-Force) : Training and quality control issues : Protrusion correction, tooth extraction or special techniques involvement : Comprehensive case and patient engagement issue (SHAP) • Impact (+ve) o Teen submission option o Lite submission option • Impact (−ve) o Scan submission process o Class II-1 o Treatment time o Treatment time o Class II-1 o Full Submission option
XGBoost	• Treatment time • No. of aligners • Submission Option • Scan process	• The higher treatment time • The higher no. of aligners	: Comprehensive case and patient engagement issue • Impact (+ve) o No. of aligner ■ U ■ L • Impact (−ve) o Treatment time

Table 6
A comprehensive insight and real-world application.

Aspect	Insight summary	Clinical implications	Research value/impact
Interpretability / Explainability	• SHAP makes XGBoost interpretable • Highlights key factors	• Emphasizes patient age in planning • Fosters trust	• Benchmark for ML in orthodontics precision plan • Bridges theory and practice
Ethical AI	• SHAP ensures transparency	• Tailors treatments o Fair o Effective	• Fair, transparent planning • Trust in AI
Clinical implications	• Consistent factor (across models) o Duration o Aligners o Compliance	• Focus on reducing o Times o aligners • Ensure adherence	• Optimizes treatment • Continuous monitoring
Predictive features	• Key features (across models) o Treatment time o Aligners o Wear time	• Prioritize for better outcomes	• Critical features for efficiency
Biological factors	• “Teen” option (higher success rate) • Suggests age advantages	• Leverage young patients’ responsiveness	• Highlights age in planning • Personalized care
Submission process	• Negative SHAP • Scan process issue	• Improve methods for accuracy	• Refine methods for better outcomes • Scan protocol revision
Model performance	• XGBoost leads, • Followed by o DT o RF o ANN	• Prefer models o Strong o Clear	• Validates ML in planning • Improves o Decision-making o Patient experience (potentially)

submission options are prone to engagement issues, further highlighting the need for tailored interventions.

The ability of our ML models to predict optimal points for intervention, such as the specific 11.25-month mark identified by the RF model, allows for more effective scheduling of adjustments and follow-ups compared with routine appointment scheduling, which in many clinical protocols occurs at fixed intervals (commonly every 6–12 weeks) based on clinician judgment and manufacturer guidance rather than individualized predictive thresholds [45]. By offering interpretable, data-driven intervention thresholds, our approach reduces the necessity for prolonged treatment durations and enhances patient adherence and satisfaction. Such precision in prediction supports improved treatment outcomes, minimizing unnecessary refinements and enhancing efficiency across the care pathway. Additionally, these insights, as shown in Table 5, assist orthodontic technicians in streamlining the planning and production processes, potentially reducing costs and improving the delivery efficiency of customized aligners.

Despite these advances, the challenge remains to make these complex ML models accessible and comprehensible to all orthodontic professionals. Continuous efforts to enhance the user-friendliness of these technologies are crucial for their ethical integration into daily clinical practice. Ensuring that these tools are used responsibly and effectively requires ongoing education and adaptation within the orthodontic community.

The methodological framework of this study also highlights the complementary roles of DSR and CRISP-DM. DSR functions as the overarching research methodology, ensuring that model development addresses clinically relevant problems and contributes to knowledge through rigorous evaluation and communication. CRISP-DM, in turn, provides the operational workflow for implementing the technical stages of data mining—data understanding, preparation, modeling, and evaluation. By embedding CRISP-DM within the broader DSR cycle, the study achieves both methodological rigor and technical reproducibility, strengthening its relevance to orthodontic practice.

Looking to the future, the potential of these models to incorporate real-time data and dynamic feedback mechanisms could revolutionize orthodontic treatment, making it not only reactive but also predictive. This evolution would enable treatments to be continuously adjusted in response to patient progress, contributing a significant leap towards truly personalized medicine in orthodontics.

This research significantly advances the field by setting new standards for orthodontic care through improved treatment planning driven by ML interpretability. As we continue to refine these models and expand their applications, the future of orthodontics potentially transforms into a precision-oriented, patient-centered practice. This shift not only promises to enhance clinical outcomes and patient satisfaction but also establishes a new benchmark for the integration of technology and ethics in healthcare, ensuring that every treatment plan is both scientifically sound and centered on the highest standards of patient care.

This study introduces an **architectural framework** that leverages machine learning models, such as XGBoost and artificial neural networks, in combination with SHapley Additive exPlanations (SHAP) to enhance the interpretability of treatment predictions. The process emphasizes the integration of ethical AI practices, focusing on transparency, accountability, and patient-centred care. This framework provides orthodontists with detailed insights into the factors influencing treatment success, offering a decision-support tool that aligns with ethical principles of fairness and explainability.

The **process-level contributions** include a proposed methodology for incorporating XAI models into clinical workflows. This involves preprocessing patient data, generating interpretable predictions, and presenting the results through clinician-friendly interfaces. The process also integrates iterative feedback loops to refine predictions based on real-time patient data, which sets a theoretical foundation for future adaptive clinical systems.

While XAI has been applied in broader medical domains, such as imaging diagnostics and predictive analytics [46,47]Its use in orthodontics, particularly for clear aligner treatment planning, represents an innovative contribution. Related work by Suh et al. demonstrates the utility of XAI for early screening of periodontitis using both deep learning and traditional ML techniques [48]. In contrast, our study applies SHAP-based explanations to orthodontic treatment planning, linking predictive insights to treatment timing and scheduling decisions rather than periodontal disease detection. The novelty of this study lies in adapting XAI to caries index detection [49], addressing the interpretability challenge posed by complex machine learning models in a domain where trust and transparency are paramount.

However, this contribution is still at the theoretical stage, as the proposed architecture and processes have not been validated in real clinical settings. Future research is needed to implement and test these concepts in practice, involving stakeholders such as orthodontists and patients, to ensure their practical feasibility and to refine the framework based on real-world feedback. This gap presents an opportunity for subsequent studies to bridge the theoretical foundation established here with practical applications that could revolutionize treatment planning in orthodontics.

5. Limitations and future Directions

While our study has revealed important insights into the predictive power of machine learning models for Invisalign treatment success, the limited scope of our dataset necessitates a cautious interpretation of its generalizability. Sourced from five private dental clinics in Thailand, our dataset, though rich in its contextual specificity, may not fully encapsulate the broader, global variability in orthodontic treatment responses. This regional dataset provides invaluable insights into the effectiveness of Invisalign within a specific cultural and demographic context, significantly contributing to the discourse on orthodontic treatment modalities. However, the socioeconomic and genetic diversity across different populations can profoundly impact the general applicability of our findings. Factors such as local dietary habits, oral hygiene practices, and access to dental care vary widely, potentially influencing the effectiveness and outcomes of orthodontic treatments like Invisalign. Acknowledging these limitations is crucial, and as such, we suggest that future research should aim to incorporate a more diverse array of datasets from various global regions. Such studies would help validate and potentially enhance the accuracy and applicability of the predictive models, ensuring that the benefits of advanced orthodontic practices can be more universally realized and adapted to meet a variety of patient needs across different geographical and cultural landscapes.

Our study has provided substantial insights into the initial outcomes of orthodontic treatments using advanced machine-learning techniques. However, a significant limitation is the focus on short-term results without an evaluation of long-term treatment stability and effectiveness. Orthodontic interventions, such as those involving Invisalign, are not only about achieving immediate alignment but also ensuring that these results are stable and enduring over time.

The predictive models currently developed base their assessments on data that capture outcomes shortly after treatment completion.

Longitudinal data are crucial in orthodontics, where the success of a treatment is measured not just by immediate alignment but by the stability of these outcomes over several years. Factors such as patient compliance with retainers, biological responses to treatment, and natural age-related changes can significantly influence long-term results.

To address this critical limitation, it is imperative to incorporate longitudinal studies into future research. We recommend the inclusion of long-term follow-up data in subsequent studies to assess the effectiveness and reliability of the predictive models over an extended period. Such data would allow us to evaluate the true efficacy of the treatments in maintaining dental health and alignment, providing a more comprehensive understanding of the models' utility in clinical practice.

In future iterations of this research, we aim to collaborate with orthodontic clinics to systematically collect follow-up data post-treatment, potentially over several years. This approach will not only enhance the validity of our predictive models but also ensure that they are robust and reliable for clinical application. By extending our research to include these longitudinal elements, we can better understand the dynamics of orthodontic treatments and refine our models to predict outcomes that genuinely reflect long-term treatment success and patient satisfaction.

6. Conclusion

This research marks a significant advancement in orthodontic treatment planning by integrating ML models. Our investigation validates the robust predictive capabilities of these models and highlights the critical role of treatment time and aligner count in determining Invisalign's success. The use of XAI through SHAP offers unprecedented clarity on influential factors, ensuring our predictive models are transparent and ethically grounded.

The high performance of XGBoost, with its strong predictive power and lower error rates, underscores its suitability for clinical applications. However, the interpretability provided by SHAP is essential for clinician trust and patient compliance. The insights gained from SHAP analysis, particularly the higher success rates in younger patients (Submission Option-Teen) and the need for improved submission processes, guide clinicians in optimizing treatment plans. Moreover, factors such as treatment time and aligner count are critical determinants of success. Notably, the analysis indicated higher success rates among younger patients, suggesting a potential area for a focused clinical strategy. These findings offer clinicians actionable guidance to optimize treatment plans, enhancing both the efficacy of the treatments and patient compliance.

For future research, we recommend exploring the application of these ML models in broader demographic and geographic contexts to enhance the generalizability of our findings. Additionally, further studies should investigate the integration of real-time data into these models to dynamically adjust treatment plans as new data becomes available. This could significantly improve the precision of treatment predictions over the course of therapy.

By aligning advanced ML technologies with personalized patient care, our study sets new benchmarks for clinical excellence. The ethical application of these technologies ensures that every therapeutic decision is transparent, understandable, and centered on the highest standards of patient care, ultimately enhancing treatment outcomes and patient satisfaction.

Summary table

What was already known on the topic?

- Machine Learning (ML) applications in orthodontics have predominantly focused on predicting treatment duration, identifying anatomical landmarks, and aiding in diagnostic plans using traditional statistical methods.
- The use of ML in the field of orthodontics, particularly with clear aligners like Invisalign, has been limited, with few studies leveraging advanced ML techniques for the orthodontic treatment planning phase.
- Existing research has rarely integrated real patient datasets from specific regions or populations in orthodontics, often relying on non-clinical records or solely sourced data.

What has this study added to our knowledge?

- This study is the first to apply SHapley Additive exPlanations (SHAP) for enhancing the interpretability and explainability of ML models in the orthodontic field, providing a pioneering method to understand

and trust the decision-making processes of complex algorithms theoretically.
- This study introduced a novel approach by incorporating a diverse range of real patient data from Thailand into Machine Learning models. This significant advancement enables highly personalized and culturally orthodontic care, tailored specifically to the unique demographic and clinical characteristics.
- Demonstrated how ethical AI considerations can be embedded in ML development for healthcare, ensuring that the use of AI in orthodontics aligns with transparent and fair treatment practices, thereby setting new ethical standards in the orthodontic field.
- Our research has proven the effectiveness of ML in predicting the success of Invisalign treatment plans, moving beyond traditional applications to address patient-specific treatment dynamics, enhancing overall treatment efficacy, and potentially improving patient satisfaction.

CRediT authorship contribution statement

Sanisa Trakulmututa: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Khin Than Win:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Dataset acquisition and composition

This research utilizes a meticulously compiled dataset consisting of anonymized orthodontic treatment records from 657 cases. These records were sourced from five private dental clinics across Thailand, selected based on their adherence to qualified provider standards, robust data management practices, and a significant volume of case histories, ensuring geographic and demographic diversity.

The dataset encompasses a comprehensive array of data points drawn from multiple sources:

1. Patient Demographics: Includes essential details such as the year treatment began, patient age, and gender.
2. Dentist Qualifications: Records the level of specialization of the dentist responsible for each treatment, differentiating between general practitioners and orthodontists.
3. Aligner Brand Information: Specifies the brand of clear aligners used, primarily focusing on Invisalign among others.
4. Dental Classification: Documents the orthodontic classification of each patient, ranging from Class I to Class III malocclusions.
5. Treatment Submission Method: Details the method used for submitting orthodontic treatment plans, which includes digital scanning, Polyvinyl Siloxane (PVS) impressions, and other traditional techniques.
6. Treatment Plan Type: Identifies the specific type of Invisalign treatment plan used, such as Full, Lite, I-7, or Teen options.
7. Duration of Treatment: Captures the treatment duration as planned by the software corresponding to the actual delivery of aligners which excludes the refinement period.
8. Aligner Count: Enumerates the number of aligners used during the treatment, detailing usage across initial and subsequent phases.
9. Wearing Duration: Measures how long each aligner was worn during the treatment phases and calculates the average wearing time.
10. Outcome of Treatment: Classifies the treatment outcomes into ’Success’ or ’Fail’, where ’Success’ aligns with the initial treatment plan and ’Fail’ indicates the necessity for additional aligners.

Appendix B. . Study design and participant selection methodology

Data collection

For this investigation, we utilized a dataset consisting of 657 anonymized patient records from individuals who underwent Invisalign treatments across various private dental practices. These records were used with appropriate patient consent for research purposes.

Selection of participants

Our study specifically included patients treated with Invisalign at multiple private clinics, adhering to the following criteria:

Criteria for inclusion

The inclusion criteria for our study were defined to ensure high data integrity and robust analysis. We only included patients with complete orthodontic records spanning from the initiation to the completed case of Invisalign treatment. This inclusion ensures comprehensive data on treatment progression, outcomes, and any necessary mid-course corrections or adjustments. Furthermore, each case had to have verifiable data points, including both planned and actual treatment durations, and the number of aligners used, ensuring that the data is both reliable and reflective of true treatment scenarios.

Clinic variability

To ensure a comprehensive representation, the dataset includes records from a variety of private dental clinics, each serving diverse patient demographics and employing distinct treatment methodologies.

Dental classification

Diversity The study encompasses patients across various dental classifications, such as Class I, Class II-1, Class II-2, and Class III, to encompass a broad spectrum of orthodontic conditions.

Appendix C. . Attribute selection

Orthodontic considerations

The chosen attributes for this study are essential for understanding the dynamics of orthodontic treatments. Patient demographics like age and gender are known to influence both the progression and outcomes of such treatments. Younger patients generally exhibit faster tooth movement due to more responsive biological processes [50,51], while gender can affect both treatment expectations and clinical outcomes [52,53].

The year of treatment initiation serves as a proxy to evaluate the evolution and adoption of new technologies and methodologies in orthodontics over time [54]. The dentist's specialization is also critical; orthodontists often handle more complex cases differently than general dentists, which can significantly influence treatment approaches and outcomes [55,56]. A study by Fabrizia et al. noted discernible differences in case management, especially in the selection for class I malocclusions with specific complications [55].

Dental classification of the patient (e.g., Class I, Class II) is crucial for assessing the complexity of cases and tailoring treatment strategies accordingly [57]. Similarly, the method of submitting treatment plans, whether through digital scans or traditional impressions, directly affects the precision of aligners and the effectiveness of treatment outcomes [58,59].

Data specific to Invisalign treatments, such as the number of aligners and duration of treatment, provide insights into case complexity. Research indicates that more complex cases often require more aligners and are prone to adjustments and refinements [60,61]. The type of Invisalign plan (Full, Lite, I-7, Teen) also reflects the tailored approach to varying case complexities and patient needs [62,63].

Treatment duration assessments are vital for validating the accuracy and effectiveness of orthodontic planning tools [64,65]. Additionally, patient compliance, reflected through aligner wear time, critically affects treatment outcomes. Non-adherence to prescribed schedules can significantly prolong treatment or diminish results [66,67].

Lastly, treatment outcomes such as success rates and the necessity for refinements serve as direct indicators of treatment efficacy. Evaluations of these outcomes are indispensable for clinicians to refine their treatment protocols and enhance the quality of care [10,139], as noted in studies by Soukaina et al. [68].

Machine learning considerations

In the field of machine learning within information technology, the attributes selected for analysis hold critical importance for several compelling reasons [69,70]. The integrity and depth of the data, which include factors like consistency and completeness, are pivotal for the effective training and predictive accuracy of machine learning models. These models depend on robust data to discern patterns effectively and apply these insights to new, unseen scenarios [71].

Attributes such as the duration of treatment and the number of aligners used are vital for enabling predictive models such as DT and RF to deliver precise forecasts regarding orthodontic treatment outcomes [7]. The variety and range of these attributes also significantly enhance the strength and dependability of the machine learning models employed. A broad dataset ensures comprehensive training and validation of these models, which, in turn, bolsters their ability to predict outcomes with greater accuracy [72,73].

Moreover, the detailed attributes under study are crucial for the customization and personalization of treatment plans. Utilizing these data points, machine learning algorithms can refine their recommendations, thereby optimizing the efficacy and efficiency of treatments like Invisalign for individual patients [62].

Appendix D. . Data preprocessing and encoding

Our research utilized advanced machine learning techniques to develop predictive models for orthodontic treatment outcomes. We chose Python for its robust ecosystem of data science libraries, making it an excellent platform for our analysis. The data, derived from de-identified records from private dental clinics across Thailand, underwent rigorous preprocessing to ensure its readiness for effective analysis and modeling. An essential part of this preprocessing was the transformation of categorical variables into numerical formats through one-hot encoding. This encoding method is crucial as machine learning algorithms require numerical input to perform optimally, ensuring both enhanced interpretability and model accuracy [32,33].

Appendix E. . Missing data handling

In our dataset, missing data was a notable challenge, particularly in the variables “Treatment Time” and “Number of Aligners.” To mitigate the potential biases and preserve the integrity of our dataset, we employed median imputation. This method was selected due to its resilience against outliers, ensuring that the extreme values did not skew the results [34]. This approach is particularly suitable for our clinical dataset, where the distribution of data can be asymmetrical and outliers are common [35].

For the variables in question, missing values constituted approximately 1.52 % for “Treatment Time” and 0.76 % for “Number of Aligners.” Handling these missing entries effectively was crucial since both attributes are significant predictors of Invisalign treatment success. The use of median imputation involved replacing missing values with the median of existing values for each attribute, thus maintaining the central tendency of the data distribution without the influence of outliers.

This method not only helped preserve the full dataset for analysis but also ensured the robustness of our statistical findings by maintaining the sample size and preventing the biases that could arise from listwise deletion. Listwise deletion could potentially lead to a reduction in sample representativeness, especially if the missingness is systematically related to other key variables [36]. By opting for median imputation, we upheld the accuracy and generalizability of our study's outcomes, adhering to recommended practices in clinical and epidemiological research for handling missing data [37].

Incorporating this method aligns with our commitment to upholding the highest standards of data integrity and reliability in our predictive modeling, essential for ensuring that our findings are both scientifically valid and practically applicable in clinical settings. This step is part of our broader data preprocessing efforts, which are critical in laying a solid foundation for the successful application of machine learning techniques in our research.

Appendix F. . Machine learning development and design science research (DSR)

This project leverages the Design Science Research (DSR) methodology to develop predictive models that refine the precision of Invisalign treatment outcomes [74]. DSR's systematic approach begins with identifying the core challenge: enhancing the accuracy of predictive models in orthodontics. A thorough review of orthodontic practices, particularly Invisalign, illuminates the existing gaps in predictability and model sophistication. Objectives are then set to improve treatment predictability and efficiency using advanced machine learning technologies, with clearly defined metrics such as accuracy, precision, and clinical applicability guiding the research.

The design and development phase transforms the established theoretical frameworks into practical applications by creating machine learning algorithms tailored for orthodontic data. Within this DSR cycle, the Cross-Industry Standard Process for Data Mining (CRISP-DM) is applied as the structured process model to manage data mining operations. This involves selecting suitable algorithms and employing the CRISP-DM to effectively manage data mining operations, ensuring robust, scalable, and real-world applicable models. Subsequently, these models undergo rigorous evaluation through demonstrations and statistical analyses to verify their accuracy in predicting treatment outcomes, employing cross-validation techniques to ensure reliability.

The final phase involves communicating the findings to both the academic community and clinical practitioners. This includes detailed reports, publications, and presentations at conferences, emphasizing the practical implications of integrating these advanced predictive models into clinical practice. By positioning CRISP-DM as the operational framework nested within the broader DSR methodology, this project ensures both methodological rigor and technical robustness. This ongoing cycle of development and refinement aims to continuously enhance the predictive accuracy and utility of the models, ensuring they meet both scientific standards and clinical needs in improving orthodontic treatment decision-making.

Appendix G. . Machine learning development and cross-industry standard process for data mining (CRISP-DM)

The CRISP-DM serves as the structural backbone for our study, specifically tailored to address the unique complexities of data from Invisalign treatments and aiming to improve predictions of treatment success. The methodology begins with a thorough business understanding phase, where the impact of predictive modeling on orthodontic practices is assessed by engaging with dental professionals and reviewing data on Invisalign outcomes. This step sets a critical business objective: enhancing the success rates of treatments through sophisticated analytics. In the data understanding phase, an exhaustive analysis of variables like treatment duration, number of aligners, and patient compliance is conducted to identify key factors influencing outcomes, thereby providing deep insights into treatment success and necessary revisions.

The subsequent data preparation phase involves cleaning and transformation of raw data to ensure it is suitable for machine learning analysis. This includes addressing incomplete patient records and converting categorical data, such as dental classifications and aligner types, into a format conducive to algorithmic processing. Modeling then proceeds with the selection of algorithms tailored for their efficacy in healthcare contexts: DT for transparency, RF, and XGBoost for advanced analysis, and neural networks to capture complex patterns. Each model undergoes rigorous evaluation using k-fold cross-validation, among other metrics like precision, recall, and AUC, to thoroughly assess and refine their predictive capabilities.

Finally, the deployment phase integrates successful models into a Clinical Decision Support System (CDSS), crafted for operational use within the orthodontic community. This system leverages predictive models to provide custom treatment plans, thereby enhancing the predictability and success rates of Invisalign treatments. The integration of CRISP-DM with CDSS exemplifies a model of enhanced clinical decision-making founded on comprehensive data analysis and aligned with strategic objectives for improved treatment outcomes. The systematic application of DSR and CRISP-DM ensures that the predictive models developed are scientifically rigorous and practically applicable, ready to be implemented in orthodontic settings to significantly advance patient care.

Appendix H. . Machine learning development and approaches

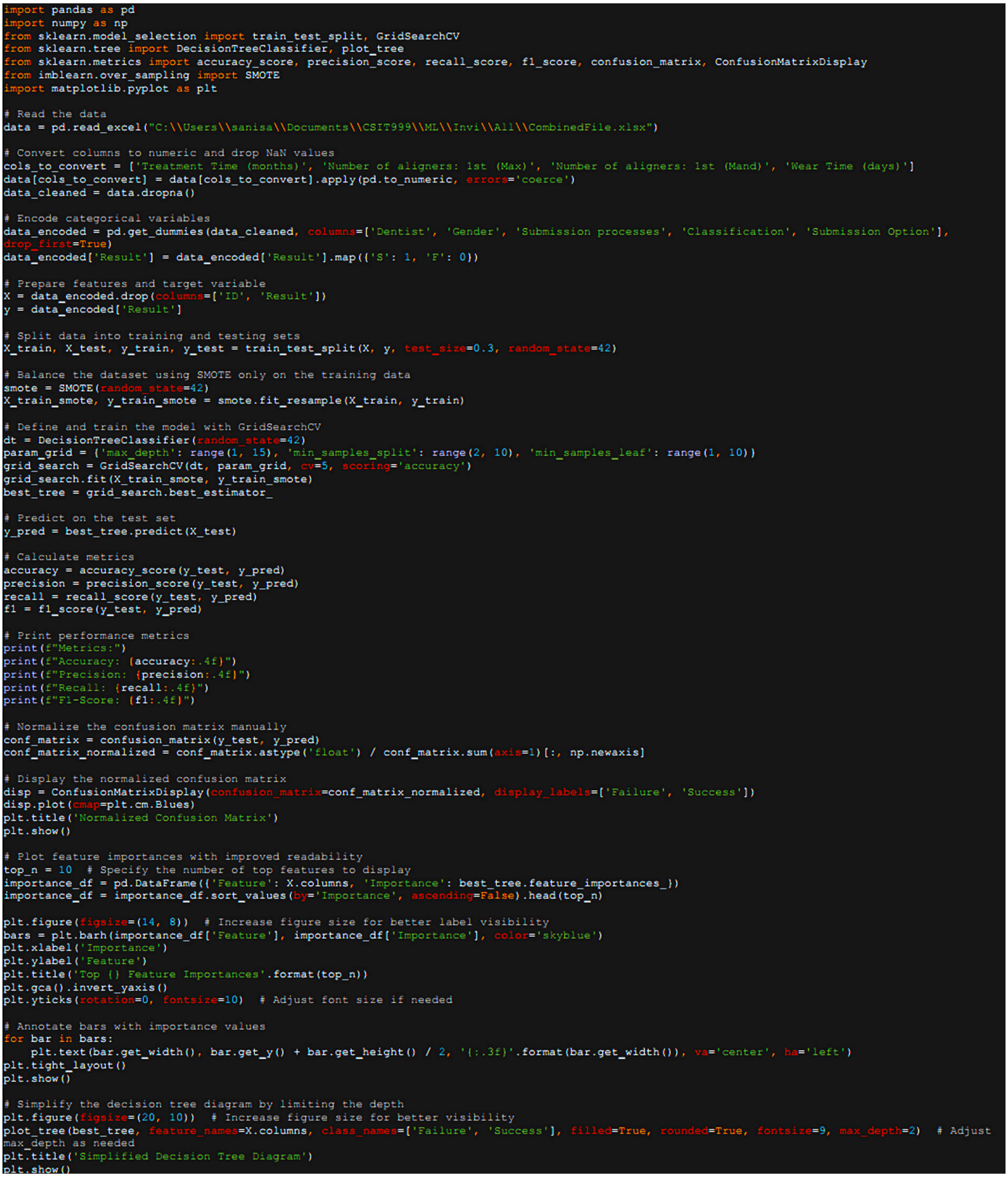
To build robust machine learning models for this project, Python was selected for its comprehensive support for data analysis and its rich ecosystem of libraries tailored for machine learning applications. This choice aligns with the methodologies of Design Science Research (DSR), which provides the overarching research framework, and the Cross-Industry Standard Process for Data Mining (CRISP-DM), which operationalizes the technical workflow. Together, they ensure a systematic and reproducible approach to model development and evaluation.Key to the data management phases of our project is the use of Pandas, a library that excels in data manipulation and cleaning via DataFrame objects [75]. This functionality is crucial for organizing complex datasets into manageable structures, ensuring the data integrity necessary for effective model development and evaluation.

Scikit-learn is another pivotal library utilized in this study, providing extensive machine-learning functionalities that support classification, regression, and clustering. It offers tools like train_test_split for segmenting data into training and testing sets, and performance metrics such as accuracy_score, precision_score, recall_score, and f1_score, which are integral for rigorous model evaluation. Scikit-learn's GridSearchCV is instrumental in optimizing model parameters and aligning model performance with the specific characteristics of our dataset.

For enhancing model transparency and accountability—critical in DSR—SHAP (SHapley Additive exPlanations) is employed. SHAP elucidates the decision-making processes of machine learning models, ensuring that the artefacts developed meet stakeholder expectations and are suitable for real-world application [76].

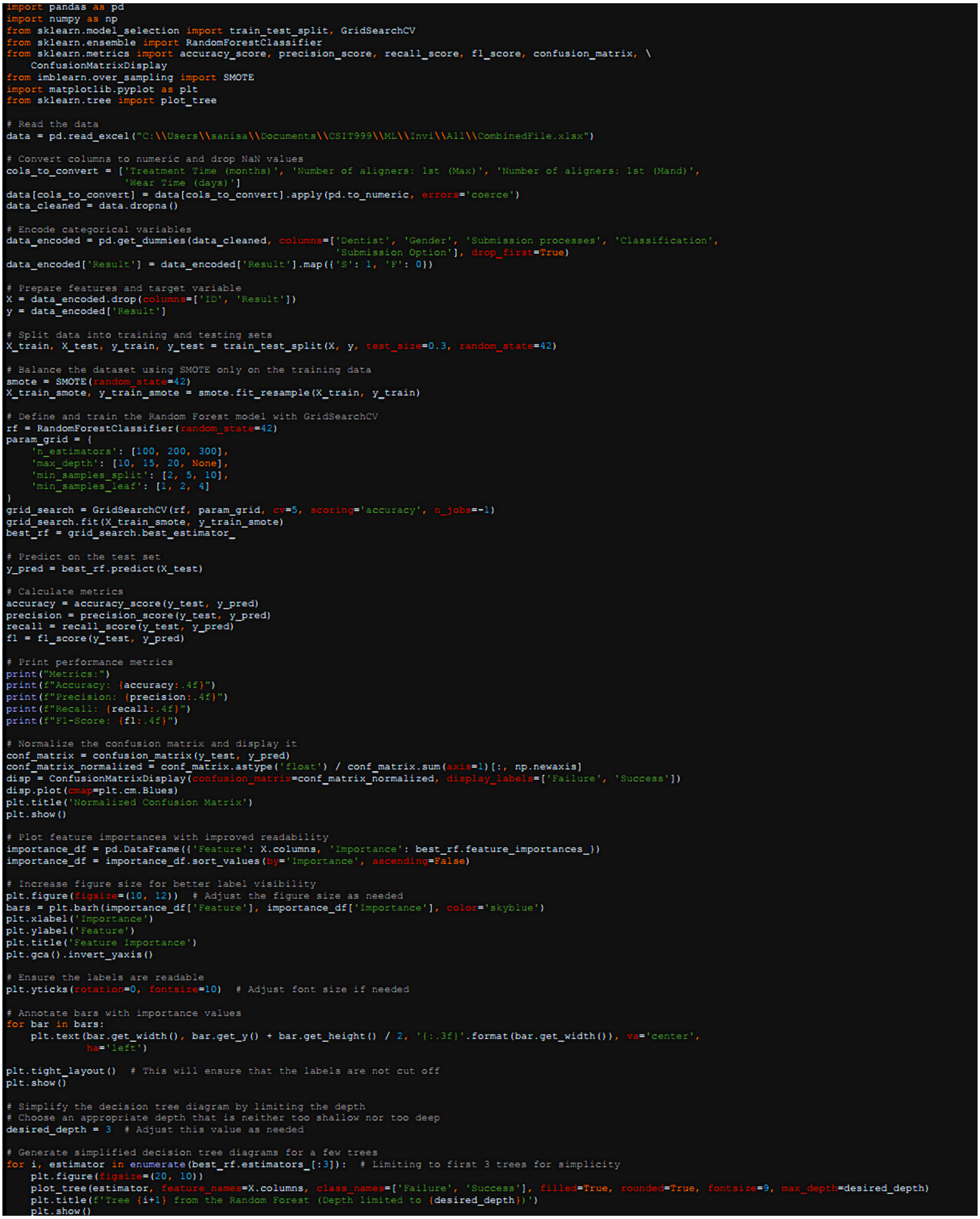
Visualization tools like Matplotlib are also integral, enhancing the presentation and understanding of research outcomes [77]. These tools facilitate the clear communication of complex results to both technical and non-technical stakeholders, supporting CRISP-DM's evaluation and deployment phases while reinforcing DSR's emphasis on transparency and communication.

Appendix I. . Python source code



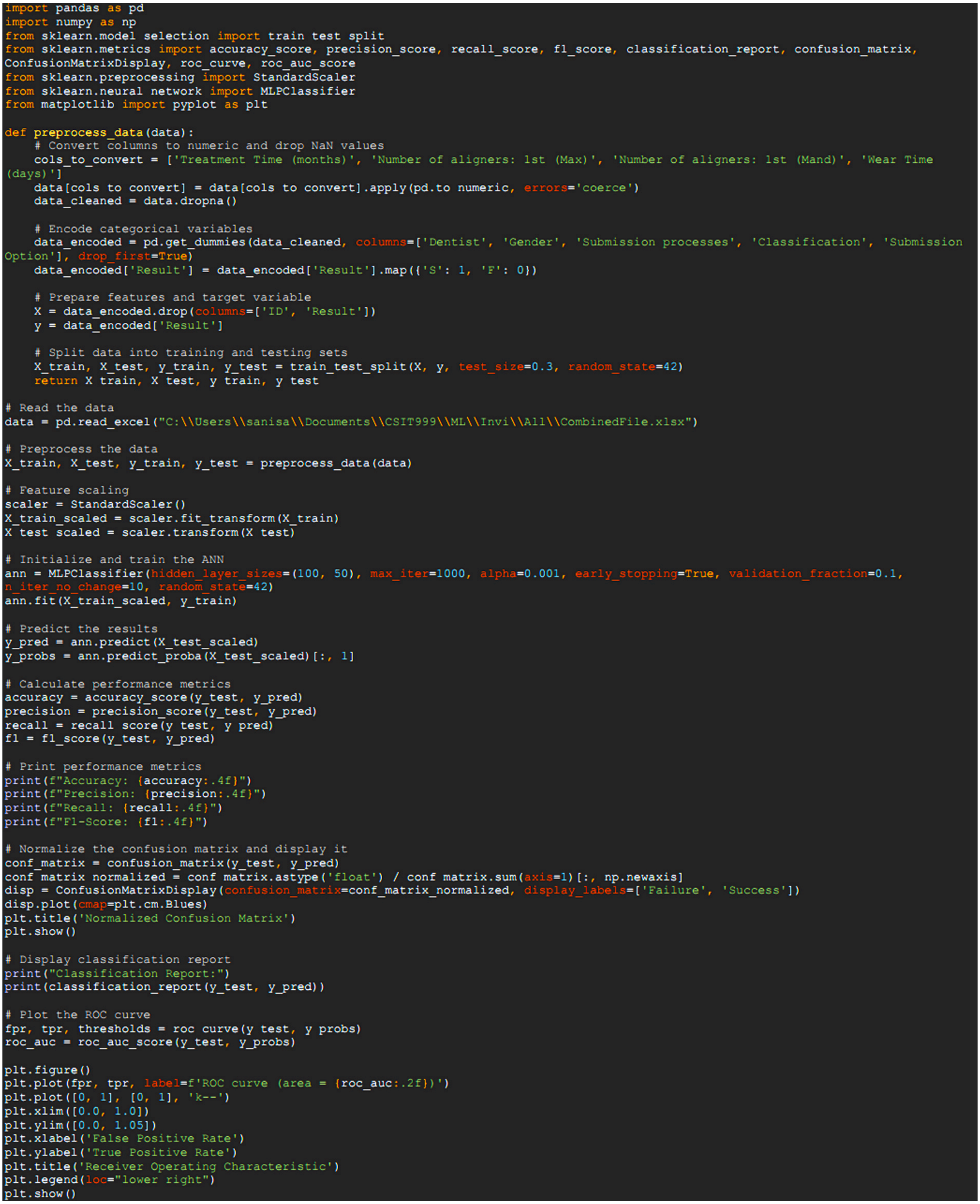
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, ConfusionMatrixDisplay
from imblearn.over_sampling import SMOTE
import matplotlib.pyplot as plt

# Read the data
data = pd.read_excel("C:\\Users\\sanisa\\Documents\\CSIT999\\ML\\Invi\\All\\CombinedFile.xlsx")

# Convert columns to numeric and drop NaN values
cols_to_convert = ['Treatment Time (months)', 'Number of aligners: 1st (Max)', 'Number of aligners: 1st (Mand)', 'Wear Time (days)']
data[cols_to_convert] = data[cols_to_convert].apply(pd.to_numeric, errors='coerce')
data_cleaned = data.dropna()

# Encode categorical variables
data_encoded = pd.get_dummies(data_cleaned, columns=['Dentist', 'Gender', 'Submission processes', 'Classification', 'Submission Option'],
drop_first=True)
data_encoded['Result'] = data_encoded['Result'].map({'S': 1, 'F': 0})

# Prepare features and target variable
X = data_encoded.drop(columns=['ID', 'Result'])
y = data_encoded['Result']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Balance the dataset using SMOTE only on the training data
smote = SMOTE(random_state=42)
X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)

# Define and train the model with GridSearchCV
dt = DecisionTreeClassifier(random_state=42)
param_grid = {'max_depth': range(1, 15), 'min_samples_split': range(2, 10), 'min_samples_leaf': range(1, 10)}
grid_search = GridSearchCV(dt, param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train_smote, y_train_smote)
best_tree = grid_search.best_estimator_

# Predict on the test set
y_pred = best_tree.predict(X_test)

# Calculate metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# Print performance metrics
print(f"Metrics:")
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-Score: {f1:.4f}")

# Normalize the confusion matrix manually
conf_matrix = confusion_matrix(y_test, y_pred)
conf_matrix_normalized = conf_matrix.astype('float') / conf_matrix.sum(axis=1)[:, np.newaxis]

# Display the normalized confusion matrix
disp = ConfusionMatrixDisplay(confusion_matrix=conf_matrix_normalized, display_labels=['Failure', 'Success'])
disp.plot(cmap=plt.cm.Blues)
plt.title('Normalized Confusion Matrix')
plt.show()

# Plot feature importances with improved readability
top_n = 10  # Specify the number of top features to display
importance_df = pd.DataFrame({'Feature': X.columns, 'Importance': best_tree.feature_importances_})
importance_df = importance_df.sort_values(by='Importance', ascending=False).head(top_n)

plt.figure(figsize=(14, 8))  # Increase figure size for better label visibility
bars = plt.barh(importance_df['Feature'], importance_df['Importance'], color='skyblue')
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.title('Top {} Feature Importances'.format(top_n))
plt.gca().invert_yaxis()
plt.yticks(rotation=0, fontsize=10)  # Adjust font size if needed

# Annotate bars with importance values
for bar in bars:
    plt.text(bar.get_width(), bar.get_y() + bar.get_height() / 2, '{:.3f}'.format(bar.get_width()), va='center', ha='left')
plt.tight_layout()
plt.show()

# Simplify the decision tree diagram by limiting the depth
plt.figure(figsize=(20, 10))  # Increase figure size for better visibility
plot_tree(best_tree, feature_names=X.columns, class_names=['Failure', 'Success'], filled=True, rounded=True, fontsize=9, max_depth=2)  # Adjust
max_depth as needed
plt.title('Simplified Decision Tree Diagram')
plt.show()
```

Fig. 9. The Python code of the decision tree developed model



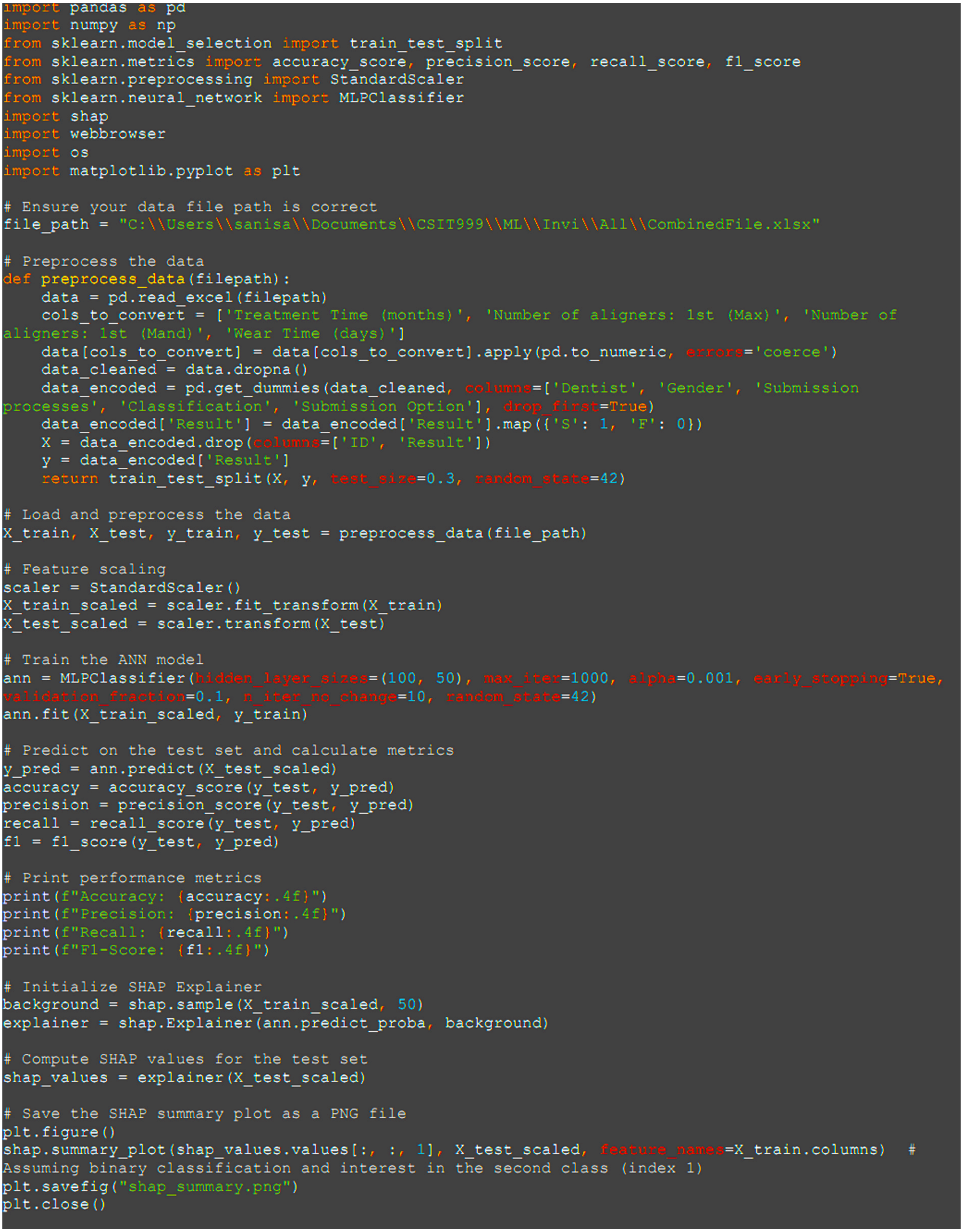
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, \
    ConfusionMatrixDisplay
from imblearn.over_sampling import SMOTE
import matplotlib.pyplot as plt
from sklearn.tree import plot_tree

# Read the data
data = pd.read_excel("C:\\Users\\sanisa\\Documents\\CSIT999\\ML\\Invi\\All\\CombinedFile.xlsx")

# Convert columns to numeric and drop NaN values
cols_to_convert = ['Treatment Time (months)', 'Number of aligners: 1st (Max)', 'Number of aligners: 1st (Mand)',
                   'Wear Time (days)']
data[cols_to_convert] = data[cols_to_convert].apply(pd.to_numeric, errors='coerce')
data_cleaned = data.dropna()

# Encode categorical variables
data_encoded = pd.get_dummies(data_cleaned, columns=['Dentist', 'Gender', 'Submission processes', 'Classification',
                                                     'Submission Option'], drop_first=True)
data_encoded['Result'] = data_encoded['Result'].map({'S': 1, 'F': 0})

# Prepare features and target variable
X = data_encoded.drop(columns=['ID', 'Result'])
y = data_encoded['Result']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Balance the dataset using SMOTE only on the training data
smote = SMOTE(random_state=42)
X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)

# Define and train the Random Forest model with GridSearchCV
rf = RandomForestClassifier(random_state=42)
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [10, 15, 20, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
grid_search = GridSearchCV(rf, param_grid, cv=5, scoring='accuracy', n_jobs=-1)
grid_search.fit(X_train_smote, y_train_smote)
best_rf = grid_search.best_estimator_

# Predict on the test set
y_pred = best_rf.predict(X_test)

# Calculate metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# Print performance metrics
print("Metrics:")
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-Score: {f1:.4f}")

# Normalize the confusion matrix and display it
conf_matrix = confusion_matrix(y_test, y_pred)
conf_matrix_normalized = conf_matrix.astype('float') / conf_matrix.sum(axis=1)[:, np.newaxis]
disp = ConfusionMatrixDisplay(confusion_matrix=conf_matrix_normalized, display_labels=['Failure', 'Success'])
disp.plot(cmap=plt.cm.Blues)
plt.title('Normalized Confusion Matrix')
plt.show()

# Plot feature importances with improved readability
importance_df = pd.DataFrame({'Feature': X.columns, 'Importance': best_rf.feature_importances_})
importance_df = importance_df.sort_values(by='Importance', ascending=False)

# Increase figure size for better label visibility
plt.figure(figsize=(10, 12))  # Adjust the figure size as needed
bars = plt.barh(importance_df['Feature'], importance_df['Importance'], color='skyblue')
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.title('Feature Importance')
plt.gca().invert_yaxis()

# Ensure the labels are readable
plt.yticks(rotation=0, fontsize=10)  # Adjust font size if needed

# Annotate bars with importance values
for bar in bars:
    plt.text(bar.get_width(), bar.get_y() + bar.get_height() / 2, '{:.3f}'.format(bar.get_width()), va='center',
             ha='left')

plt.tight_layout()  # This will ensure that the labels are not cut off
plt.show()

# Simplify the decision tree diagram by limiting the depth
# Choose an appropriate depth that is neither too shallow nor too deep
desired_depth = 3  # Adjust this value as needed

# Generate simplified decision tree diagrams for a few trees
for i, estimator in enumerate(best_rf.estimators_[:3]):  # Limiting to first 3 trees for simplicity
    plt.figure(figsize=(20, 10))
    plot_tree(estimator, feature_names=X.columns, class_names=['Failure', 'Success'], filled=True, rounded=True, fontsize=9, max_depth=desired_depth)
    plt.title(f'Tree {i+1} from the Random Forest (Depth limited to {desired_depth})')
    plt.show()
```

Fig. 10. The Python code of the random forest developed model



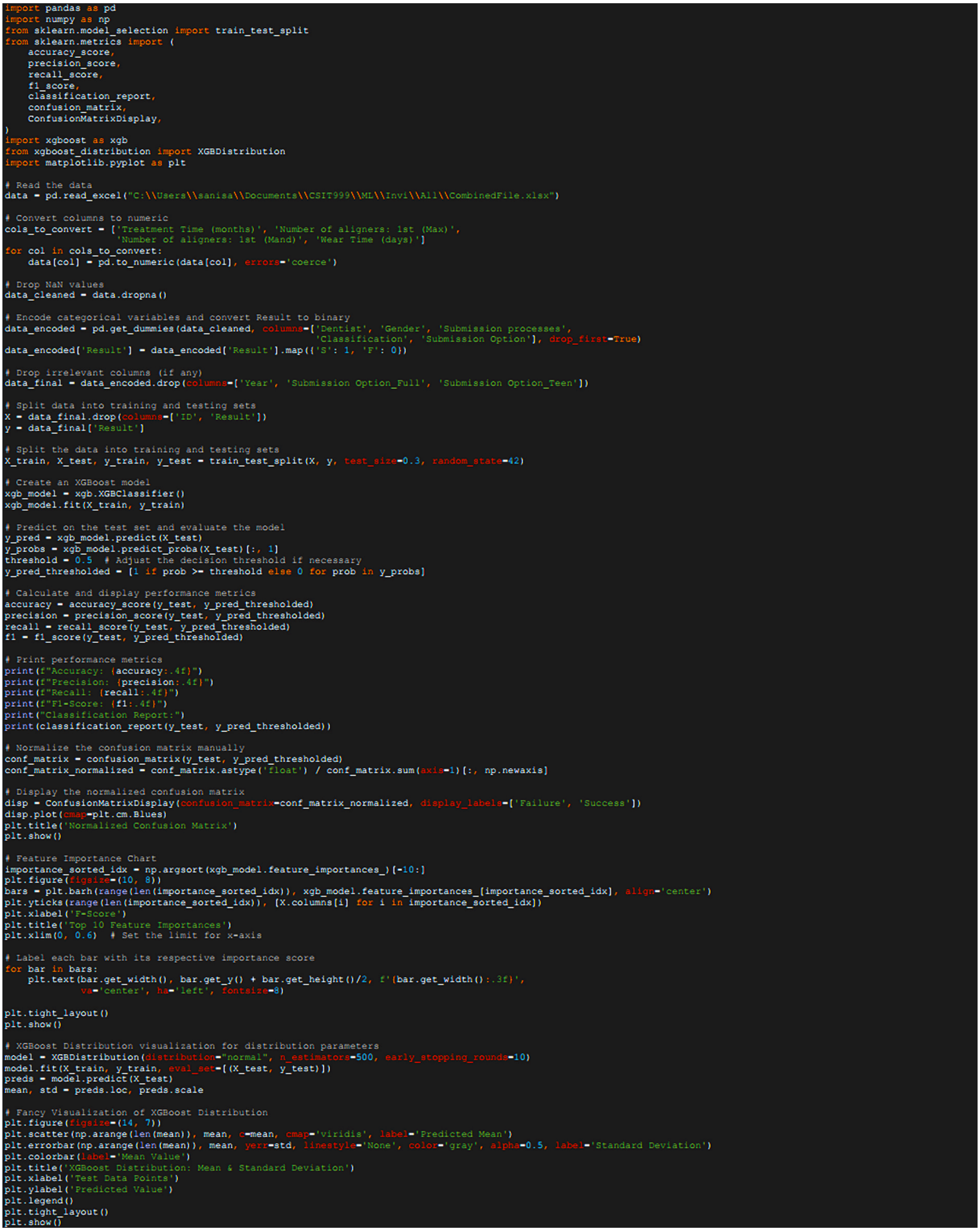
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report, confusion_matrix,
ConfusionMatrixDisplay, roc_curve, roc_auc_score
from sklearn.preprocessing import StandardScaler
from sklearn.neural_network import MLPClassifier
from matplotlib import pyplot as plt

def preprocess_data(data):
    # Convert columns to numeric and drop NaN values
    cols_to_convert = ['Treatment Time (months)', 'Number of aligners: 1st (Max)', 'Number of aligners: 1st (Mand)', 'Wear Time
(days)']
    data[cols_to_convert] = data[cols_to_convert].apply(pd.to_numeric, errors='coerce')
    data_cleaned = data.dropna()

    # Encode categorical variables
    data_encoded = pd.get_dummies(data_cleaned, columns=['Dentist', 'Gender', 'Submission processes', 'Classification', 'Submission
Option'], drop_first=True)
    data_encoded['Result'] = data_encoded['Result'].map({'S': 1, 'F': 0})

    # Prepare features and target variable
    X = data_encoded.drop(columns=['ID', 'Result'])
    y = data_encoded['Result']

    # Split data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
    return X_train, X_test, y_train, y_test

# Read the data
data = pd.read_excel("C:\\Users\\sanisa\\Documents\\CSIT999\\ML\\Invi\\All\\CombinedFile.xlsx")

# Preprocess the data
X_train, X_test, y_train, y_test = preprocess_data(data)

# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Initialize and train the ANN
ann = MLPClassifier(hidden_layer_sizes=(100, 50), max_iter=1000, alpha=0.001, early_stopping=True, validation_fraction=0.1,
n_iter_no_change=10, random_state=42)
ann.fit(X_train_scaled, y_train)

# Predict the results
y_pred = ann.predict(X_test_scaled)
y_probs = ann.predict_proba(X_test_scaled)[:, 1]

# Calculate performance metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# Print performance metrics
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-Score: {f1:.4f}")

# Normalize the confusion matrix and display it
conf_matrix = confusion_matrix(y_test, y_pred)
conf_matrix_normalized = conf_matrix.astype('float') / conf_matrix.sum(axis=1)[:, np.newaxis]
disp = ConfusionMatrixDisplay(confusion_matrix=conf_matrix_normalized, display_labels=['Failure', 'Success'])
disp.plot(cmap=plt.cm.Blues)
plt.title('Normalized Confusion Matrix')
plt.show()

# Display classification report
print("Classification Report:")
print(classification_report(y_test, y_pred))

# Plot the ROC curve
fpr, tpr, thresholds = roc_curve(y_test, y_probs)
roc_auc = roc_auc_score(y_test, y_probs)

plt.figure()
plt.plot(fpr, tpr, label=f'ROC curve (area = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()
```

Fig. 11. The Python code of the ANN-developed model



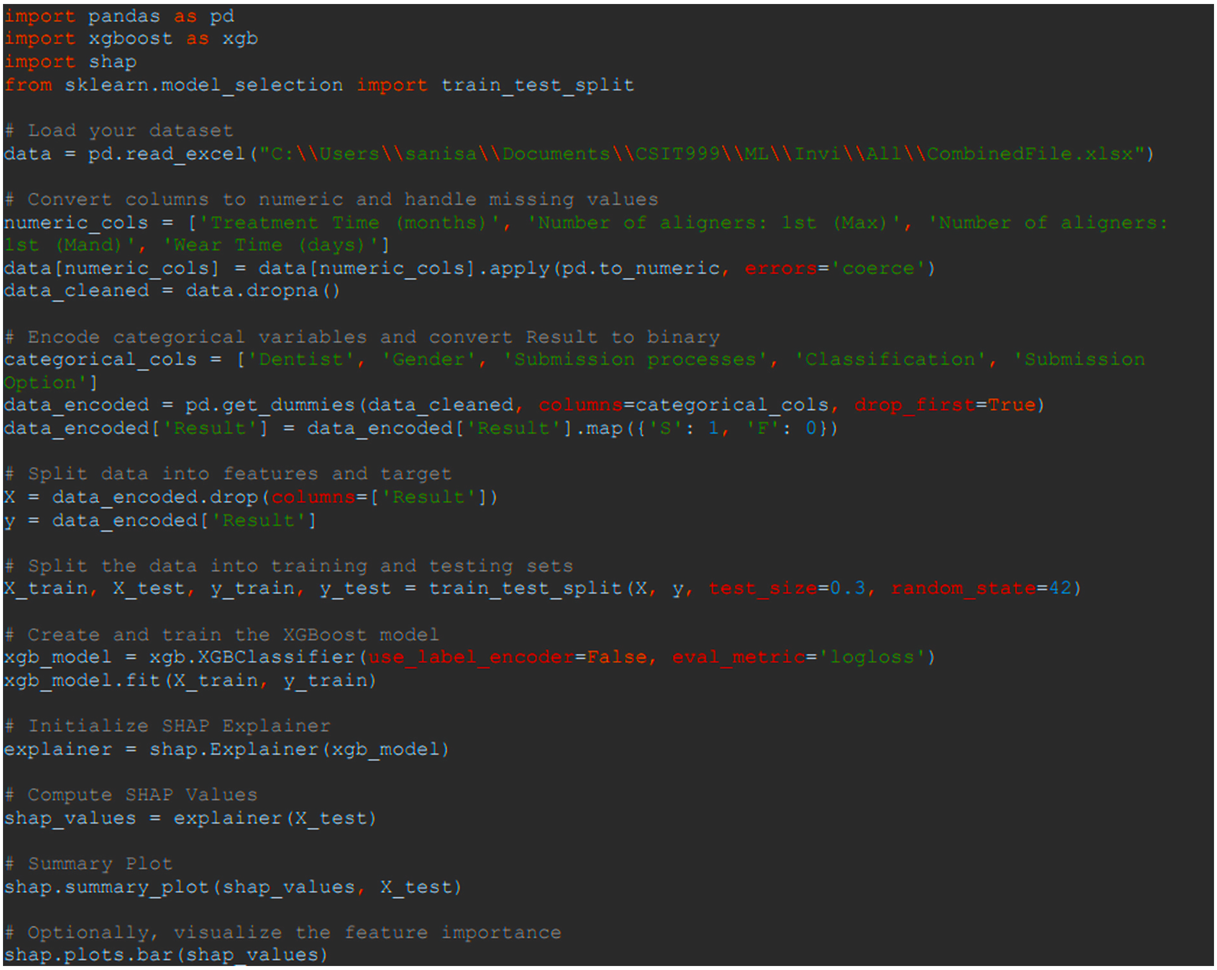
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.preprocessing import StandardScaler
from sklearn.neural_network import MLPClassifier
import shap
import webbrowser
import os
import matplotlib.pyplot as plt

# Ensure your data file path is correct
file_path = "C:\\Users\\sanisa\\Documents\\CSIT999\\ML\\Invi\\All\\CombinedFile.xlsx"

# Preprocess the data
def preprocess_data(filepath):
    data = pd.read_excel(filepath)
    cols_to_convert = ['Treatment Time (months)', 'Number of aligners: 1st (Max)', 'Number of
aligners: 1st (Mand)', 'Wear Time (days)']
    data[cols_to_convert] = data[cols_to_convert].apply(pd.to_numeric, errors='coerce')
    data_cleaned = data.dropna()
    data_encoded = pd.get_dummies(data_cleaned, columns=['Dentist', 'Gender', 'Submission
processes', 'Classification', 'Submission Option'], drop_first=True)
    data_encoded['Result'] = data_encoded['Result'].map({'S': 1, 'F': 0})
    X = data_encoded.drop(columns=['ID', 'Result'])
    y = data_encoded['Result']
    return train_test_split(X, y, test_size=0.3, random_state=42)

# Load and preprocess the data
X_train, X_test, y_train, y_test = preprocess_data(file_path)

# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train the ANN model
ann = MLPClassifier(hidden_layer_sizes=(100, 50), max_iter=1000, alpha=0.001, early_stopping=True,
validation_fraction=0.1, n_iter_no_change=10, random_state=42)
ann.fit(X_train_scaled, y_train)

# Predict on the test set and calculate metrics
y_pred = ann.predict(X_test_scaled)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# Print performance metrics
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-Score: {f1:.4f}")

# Initialize SHAP Explainer
background = shap.sample(X_train_scaled, 50)
explainer = shap.Explainer(ann.predict_proba, background)

# Compute SHAP values for the test set
shap_values = explainer(X_test_scaled)

# Save the SHAP summary plot as a PNG file
plt.figure()
shap.summary_plot(shap_values.values[:, :, 1], X_test_scaled, feature_names=X_train.columns)  #
Assuming binary classification and interest in the second class (index 1)
plt.savefig("shap_summary.png")
plt.close()
```

Fig. 12. The Python code of the ANN-SHAP visualization



```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import (
    accuracy_score,
    precision_score,
    recall_score,
    f1_score,
    classification_report,
    confusion_matrix,
    ConfusionMatrixDisplay,
)
import xgboost as xgb
from xgboost_distribution import XGBDistribution
import matplotlib.pyplot as plt

# Read the data
data = pd.read_excel("C:\\Users\\sanisa\\Documents\\CSIT999\\ML\\Invi\\All\\CombinedFile.xlsx")

# Convert columns to numeric
cols_to_convert = ['Treatment Time (months)', 'Number of aligners: 1st (Max)',
                   'Number of aligners: 1st (Mand)', 'Wear Time (days)']
for col in cols_to_convert:
    data[col] = pd.to_numeric(data[col], errors='coerce')

# Drop NaN values
data_cleaned = data.dropna()

# Encode categorical variables and convert Result to binary
data_encoded = pd.get_dummies(data_cleaned, columns=['Dentist', 'Gender', 'Submission processes',
                                                     'Classification', 'Submission Option'], drop_first=True)
data_encoded['Result'] = data_encoded['Result'].map({'S': 1, 'F': 0})

# Drop irrelevant columns (if any)
data_final = data_encoded.drop(columns=['Year', 'Submission Option_Full', 'Submission Option_Teen'])

# Split data into training and testing sets
X = data_final.drop(columns=['ID', 'Result'])
y = data_final['Result']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Create an XGBoost model
xgb_model = xgb.XGBClassifier()
xgb_model.fit(X_train, y_train)

# Predict on the test set and evaluate the model
y_pred = xgb_model.predict(X_test)
y_probs = xgb_model.predict_proba(X_test)[:, 1]
threshold = 0.5  # Adjust the decision threshold if necessary
y_pred_thresholded = [1 if prob >= threshold else 0 for prob in y_probs]

# Calculate and display performance metrics
accuracy = accuracy_score(y_test, y_pred_thresholded)
precision = precision_score(y_test, y_pred_thresholded)
recall = recall_score(y_test, y_pred_thresholded)
f1 = f1_score(y_test, y_pred_thresholded)

# Print performance metrics
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-Score: {f1:.4f}")
print("Classification Report:")
print(classification_report(y_test, y_pred_thresholded))

# Normalize the confusion matrix manually
conf_matrix = confusion_matrix(y_test, y_pred_thresholded)
conf_matrix_normalized = conf_matrix.astype('float') / conf_matrix.sum(axis=1)[:, np.newaxis]

# Display the normalized confusion matrix
disp = ConfusionMatrixDisplay(confusion_matrix=conf_matrix_normalized, display_labels=['Failure', 'Success'])
disp.plot(cmap=plt.cm.Blues)
plt.title('Normalized Confusion Matrix')
plt.show()

# Feature Importance Chart
importance_sorted_idx = np.argsort(xgb_model.feature_importances_)[-10:]
plt.figure(figsize=(10, 8))
bars = plt.barh(range(len(importance_sorted_idx)), xgb_model.feature_importances_[importance_sorted_idx], align='center')
plt.yticks(range(len(importance_sorted_idx)), [X.columns[i] for i in importance_sorted_idx])
plt.xlabel('F-Score')
plt.title('Top 10 Feature Importances')
plt.xlim(0, 0.6)  # Set the limit for x-axis

# Label each bar with its respective importance score
for bar in bars:
    plt.text(bar.get_width(), bar.get_y() + bar.get_height()/2, f'{bar.get_width():.3f}',
             va='center', ha='left', fontsize=8)

plt.tight_layout()
plt.show()

# XGBoost Distribution visualization for distribution parameters
model = XGBDistribution(distribution="normal", n_estimators=500, early_stopping_rounds=10)
model.fit(X_train, y_train, eval_set=[(X_test, y_test)])
preds = model.predict(X_test)
mean, std = preds.loc, preds.scale

# Fancy Visualization of XGBoost Distribution
plt.figure(figsize=(14, 7))
plt.scatter(np.arange(len(mean)), mean, c=mean, cmap='viridis', label='Predicted Mean')
plt.errorbar(np.arange(len(mean)), mean, yerr=std, linestyle='None', color='gray', alpha=0.5, label='Standard Deviation')
plt.colorbar(label='Mean Value')
plt.title('XGBoost Distribution: Mean & Standard Deviation')
plt.xlabel('Test Data Points')
plt.ylabel('Predicted Value')
plt.legend()
plt.tight_layout()
plt.show()
```

Fig. 13. The Python code of the XGBoost-developed model



```
import pandas as pd
import xgboost as xgb
import shap
from sklearn.model_selection import train_test_split

# Load your dataset
data = pd.read_excel("C:\\Users\\sanisa\\Documents\\CSIT999\\ML\\Invi\\All\\CombinedFile.xlsx")

# Convert columns to numeric and handle missing values
numeric_cols = ['Treatment Time (months)', 'Number of aligners: 1st (Max)', 'Number of aligners:
1st (Mand)', 'Wear Time (days)']
data[numeric_cols] = data[numeric_cols].apply(pd.to_numeric, errors='coerce')
data_cleaned = data.dropna()

# Encode categorical variables and convert Result to binary
categorical_cols = ['Dentist', 'Gender', 'Submission processes', 'Classification', 'Submission
Option']
data_encoded = pd.get_dummies(data_cleaned, columns=categorical_cols, drop_first=True)
data_encoded['Result'] = data_encoded['Result'].map({'S': 1, 'F': 0})

# Split data into features and target
X = data_encoded.drop(columns=['Result'])
y = data_encoded['Result']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Create and train the XGBoost model
xgb_model = xgb.XGBClassifier(use_label_encoder=False, eval_metric='logloss')
xgb_model.fit(X_train, y_train)

# Initialize SHAP Explainer
explainer = shap.Explainer(xgb_model)

# Compute SHAP Values
shap_values = explainer(X_test)

# Summary Plot
shap.summary_plot(shap_values, X_test)

# Optionally, visualize the feature importance
shap.plots.bar(shap_values)
```

Fig. 14. The Python code of the XGBoost-SHAP visualization

Appendix J. . Related work reviews

Table 7
Related revision.

Study	Clinical Problem	ML Problem	Dataset (source, size, context)	Prediction Target	Key Predictors	Algorithms	Performance Evaluation	Limitations
Imangaliyev et al. (2017) [10]	Dental plaque assessment using Quantitative Light-Induced Fluorescence (QLF) images	Classification (Supervised)	QLF image dataset with multi-channel data (RGB channels)	Dental plaque levels (derived from three plaque scores)	Colour channel intensities (Red, Green, Blue)	Convolutional Neural Network (CNN)	Improved accuracy over shallow models; evaluation metrics not specified	Limited generalizability due to reliance on specific imaging techniques; potential issues in diverse clinical settings
Rana et al. (2017) [78]	Early detection of periodontal disease through gingival inflammation	Classification (Supervised)	Intraoral images with expert annotations, size not specified	Pixel-wise segmentation of inflamed vs. healthy gingiva	Image colour channels, spatial features	Convolutional Neural Networks (CNNs)	AUC: 0.746, Precision: 0.347, Recall: 0.621	Moderate precision and recall; limited generalizability to other imaging devices
Lee et al. (2018) [79]	Diagnosis and prediction of periodontally	Classification (Supervised)	Periapical radiographic images, split	Diagnosis and extraction	Image features from radiographs	Deep CNN (Keras framework)	Accuracy: 81 % (premolars), 76.7 %	Limited dataset diversity, generalizability

(continued on next page)

Table 7 (*continued*)

Study	Clinical Problem	ML Problem	Dataset (source, size, context)	Prediction Target	Key Predictors	Algorithms	Performance Evaluation	Limitations
	compromised teeth (PCT)		into training (1,044), validation (348), and test (348)	prediction of PCT			(molars); Extraction prediction accuracy: 82.8 % (premolars), 73.4 % (molars)	to other imaging contexts
Nakhleh et al. (2018) [80]	Detection of halitosis	Classification (Supervised)	Exhaled breath samples analyzed for volatile sulfur compounds (VSCs)	Presence of halitosis	VSC levels, presence of non-sulfuric compounds	Statistical pattern recognition, sensor fusion algorithms	Not specified in the study	The absence of VSCs does not rule out halitosis; limited in detecting systemic disease indicators
Lee et al. (2017) [81]	Cephalometric landmark detection for orthodontic planning	Regression (Coordinate-wise prediction)	Public dataset, Grand Challenges in Dental X-ray Image Analysis (ISBI 2015)	Coordinates of 19 cephalometric landmarks	X-ray image pixel data	Convolutional Neural Network (CNN)-based coordinate regression	Detected landmarks within close margins to ground truths	Complex setup requiring 38 CNN systems for 19 landmarks; limited generalizability to other image types
Niño-Sandoval et al. (2017) [82]	Predicting mandibular morphology in facial reconstruction	Regression (Prediction of linear and angular mandibular measures)	229 lateral radiographs from Colombian patients (ages 18–25, skeletal classes I, II, III)	Mandibular morphology (17 linear and angular measures)	Craniomaxillary landmarks (coordinates)	Artificial Neural Networks (ANN), Support Vector Regression (SVR)	Correlation coefficient: 0.84–0.99 (ANN), >0.7 (SVR)	Small dataset; limited to specific demographic (Colombian patients); potential overfitting due to high correlation with limited sample
Allareddy et al. (2019) [83]	Leveraging big data for personalized orthodontics	Classification, Estimation	Omics data, CBCT imaging, centralized clinical repositories	Treatment outcomes, patient-specific factors	Genetic markers, radiomic features, demographic data	Various ML approaches (e.g., supervised/unsupervised learning)	Contextual discussion of potential applications rather than specific metrics	Lack of specific data, general discussion of big data use in orthodontics
Zhang et al. (2018) [84]	Genetic risk assessment for non-syndromic orofacial cleft in infants	Classification	Blood samples from Han and Uyghur Chinese populations, 43 SNPs validated by GWAS	Risk of NSCL/P	SNPs from MTHFR, RBP4 genes	Logistic Regression	Area Under Curve (AUC), feature importance ranking	Limited ethnic diversity in data; moderate predictive power for Uyghur group
Bianchi et al. (2019) [85]	Bone analysis of mandibular condyles in TMJ disorders	Radiomic feature extraction	HR-CBCT scans of 66 mandibular condyles with 0.08 mm^3 voxel resolution	Bone morphometric and textural feature correlation	Grey Level Non-Uniformity, Long Run Emphasis	BoneTexture, Ibex, BoneJ	Spearman correlation (r = 0.7–1), Bland–Altman analysis	Limited to software comparison; single modality (CBCT)
Lee et al. (2018) [86]	Detection of dental caries	Classification	3000 periapical radiographs	Presence of dental caries	Pixel intensity, image features	Deep CNN (GoogLeNet Inception v3)	Accuracy: 89 % (premolar), AUC: 0.917 (premolar), 0.890 (molar)	Limited to periapical radiographs; no multi-view validation
Saghiri et al. (2012) [11]	Locating minor apical foramen in endodontics	Classification	50 single-rooted teeth from 19 cadavers	Accurate working length determination	File position relative to foramen	ANN	Accuracy: 96 %, higher than endodontists' 76 % accuracy	Small dataset; cadaver study may differ from live cases
Poswar et al. (2015) [87]	Characterization of radicular cysts (RCs) and periapical granulomas (PGs) gene	Classification	Gene databases (PubMed, GenBank, STRING)	Leader genes for RC and PG	Gene expressions for RCs, PGs	Bioinformatics algorithms, MLP neural network	Identification of leader genes based on links	Focused on gene data; not clinical samples

(continued on next page)

Table 7 (*continued*)

Study	Clinical Problem	ML Problem	Dataset (source, size, context)	Prediction Target	Key Predictors	Algorithms	Performance Evaluation	Limitations
	expression							
Johari et al. (2017) [88]	Detection of Vertical Root Fractures (VRFs)	Classification	240 radiographs (120 intact, 120 fractured)	Presence of VRFs	Image analysis coefficients (wavelet and Gabor)	Probabilistic Neural Network (PNN)	Accuracy: 96.6 %, Sensitivity: 93.3 %, Specificity: 100 % (CBCT)	Limited to premolar teeth without caries or fillings
Yang et al. (2018) [89]	Post-treatment Quality Evaluation	Classification	196 periapical radiographs	Treatment outcome	Regions of Interest (ROIs) in apical region	CNN	F1 Score: 0.749	Small dataset, limited clinical diversity
Hiraiwa et al. (2019) [90]	Root morphology classification in mandibular molars	Classification	760 mandibular first molars (panoramic radiographs, CBCT)	Root type (single/extra root)	Distal root patches from panoramic radiographs	Deep Learning (CNN)	Accuracy: 86.9 %	Limited to mandibular first molars; generalization across other tooth types not verified
De Tobel et al. (2017) [91]	Age estimation from third molar development	Classification	20 panoramic radiographs per stage per gender	Development stage of third molar	Image contrast, bounding box	Transfer Learning (CNN)	Accuracy: 51 %, Kappa: 0.82	Moderate accuracy, limited sample size
Miki et al. (2017) [92]	Tooth classification for forensic use	Classification	52 CBCT volumes (CT slices)	Tooth type classification	ROIs from CT slices	DCNN (AlexNet)	Accuracy: 88.8 % (with augmentation)	Small dataset, limited diversity
O'Sullivan et al. (2019) [93]	AI in robotic surgery	Framework design	N/A	Procedure/ skill template	Anatomical models, medical imaging	Explainable AI, ML	Error reduction > 40 %	Limited to surgery, theoretical basis
Poedjiastoeti & Suebnukarn (2018) [94]	Ameloblastoma and keratocystic odontogenic tumors (KCOT) detection	Classification	500 panoramic radiographs	Tumor type	Preprocessed image features	CNN (VGG-16 with transfer learning)	Accuracy: 83.0 %, Sensitivity: 81.8 %	Small dataset, specific tumor types only
Patcas et al. (2018) [95]	Impact of orthognathic surgery on attractiveness	Regression (Scoring)	2164 images, 146 patients	Facial attractiveness, age appearance	Pre- and post-treatment photographs	CNN (trained on large-scale image data)	Attractiveness ↑ in 74.7 %, Age ↓ by 0.93 years ($p < 0.001$)	Single-center, specific to orthognathic surgery
Tarassoli et al. (2019) [96]	Surgical planning improvements	Predictive Analysis	Big data projections	Surgery outcomes with predictive analysis	Image processing, patient records	Quantum computing, predictive algorithms	Speculative due to future-oriented approach	Predictive models theoretical; data implementation unverified
Wirtz et al. (2018) [97]	Teeth segmentation in panoramic X-rays	Image segmentation	14 panoramic X-ray images	Individual teeth segmentation	Gradient image features, spatial relations	Coupled shape model, Neural Network	Precision: 0.790, Recall: 0.827, DICE: 0.744	Small dataset, image quality variability
Torosdagli et al. (2019) [98]	Mandible segmentation and landmarking	Segmentation, Landmarking	CBCT scans (50 patients for training, 250 for testing)	Mandible segmentation and landmarking	Geodesic distance, anatomical variability	Deep neural network (DNN), LSTM	Superior to state-of-the-art, visual inspection, MICCAI Head-Neck Challenge dataset	Limited to mandible and CBCT data only
Egger et al. (2018) [99]	Mandible segmentation in CT images	Segmentation	CT images of mandible; strict inclusion/ exclusion criteria	Mandible segmentation	Image pixel values	Fully Convolutional Network (FCN)	Qualitative and quantitative agreement between experts	Limited dataset availability; excludes artifacts
Du et al. (2018) [100]	Dental arch positioning in DPR	Positioning correction	Dental panoramic radiographs (DPR)	Positioning error in the dental arch	Patient posture, jaw morphology	CNN	Improved image quality for stable diagnosis	Limited to DPR; does not address all image artifacts

(*continued on next page*)

Table 7 (*continued*)

Study	Clinical Problem	ML Problem	Dataset (source, size, context)	Prediction Target	Key Predictors	Algorithms	Performance Evaluation	Limitations
Park et al. (2018) [101]	CT image resolution enhancement	Super-resolution	52 CT scans for training, 13 for testing	High-res CT images	Low-res CT image slices	Modified U-Net CNN	Peak SNR, NRMSE	Limited to CT images; not specific to dental care
Lee et al. (2018) [102]	Osteoporosis detection	Classification	1268 panoramic radiographs of females	Osteoporosis presence	Mandibular cortical erosion	Single & Multi-Column DCNN	AUC: SC-DCNN 0.9763; MC-DCNN 0.9987	Limited to female patients; small testing set
Zhang et al. (2018) [103]	Teeth recognition from periapical X-rays	Multi-class classification	Limited training dataset	Tooth position recognition (32 classes)	Pixel intensities, tooth features	Cascade CNN with label tree	Precision: 95.8 %; Recall: 96.1 %	Small training dataset; limited to specific task
Tuzoff et al. (2019) [104]	Teeth detection and numbering in radiographs	Object detection and classification	1352 panoramic radiographs (training), 222 (testing)	Tooth boundaries and FDI numbering	Radiograph pixel intensity	Faster R-CNN (detection), VGG-16 (classification)	Sensitivity: 0.9941 (detection), 0.9800 (numbering); Specificity: 0.9994 (numbering)	Errors due to similar factors as experts; limited to panoramic images
Ariji et al. (2018) [105]	Diagnosis of lymph node metastasis in oral cancer	Classification	441 CT images (127 positive, 314 negative nodes)	Metastatic vs. non-metastatic lymph nodes	CT image features	Deep Learning (Image Classification)	Accuracy: 78.2 %, Sensitivity: 75.4 %, Specificity: 81.0 %, AUC: 0.80	Small dataset; results not significantly better than radiologists
Kann et al. (2018) [106]	Identification of nodal metastasis and ENE	Classification	CT-segmented lymph nodes (2,875 samples; pathology labels)	ENE and nodal metastasis	CT imaging features	3D Convolutional Neural Network	AUC: 0.91 (95 % CI: 0.85–0.97)	Focused on radiographic data; requires clinical integration
Kats et al. (2019) [107]	Detection of atherosclerotic carotid plaques (ACP)	Classification	65 panoramic radiographs (small dataset)	Presence of ACP	Radiographic features	Faster R-CNN (Deep Learning)	Accuracy: 83 %, Sensitivity: 75 %, Specificity: 80 %, AUC (significant)	Small dataset, needs further validation for clinical integration
Murata et al. (2018) [108]	Maxillary sinusitis diagnosis	Classification (Supervised)	Panoramic radiographs: 6000 augmented samples for training, 120 samples for testing	Maxillary sinus condition (healthy/ inflamed)	Image patches from panoramic radiographs	Convolutional Neural Network (CNN)	Accuracy: 87.5 %, Sensitivity: 86.7 %, Specificity: 88.3 %, AUC: 0.875	Data augmentation may introduce biases; limited diversity in real-world testing datasets
Kositbowornchai et al. (2012) [109]	Vertical root fracture detection	Classification (Supervised)	200 digital radiography images (50 sound, 150 fractured)	Presence of vertical root fracture	Grey-scale data per line through the root	Probabilistic Neural Network (PNN)	Sensitivity: 98 %, Specificity: 90.5 %, Accuracy: 95.7 %	Small dataset; ex vivo study limits generalizability to clinical environments.
Kise et al. (2019) [110]	Diagnosis of Sjögren's syndrome (SjS) on CT images	Classification (Supervised)	CT images (500 images from 50 patients: 25 SjS cases, 25 controls)	Presence of SjS	CT image features	Deep Learning System	Accuracy: 96.0 %, Sensitivity: 100 %, Specificity: 92.0 %	Limited dataset size; specific to CT imaging and SjS cases.
Jung & Kim (2015) [111]	Diagnosis of tooth extraction in orthodontics	Classification (Supervised)	Clinical data from 156 patients (cephalometric variables and indexes)	Extraction vs non-extraction; Extraction patterns	12 cephalometric variables, 6 indexes	Neural Networks (Back-propagation)	Accuracy: 93 % (extraction vs non-extraction), 84 % (extraction patterns)	Small dataset; limited generalizability due to dataset size and diversity

References

[1] C.M. Alexander, R.G. Alexander, J.C. Gorman, J. Hilgers, C. Kurz, R.P. Scholz, et al., Lingual orthodontics. a status report, Journal of Clinical Orthodontics: JCO. 16 (4) (1982) 255–262.

[2] M. Upadhyay, S.A. Arqub, Biomechanics of clear aligners: hidden truths & first principles, Journal of the World Federation of Orthodontists. 11 (1) (2022) 12–21.

[3] N. Haouili, N.D. Kravitz, N.R. Vaid, D.J. Ferguson, L. Makki, Has Invisalign improved? a prospective follow-up study on the efficacy of tooth movement with Invisalign, Am. J. Orthod. Dentofac. Orthop. 158 (3) (2020) 420–425.

[4] O.M. Teeth, The story of beauty, inequality, and the struggle for oral health in America, The New Press, 2017.

[5] L. Wong, F.S. Ryan, L.R. Christensen, S.J. Cunningham, Factors influencing satisfaction with the process of orthodontic treatment in adult patients, Am. J. Orthod. Dentofac. Orthop. 153 (3) (2018) 362–370.

[6] Y.-w. Chen, K. Stanley, W. Att, Artificial intelligence in dentistry: current applications and future perspectives, Quintessence Int. 51 (3) (2020) 248–257.

[7] Y.M. Bichu, I. Hansa, A.Y. Bichu, P. Premjani, C. Flores-Mir, N.R. Vaid, Applications of artificial intelligence and machine learning in orthodontics: a scoping review, Prog. Orthod. 22 (2021) 1–11.

[8] K.F. Wong, X.Y. Lam, Y. Jiang, A.W.K. Yeung, Y. Lin, Artificial intelligence in orthodontics and orthognathic surgery: a bibliometric analysis of the 100 most-cited articles, Head Face Med. 19 (1) (2023) 38.

[9] M.A. Predicting, Invisalign Treatment Time using Machine Learning, The Ohio State University, 2023.

[10] Imangaliyev S, van der Veen MH, Volgenant C, Loos BG, Keijser BJ, Crielaard W, et al. Classification of quantitative light-induced fluorescence images using convolutional neural network. arXiv preprint arXiv:170509193. 2017.

[11] M.A. Saghiri, F. Garcia-Godoy, J.L. Gutmann, M. Lotfi, K. Asgar, The reliability of artificial neural network in locating minor apical foramen: a cadaver study, J. Endod. 38 (8) (2012) 1130–1134.

[12] F. Magrabi, E. Ammenwerth, J.B. McNair, N.F. De Keizer, H. Hyppönen, P. Nykänen, et al., Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications, Yearb. Med. Inform. 28 (01) (2019) 128–134.

[13] H. Kök, A.M. Acilar, M.S. İzgi, Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics, Prog. Orthod. 20 (2019) 1–10.

[14] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (2008) 1–37.

[15] Tejera BC. Quantification of Cleft Volume in Patients With Unilateral Cleft Lip and Palate: A Deep Learning-Based 3D-Morphometric Analysis: Nova Southeastern University; 2023.

[16] H.-I. Choi, S.-K. Jung, S.-H. Baek, W.H. Lim, S.-J. Ahn, I.-H. Yang, et al., Artificial intelligent model with neural network machine learning for the diagnosis of orthognathic surgery, Journal of Craniofacial Surgery. 30 (7) (2019) 1986–1989.

[17] J. Brownlee, XGBoost with python: Gradient boosted trees with XGBoost and scikit-learn, Machine Learning Mastery (2016).

[18] X. Wang, X. Zhao, G. Song, J. Niu, T. Xu, Machine learning-based evaluation on Craniodentofacial morphological harmony of patients after orthodontic treatment, Front. Physiol. 13 (2022) 862847.

[19] L. Xing, X. Zhang, Y. Guo, D. Bai, H. Xu, XGBoost-aided prediction of lip prominence based on hard-tissue measurements and demographic characteristics in an asian population, Am. J. Orthod. Dentofac. Orthop. 164 (3) (2023) 357–367.

[20] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[21] W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, Müller K.-R. Explainable, Ai,, interpreting, explaining and visualizing deep learning, Springer Nature, 2019.

[22] N. Kazimierczak, W. Kazimierczak, Z. Serafin, P. Nowicki, J. Nożewski, J. Janiszewska-Olszowska, AI in Orthodontics: Revolutionizing Diagnostics and Treatment Planning—A Comprehensive Review, J. Clin. Med. 13 (2) (2024) 344.

[23] J. Liu, Y. Chen, S. Li, Z. Zhao, Z. Wu, Machine learning in orthodontics: challenges and perspectives, Adv. Clin. Exp. Med. 30 (10) (2021) 1065–1074.

[24] S. Fa, W. Samek, J. Krois, Artificial intelligence in dentistry: chances and challenges, J. Dent. Res. 99 (7) (2020) 769–774.

[25] R.-K. Sheu, M.S. Pardeshi, A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system, Sensors 22 (20) (2022) 8068.

[26] Lee MK, Allareddy V, Rampa S, Elnagar MH, Oubaidin M, Yadav S, et al., editors. Applications and challenges of implementing artificial intelligence in orthodontics: A primer for orthodontists. Seminars in Orthodontics; 2024: Elsevier.

[27] I. Gomez-Rios, E. Egea-Lopez, A.J. Ortiz Ruiz, ORIENTATE: automated machine learning classifiers for oral health prediction and research, BMC Oral Health 23 (1) (2023) 408.

[28] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Proces. Syst. 30 (2017).

[29] B. Köktürk, H. Pamukçu, Ö. Gözüaçık, Evaluation of different machine learning algorithms for extraction decision in orthodontic treatment, Orthod. Craniofac. Res. (2024).

[30] S.B. Khanagar, A. Al-Ehaideb, S. Vishwanathaiah, P.C. Maganur, S. Patil, S. Naik, et al., Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making-a systematic review, Journal of Dental Sciences. 16 (1) (2021) 482–492.

[31] Wirth R, Hipp J, editors. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining; 2000: Manchester.

[32] J.T. Hancock, T.M. Khoshgoftaar, Survey on categorical data for neural networks, Journal of Big Data. 7 (1) (2020) 28.

[33] H. Nugroho, N.P. Utama, K. Surendro, Smoothing target encoding and class center-based firefly algorithm for handling missing values in categorical variable, Journal of Big Data. 10 (1) (2023) 10.

[34] Phung S, Kumar A, Kim J, editors. A deep learning technique for imputing missing healthcare data. 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2019: IEEE.

[35] Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. Theoretical epidemiology: principles of occurrence research in medicine1985. p. xxii, 359-xxii, .

[36] M.H. Gorelick, Bias arising from missing data in predictive models, J. Clin. Epidemiol. 59 (10) (2006) 1115–1123.

[37] J.A. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, et al., Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, BMJ 338 (2009).

[38] J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psychol. Methods 7 (2) (2002) 147.

[39] P. Kokol, S. Pohorec, G. Štiglic, V. Podgorelec, Evolutionary design of decision trees for medical application, Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 2 (3) (2012) 237–254.

[40] M. Belgiu, L. Drăguţ, Random forest in remote sensing: a review of applications and future directions, ISPRS J. Photogramm. Remote Sens. 114 (2016) 24–31.

[41] H. Han, H. Neira-Molina, A. Khan, M. Fang, H.A. Mahmoud, E.M. Awwad, et al., Advanced series decomposition with a gated recurrent unit and graph convolutional neural network for non-stationary data patterns, Journal of Cloud Computing. 13 (1) (2024) 20.

[42] Y. Yu, M. Li, L. Liu, Y. Li, J. Wang, Clinical big data and deep learning: applications, challenges, and future outlooks, Big Data Min. Anal. 2 (4) (2019) 288–305.

[43] Y. Shao, Imbalance Learning and its Application on Medical Datasets: Dissertation, Georg-August Universität, Göttingen, 2021, p. 2021.

[44] S. Sharma, Heart diseases prediction using hybrid ensemble learning, Dublin Business School (2020).

[45] İ. Tamer, E. Öztaş, G. Marşan, Orthodontic treatment with clear aligners and the scientific reality behind their marketing: a literature review, Turkish Journal of Orthodontics. 32 (4) (2019) 241.

[46] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 9 (4) (2019) e1312.

[47] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE Trans. Neural Networks Learn. Syst. 32 (11) (2020) 4793–4813.

[48] B. Suh, H. Yu, J.-K. Cha, J. Choi, J.-W. Kim, Explainable Deep Learning Approaches for Risk Screening of Periodontitis, J. Dent. Res. 00220345241286488 (2024).

[49] N. Adnan, S.M. Faizan Ahmed, J.K. Das, S. Aijaz, R.H. Sukhia, Z. Hoodbhoy, et al., Developing an AI-based application for caries index detection on intraoral photographs, Sci. Rep. 14 (1) (2024) 26752.

[50] G. Kummer, T. Eliades, D. Koletsi, Gender-specific treatment effects and outcomes reported in orthodontic research. a cross-sectional empirical study, Eur. J. Orthod. 46(1):cjad073 (2024).

[51] A. Lagorsse, S. Gebeile-Chauty, Does gender make a difference in orthodontics? A Literature Review. L'orthodontie Francaise. 89 (2) (2018) 157–168.

[52] J. Yao, D.-D. Li, Y.-Q. Yang, C.P.J. McGrath, N. Mattheos, What are patients' expectations of orthodontic treatment: a systematic review, BMC Oral Health 16 (2016) 1–8.

[53] F. Keles, A. Bos, Satisfaction with orthodontic treatment, Angle Orthod. 83 (3) (2013) 507–511.

[54] D. Mavreas, A.E. Athanasiou, Factors affecting the duration of orthodontic treatment: a systematic review, The European Journal of Orthodontics. 30 (4) (2008) 386–395.

[55] F. d'Apuzzo, L. Perillo, C.K. Carrico, T. Castroflorio, V. Grassia, S.J. Lindauer, et al., Clear aligner treatment: different perspectives between orthodontists and general dentists, Prog. Orthod. 20 (2019) 1–9.

[56] S.A. Alsaeed, D.B. Kennedy, J. Aleksejuniene, E.H. Yen, B.T. Pliska, D. C. Flanagan, Outcomes of orthodontic treatment performed by individual orthodontists vs 2 orthodontists collaborating on treatment, Am. J. Orthod. Dentofac. Orthop. 158 (1) (2020) 59–67.

[57] E.M. Heath, J.D. English, C.D. Johnson, E.B. Swearingen, S. Akyalcin, Perceptions of orthodontic case complexity among orthodontists, general practitioners, orthodontic residents, and dental students, Am. J. Orthod. Dentofac. Orthop. 151 (2) (2017) 335–341.

[58] J.-P. Houle, L. Piedade, R. Todescan Jr, F.H.L. Pinheiro, The predictability of transverse changes with Invisalign, Angle Orthod. 87 (1) (2017) 19–24.

[59] S. Saccomanno, S. Saran, V. Vanella, R.F. Mastrapasqua, L. Raffaelli, L. Levrini, The potential of digital impression in orthodontics, Dentistry Journal. 10 (8) (2022) 147.

[60] S.A. Arqub, S. Banankhah, R. Sharma, L.D.C. Godoy, C.-L. Kuo, M. Ahmed, et al., Association between initial complexity, frequency of refinements, treatment duration, and outcome in Invisalign orthodontic treatment, Am. J. Orthod. Dentofac. Orthop. 162 (3) (2022) e141–e155.

[61] N.D. Kravitz, B. Dalloul, Y.A. Zaid, C. Shah, N.R. Vaid, What percentage of patients switch from Invisalign to braces? a retrospective study evaluating the

conversion rate, number of refinement scans, and length of treatment, Am. J. Orthod. Dentofac. Orthop. 163 (4) (2023) 526–530.

[62] A. Papadimitriou, S. Mousoulea, N. Gkantidis, D. Kloukos, Clinical effectiveness of Invisalign® orthodontic treatment: a systematic review, Prog. Orthod. 19 (2018) 1–24.

[63] D.A. Kuncio, Invisalign: current guidelines for effective treatment, N. Y. State Dent. J. 80 (2) (2014) 11.

[64] T. Castroflorio, A. Sedran, S. Parrini, F. Garino, M. Reverdito, R. Capuozzo, et al., Predictability of orthodontic tooth movement with aligners: effect of treatment design, Prog. Orthod. 24 (1) (2023) 2.

[65] B. Sereewisai, R. Chintavalakorn, P. Santiwong, T. Nakornnoi, S.P. Neoh, K. Sipiyaruk, The accuracy of virtual setup in simulating treatment outcomes in orthodontic practice: a systematic review, BDJ Open. 9 (1) (2023) 41.

[66] L. Xu, H. Li, L. Mei, Y. Li, P. Wo, Y. Li, Aligner treatment: Patient experience and influencing factors, Australasian Orthodontic Journal. 38 (1) (2022) 88–95.

[67] L.H. Timm, G. Farrag, M. Baxmann, F. Schwendicke, Factors influencing patient compliance during clear aligner therapy: a retrospective cohort study, J. Clin. Med. 10 (14) (2021) 3103.

[68] S. Sahim, F. El Quars, Effectiveness and Stability of Treatment with Orthodontics Clear Aligners, What Evidence? Current Trends in Orthodontics. (2018).

[69] R.-C. Chen, C. Dewi, S.-W. Huang, R.E. Caraka, Selecting critical features for data classification based on machine learning methods, Journal of Big Data. 7 (1) (2020) 52.

[70] D. Rengasamy, J.M. Mase, A. Kumar, B. Rothwell, M.T. Torres, M.R. Alexander, et al., Feature importance in machine learning models: a fuzzy information fusion approach, Neurocomputing 511 (2022) 163–174.

[71] Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, et al. The effects of data quality on machine learning performance. arXiv preprint arXiv: 220714529. 2022.

[72] N. Arya, S. Saha, A. Mathur, S. Saha, Improving the robustness and stability of a machine learning model for breast cancer prognosis through the use of multi-modal classifiers, Sci. Rep. 13 (1) (2023) 4079.

[73] K. Li, B. DeCost, K. Choudhary, M. Greenwood, J. Hattrick-Simpers, A critical examination of robustness and generalizability of machine learning prediction of materials properties, npj Comput. Mater. 9(1):55 (2023).

[74] A. Veglis, T.A. Maniou, The mediated data model of communication flow: big data and data journalism. KOME: an International Journal of Pure Communication, Inquiry 6 (2) (2018) 32–43.

[75] S. Molin, Hands-on Data Analysis with Pandas: Efficiently perform data collection, wrangling, analysis, and visualization using Python, Packt Publishing Ltd (2019).

[76] I. Ekanayake, D. Meddage, U. Rathnayake, A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP), Case Stud. Constr. Mater. 16 (2022) e01059.

[77] O. Kharakhash, Data visualization: transforming complex data into actionable insights, Automation of Technological and Business Processes. 15 (2) (2023) 4–12.

[78] Rana A, Yauney G, Wong LC, Gupta O, Muftu A, Shah P, editors. Automated segmentation of gingival diseases from oral images. 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT); 2017: IEEE.

[79] J.-H. Lee, D.-h. Kim, S.-N. Jeong, S.-H. Choi, Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm, Journal of Periodontal & Implant Science. 48 (2) (2018) 114–123.

[80] M. Nakhleh, M. Quatredeniers, H. Haick, Detection of halitosis in breath: between the past, present, and future, Oral Dis. 24 (5) (2018) 685–695.

[81] Lee H, Park M, Kim J, editors. Cephalometric landmark detection in dental x-ray images using convolutional neural networks. Medical imaging 2017: Computer-aided diagnosis; 2017: SPIE.

[82] Niño-Sandoval TC, Pérez SVG, González FA, Jaque RA, Infante-Contreras C. Use of automated learning techniques for predicting mandibular morphology in skeletal class I, II and III. Forensic science international. 2017;281:187. e1-. e7.

[83] V. Allareddy, S. Rengasamy Venugopalan, R.P. Nalliah, J.L. Caplin, M.K. Lee, V. Allareddy, Orthodontics in the era of big data analytics, Orthod. Craniofac. Res. 22 (2019) 8–13.

[84] S.-J. Zhang, P. Meng, J. Zhang, P. Jia, J. Lin, X. Wang, et al., Machine learning models for genetic risk assessment of infants with non-syndromic orofacial cleft, Genomics, Proteomics and Bioinformatics. 16 (5) (2018) 354–364.

[85] J. Bianchi, J.R. Gonçalves, A.C.O. Ruellas, J.-B. Vimort, M. Yatabe, B. Paniagua, et al., Software comparison to analyze bone radiomics from high resolution CBCT scans of mandibular condyles, Dentomaxillofacial Radiology. 48(6):20190049 (2019).

[86] J.-H. Lee, D.-H. Kim, S.-N. Jeong, S.-H. Choi, Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm, J. Dent. 77 (2018) 106–111.

[87] P.F. de Oliveira, L.C. Farias, C.A. de Carvalho Fraga, W. Bambirra Jr, M. Brito-Júnior, M.D. Sousa-Neto, et al., Bioinformatics, interaction network analysis, and neural networks to characterize gene expression of radicular cyst and periapical granuloma, J. Endod. 41 (6) (2015) 877–883.

[88] M. Johari, F. Esmaeili, A. Andalib, S. Garjani, H. Saberkari, Detection of vertical root fractures in intact and endodontically treated premolar teeth by designing a probabilistic neural network: an ex vivo study, Dentomaxillofacial Radiology. 46 (2) (2017) 20160107.

[89] Yang J, Xie Y, Liu L, Xia B, Cao Z, Guo C, editors. Automated dental image analysis by deep learning on small dataset. 2018 IEEE 42nd annual computer software and applications conference (COMPSAC); 2018: IEEE.

[90] T. Hiraiwa, Y. Ariji, M. Fukuda, Y. Kise, K. Nakata, A. Katsumata, et al., A deep-learning artificial intelligence system for assessment of root morphology of the mandibular first molar on panoramic radiography, Dentomaxillofacial Radiology. 48 (3) (2019) 20180218.

[91] J. De Tobel, P. Radesh, D. Vandermeulen, P.W. Thevissen, An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study, J. Forensic Odontostomatol. 35 (2) (2017) 42.

[92] Y. Miki, C. Muramatsu, T. Hayashi, X. Zhou, T. Hara, A. Katsumata, et al., Classification of teeth in cone-beam CT using deep convolutional neural network, Comput. Biol. Med. 80 (2017) 24–29.

[93] S. O'Sullivan, S. Leonard, A. Holzinger, C. Allen, F. Battaglia, N. Nevejans, et al., Operational framework and training standard requirements for AI-empowered robotic surgery, The International Journal of Medical Robotics and Computer Assisted Surgery. 16 (5) (2020) 1–13.

[94] W. Poedjiastoeti, S. Suebnukarn, Application of convolutional neural network in the diagnosis of jaw tumors, Healthcare Informatics Research. 24 (3) (2018) 236–241.

[95] R. Patcas, D.A. Bernini, A. Volokitin, E. Agustsson, R. Rothe, R. Timofte, Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age, Int. J. Oral Maxillofac. Surg. 48 (1) (2019) 77–83.

[96] S.P. Tarassoli, Artificial intelligence, regenerative surgery, robotics? what is realistic for the future of surgery? Ann. Med. Surg. 41 (2019) 53–55.

[97] Wirtz A, Mirashi SG, Wesarg S, editors. Automatic teeth segmentation in panoramic X-ray images using a coupled shape model in combination with a neural network. Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11; 2018: Springer.

[98] N. Torosdagli, D.K. Liberton, P. Verma, M. Sincan, J.S. Lee, U. Bagci, Deep geodesic learning for segmentation and anatomical landmarking, IEEE Trans. Med. Imaging 38 (4) (2018) 919–931.

[99] Egger J, Pfarrkirchner B, Gsaxner C, Lindner L, Schmalstieg D, Wallner J, editors. Fully convolutional mandible segmentation on a valid ground-truth dataset. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018: IEEE.

[100] Du X, Chen Y, Zhao J, Xi Y, editors. A convolutional neural network based auto-positioning method for dental arch in rotational panoramic radiography. 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2018: IEEE.

[101] J. Park, D. Hwang, K.Y. Kim, S.K. Kang, Y.K. Kim, J.S. Lee, Computed tomography super-resolution using deep convolutional neural network, Phys. Med. Biol. 63 (14) (2018) 145011.

[102] J.-S. Lee, S. Adhikari, L. Liu, H.-G. Jeong, H. Kim, S.-J. Yoon, Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study, Dentomaxillofacial Radiology. 48 (1) (2019) 20170344.

[103] K. Zhang, J. Wu, H. Chen, P. Lyu, An effective teeth recognition method using label tree with cascade network structure, Comput. Med. Imaging Graph. 68 (2018) 61–70.

[104] D.V. Tuzoff, L.N. Tuzova, M.M. Bornstein, A.S. Krasnov, M.A. Kharchenko, S. I. Nikolenko, et al., Tooth detection and numbering in panoramic radiographs using convolutional neural networks, Dentomaxillofacial Radiology. 48 (4) (2019) 20180051.

[105] Y. Ariji, M. Fukuda, Y. Kise, M. Nozawa, Y. Yanashita, H. Fujita, et al., Contrast-enhanced computed tomography image assessment of cervical lymph node metastasis in patients with oral cancer by using a deep learning system of artificial intelligence, Oral Surg Oral Med Oral Pathol Oral Radiol 127 (5) (2019) 458–463.

[106] B.H. Kann, S. Aneja, G.V. Loganadane, J.R. Kelly, S.M. Smith, R.H. Decker, et al., Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks, Sci. Rep. 8 (1) (2018) 14036.

[107] L. Kats, M. Vered, A. Zlotogorski-Hurvitz, I. Harpaz, Atherosclerotic carotid plaque on panoramic radiographs: neural network detection, Int. J. Comput. Dent. 22 (2) (2019).

[108] M. Murata, Y. Ariji, Y. Ohashi, T. Kawai, M. Fukuda, T. Funakoshi, et al., Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography, Oral Radiol. 35 (2019) 301–307.

[109] S. Kositbowornchai, S. Plermkamon, T. Tangkosol, Performance of an artificial neural network for vertical root fracture detection: an ex vivo study, Dent. Traumatol. 29 (2) (2013) 151–155.

[110] Y. Kise, H. Ikeda, T. Fujii, M. Fukuda, Y. Ariji, H. Fujita, et al., Preliminary study on the application of deep learning system to diagnosis of Sjögren's syndrome on CT images, Dentomaxillofacial Radiology. 48 (6) (2019) 20190019.

[111] S.-K. Jung, T.-W. Kim, New approach for the diagnosis of extractions with neural network machine learning, Am. J. Orthod. Dentofac. Orthop. 149 (1) (2016) 127–133.