Predicting Tooth Mobility and Implant Stability using Periapical Radiographic Features and Implant Stability Test Data

Li, Zhilin

Department of Dentistry
Graduate School
Yonsei University

Predicting Tooth Mobility and Implant Stability using Periapical Radiographic Features and Implant Stability Test Data

Advisor Kim, Jong-Eun

A Master's Thesis Submitted
to the Department of Dentistry
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Master of Dental Science

Li, Zhilin

June 2025

Predicting Tooth Mobility and Implant Stability
using Periapical Radiographic Features
and Implant Stability Test Data

This certifies that the Master's Thesis
of Li, Zhilin is approved

_____

Committee Chair **Oh, Kyung Chul**

_____

Committee Member **Kim, Jong-Eun**

_____

Committee Member **Lee, Hyeonjong**

Department of Dentistry
Graduate School
Yonsei University
June 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Predicting Tooth Mobility and Implant Stability using Periapical Radiographic Features and Implant Stability Test Data

This study proposes a machine learning framework for predicting tooth mobility and implant stability by integrating anatomical features extracted from periapical radiographs with biomechanical measurements (IST values). A total of 407 annotated radiographs were expanded to 2,038 via geometric augmentation. Structural indices—such as head-to-root area ratios, periodontal ligament visibility, and root morphology—were engineered into composite features. A stacked ensemble model, incorporating LightGBM, XGBoost, and Random Forest with a Ridge Regression meta-learner, was trained on these features.

The best-performing model achieved an R^2 of 0.6840, MAE of 4.0132, and MSE of 46.6392, demonstrating robust alignment between predicted and actual IST values. SHAP analysis revealed that root type and crown-root ratios were the most influential predictors. Although ligament annotations were sparse, their inclusion improved model accuracy in well-annotated cases.

These findings highlight the potential of anatomy-aware, image-based regression models to non-invasively assess periodontal support and implant stability. The proposed framework bridges radiographic morphology and objective biomechanics, offering a reproducible, data-driven approach for clinical decision support in dentistry.

———

Key words: Tooth mobility, Implant stability, IST value, Periapical radiograph, Machine learning, Periodontal assessment

1. INTRODUCTION

Periodontal disease is a prevalent, chronic inflammatory condition that compromises the supporting structures of the teeth, including the periodontal ligament and alveolar bone (Lang & Bartold, 2018). Characterized by progressive tissue degradation, the disease leads to increased tooth mobility as the periodontal ligament fibers loosen and vertical bone height diminishes (Elemek, 2022). Clinically, it manifests through sustained inflammation, connective tissue attachment loss, and alveolar bone resorption—pathological processes that collectively increase tooth mobility. These changes adversely affect masticatory function, and may result in pain, discomfort, and eventual tooth loss, thereby impairing both oral and systemic quality of life.

Radiographic imaging is essential for the diagnosis and longitudinal assessment of periodontal disease (Hoss et al., 2023). Among available modalities, periapical radiography remains a widely adopted tool due to its high resolution and ability to visualize fine anatomical details of teeth and their surrounding structures. When combined with clinical indicators—such as probing depth and tooth mobility—periapical images offer a comprehensive diagnostic framework that enhances the accuracy of disease evaluation and informs treatment strategies (Elemek, 2022).

In this context, tooth mobility serves as a critical parameter in assessing periodontal health (Kim et al., 2023), providing valuable insight into disease severity and progression. It also plays a central role in diagnostic decisions, treatment planning, and outcome monitoring. Conventionally, mobility is evaluated manually through controlled force application, with displacement assessed subjectively (Meirelles et al., 2020). The Miller Classification System, one of the most widely used frameworks, grades mobility into three categories based on the extent of horizontal and vertical movement. However, these methods are inherently subjective and susceptible to inter-examiner variability, which compromises reproducibility and limits research applicability.

To address the limitations of subjective mobility grading systems, objective and non-invasive diagnostic tools have been developed (Okuhama et al., 2022). Among them, the stability measuring instrument (AnyCheck, Neo Biotech, South Korea) has gained clinical acceptance for their ability to standardize mobility assessments through vibrational impulse testing. The instrument quantifies the biomechanical response of a tooth by delivering controlled vibrational impulses (D.-H. Lee et al., 2020). This procedure yields an Implant Stability Test (IST) value that

inversely correlates with mobility (J. Lee et al., 2020). Higher IST values indicate greater structural stability, while lower values suggest compromised periodontal support.

Although the IST values obtained from the instrument offer improved reproducibility and diagnostic precision, they represent a purely mechanical measurement and do not incorporate anatomical variations that may critically influence overall stability. Therefore, IST values alone may not capture the full biomechanical context necessary for accurate clinical interpretation.

Recent advances in digital imaging and artificial intelligence (AI) (Ari et al., 2022) have further enhanced the precision and depth of periodontal assessment (Medina-Sotomayor et al., 2019). They have enabled automated extraction of structural features from dental radiographs, supporting enhanced diagnostic workflows (Çelik et al., 2023). However, existing AI applications in dentistry have largely focused on classification tasks (Benakatti et al., 2022), such as caries detection or bone loss segmentation, rather than the prediction of quantitative biomechanical indicators like IST values. Furthermore, few models integrate radiographic anatomy with clinical mechanical measurements (Cha et al., 2021) to evaluate tooth mobility or implant stability. This gap limits the potential for fully automated, anatomy-aware diagnostic systems.

To address this, I propose a data-driven machine learning framework that integrates radiographic structure and clinical measurement by combining image-derived support-related features with IST values obtained from the instrument. By combining objective radiographic and clinical measurement, the framework enables more precise, individualized assessment of dental support conditions and may facilitate earlier intervention to prevent structural deterioration.

To approximate clinically relevant indicators of mobility, I extracted image-based structural features from periapical radiographs that reflect anatomical support integrity, such as head-to-root proportions and ligament visibility, both of which are well-established correlates of biomechanical stability in periodontics.

2. MATERIALS AND METHODS

2.1 Data Pre-processing

2.1.1 Data Annotation

This study developed a quantitative index system based on anatomical segmentation and weighted area ratios to assess the contribution of distinct dental regions to overall tooth mobility. This system enables objective measurement and facilitates multivariate statistical analysis of biomechanical stability.

The dataset comprises 407 periapical radiographs, approved by the Institutional Review Board of Dental Hospital, Yonsei University, South Korea (IRB No. 2-2025-0018). All images were selected to preserve complete tooth morphology, ensuring morphological completeness and annotation consistency. Corresponding IST values were acquired during clinical procedures. These values served as surrogate indicators of tooth mobility.

Each radiograph was annotated using the open-source labeling platform Computer Vision Annotation Tool (CVAT). The following three anatomical regions were defined:

Head: The supragingival portion of the tooth above the bone line.

Root: The subgingival portion of the tooth below the bone line (each tooth could contain one or more root labels).

Ligament: The visible periodontal ligament (PDL) region, annotated selectively based on visibility thresholds.

2.1.2 Data Augmentation Specifics

To enhance model robustness and mitigate overfitting, a series of geometric data augmentation techniques were applied to simulate anatomical variability commonly observed in periapical radiographs. The augmentation process included the following transformations:

Rotation: Random rotation within a range of ±30 degrees

Scaling: Resizing within a range of 80% to 120% of the original image size

Shearing: Affine shear transformations along both the horizontal (X-axis) and vertical (Y-axis) directions to mimic non-uniform morphological distortions

Interpolation: Applied to maintain image continuity during spatial transformations

A total of 1,628 augmented samples were generated through this process, increasing the dataset from 407 original instances to approximately 2,038 samples. Among these, 1,160 samples belong to the Tooth category and 878 to the Implant category. This expanded dataset contributed to reducing model variance and improving generalization performance, especially under conditions of limited annotated data.

2.1.3 Data Standardization and Splitting

All input features were standardized using the StandardScaler method from Scikit-learn to achieve zero mean and unit variance, ensuring scale uniformity across predictors. The dataset was randomly split into training and validation sets in an 80:20 ratio, the random seed was fixed at 42 to ensure reproducibility.

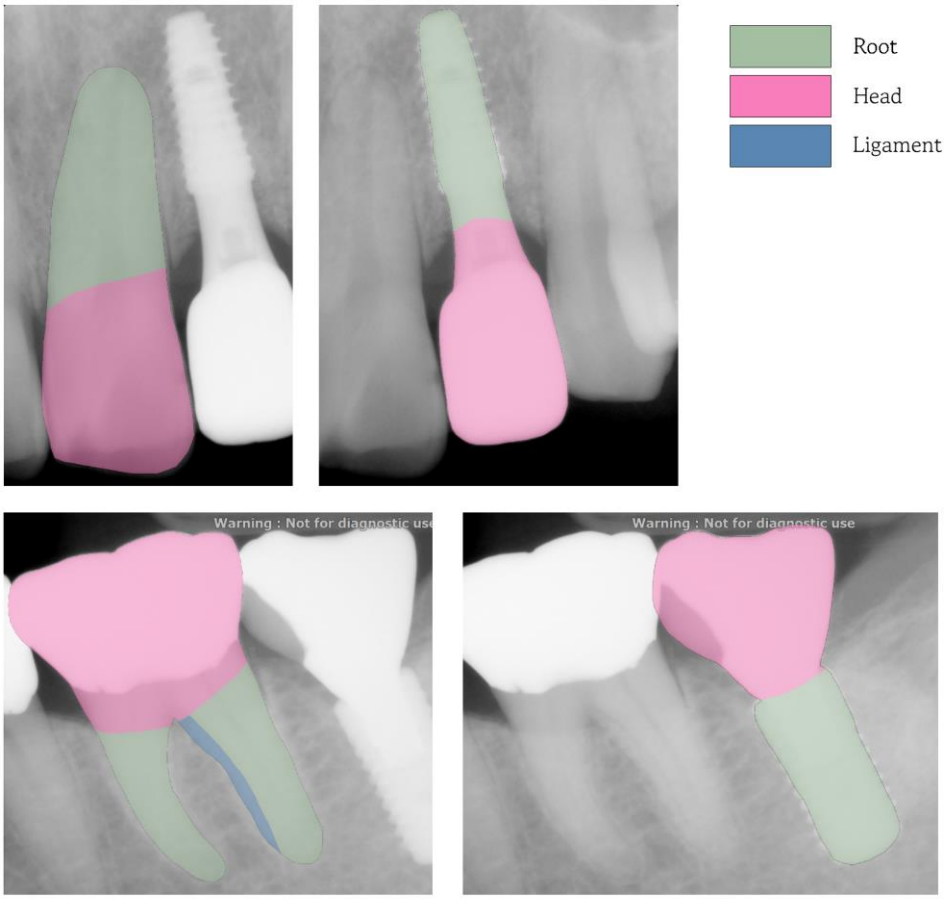

Figure 1. Periapical & Annotation image sample

2.2 Computational Methods

2.2.1 Feature Definition and Engineering

The selected structural features were designed to approximate clinically recognized determinants of tooth mobility. Specifically, head-to-root ratios have long been associated with periodontal support capacity (Tada et al., 2015; Hartmann et al., 2017), while the visibility of the periodontal ligament reflects surrounding soft tissue integrity. By quantifying these anatomical indicators from radiographs, the model aims to emulate clinical judgment within an interpretable, image-based framework, using IST values as an objective surrogate for mobility grading.

The dataset was composed of two distinct categories: Tooth and Implant. During modeling, separate models were trained for each category to account for their specific characteristics.

Based on the annotated regions, three primary structural metrics were derived to quantify region-specific contributions to overall tooth mobility:

Head_Area_Ratio: The ratio of the crown area to the total object area.

$$Head\_Area\_Ratio = Head\_Area / Total\_Area$$

A higher value suggests a larger exposed crown relative to root support, potentially indicating higher mobility.

Root_Area_Ratio: The ratio of the root area to the total object area.

$$Root\_Area\_Ratio = Root\_Area / Total\_Area$$

Higher values generally reflect greater subgingival anchorage, implying increased stability.

Ligament_Weight: A categorical variable (ranging from 0 to 3) representing the proportion of the visible PDL region.

For the Implant category, Ligament_Weight is consistently set to 0, while for the Tooth category, weights ranging from 1 to 3 are assigned based on the relative area of the visible PDL.

The Ligament_Weight feature, unique to the Tooth category, was present in only 24.14% of annotated samples.

This reflects the hypothesis that a more extensive ligament area, when present, may correlate with reduced stability and increased mobility.

In addition to annotated regions, a binary classification of root morphology—referred to as Root_Count—was introduced to further distinguish between tooth types:

Root_Count: Categorized based on the number of tooth roots.

Teeth with FDI numbers 1–5 were labeled as single-rooted (Root_Count_Single), while teeth numbered 6–7 were labeled as multi-rooted (Root_Count_Multi) (Ziegler et al., 2005).

To enhance feature expressiveness and capture second-order relationships, I introduced composite features that reflect both physiological interactions and statistical dependencies between anatomical regions.

These composite features including Multiplicative and Division Interaction between Head_Area_Ratio and Root_Area_Ratio:

$$Head\_mul\_Root = Head\_Area\_Ratio \times Root\_Area\_Ratio$$

$$Head\_div\_Root = Head\_Area\_Ratio \div Root\_Area\_Ratio$$

Given that IST values were predominantly concentrated within the range of 40 to 99 and exhibited a slight right-skew, using a log transformation can help improve model performance by providing a smoother compression and enhancing generalization. This transformation was applied to stabilize variance and mitigate the influence of outliers, as formulated below:

Log Transformation

$$IST_{transformed} = log(IST + 1)$$

For model training, the regression target IST underwent a logarithmic transformation to stabilize variance and reduce right-skewness. For final evaluation, predictions were transformed back to the original scale to facilitate clinical interpretation.

These variables collectively form the basis of the proposed index system, incorporating both geometric and biological factors into a unified predictive model of tooth mobility.

For the Implant category, only “Head_Area_Ratio,” “Root_Area_Ratio,” “Multiplicative Interaction” and “Division Interaction” were utilized, whereas the Tooth category additionally included “Ligament_Weight” and “Root_Count” to better capture category-specific anatomical features.

Table 1. Independent variables & Dependent variable Sample

Number	Head_Area_Ratio	Root_Area_Ratio	Ligament_Weight	Root_Count	IST
234_1_36	0.633610452514933	0.331466425162557	2	Multi	85
212_3_45	0.621939166530842	0.378060833469159	1	Single	64
178_1_25	0.793570225976259	0.206429774023741	1	Single	52
91_2_21	0.662396467998998	0.296749255855085	2	Single	72
59_2_46	0.585527248193002	0.366902942098214	3	Multi	91

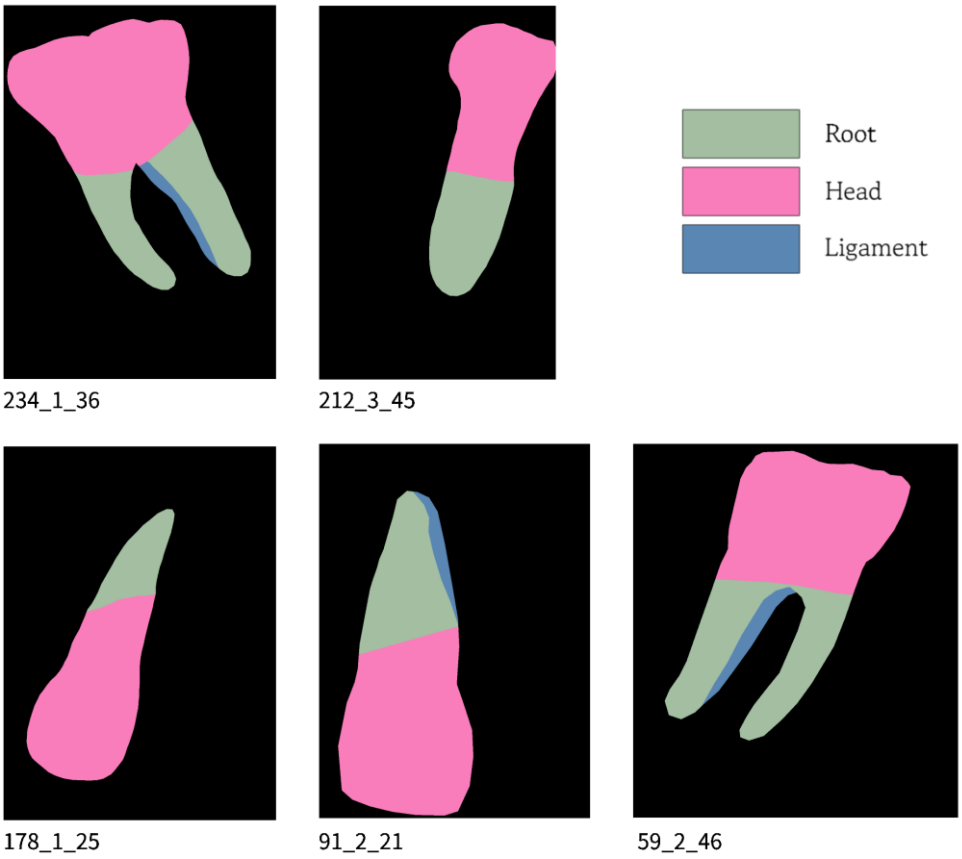

Figure 2. Independent variables & Dependent variable image sample

2.2.2 Model Architecture and Implementation

All computational procedures were implemented using Python 3.11, supported by PyTorch, OpenCV, and Scikit-learn libraries.The overall workflow included preliminary model exploration, features preprocessing, stacked ensemble design and hyperparameter optimization.

Prior to developing the final ensemble model, several baseline regressors were tested to assess the predictive potential of the data. Ridge Regression and Lasso were initially employed due to their built-in regularization properties, which are particularly effective on small datasets with potential multicollinearity. When linear models demonstrated insufficient performance, non-linear alternatives such as Random Forest, Support Vector Regression (SVR), and Polynomial Regression were explored.

During this phase, standardization was selectively applied to mitigate distributional disparities across features and facilitate fair model comparison. These preliminary experiments served to identify effective modeling approaches and guide subsequent ensemble design.

Based on insights from the exploratory stage, a two-layer stacked regression model was constructed to enhance predictive performance. The ensemble consisted of three diverse base learners:

Random Forest Regressor – Robust to noise with strong interpretability.

XGBoost Regressor– High predictive accuracy with effective overfitting control.

LightGBM Regressor – Memory-efficient and well-suited for large-scale data.

These models were selected for their capacity to handle non-linear relationships and feature interactions without requiring extensive preprocessing. Their predictions were combined via a meta-regressor implemented using LightGBM, forming a standard level-1 stacking architecture, which allows the ensemble to leverage the strengths of each base learner—combining the robustness and interpretability of Random Forests, the precision and regularization of XGBoost, and the scalability of LightGBM.

This stacking approach enhances generalization by learning from the unique error patterns of each model via a meta-learner, making the ensemble more adaptable to diverse feature types and data distributions.

To ensure that all models operated under comparable conditions, input features were standardized using z-score normalization. Categorical variables (e.g., Root_Count from root morphology) were one-hot encoded, and derived composite features (e.g., multiplicative and divisive interactions) were retained in the final feature matrix. The target variable (IST) was log-transformed prior to training to reduce skewness and stabilize model learning.

For the Implant category, features irrelevant to implants—such as Ligament_Weight and Root_Count, which pertain exclusively to natural dentition—were excluded to prevent information leakage and model bias. By tailoring the feature set in this way, the modeling strategy ensured that the learned representations were both physiologically meaningful and statistically robust across categories.

To improve the generalization capacity of the stacking ensemble, the LightGBM model—used as both a base learner and the meta-learner—was fine-tuned using Bayesian optimization via Optuna.

The optimization process yielded a set of finely tuned hyperparameters that balanced model complexity and regularization. The best configuration identified by Optuna included a learning rate of 0.02 and 104 boosting iterations (n_estimators), supporting gradual convergence and enhanced stability. Tree structure was controlled through a maximum depth of 6, 139 leaves, and a minimum of 47 samples per leaf, which effectively mitigated overfitting risks. In terms of sampling strategies, the model employed 62.73% feature subsampling (colsample_bytree) and 74.14% instance subsampling (subsample), contributing to model robustness and generalization. Regularization was applied via L1 (reg_alpha = 0.6562) and L2 (reg_lambda = 0.2589) penalties, promoting sparsity and reducing variance. These hyperparameters collectively improved the ensemble's ability to generalize across unseen data while maintaining high predictive performance.

The optimal hyperparameters derived from this process were directly applied to the LightGBM components of the final ensemble. Model fitting was conducted on 80% of the data, with the remaining 20% reserved for independent performance evaluation.

2.3 Model Evaluation

2.3.1 Statistical Analysis of Model Performance

Model performance was quantitatively assessed using three widely recognized regression metrics:

R^2 (Coefficient of Determination): Measures the proportion of variance in the dependent variable explained by the model.

MAE (Mean Absolute Error): Represents the average magnitude of absolute prediction errors, providing a direct measure of accuracy.

MSE (Mean Squared Error): Penalizes larger errors more heavily, offering sensitivity to outliers due to the squaring operation.

In addition to numerical evaluation, two forms of visual analysis were conducted to further inspect model behavior:

Residual Distribution Plots: Used to assess the spread and symmetry of residuals, supporting the detection of bias and heteroskedasticity.

Scatter Plots of Predicted vs. Actual IST Values: Used to visualize the alignment between predicted outcomes and ground truth labels, offering insight into the consistency and precision of model predictions.

2.3.2 Evaluation Strategy

To assess the predictive performance of the proposed model, I employed a stratified 80/20 hold-out validation strategy based on the log-transformed IST values. This approach ensured that the held-out test set preserved the overall distribution of the target variable, thereby providing a stable and interpretable estimate of model generalization.

While cross-validation (e.g., 5-fold) is commonly used in small-sample machine learning settings, it was not adopted in this study due to several key limitations observed during preliminary evaluation:

Limited sample size: Although data augmentation increased the total sample count from 407 to approximately 2,038, the augmented instances were derived from existing images and thus lacked independent diversity. As a result, partitioning the dataset into five folds yielded validation subsets that were both small in size and compositionally similar, undermining the stability and reliability of cross-validation estimates.

Non-uniform target distribution with outliers: The IST values were concentrated in the 40–99 range but included a small number of outliers. These outliers disproportionately influenced individual folds, resulting in unstable and unrepresentative performance estimates.

Low variance in target values across folds: A substantial number of samples shared identical IST values despite having distinct structural features. This redundancy reduced the model's capacity to generalize within each fold. It often led to negative R^2 values in cross-validation, which did not reflect the model's true predictive capacity.

Given these challenges, the stratified hold-out strategy was determined to be a more reliable alternative. It allowed for a distributionally representative and independent test set, offering a more reliable estimate of the model's real-world predictive utility.

All feature engineering and model configurations remained consistent across evaluation strategies.

It is important to note that k-fold cross-validation does not inherently cause overfitting. On the contrary, it is typically used as a safeguard against overfitting by assessing model performance across multiple subsets of the data. However, in this study, the issues related to sample redundancy, limited diversity, and outlier sensitivity impaired its effectiveness. Consequently, the poor cross-validation results observed were not due to overfitting induced by the method itself, but rather due to the unsuitability of cross-validation under the given data constraints.

3. RESULTS

3.1 Data Characterization

SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explain the output of machine learning models by assigning each feature an importance value for a particular prediction. It is based on Shapley values from cooperative game theory, which fairly attribute the contribution of each feature by considering all possible feature combinations. SHAP ensures consistency and local accuracy, and it provides a unified framework for interpreting model predictions across various algorithms (Lundberg & Lee, 2017).

Based on SHAP global importance analysis, Root_Count_Single emerged as the most influential feature in the final stacked model, followed by Head_Area_Ratio. Features such as Head_mul_Root, Head_div_Root, and Root_Area_Ratio contributed modestly to the model's predictions, suggesting that interactions among anatomical ratios still provide some predictive information. Although Ligament_Weight had low global SHAP importance, its inclusion improved performance in specific well-annotated cases. Conversely, features like Root_Count_Multi showed relatively lower SHAP values, indicating limited global influence on prediction outcomes.

Table 2. The quantitative feature contributions

Head_Area_Ratio	0.026463092252297
Root_Area_Ratio	0.00811624670968596
Ligament_Weight	0.0106117627431046
Head_div_Root	0.00820128856193532
Head_mul_Root	0.00907187055508318
Root_Count_Single	0.0342623996169037
Root_Count_Multi	0.011981388153502

3.2 Model evaluation results

To evaluate the predictive efficacy of the proposed framework, a series of regression models were trained and tested on both baseline and augmented datasets. A comprehensive comparison across all model configurations is summarized in Table 3 and Figure 3.

Among the tested configurations, the best-performing model was the Stacked Ensemble Regressor with Optuna-Tuned Base Learners, which achieved $R^2 = 0.6840$, MAE = 4.0132, and MSE = 46.6392. As shown in the scatter plot (Figure 4) reveals a strong linear relationship between predicted and actual IST values. The majority of data points distributed closely around the identity line (y=x), indicating a high level of consistency and predictive reliability. The scatter plot indicates high predictive consistency across the dataset, although slight deviations appear at the lower and upper bounds.

This model exhibited greater deviation from the identity line(Figure 5), particularly at the distribution extremes—suggesting lower predictive stability for high and low IST values. The residuals are approximately normally distributed, centered near zero, suggesting that the model does not exhibit systematic over- or under-prediction. A minor left-skew was observed, corresponding to a slight tendency to overestimate IST values in a subset of cases. Most residuals fell within the acceptable range of -15 to +15, which aligns well with the target IST range (~40–99).

A slightly less performant, but still competitive, configuration was the Stacked Ensemble with Next-Best Parameters, which yielded an $R^2 = 0.6712$, MAE = 3.9735, and MSE = 48.5285. Although this variant used a different parameter set, the ensemble structure still contributed to superior performance relative to standalone models.

In contrast, the single-model LightGBM regressor, trained using the best Optuna-tuned parameters, achieved an $R^2 = 0.5443$, MAE = 4.9295, and MSE = 52.6795. Another LightGBM variant with an alternative tuning set performed $R^2 = 0.5928$, though with marginal improvements in MSE.

These findings highlight the superiority of ensemble learning—particularly stacking approaches—in modeling the complex relationship between anatomical features and IST values.

The added diversity and regularization inherent in the ensemble framework appear to contribute to both improved generalization and resilience against overfitting, especially in data-constrained settings.

Table 3. Performance comparison of regression models on IST prediction

Model name	Description	R^2	MAE	MSE
LGBM with best Optuna-tuned	LGBM with Optuna-tuned parameters	0.5443	4.9295	52.6795
LGBM with next best parameters	LGBM with alternative tuning	0.5928	4.5154	47.0762
Stack with next best parameters	Stacking ensemble with alternative tuning	0.6712	3.9735	48.5285
Stack with best Optuna-tuned	Stacking ensemble with optimal tuning	0.6840	4.0132	46.6392

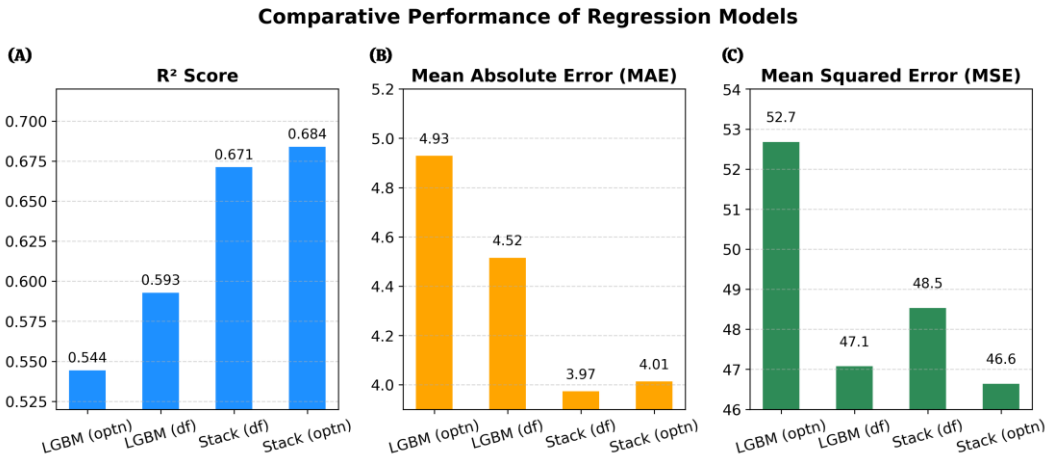

Figure 3. Comparative Performance of Regression Models

(A) The R^2 scores for each model variant demonstrate the increasing explanatory power from single LightGBM models to stacked ensembles. (B) The MAE comparison highlights improved prediction accuracy in ensemble models, with reduced average absolute error. (C) MSE values further confirm the robustness of the best-performing ensemble, showing minimized squared deviations from actual IST values.

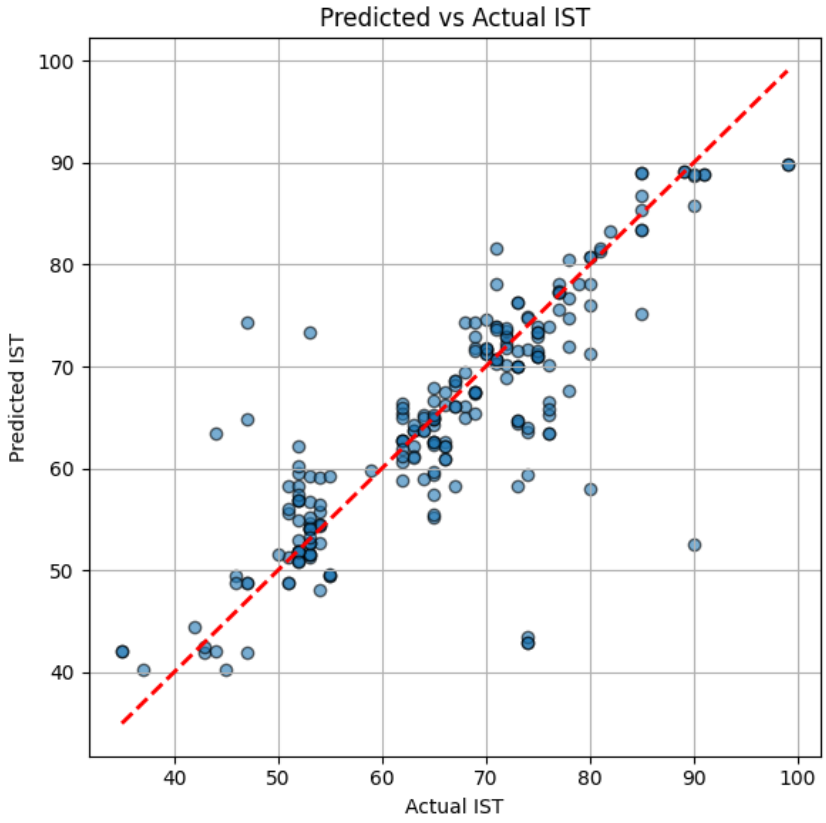

Figure 4. Scatter plot for the Best stacking ensemble model

Figure 4 illustrates the relationship between the predicted and actual IST values using the best-performing stacking ensemble model.

The red dashed line represents the ideal prediction line (i.e., y = x), where predicted values perfectly match the actual values. Most points are closely clustered around this line, indicating that the model achieves good predictive accuracy across the IST range. Some deviations are observed at the lower and higher ends of the spectrum, which may be attributed to either natural data variability or feature distribution sparsity in those regions.

Overall, the alignment between predicted and actual values confirms that the model generalizes well and captures the underlying trends in the dataset.

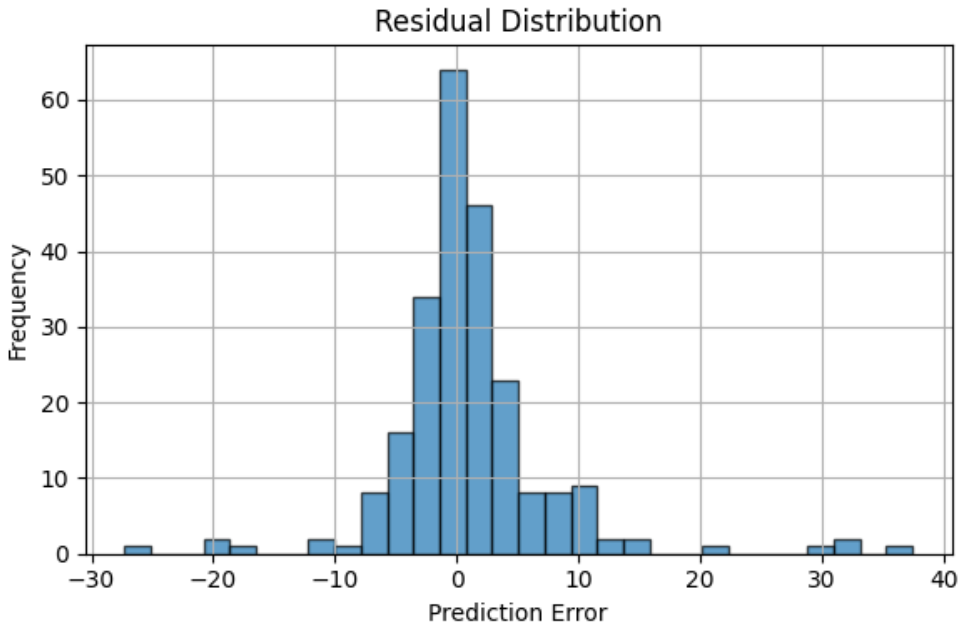

Figure 5. Residual error distribution for the Best stacking ensemble model

Figure 5 shows the residual distribution of the model's predictions, calculated as the difference between predicted and actual IST values.

The distribution is approximately symmetric and centered around zero, suggesting that the model does not exhibit significant bias toward under- or over-prediction. The sharp peak near zero indicates that a large number of predictions are very close to the true values. While there are a few outliers on both sides, their frequency remains relatively low, further supporting the model's stability and reliability.

This residual analysis provides evidence that the stacking ensemble is both accurate and consistent in its performance.

Table 4. Predict Sample of regression models

NO	Ground Truth IST	single-model LightGBM regressor	Stacked Ensemble Regressor	Best single-model LightGBM regressor	Best Stacked Ensemble Regressor
1	89	67.0678751640609	71.1676402452464	77.1332373477588	84.3374702690263
2	85	60.2180043276674	62.586692774391	76.7609012128253	77.5427520945426
3	82	61.4157276436492	63.5794775412525	70.3452923363921	70.3753408507754
4	71	66.9118596274807	67.7177509260275	77.9459981014538	72.9760140305359
5	68	82.2642009499803	76.3036935152867	72.5085160633591	71.3952839030461
6	65	68.2898967789879	69.0995562634188	65.4657097072644	65.3433472116171
7	65	68.7817063962947	67.2148531529094	70.2864725056201	67.9001878907733
8	63	63.3571639179658	66.7881551038627	67.3833460130366	63.1097995392522
9	56	63.6274764397264	64.0688536739561	61.9850946935996	61.6014424055396
10	43	76.3857699636348	75.1367352801903	59.0547807673703	58.9560645537547

3.3 Hold-Out Validation Result

The final model, previously identified as the stacking ensemble with optimal Optuna tuning, was evaluated on the hold-out set to assess generalization performance. The stacking ensemble achieved $R^2 = 0.6840$, MAE = 4.0132, and MSE = 46.6392. These results, though slightly lower than those from the augmented training set, confirm the model's generalizability and practical viability in clinical prediction tasks.

4. DISCUSSION

4.1 Impact of Sample Size and Data Sparsity

The limited sample size (n = 407) constrained model generalization and reduced statistical power. To mitigate this, geometric data augmentation was applied to simulate anatomical variability, expanding the dataset to 2,038 instances. This enrichment improved model stability—particularly for implants, where greater heterogeneity exists (Park et al., 2023; Pedram Pakravan et al., 2024).

Sparse annotation of ligament structures posed another challenge. The Ligament_Weight feature was present in only 24.14% of annotated cases, yet its inclusion yielded noticeable gains in specific scenarios where the periodontal ligament was well-defined. Although SHAP analysis indicated low global importance, its localized predictive value underscores the diagnostic potential of ligament-related features.

These findings suggest that improving segmentation accuracy could enhance ligament-aware modeling. Future work should incorporate more precise ligament annotations to fully leverage this feature's utility.

4.2 Model Performance Metrics and Interpretation

An R^2 value of 0.6840 indicates that the model explains approximately 68.4% of the variance in IST values. While not exceptionally high, this level of performance is acceptable considering the limited dataset size and the restricted scope of available features.

Such values are typical in biomedical regression tasks involving small or noisy data, especially in scenarios where ligament-related samples are underrepresented, limiting the model’s ability to learn biomechanical associations.

However, it also reflects the model's inability to fully capture latent biomechanical or clinical determinants of mobility, underscoring the need for richer and more diverse input data.

The mean absolute error (MAE) of 4.0132 should be interpreted in the context of a typical IST range of 40 to 99. This suggests that the model may be adequate for screening purposes or as an adjunct to clinical assessment, though it may not yet meet the precision standards required for high-stakes diagnostic decisions.

The mean squared error (MSE) of 46.6392 further highlights the presence of occasional large prediction deviations, indicating the need to control for outliers and reduce predictive variance.

Future improvements should aim to further enhance R^2 and minimize large errors by incorporating more diverse and semantically rich input features—particularly those capturing ligament-specific biomechanical and anatomical characteristics.

4.3 Model Performance and Feature Utility

The stacking ensemble model with Optuna-tuned base learners exhibited the best predictive performance across all tested configurations. By integrating LightGBM, XGBoost, and Random Forest as base learners—and employing Ridge Regression as the meta-learner—the ensemble effectively captured both nonlinear and linear feature interactions. This architecture enhanced generalization capacity and mitigated the risk of overfitting.

Among the input features, Head_Area_Ratio displayed a strong inverse association with IST values, aligning with clinical expectations that a higher head-to-root ratio suggests diminished periodontal support.

In contrast, Root_Area_Ratio was positively correlated with IST values, likely indicating the stabilizing biomechanical influence of the subgingival root structure.

To better capture complex anatomical interactions, two nonlinear composite features—Head_Area_Ratio × Root_Area_Ratio and Head_Area_Ratio ÷ Root_Area_Ratio—were introduced to capture potential interactions between anatomical regions:

Head_Area_Ratio × Root_Area_Ratio (Multiplicative Interaction): Reflects the joint contribution of crown and root area, emphasizing configurations where both regions are

simultaneously large. This can signal enhanced or reduced biomechanical stability depending on structural coupling, and is especially informative for linear models such as the RidgeCV meta-learner used in the stacking framework.

Head_Area_Ratio ÷ Root_Area_Ratio (Division Interaction): Captures the relative balance between the crown and root, e.g., "large head with small root" or vice versa. This proportion may reflect morphological abnormalities or stress concentration zones, providing additional shape-based cues for IST prediction.

Table 5 summarizes the contribution of each composite feature within the stacked model. These results confirm that both features contribute positively to predictive accuracy, particularly in capturing complex biomechanical interactions.

Table 5. Each Features Influence in Stack model

Configuration	R^2	MAE	MSE
Best result (All Features)	0.6840	4.0132	46.6392
Exception Area_Ratio_Head ÷Area_Ratio_Root	0.4978	5.1276	58.0490
Exception Area_Ratio_Head ×Area_Ratio_Root	0.5237	5.2649	55.0646

4.4 Limitations and Future Directions

Achieving an R^2 of 0.6840, suggesting moderate predictive capacity within the current dataset and feature scope, several limitations merit discussion:

Limited Sample Size: Despite a fivefold increase in dataset size through augmentation, the lack of genuinely independent cases limited the effective variability of the training data. In particular, multiple samples sharing identical IST values but differing anatomical features may have confused the model's learning process. Expanding the dataset to 3,000–5,000 radiographically distinct cases would likely enhance generalizability and improve the stability of model predictions.

Feature Redundancy and Low Interaction Diversity: The current features and their combinations capture structural ratios but may lack sufficient heterogeneity. Inclusion of global metrics such as Area_Ligament / Total, Total Area, or morphological complexity indices could enrich the feature space.

Absence of Clinical Metadata: Potential confounders such as age group (e.g., young vs. elderly), periodontal status, or trauma history of periodontal trauma were not included. Incorporating such non-image features could enhance model interpretability and clinical relevance through stratified modeling.

4.5 Category-Specific Validation

The final model was trained on a combined dataset composed of both tooth-only and implant-only categories, using a unified modeling framework. Performance across these subsets was:

Tooth-only Category: $R^2 = 0.5526$, MAE = 4.9249, MSE = 51.7141

Implant-only Category: $R^2 = 0.3957$, MAE = 6.2847, MSE = 66.6596

These results reveal higher predictive accuracy in tooth-only category compared to implant-only category. The lower performance in the implant category may stem from greater anatomical

variability or data imbalance, and highlights an area for future model refinement or stratified training.

4.6 Lower Predictive Performance in Implant-only Category

Subgroup analysis revealed that model performance was substantially lower for implant-only category, compared to tooth-only category. This discrepancy likely reflects greater anatomical and biomechanical heterogeneity in implant-supported structures, which are less biologically standardized than natural teeth.

Future improvements could include:

Incorporating implant-specific features, such as implant length, material type, crown-abutment design, or peri-implant bone density (Pedram Pakravan et al., 2024).

Including peri-implant morphological characteristics (Jang et al., 2022) (e.g., bone-implant contact ratio, surrounding bone remodeling patterns).

Using separate sub-models or ensemble components tuned specifically for implant cases to reduce modeling noise from heterogeneity.

These implant-specific enhancements may help close the performance gap and ensure more balanced predictive capacity across clinical scenarios.

5.CONCLUSION

This study demonstrated the feasibility of predicting tooth mobility by integrating radiographic anatomical features with objective stability measurements derived from the device. By aligning clinical insight with image-based structural analysis, I developed a supervised learning framework capable of estimating IST values—serving as a surrogate marker for periodontal support—based on periapical radiographs (Özbay et al., 2024).

The proposed stacking ensemble model, incorporating LightGBM, XGBoost, and Random Forest as base learners with a Ridge Regression meta-learner, achieved the highest predictive performance across all tested configurations. Feature analysis revealed that head-to-root ratios and composite morphological interactions significantly contributed to prediction accuracy (Park et al., 2023), while even sparsely annotated periodontal ligament features provided meaningful signal when appropriately embedded within the ensemble architecture.

Looking forward, future studies should expand the dataset to improve statistical power, incorporate clinical metadata to enhance context sensitivity, and explore multi-modal frameworks that integrate imaging, patient history, and biomechanical data (Huang et al., 2022). These directions will be essential for refining the predictive capacity of AI-based diagnostic systems and bring them closer to real-world clinical integration.

REFERENCES

Ari, T., Saglam, H., Öksüzoğlu, H., Kazan, O., Bayrakdar, I. S., et al. (2022). Automatic feature segmentation in dental periapical radiographs. Diagnostics, 12(9), 2211. https://doi.org/10.3390/diagnostics12123081

Benakatti, V. B., Nayakar, R. P., Anandhalli, M., & Lagali-Jirge, V. (2022). Accuracy of machine learning in identification of dental implant systems in radiographs: A systematic review and meta-analysis. Journal of Indian Academy of Oral Medicine and Radiology, 34(3), 354–358. https://doi.org/10.4103/jiaomr.jiaomr_86_22

Celik, B., Savaştar, E. F., Kaya, H. İ., & Celik, M. (2023). The role of deep learning for periapical lesion detection on panoramic radiographs. Dentomaxillofacial Radiology, 52(1), 20220180. https://doi.org/10.1259/dmfr.20230118

Cha, J.-Y., Yoon, H.-I., Yeo, I.-S., Huh, K.-H., & Han, J.-S. (2021). Peri-implant bone loss measurement using a region-based convolutional neural network. Journal of Clinical Medicine, 10(18), 4231. https://doi.org/10.3390/JCM10051009

Elemek, E. (2022). Periodontal disease severity, tooth loss, and periodontal stability in private practice. Nigerian Journal of Clinical Practice, 25(6), 931–937. https://doi.org/10.4103/njcp.njcp_1952_21

Hartmann, M., Dirk, C., Reimann, S., Keilig, L., Konermann, A., Jäger, A., & Bourauel, C. (2017). Influence of tooth dimension on the initial mobility based on plaster casts and X-ray images. Journal of Orofacial Orthopedics, 78(4), 285–292. https://doi.org/10.1007/s00056-016-0082-9

Hoss, P., Meyer, O., Wölfle, U. C., Wülk, A., & Meusburger, T. (2023). Detection of periodontal bone loss on periapical radiographs. Journal of Clinical Medicine, 12(3), 834. https://doi.org/10.3390/jcm12227189

Huang, Z., Zheng, H., Huang, J., Yang, Y., Wu, Y., Ge, L., & Wang, L. (2022). The construction and evaluation of a multi-task convolutional neural network for a cone-beam computed-tomography-based assessment of implant stability. Diagnostics, 12(11), 2673. https://doi.org/10.3390/diagnostics12112673

Jang, W.-S., Kim, S., Yun, P., Jang, H.-S., Seong, Y. W., Yang, H. S., & Chang, J. S. (2022). Accurate detection for dental implant and peri-implant tissue by transfer learning. BMC Oral Health, 22(1), 174. https://doi.org/10.1186/s12903-022-02539-x

Kim, G. Y., Kim, S., Chang, J.-S., & Pyo, S.-W. (2024). Advancements in methods of classification and measurement used to assess tooth mobility: A narrative review. Journal of Clinical Medicine, 13(1), 142. https://doi.org/10.3390/jcm13010142

Lang, N. P., & Bartold, P. M. (2018). Periodontal health. Journal of Periodontology, 89(Suppl 1), S9–S16. https://doi.org/10.1002/JPER.16-0517

Lee, D. H., Shin, Y. H., Park, J. H., Shim, J. S., Shin, S. W., & Lee, J. Y. (2020). The reliability of AnyCheck device related to healing abutment diameter. Journal of Advanced Prosthodontics, 12(2), 83–88. https://doi.org/10.4047/jap.2020.12.2.83

Lee, J., Pyo, S. W., Cho, H. J., An, J. S., Lee, J. H., Koo, K. T., & Lee, Y. M. (2020). Comparison of implant stability measurements between a resonance frequency analysis device and a modified damping capacity analysis device: An in vitro study. Journal of Periodontal & Implant Science, 50(1), 56–66. https://doi.org/10.5051/jpis.2020.50.1.56

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). https://doi.org/10.48550/arXiv.1705.07874

Medina-Sotomayor, P., Pascual-Moscardo, A., & Camps, A. I. (2019). Accuracy of 4 digital scanning systems on prepared teeth digitally isolated from a complete dental arch. Journal of Prosthetic Dentistry, 121(5), 811–820. https://doi.org/10.1016/j.prosdent.2018.08.020

Meirelles, L., Siqueira, R., Garaicoa-Pazmino, C., Yu, S. H., Chan, H. L., & Wang, H. L. (2020). Quantitative tooth mobility evaluation based on intraoral scanner measurements. Journal of Periodontology, 91(2), 202–208. https://doi.org/10.1002/JPER.19-0282

Okuhama, Y., Nagata, K., Kim, H., Tsuruoka, H., Atsumi, M., & Kawana, H. (2022). Validation of an implant stability measurement device using the percussion response: A clinical research study. BMC Oral Health, 22(1), 286. https://doi.org/10.1186/s12903-022-02320-0

Özcan, C. (2023). Detection of separated endodontic instruments on periapical radiographs. Australian Endodontic Journal. https://doi.org/10.1111/aej.12822

Pakravan, P., Behazin, S., Mohammadpour, M., Soorati, A. B., & Rahimipour, E. (2024). Predicting implant success using AI for peri-implant bone loss. World Journal of Biology Pharmacy and Health Sciences. https://doi.org/10.30574/wjbphs.2024.20.3.0958

Park, J.-H., Moon, H. S., Jung, H.-I., Hwang, J., Choi, Y.-H., & Kim, J.-E. (2023). Deep learning and clustering approaches for dental implant size classification. Scientific Reports, 13, 924. https://doi.org/10.1038/s41598-023-42385-7

Tada, S., Allen, P. F., Ikebe, K., Zheng, H., Shintani, A., & Maeda, Y. (2015). The Impact of the Crown-Root Ratio on Survival of Abutment Teeth for Dentures. Journal of dental research, 94(9 Suppl), 220S–5S. https://doi.org/10.1177/0022034515589710

Ziegler, A., Keilig, L., Kawarizadeh, A., Jäger, A., & Bourauel, C. (2005). Numerical simulation of the biomechanical behaviour of multi-rooted teeth. European Journal of Orthodontics, 27(4), 333–339. https://doi.org/10.1093/ejo/cji020

ABSTRACT IN KOREAN

치근단 방사선 사진과 임플란트 안정성 테스트 데이터를 이용한 치아의 동요도 및 임플란트 안정성 예측 연구

본 연구는 치근단 방사선 영상에서 추출한 해부학적 특징과 AnyCheck 장치를 통해 측정한 초기 안정성 테스트(IST) 값을 통합하여, 치아 동요도와 임플란트 안정성을 예측하는 머신러닝 프레임워크를 제안한다. 총 407 개의 주석된 방사선 영상은 기하학적 증강을 통해 2,038 개로 확장되었으며, 치관-치근 면적 비율, 치주 인대 가시성, 치근 형태와 같은 구조적 지표를 복합 특징으로 재구성하였다. LightGBM, XGBoost, Random Forest 기반 학습기와 Ridge Regression 메타 학습기로 구성된 스태킹 앙상블 회귀 모델을 구축하였다.

최종 모델은 R^2 = 0.6840, MAE = 4.0132, MSE = 46.6392 의 성능을 보였으며, 예측된 IST 값과 실제 측정값 간의 높은 일치도를 나타냈다. SHAP 분석에 따르면, 치근 유형과 치관-치근 비율이 가장 영향력 있는 예측 변수로 확인되었으며, 드물게 주석된 인대 정보도 모델 성능 향상에 기여하였다.

이러한 결과는 해부학적 구조를 고려한 영상 기반 회귀 모델이 치주 지지력 및 임플란트 안정성을 비침습적으로 평가할 수 있는 가능성을 보여준다. 제안된 프레임워크는 방사선 형태학과 생역학적 지표를 연결하여, 치의학 임상에서 신뢰할 수 있는 진단 보조 도구로 활용될 수 있다.

핵심 되는 말: 치아 이동도, IST 값, 치근단 방사선 사진, 머신러닝, 치주 평가