Predicting risk indicators for periodontitis in a Korean population: Machine learning algorithms approach

Hee-Jung Park, Young-Ro Lee, Xianhua Che, Jun-Min Kim

Submitted to: JMIR Public Health and Surveillance
on: June 25, 2024

Table of Contents

Predicting risk indicators for periodontitis in a Korean population: ?Machine learning algorithms approach

Hee-Jung Park[1]; ?Young-Ro ?Lee[2]; ?Xianhua ?Che[3]; ?Jun-Min ?Kim[4]

[1]?Kangwon National University Samcheok-si KR
[2]?Seoul National University ?Seoul KR
[3]?Department of Health Policy Research, Daejeon Public Health Policy Institute ?Daejeon KR
[4]?Hansung University ?Seoul KR

Corresponding Author:
Hee-Jung Park
?Kangwon National University
Department of Dental Hygiene, College of Health Science, Kangwon National University
Samcheok-si
KR

Abstract

Background: ?Periodontitis is a multifactorial disease that involves numerous risk factors and indicators. While a few factors, such as age, uncontrolled diabetes, and smoking, have been well established as true risk factors for periodontitis, many risk factors and indicators remain debatable due to conflicting data from previous studies. This calls for a novel approach to data analysis to improve the accuracy of risk assessments and disease predictions.

Objective: ?This study aimed to assess the ability of machine learning approaches to identify important risk indicators for periodontitis using data from the 2015–2018 Korea National Health and Nutrition Examination Survey of 13,946 subjects.

Methods: ?The severity of periodontitis was categorized as non-severe and severe according to the community periodontal index. Machine learning models such as classification and regression tree, gradient boosting machine, random forest, extareme gradient boost (XGBoost), and multilayer perceptron (MLP) were developed, and their performance based on the area under the receiver operating characteristic curve (AUC) was compared with that of the conventional logistic regression analysis.

Results: ?XGBoost and MLP showed higher performance than the logistic regression model. The important risk indicators for periodontitis were age, sex, education, smoking, blood pressure, use of interdental cleaning aids, and glycated hemoglobin. Interestingly, further analysis showed that glycated hemoglobin and frequency of binge drinking were significant indicators of severe periodontitis. While the findings of this study showed moderate AUC in machine learning models, it confirmed consistent risk indicator rank, which was derived from feature importance analysis such as Gini impurity, permutation importance, and Shapley additive explanations.

Conclusions: ?Collectively, our findings suggest the importance of considering a novel data analysis methodology, such as machine learning, for the better management of periodontitis.

(JMIR Preprints 25/06/2024:63621)
DOI: https://doi.org/10.2196/preprints.63621

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

✓

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a h

Original Manuscript

Machine learning algorithms predicting risk indicators for periodontitis in a Korean population

Young-Ro Lee[1], Xianhua Che[2], Jun-Min Kim[3], and Hee-Jung Park[4,*]
[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul, 08826, Republic of Korea
[2]Department of Health Policy Research, Daejeon Public Health Policy Institute, Daejeon, Republic of Korea
[3]Department of Electronics and Information Engineering, Hansung University, Seoul, Republic of Korea
[4]Department of Dental Hygiene, Kangwon National University, Samcheok, Republic of Korea

[*] Correspondences: Hee-Jung Park (phealth172@kangwon.ac.kr)

Abstract

Background: Periodontitis is a multifactorial disease that involves numerous risk factors and indicators. While a few factors, such as age, uncontrolled diabetes, and smoking, have been well established as true risk factors for periodontitis, many risk factors and indicators remain debatable due to conflicting data from previous studies. This calls for a novel approach to data analysis to improve the accuracy of risk assessments and disease predictions.

Objective: This study aimed to assess the ability of machine learning approaches to identify important risk indicators for periodontitis using data from the 2015–2018 Korea National Health and Nutrition Examination Survey of 13,946 subjects.

Methods: The severity of periodontitis was categorized as non-severe and severe according to the community periodontal index. Machine learning models such as classification and regression tree, gradient boosting machine, random forest, extareme gradient boost (XGBoost), and multilayer perceptron (MLP) were developed, and their performance based on the area under the receiver operating characteristic curve (AUC) was compared with that of the conventional logistic regression analysis.

Results: XGBoost and MLP showed higher performance than the logistic regression model. The important risk indicators for periodontitis were age, sex, education, smoking, blood pressure, use of interdental cleaning aids, and glycated hemoglobin. Interestingly, further analysis showed that glycated hemoglobin and frequency of binge drinking were significant indicators of severe periodontitis. While the findings of this study showed moderate AUC in machine learning models, it confirmed consistent risk indicator rank, which was derived from feature importance analysis such as Gini impurity, permutation importance, and Shapley additive explanations.

Conclusion: Collectively, our findings suggest the importance of considering a novel data analysis methodology, such as machine learning, for the better management of periodontitis.

Keywords: risk indicators, periodontitis, machine learning, feature importance, SHAP

Introduction

Despite its complex and multifactorial pathogenesis, periodontitis is mostly driven by the presence of dental plaque[1]. A host in a periodontically vulnerable state is susceptible to the induction of antimicrobial inflammation and disruption of microbial homeostasis[2]. The onset and progression of periodontitis vary depending on the predisposing factors of the host. In addition, given its chronic nature, periodontitis cannot be fully cured, and a prevention-oriented approach and continuous control of undesirable factors are necessary[3].

Many risk assessment tools[4] and prediction models[5,6] have been developed to determine individual susceptibility to periodontitis. However, except for age, uncontrolled diabetes, and smoking, which are true risk factors with an established causal relationship[7], parameters that involve categories such as sociodemographic factors, pathogens, oral examination, oral hygiene, and lifestyle have been applied inconsistently[5]. Numerous factors have been associated with periodontitis[4-7]. However, we cannot prematurely assume that all factors have the same effect, and we need to establish the relative importance of each factor for accurate risk assessment. This change in approach will allow individuals with important risk indicators to be screened and provided with appropriate and efficient interventions. Unfortunately, a significant knowledge gap remains regarding the relative importance of the various risk factors of periodontitis[8].

Machine learning has been widely applied to the diagnosis and risk prediction of diseases. Machine learning can create a unique algorithm by recognizing the interactions among multiple variables[9]. In the study and treatment of periodontitis, these technologies have been applied to periodontal diagnosis using radiographic imaging[10-12] or immunologic parameters[13], as well as for developing predictive models for tooth loss[14]. However, while regression analysis has been extensively applied to ascertain the risk factors of periodontitis, the importance of machine learning as a methodology has been poorly explored. This is because of the difficulty driven by the 'black box' characteristic, where the complicated inner workings hinder the interpretation of the results for prediction. To overcome this lack of interpretability, recent studies have suggested a model-specific method driven by the specific model algorithm and model-agnostic method, which is independent of the prediction model algorithm. Shapley additive explanations (SHAP) is an approach currently used as a representative interpretable artificial intelligence method because it is model agnostic and provides a deeper analysis of feature importance—probabilistic distribution of feature importance and interaction between features.

Therefore, the objectives of this study were to 1) identify critical risk indicators for periodontitis via machine learning approaches according to the severity of periodontitis, 2) compare the outcome with the results of conventional statistical analysis, and 3) analyze the feature importance regarding feature effects and interaction between risk indicators using a SHAP summary plot and SHAP dependence plot.

Methods

Data source

A nationally representative database obtained from the sixth Korea national health and nutrition examination survey (KNHANES) (2015) and seventh KNHANES (2016–2018) by the Korean Disease Control and Prevention Agency (KDCA) (https://knhanes.kdca.go.kr/knhanes) was used. A total of 23,466 participants were recruited, and 17,530 of them completed the KNHANES (2015–2018). The inclusion criterion was an age greater than 30 years, and 13,946 subjects were included in the data. The methodology used in this study is illustrated in Figure 4. The KNHANES protocols were approved by the Institutional Review Board of the KDCA, and the participants provided written informed consent at baseline. The Institutional Review Board of Seoul National University School of Dentistry approved this study, which qualified for exempt status due to the analysis of de-identified secondary data (exemption number: S-D20210001).

Definition of periodontitis

The community periodontal index (CPI) developed by the World Health Organization (WHO) was used to assess periodontal conditions, and a CPI probe that met the WHO guidelines was used[15]. Oral health examinations were conducted by public health dentists who had received training twice a year to ensure the reliability of the periodontal health survey. The mouth was divided into three sextants in each arch, and the presence of permanent teeth was included in the examination. The CPI was scored on a scale of 0 to 4 as follows: 0 points for healthy periodontal tissue, 1 point for bleeding only with probing, 2 points for periodontal tissue with calculus or plaque retentive factors, 3 points for periodontal tissue with shallow periodontal pockets (pocket depth 3.5–5.5 mm), and 4 points for periodontal tissue with deep periodontal pockets (pocket depth > 5.5 mm)[16]. After assigning the scores, the highest CPI score among the six sextants was chosen. In this study, a score of 0–2 was defined as no periodontitis (NoP), and a score of 3 or 4 was defined as total periodontitis

(TP). A CPI score of 3 was defined as non-severe periodontitis (NSP), and a score of 4 was defined as severe periodontitis (SP)[17]. The important risk indicators in predicting TP against NoP were identified, and the important risk indicators in predicting SP against NSP were also identified.

Risk indicators

The risk predictors used in the prediction models of periodontitis published so far were summarized in a review paper by Mi et al.[5], among which 20 variables from the KNHANES data were selected as variables for this study. In addition, high-sensitivity C-reactive protein (hs-CRP) was added because it is a systemic inflammatory mediator and has been reported as a potential diagnostic marker while also having an association with periodontitis[18].

Age, sex, income (low, lower-middle, upper-middle, and upper), and education level (elementary school, middle school, high school, and college or higher) were selected as risk predictors for this study. Income was defined as the average monthly household income adjusted for the number of family members in the household (household income divided by the number of family members) and was categorized in terms of household income quartiles (low, lower-middle, upper-middle, and upper).

General health status included glycated hemoglobin (HbA1c), fasting blood glucose (FBG) (< 100 mg/dL or ≥ 100 mg/dL), diabetes (normal: FBG < 100 mg/dL or HbA1c < 5.7%, pre-diabetes: 100 mg/dL ≤ FBG < 126 mg/dL or 5.7% ≤ HbA1c < 6.5%, and diabetes: FBG ≥ 126 mg/dL or HbA1c ≥ 6.5%), triglycerides (< 200 mg/dL or ≥ 200 mg/dL), high-density lipoprotein (HDL) cholesterol (< 240 mg/dL or ≥ 240 mg/dL), and abdominal obesity (no or yes). Metabolic syndrome was classified as positive if three or more of the following five criteria were met[19]: (1) abdominal obesity (the Asia-Pacific criteria for obesity based on waist circumference ≥ 90 cm in males or ≥ 80 cm in females); (2) triglycerides ≥ 150 mg/dL (1.7 mmol/L); (3) HDL cholesterol < 40 mg/dL (1.03 mmol/L) in males or < 50 mg/dL (1.3 mmol/L) in females; (4) hypertension as blood pressure ≥ 130/85 mmHg; (5) fasting glucose levels ≥ 100 mg/dL (5.6 mmol/L). Blood pressure was defined as normal, prehypertensive, or hypertensive. Normal is a systolic blood pressure < 120 mmHg and a diastolic blood pressure < 80 mmHg, and pre-hypertension is a systolic blood pressure of 120–140 mmHg and a diastolic blood pressure of 80–90 mmHg. Hypertension was defined as a systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg. Hs-CRP levels were classified as < 1 mg/L, 1–3 mg/L, and ≥ 3 mg/L[20]. Body mass index was classified as < 23 kg/m2, 23–25 kg/m2, and ≥ 25 kg/m2[21].

Lifestyle habits included smoking, quantity of alcohol consumed per occasion, and frequency of binge drinking. Smoking status was classified as never smoker, former smoker, or current smoker (current smoker at the time of the interview). The drinking quantity per occasion was classified as ≤ 2, 3–4, 5–9, and ≥ 10 glasses. Binge drinking was defined as consuming ≥ 5 standard drinks (≥ 4 drinks for women) consecutively on one occasion. We classified the frequency of binge drinking into three categories: none (non-binge drinking), < 1 time/month, and ≥ 1 time/month. Self-reported stress levels were divided into very stressed, stressed, slightly stressed, and stress free. Additionally, oral health behaviors included the use of interdental cleaning aids such as dental floss, interdental brushes (no or yes), frequency of toothbrushing per day (1–2 times or ≥ 3 times), and dental visits within a year (no or yes).

Statistical analysis

Statistical analyses for the complex sampling design were performed by applying stratum variance estimates, stratification variables, and sampling weights in Stata Version 15. According to the

statistical guidelines of the KDCA, our analyses incorporated sampling weights to obtain nationally representative estimations and to consider the annual weight for four years. We compared the weighted frequencies of the general sample characteristics according to periodontal status using the chi-squared test. One-way analysis of variance was performed for continuous variables. Multivariate logistic regression analyses were performed to examine the factors associated with TP and SP. The odds ratios (ORs) and 95% confidence intervals (CIs) were estimated using a logistic model.

Data distribution

The overall data distribution was visualized using T-distributed stochastic neighbor embedding (T-SNE), which embeds multidimensional data into a two-dimensional visualization while preserving the T-distribution distance between neighbors. The embedding transformation is optimized using the gradient descent algorithm with the cost function of the Kullback-Leibler divergence[22].

Model training

As a reliable verification method, 12-fold cross-validation was applied. The dataset was divided into 12 sets, and each set was used as the test set for each iteration. The final output regarding performance and interpretation was deduced from the average of the 12 experiments for each model[23]. Each training dataset was normalized using the min-max scaler, which subtracts each feature from its minimum and divides it with the maximum-minimum. With the min-max scaler obtained in the training dataset, the test dataset was also normalized to be better inserted into the prediction model.

Seven different types of predictive models were tested using two regression and five machine learning algorithms. The scikit-learn library was used to build the predictive models. Specific hyperparameters in the experiment can be found in the Supplementary Notes online[24]. The linear regression model utilizes a linear combination to predict the dependent variable. Logistic regression uses a logit function and has an outcome between 0 and 1.

For machine learning, classification and regression tree (CART), Gradient Boost Machine (GBM), random forest, extareme gradient boost (XGBoost), and multilayer perceptron (MLP) were tested. The CART algorithm predicts outcomes by building a binary decision tree. Unlike conventional decision tree algorithms, an entropy matrix is used to determine split points[22,25]. GBM updates the weighting errors using gradient descent methods for optimizing training[26]. Random forest is an ensemble machine learning model that uses majority voting among multiple decision trees to overcome the overfitting problem. Random forest, a representative bagging algorithm, is known to outperform decision tree methods in models with high variance[22,27,28]. XGBoost improves the overfitting and slow training problem of GBM by parallel processing and regularized boosting while maintaining the gradient boosting algorithm[22,29,30]. MLP utilizes a large number of nodes and connections between layers to embody the complexity in the level of nonlinearity, particularly as the layer becomes deeper. To prevent overfitting resulting from complex prediction, a regularization term is added to the cost function for the gradient descent[24,31,32].

Performance comparison

To decide which predictive model should be used for feature importance analysis, an appropriate performance standard is required. The performances, except area under the receiver operating characteristic curve (AUC), are dependent on the gold standard between 0 and 1 to decide whether the predictive value indicates positive (TP, SP) or negative (NoP, NSP). To reflect both negative and positive performance, AUC was implemented to decide which model to analyze[33].

Variable importance

For the analysis of variable importance, Gini impurity for XGBoost and permutation for MLP were tested. To observe the direction of feature importance and the interaction between features, SHAP was additionally applied to XGBoost and MLP.

The Gini impurity is used for feature importance in decision trees in many areas of research, including medicine. Purity is proportional to the performance of the classification because the output for each class is composed purely of its class data if the classification is performed perfectly. Each node in a decision tree classifies the input data using a feature to increase the overall classification performance and decrease the Gini impurity. The decrease in the Gini impurity in each node represents the importance of a feature in that node[22,34-36].

The permutation importance estimates the change in the accuracy after permuting the data of a single feature to infer the importance of the feature in the model, as the change would be proportionate to the importance. Many analyses of medical big data utilize this mechanism for various networks, including tree-based machine learning models and MLP[37-40]. There are various ways to define change, and this study defined it as follows[41,42]. The score was obtained by permuting the value of a target feature in the test dataset. Because permutation is a random process, the permutation importance can be varied in each trial. To supplement this variance, the permutations were performed 100 times and averaged for each feature.

SHAP explains the inner workings of a “black box” using the Shapley value from game theory. Using SHAP, more than the size of the feature importance can be calculated. The distribution and sign of the SHAP values show the importance distribution and direction of each data point, which is depicted in the SHAP summary plot. By observing the SHAP values with respect to other feature values, the interaction between features in machine learning can be understood. The SHAP interaction index from game theory indicates the level of interaction between features, and the SHAP dependence plot depicts how the interaction works. Tree Explainer for XGBoost and Kernel Explainer for MLP were also used[43,44].

Data availability

The datasets generated or analyzed during the current study are available from the KNHANES by the KDCA (https://knhanes.kdca.go.kr/knhanes). The datasets and codes used in this study are available at https://github.com/junmin83/indicators-for-periodontitis.

Results

The T-SNE graphs in Figure 1 demonstrate the 21-dimensional characteristics of the subjects analyzed with 21 risk indicators in two dimensions[45]. When all the subjects were classified as (NoP or TP, and the subjects with periodontitis were classified as having NSP or SP, they were divided into various clusters rather than two groups. Each cluster showed a different pattern; some clusters had a high percentage of one disease state, whereas the two disease states overlapped in many other clusters. That is, the disease manifests differently in individuals with similar risk indicators.

Table 1 presents the detailed characteristics of the study subjects categorized according to their periodontal status. The prevalences of NSP and SP were 24.3% and 9.4%, respectively. The mean

ages of participants with NSP or SP (57.1±0.3 and 57.8±0.4 years) were significantly higher than that of participants without periodontitis (51.2±0.2 years) ($P < 0.001$). Overall, participants with varying periodontal statuses showed significant differences in all variables tested, except for HDL cholesterol.

Univariate logistic regression analysis demonstrated that old age, male sex, low education level, low income level, high triglyceride level, glycated hemoglobin level, metabolic syndrome, high body mass index, smoking, and use of interdental cleaning aids were significantly associated with TP (Table 2). Glycated hemoglobin, metabolic syndrome, and frequency of binge drinking (≥ 1 time/week) were significantly associated with SP.

To determine which machine learning algorithm had the best performance in observing feature importance, seven predictive models were built and tested. The performance levels, such as the AUC, sensitivity, specificity, positive predictive values, negative predictive values, and accuracy, were compared in the prediction of TP and SP (Supplementary Tables S1 and S2 online). The AUC was used to select the best model, as the AUC values reflect performance regarding predictiveness for negative and positive together. Our results showed that XGBoost and MLP had superior performance compared to other machine learning models and were marginally better than regression models. For predicting TP against NoP, the AUCs of XGBoost and MLP were the highest. In predicting SP against NSP, XGBoost had the highest performance, followed by regression and MLP.

In Tables 3 and 4, the importance rank of the risk indicators obtained from the XGBoost and MLP models are shown from the highest to lowest rank[45]. Risk indicators with high rankings in all four analysis methods were considered to be important. Age, sex, education, smoking, blood pressure, use of interdental cleaning aids, and glycated hemoglobin were important risk indicators for TP in both XGBoost and MLP (Table 3). Glycated hemoglobin and frequency of binge drinking showed a substantial increase in importance in SP compared with TP in both models (Table 4). Age, sex, education, and smoking were important risk indicators in SP using MLP and similar in TP and SP using XGBoost.

In Figure 2, the importance of the variables applied with SHAP analysis in the XGBoost model is shown in the order of relative importance. Although the performances of XGBoost and MLP in TP were similar, XGBoost exhibited slightly higher performance in SP. Therefore, SHAP analysis was performed using XGBoost. By applying SHAP, information about the degree of importance of each feature and the direction in which it acts as important can be obtained. Figure 2a and 2b show the relative importance of the factors in predicting TP and SP, respectively, and Figure 2c shows the classification and color labeling of the variables used in the analysis. As shown in Figure 2a, the influence of age in predicting TP was substantial, and in particular, with younger age, there was a tendency to predict NoP. As shown in Figure 2b, the influence of glycated hemoglobin in predicting SP was profound, and in particular, a state of high glycated hemoglobin was more likely to be predicted as SP than low glycated hemoglobin.

Sex, smoking, and education were important factors not only in TP but also in SP, but not in the logistic regression analysis. The SHAP dependence plot confirmed which interaction the XGBoost model used for the prediction. The risk indicator automatically extracted as having the highest interaction with smoking was education level. In Figure 3a, the absolute value of SHAP was larger in the case of low education level than in the case of high education level, which indicated that smoking was more predictive when accompanied by a low education level. Sex was the factor automatically extracted from the algorithm as having the highest interaction with education level. In Figure 3b, the absolute SHAP values for males were larger than those for females at all educational levels, which

showed that when paired with males, educational level was used more importantly in making predictions.

Discussion

In this study, we aimed to evaluate the relative importance of 21 risk indicators of periodontitis via machine learning for disease severity prediction and compare the outcomes obtained by machine learning and logistic regression analysis. Unlike logistic regression analysis, artificial intelligence can recognize nonlinear relationships, such as interactions, and the more important a variable is, the more weight it is given in a prediction; therefore, it is suitable to analyze the relative importance of variables. Among the algorithms, XGBoost and MLP performed the best and were thus used for the assessment of the variables. The important risk indicators for TP in both XGBoost and MLP were age, sex, education, smoking, blood pressure, use of interdental cleaning aids, and glycated hemoglobin. In addition, the risk indicators with increased importance in SP were glycated hemoglobin and frequency of binge drinking.

Age was extremely important in TP, and its importance decreased in SP. Grossi et al. classified periodontitis into five categories according to attachment loss, and age was the most strongly associated factor;OR was 1.72 at 35–44 years old, increasing to 9.01 at 65–74 years old[46]. A study by Eke et al. used the Centers for Disease Control and Prevention/American Academy of Periodontology case definition and the 2009–2012 National Health and Nutrition Examination Survey (NHANES) data, and the prevalence of TP increased by 1.52 times for those aged 45–54 years, 1.79 times for those aged 55–64 years, and 2.71 times for those aged 65 years and older, using 30–44 years of age as a reference. The prevalence of SP did not show such a tendency, increasing by 1.52 times for those aged 45–54 years, 2.58 times for those aged 55–64 years, and 2.26 times for those aged 65 years and older, with 30–44 years of age as a reference[8].

Sex was not statistically significant for SP in the logistic regression analysis, but it was an influential factor in both TP and SP using both MLP and XGBoost. Furthermore, the rank increased in SP. Montero et al. analyzed the 2011–2012 NHANES and reported that the OR for 4–6 mm of clinical attachment loss was 2.58, and the OR for ≥ 6 mm of clinical attachment loss was 5.98 in men when women were set as a reference[6]. In the 2009–2012 NHANES, the prevalence of TP in men increased by 1.47 times and the prevalence of SP by 2.84 times[17].

Glycated hemoglobin was the most influential factor in the prediction of SP and was ranked higher than fasting blood glucose. It is well known that the prevalence of SP is high in diabetic patients. Glycated hemoglobin reflects glycemic control for the previous 2–3 months and is highly correlated with diabetes complications[47]. In particular, it has been reported that uncontrolled diabetes, defined as more than 7.0% of glycated hemoglobin, accelerated periodontal attachment loss, and tooth loss[48].

Metabolic syndrome was more important in SP using MLP and similar in TP and SP using XGBoost. Several studies have reported an association between metabolic syndrome and SP[49-51], and metabolic syndrome is considered to contribute to the further destruction of periodontal tissue through upregulated inflammatory responses[52].

The importance of the frequency of binge drinking was higher in SP than in TP using both XGBoost and MLP, and this was consistent with the logistic regression analysis. The importance was higher than that of drinking quantity and even more so than smoking. Most studies on the association of alcohol with periodontitis have focused on drinking quantity[53], while the frequency of binge drinking has not been used as a variable in such studies. A recent study reported that aggravated

alveolar bone loss was linked to binge alcohol intake in rats[54]. Further epidemiological studies are necessary to investigate the effect of binge drinking frequency on periodontitis severity.

Although several important variables identified by machine learning were generally similar to the results of the logistic regression analysis, several differences were observed. Blood pressure was considered an important factor for TP by machine learning, but it was not significant in the logistic regression analysis. Sex, smoking, education, and age were not significant for SP in the logistic regression, but analysis using machine learning showed that they were important in SP. Figure 3 indicates that the interaction between the variables was confirmed through the SHAP dependence plot. A low education level is a widely accepted risk factor for periodontitis[55]. It is well demonstrated that cigarette smoke exposure is also a strong risk factor for periodontitis due to the increased presence of T cells in the oral cavity and decreased elastase and neutrophil levels[56]. Our results support the hypothesis that low education level is associated with smoking status and aggravated periodontal conditions. Additionally, current research also indicates males with lower education levels are at a higher risk of severe periodontal disease.

This study has several limitations. The AUC values remained relatively low for all models, which is most likely due to the KNHANES data having missing information about select pathogens and diseases. Risk indicators such as periodontal pathogens, innate immune disorders, nutritional deficiencies, osteoporosis, and cognitive disorders have been considered important for periodontitis in previous literature[7,52]. However, these indicators were not measured in the KNHANES data; therefore, they were not included in this study. The small sample size of SP may have also contributed to low AUC values. Moreover, the T-SNE images demonstrate the complex composition of a variety of groups with overlapping patterns, thereby affecting accurate prediction. Nevertheless, combining MLP with deep layers and XGBoost, an advanced machine learning algorithm, may help overcome the complexity of the data and obtain an improved AUC compared to the other algorithms. This study may also be limited by its cross-sectional design, which could not determine the causality for the occurrence and progression of periodontitis and the use of the CPI score recorded in KNHANES instead of the revised classification of periodontitis[15]. Thus, further investigations are required to include other predictors that have a profound impact on periodontitis, increase the sample size, and use the updated classification of periodontitis.

Overall, our findings revealed that important risk indicators were consistently confirmed in the MLP and XGBoost models, and superior performance was observed compared to the results obtained using the conventional statistical method. The important risk indicators for periodontitis were age, sex, education, smoking, blood pressure, use of interdental cleaning aids, and glycated hemoglobin. Glycated hemoglobin and the frequency of binge drinking were important indicators of SP. Interestingly, the effect of controllable factors on the progression to SP increased, although unchangeable factors such as age, sex, and education were significant in the onset of periodontitis. These results collectively suggest that a novel approach to analyzing periodontitis risk indicators and factors via machine learning is needed for efficient disease management.

References

1. Persson, G. Rutger, Lloyd A. Mancl, John Martin, and Roy C. Page. 2003. Assessing Periodontal Disease Risk: A Comparison of Clinicians' Assessment Versus a Computerized Tool. Journal of the American Dental Association 134, no. 5: 575–82. DOI: 10.14219/jada.archive.2003.0224, PMID: 12785492.

2. Graves, D. T., J. D. Corrêa, and T. A. Silva. 2019. The Oral Microbiota Is Modified by Systemic Diseases. Journal of Dental Research 98, no. 2: 148–56. DOI: 10.1177/0022034518805739, PMID: 30359170.

3. Bartold, P. Mark. 2018. Lifestyle and Periodontitis: The Emergence of Personalized Periodontics. Periodontology 2000 78, no. 1: 7–11. DOI: 10.1111/prd.12237, PMID: 30198129.
4. Lang, Niklaus P., Jean E. Suvan, and Maurizio S. Tonetti. 2015. Risk Factor Assessment Tools for the Prevention of Periodontitis Progression a Systematic Review. Journal of Clinical Periodontology 42 Suppl. 16: S59–S70. DOI: 10.1111/jcpe.12350, PMID: 25496279.
5. Du, M., Tao Bo, Kostas Kapellas, and Marco A. Peres. 2018. Prediction Models for the Incidence and Progression of Periodontitis: A Systematic Review. Journal of Clinical Periodontology 45, no. 12: 1408–20. DOI: 10.1111/jcpe.13037, PMID: 30394558.
6. Montero, Eduardo, David Herrera, Mariano Sanz, Sangeeta Dhir, Thomas Van Dyke, and Corneliu Sima. 2019. Development and Validation of a Predictive Model for Periodontitis Using NHANES 2011–2012 Data. Journal of Clinical Periodontology 46, no. 4: 420–9. DOI: 10.1111/jcpe.13098, PMID: 30891834.
7. Bouchard, Philippe, Maria Clotilde Carra, Adrien Boillot, Francis Mora, and H. élène Rangé. 2017. Risk Factors in Periodontology: A Conceptual Framework. Journal of Clinical Periodontology 44, no. 2: 125–31. DOI: 10.1111/jcpe.12650, PMID: 27862138.
8. Eke, Paul I., Wenche S. Borgnakke, and Robert J. Genco. 2020. Recent Epidemiologic Trends in Periodontitis in the USA. Periodontology 2000 82, no. 1: 257–67. DOI: 10.1111/prd.12323, PMID: 31850640.
9. Zhang, Liying, Yikang Wang, Miaomiao Niu, Chongjian Wang, and Zhenfei Wang. 2020. Machine Learning for Characterizing Risk of Type 2 Diabetes Mellitus in a Rural Chinese Population: The Henan Rural Cohort Study. Scientific Reports 10, no. 1: 4406. DOI: 10.1038/s41598-020-61123-x, PMID: 32157171.
10. Chang, Hyuk-Joon, Sang-Jeong Lee, Tae-Hoon Yong, Nan-Young Shin, Bong-Geun Jang, Jo-Eun Kim, Kyung-Hoe Huh, Sam-Sun Lee, Min-Suk Heo, Soon-Chul Choi, Tae-Il Kim, and Won-Jin Yi. 2020. Deep Learning Hybrid Method to Automatically Diagnose Periodontal Bone Loss and Stage Periodontitis. Scientific Reports 10, no. 1: 7531. DOI: 10.1038/s41598-020-64509-z, PMID: 32372049.
11. Krois, Joachim, Thomas Ekert, Leonie Meinhold, Tatiana Golla, Basel Kharbot, Agnes Wittemeier, Christof Dörfer, and Falk Schwendicke. 2019. Deep Learning for the Radiographic Detection of Periodontal Bone Loss. Scientific Reports 9, no. 1: 8495. DOI: 10.1038/s41598-019-44839-3, PMID: 31186466.
12. Lee, Jae-Hong, Do-Hyung Kim, Seong-Nyum Jeong, and Seong-Ho Choi. 2018. Diagnosis and Prediction of Periodontally Compromised Teeth Using a Deep Learning-Based Convolutional Neural Network Algorithm. Journal of Periodontal & Implant Science 48, no. 2: 114–23. DOI: 10.5051/jpis.2018.48.2.114, PMID: 29770240.
13. Papantonopoulos, Georgios, Keiso Takahashi, Tasos Bountis, and Bruno G. Loos. 2014. Artificial Neural Networks for the Diagnosis of Aggressive Periodontitis Trained by Immunologic Parameters. PLOS ONE 9, no. 3: e89757. DOI: 10.1371/journal.pone.0089757, PMID: 24603408.
14. Krois, J., C. Graetz, B. Holtfreter, P. Brinkmann, T. Kocher, and F. Schwendicke. 2019. Evaluating Modeling and Validation Strategies for Tooth Loss. Journal of Dental Research 98, no. 10: 1088–95. DOI: 10.1177/0022034519864889, PMID: 31361174.
15. Tonetti, Maurizio S., Henry Greenwell, and Kenneth S. Kornman. 2018. Staging and Grading of Periodontitis: Framework and Proposal of a New Classification and Case Definition. Journal of Clinical Periodontology 45 Suppl. 20: S149–61. DOI: 10.1111/jcpe.12945, PMID: 29926495.
16. Petersen, P. E., 2013. Baez, R.J. and World Health Organization Assessment of Oral Health Status in Oral Health Surveys: Basic Methods 47–51. World Health Organization.
17. Eke, Paul I., Liang Wei, Gina O. Thornton-Evans, Luisa N. Borrell, Wenche S. Borgnakke, Bruce Dye, and Robert J. Genco. 2016. Risk Indicators for Periodontitis in US Adults: NHANES 2009 to 2012. Journal of Periodontology 87, no. 10: 1174–85. DOI: 10.1902/jop.2016.160013, PMID: 27367420.

18. Gupta, Shivangi, Prerna Suri, Pankaj Bajirao Patil, Jagadish Prasad Rajguru, Palak Gupta, and Niraliben Patel. 2020. Comparative Evaluation of Role of Hs C-Reactive Protein as a Diagnostic Marker in Chronic Periodontitis Patients. Journal of Family Medicine & Primary Care 9, no. 3: 1340–7. DOI: 10.4103/jfmpc.jfmpc_1063_19, PMID: 32509613.
19. Grundy, Scott M., James I. Cleeman, Stephen R. Daniels, Karen A. Donato, Robert H. Eckel, Barry A. Franklin, David J. Gordon, Ronald M. Krauss, Peter J. Savage, Sidney C. Smith, John A. Spertus, Fernando Costa, American Heart Association, and National Heart, Lung, and Blood Institute. 2005. Diagnosis and Management of the Metabolic Syndrome: An American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. Circulation 112, no. 17: 2735–52. DOI: 10.1161/CIRCULATIONAHA.105.169404, PMID: 16157765.
20. Ridker, Paul M. 2016. A Test in Context: High-Sensitivity C-Reactive Protein. Journal of the American College of Cardiology 67, no. 6: 712–23. DOI: 10.1016/j.jacc.2015.11.037, PMID: 26868696.
21. Pan, Wen-Harn, and Wen-Ting Yeh. 2008. How to Define Obesity? Evidence-Based Multiple Action Points for Public Awareness, Screening, and Treatment: An Extension of Asian-Pacific Recommendations. Asia Pacific Journal of Clinical Nutrition 17, no. 3: 370–4. PMID: 18818155.
22. van der Maaten, L., and G. Hinton. 2008. Visualizing Data Using t-SNE. Journal of Machine Learning Research 9: 2579–605.
23. Mohri, M., A. Rostamizadeh, and A. Talwalkar, 2018. Introduction in Foundations of Machine Learning 1–9. MIT Press.
24. Pedregosa, F. et al. 2011. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research 12: 2825–30.
25. Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984. Splitting Rules in Classification and Regression Trees 93–126. Chapman & Hall/CRC.
26. Ridgeway, G. 2005. Generalized Boosted Models: A Guide to the GBM Package. Computer Engineering 1: 1–12.
27. Breiman, L. 2001. Random Forests. Machine Learning 45, no. 1: 5–32. DOI: 10.1023/A:1010933404324.
28. Alam, M. Z., M. S. Rahman, and M. S. Rahman. 2019. A Random Forest Based Predictor for Medical Data Classification Using Feature Ranking. Informatics in Medicine Unlocked 15. DOI: 10.1016/j.imu.2019.100180.
29. Chen, T., and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785–94, 2016. DOI: 10.1145/2939672.2939785.
30. Chang, Wenbing, Yinglai Liu, Yiyong Xiao, Xinglong Yuan, Xingxing Xu, Siyue Zhang, and Shenghan Zhou. 2019. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. Diagnostics 9, no. 4: 178. DOI: 10.3390/diagnostics9040178, PMID: 31703364.
31. Tian, Y., and Y. Zhang. 2022. A Comprehensive Survey on Regularization Strategies in Machine Learning. Information Fusion 80: 146–66. DOI: 10.1016/j.inffus.2021.11.005.
32. Paszke, A. et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.
33. Šimundić, Ana-Maria. 2009. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC 19, no. 4: 203–11. PMID: 27683318.
34. Menze, Bjoern H., B. Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A. Hamprecht. 2009. A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. BMC Bioinformatics 10: 213. DOI: 10.1186/1471-2105-10-213, PMID: 19591666.
35. Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts. 2013. Understanding Variable Importances in

Forests of Randomized Trees. Advances in Neural Information Processing Systems 26. https://proceedings.neurips.cc/paper/2013/hash/e3796ae838835da0b6f6ea37bcf8bcb7-Abstract.html.
36. Khalilia, Mohammed, Sounak Chakraborty, and Mihail Popescu. 2011. Predicting Disease Risks from Highly Imbalanced Data Using Random Forest. BMC Medical Informatics & Decision Making 11: 51. DOI: 10.1186/1472-6947-11-51, PMID: 21801360.
37. Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2019. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research 20: 1–81. PMID: 34335110.
38. Atabaki-Pasdar, Naeimeh, Mattias Ohlsson, Ana Viñuela, Francesca Frau, Hugo Pomares-Millan, Mark Haid, Angus G. Jones, E. Louise Thomas, Robert W. Koivula, Azra Kurbasic, Pascal M. Mutie, Hugo Fitipaldi, Juan Fernandez, Adem Y. Dawed, Giuseppe N. Giordano, Ian M. Forgie, Timothy J. McDonald, Femke Rutters, Henna Cederberg, Elizaveta Chabanova, Matilda Dale, Federico De Masi, Cecilia Engel Thomas, Kristine H. Allin, Tue H. Hansen, Alison Heggie, Mun-Gwan Hong, Petra J. M. Elders, Gwen Kennedy, Tarja Kokkola, Helle Krogh Pedersen, Anubha Mahajan, Donna McEvoy, Francois Pattou, Violeta Raverdy, Ragna S. Häussler, Sapna Sharma, Henrik S. Thomsen, Jagadish Vangipurapu, Henrik Vestergaard, Leen M. 't Hart, Jerzy Adamski, Petra B. Musholt, Soren Brage, S. øren Brunak, Emmanouil Dermitzakis, Gary Frost, Torben Hansen, Markku Laakso, Oluf Pedersen, Martin Ridderstråle, Hartmut Ruetten, Andrew T. Hattersley, Mark Walker, Joline W. J. Beulens, Andrea Mari, Jochen M. Schwenk, Ramneek Gupta, Mark I. McCarthy, Ewan R. Pearson, Jimmy D. Bell, Imre Pavo, and Paul W. Franks. 2020. Predicting and Elucidating the Etiology of Fatty Liver Disease: A Machine Learning Modeling and Validation Study in the IMI DIRECT Cohorts. PLOS Medicine 17, no. 6: e1003149. DOI: 10.1371/journal.pmed.1003149, PMID: 32559194.
39. Hallett, M. J., J. J. Fan, X. G. Su, R. A. Levine, and M. E. Nunn. 2014. Random Forest and Variable Importance Rankings for Correlated Survival Data, with Applications to Tooth Loss. Statistical Modelling 14, no. 6: 523–47. DOI: 10.1177/1471082X14535517.
40. Galkin, Fedor, Polina Mamoshina, Alex Aliper, Evgeny Putin, Vladimir Moskalev, Vadim N. Gladyshev, and Alex Zhavoronkov. 2020. Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. iScience 23, no. 6: 101199. DOI: 10.1016/j.isci.2020.101199, PMID: 32534441.
41. Janitza, Silke, Carolin Strobl, and Anne-Laure Boulesteix. 2013. An AUC-Based Permutation Variable Importance Measure for Random Forests. BMC Bioinformatics 14: 119. DOI: 10.1186/1471-2105-14-119, PMID: 23560875.
42. Yang, Jian-Bo, Kai-Quan Shen, Chong-Jin Ong, and Xiao-Ping Li. 2009. Feature Selection for MLP Neural Network: The Use of Random Permutation of Probabilistic Outputs. IEEE Transactions on Neural Networks 20, no. 12: 1911–22. DOI: 10.1109/TNN.2009.2032543, PMID: 19822474.
43. Lundberg, S. M., and S. I. Lee. 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
44. Molnar, C. Interpretable Models in Interpretable Machine Learning 49–142 (Leanpub, 2020).
45. Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9, no. 3: 90–5. DOI: 10.1109/MCSE.2007.55.
46. Grossi, S. G., J. J. Zambon, A. W. Ho, G. Koch, R. G. Dunford, E. E. Machtei, O. M. Norderyd, and R. J. Genco. 1994. Assessment of Risk for Periodontal Disease. I. Risk Indicators for Attachment Loss. Journal of Periodontology 65, no. 3: 260–7. DOI: 10.1902/jop.1994.65.3.260, PMID: 8164120.
47. Sherwani, Shariq I., Haseeb A. Khan, Aishah Ekhzaimy, Afshan Masood, and Meena K. Sakharkar. 2016. Significance of HbA1C Test in Diagnosis and Prognosis of Diabetic Patients. Biomarker Insights 11: 95–104. DOI: 10.4137/BMI.S38440, PMID: 27398023.
48. Demmer, Ryan T., Birte Holtfreter, Moïse Desvarieux, David R. Jacobs, Wolfgang Kerner,

Matthias Nauck, Henry Völzke, and Thomas Kocher. 2012. The Influence of Type 1 and Type 2 Diabetes on Periodontal Disease Progression: Prospective Results from the Study of Health in Pomerania (SHIP). Diabetes Care 35, no. 10: 2036–42. DOI: 10.2337/dc11-2453, PMID: 22855731.
49. D'Aiuto, Francesco, Wael Sabbah, Gopalakrishnan Netuveli, Nikos Donos, Aroon D. Hingorani, John Deanfield, and Georgios Tsakos. 2008. Association of the Metabolic Syndrome with Severe Periodontitis in a Large U.S. Population-Based Survey. Journal of Clinical Endocrinology & Metabolism 93, no. 10: 3989–94. DOI: 10.1210/jc.2007-2522, PMID: 18682518.
50. Gomes-Filho, Isaac Suzart, I. D. S. C. E. Balinha, Simone S. da Cruz, Soraya C. Trindade, E. M. M. Cerqueira, J. S. Passos-Soares, Julita Maria F. Coelho, Ana Marice T. Ladeia, Maria Isabel P. Vianna, Alexandre M. Hintz, Teresinha C. de Santana, Pedro P. Dos Santos, Ana Claúdia M. G. Figueiredo, Ivana C. O. da Silva, Frank A. Scannapieco, Maurício L. Barreto, and Peter M. Loomer. 2021. Moderate and Severe Periodontitis Are Positively Associated with Metabolic Syndrome. Clinical Oral Investigations 25, no. 6: 3719–27. DOI: 10.1007/s00784-020-03699-2, PMID: 33226499.
51. Kim, O. S., M. H. Shin, S. S. Kweon, Y. H. Lee, O. J. Kim, Y. J. Kim, and H. J. Chung. 2018. The Severity of Periodontitis and Metabolic Syndrome in Korean Population: The Dong-Gu Study. Journal of Periodontal Research 53, no. 3: 362–8. DOI: 10.1111/jre.12521, PMID: 29226321.
52. Genco, Robert J., and Wenche S. Borgnakke. 2013. Risk Factors for Periodontal Disease. Periodontology 2000 62, no. 1: 59–94. DOI: 10.1111/j.1600-0757.2012.00457.x, PMID: 23574464.
53. Wang, Jiantao, Jian Lv, Wanchun Wang, and Xiubo Jiang. 2016. Alcohol Consumption and Risk of Periodontitis: A Meta-analysis. Journal of Clinical Periodontology 43, no. 7: 572–83. DOI: 10.1111/jcpe.12556, PMID: 27029013.
54. Frazão, Deborah Ribeiro, C. D. S. F. Maia, Victória Dos Santos Chemelo, Deiweson Monteiro, R. O. Ferreira, Leonardo Oliveira Bittencourt, G. S. Balbinot, Fabrício Mezzomo Collares, Cassiano Kuchenbecker Rösing, Manoela Domingues Martins, and Rafael Rodrigues Lima. 2020. Ethanol Binge Drinking Exposure Affects Alveolar Bone Quality and Aggravates Bone Loss in Experimentally-Induced Periodontitis. PLOS ONE 15, no. 7: e0236161. DOI: 10.1371/journal.pone.0236161, PMID: 32730269.
55. Park, Hee-Jung, Jun Hyup Lee, Sujin Park, and Tae-Il Kim. 2016. Changes in Dental Care Access upon Health Care Benefit Expansion to Include Scaling. Journal of Periodontal & Implant Science 46, no. 6: 405–14. DOI: 10.5051/jpis.2016.46.6.405, PMID: 28050318.
56. Petropoulos, Georgios, Ian J. McKay, and Francis J. Hughes. 2004. The Association Between Neutrophil Numbers and Interleukin-1α Concentrations in Gingival Crevicular Fluid of Smokers and Non-smokers with Periodontal Disease. Journal of Clinical Periodontology 31, no. 5: 390–5. DOI: 10.1111/j.1600-051x.2004.00489.x, PMID: 15086622.

Abbreviations

AUC, Area under the receiver operating characteristic curve; CART, Classification and regression tree; CI, Confidence interval; CPI, Community periodontal index; GBM; Gradient Boost Machine; HDL, High-density lipoprotein; KDCA, Korean Disease Control and Prevention Agency; MLP, Multilayer perceptron; NHANES, National Health and Nutrition Examination Survey; NoP, No periodontitis; NSP, Non-severe periodontitis; OR, Odds ratio; SHAP, Shapley additive explanations; SP, Severe periodontitis; TP, Total periodontitis; WHO, World Health Organization;

Acknowledgments

J.K. was partially supported by Hansung University.

Author contributions

XH C., Y.L., H.P., conceived and designed the study. J.K. designed the algorithm and conducted experiments. XH C., Y.L., J.L. interpreted data. XH C., H.P. advised on data analysis. Y. L., J.K., H.P., drafted the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The funding organizations had no role in the design or conduct of this study, and the authors declare no competing interests.

Figure legends

Figure 1. T-distributed stochastic neighbor embedding graphs showing the distribution of groups (a) between no periodontitis (NoP) and total periodontitis (TP) and (b) between non-severe periodontitis (NSP) and severe periodontitis (SP).
Figure 2. Shapley additive explanations (SHAP) summary plots (a) for prediction of TP and (b) for prediction of SP using extreme gradient boost (XGBoost). (c) The classification and color labeling of the variables used in the analysis. Dots with positive values have importance for predicting TP or SP, while dots with negative values have importance for predicting NoP or NSP.
Figure 3. SHAP dependence plots of (a) smoking and (b) education level for prediction of SP using XGBoost.
Figure 4. Diagram of the methodology**.** *KNHANES* Korea national health and nutrition examination survey, *T-SNE* t-distributed stochastic neighbor embedding, *CART* classification and regression tree, *GBM* gradient boosting machine, *XGBoost* extreme gradient boost, *MLP* multilayer perceptron, *AUC* area under the receiver operating characteristic curve, *NoP* no periodontitis, *TP* total periodontitis, *NSP* non-severe periodontitis, *SP* severe periodontitis.

Tables

Table 1. Characteristics of the subjects according to periodontitis. For discrete type variables, N (%) is illustrated, and *P*-values were obtained by chi-squared tests. For continuous type variables (age in years, glycated hemoglobin), the mean±standard deviation is illustrated, and the *P*-values were calculated by one-way analysis of variance.

Variable	No periodontitis	Non-severe periodontitis	Severe periodontitis	*P*
Age	51.2±0.2	57.1±0.3	57.8±0.4	< 0.001
Sex				
Male	3,526 (40.6)	1,778 (52.3)	792 (59.3)	< 0.001
Female	5,594 (59.4)	1,722 (47.7)	534 (40.7)	
Education				
Elementary school	1,842 (17.6)	1,034 (26.6)	420 (28.0)	< 0.001
Middle school	891 (9.2)	494 (14.6)	211 (15.5)	
High school	2,642 (30.2)	1,065 (32.4)	390 (30.5)	
College or higher	3,745 (43.0)	907 (26.4)	305 (26.0)	
Household income				
Low	2,068 (22.0)	959 (28.0)	360 (27.7)	< 0.001
Middle-low	2,252 (24.0)	909 (25.8)	352 (25.1)	
Middle-high	2,380 (27.0)	844 (23.8)	321 (23.8)	
High	2,420 (27.0)	788 (22.4)	293 (23.4)	
Glycated hemoglobin[*]	5.7±0.1	5.8±0.1	6.0±0.1	< 0.001
Fasting blood glucose				
< 100 mg/dL	5,830 (65.3)	1,785 (53.6)	599 (46.9)	< 0.001
≥ 100 mg/dL	3,290 (34.7)	1,715 (46.4)	727 (53.1)	
Diabetes				
Normal	4,574 (52.1)	1,288 (39.5)	1,001 (34.3)	< 0.001
Pre-diabetes	3,545 (38.2)	1,616 (45.0)	596 (44.5)	
Diabetes	1,001 (9.7)	599 (15.5)	301 (21.2)	
Triglycerides				
< 200 mg/dL	7,801 (85.5)	2,788 (78.9)	1,034 (75.5)	< 0.001
≥ 200 mg/dL	1,319 (14.5)	712 (21.1)	292 (24.5)	
HDL cholesterol				
< 240 mg/dL	8,135 (89.9)	3,106 (88.8)	1,189 (88.4)	0.202
≥ 240 mg/dL	985 (10.1)	394 (11.2)	137 (11.6)	
Abdominal obesity				
No	5,488 (61.0)	1,806 (52.8)	688 (52.1)	< 0.001
Yes	3,672 (39.0)	1,694 (47.2)	638 (47.9)	
Metabolic syndrome				
No	6,960 (77.7)	2,323 (67.9)	837 (60,9)	< 0.001
Yes	2,160 (22.3)	1,177 (32.1)	489 (39.1)	
Blood pressure				
Normal	4,054 (46.8)	1,075 (34.0)	384 (30.0)	< 0.001

Pre-hypertension	2,202 (24.8)	868 (24.3)	314 (24.3)	
Hypertension	2,864 (28.4)	1,557 (41.7)	628 (45.7)	
Hs-CRP				
< 1 mg/L	6,602 (73.3)	2,337 (67.3)	855 (65.5)	< 0.001
1–3 mg/L	1,783 (18.9)	823 (23.1)	342 (25.3)	
≥ 3 mg/L	735 (7.8)	340 (9.6)	129 (9.2)	
Body mass index				
< 23 kg/m^2	3,873 (43.3)	1,162 (33.3)	411 (31.4)	< 0.001
23-25 kg/m^2	2,180 (24.4)	881 (25.1)	344 (25.4)	
≥ 25 kg/m^2	3,067 (32.3)	1,457 (41.6)	571 (43.2)	
Smoking				
Never-smoker	1,259 (15.2)	796 (24.4)	350 (27.0)	< 0.001
Former smoker	1,966 (21.5)	865 (25.3)	387 (28.5)	
Current smoker	5,895 (63.3)	1,839 (50.3)	589 (44.5)	
Drinking quantity per occasion				
≤ 2 glasses	5,408 (56.5)	1,967 (53.6)	688 (49.6)	< 0.001
3-4 glasses	1,476 (17.0)	521 (14.7)	186 (14.2)	
5- 9 glasses	1,620 (19.1)	734 (22.5)	344 (26.8)	
≥ 10 glasses	616 (7.4)	278 (9.2)	108 (9.4)	
Frequency of binge drinking				
Never	5,540 (58.1)	2,056 (56.3)	705 (50.3)	< 0.001
< 1 time/month	1,301 (15.0)	415 (12.2)	167 (12.8)	
≥ 1 time/week	2,279 (26.9)	1,029 (31.5)	454 (36.9)	
Self-rated stress level				
Very stressed	402 (4.7)	177 (5.5)	45 (3.8)	< 0.001
Stressed	1,998 (22.9)	699 (20.2)	232 (17.3)	
Slightly stressed	5,186 (56.8)	1,893 (53.9)	761 (57.7)	
Stress free	1,534 (15.6)	731 (20.4)	288 (21.2)	
Use of interdental cleaning aids				
No	5,550 (59.0)	2,590 (73.6)	993 (73.0)	< 0.001
Yes	3,570 (41.0)	910 (26.4)	333 (27.0)	
Frequency of toothbrushing per day				
1-2 times	4,033 (43.2)	1,847 (52.2)	741 (52.3)	< 0.001
≥ 3 times	5,087 (56.8)	1,653 (47.8)	585 (47.7)	
Dental visit within a year				
No	5,743 (63.1)	2,426 (69.1)	899 (67.2)	< 0.001
Yes	3,377 (26.9)	1,074 (30.9)	427 (32.8)	

HDL high-density lipoprotein, *Hs-CRP* high-sensitivity C-reactive protein.

Table 2. Univariate logistic regression analyses of the association between the risk indicators and periodontitis.

Variable	Total periodontitis (vs. No	Severe periodontitis (vs. Non-severe

	periodontitis)			periodontitis)		
	OR	95% CI	*P*	OR	95% CI	*P*
Age	1.03	1.02-1.03	*< 0.001*	1.01	0.99-1.01	*0.288*
Sex						
Male	1.00			1.00		
Female	0.83	0.72-0.96	*0.013*	0.86	0.66-1.11	*0.234*
Education						
College or higher	1.00			1.00		
High school	1.34	1.18-1.53	*< 0.001*	0.97	0.76-1.22	*0.779*
Middle school	1.62	1.36-1.92	*< 0.001*	1.11	0.83-1.48	*0.485*
Elementary school	1.18	0.99-1.40	*0.060*	1.13	0.84-1.51	*0.428*
Household income						
High	1.00			1.00		
Middle-high	1.01	0.88-1.16	*0.863*	0.98	0.77-1.24	*0.835*
Middle-low	1.15	1.00-1.32	*0.046*	0.95	0.75-1.21	*0.684*
Low	1.28	1.11-1.47	*0.001*	0.97	0.76-1.24	*0.814*
Glycated hemoglobin	1.15	1.06-1.25	*< 0.001*	1.17	1.04-1.32	*0.012*
Fasting blood glucose						
< 100 mg/dL	1.00			1.00		
≥ 100 mg/dL	1.07	0.92-1.25	*0.371*	1.03	0.79-1.35	*0.823*
Diabetes						
Normal	1.00			1.00		
Pre-diabetes	0.93	0.79-1.08	*0.340*	0.92	0.70-1.21	*0.556*
Diabetes	0.83	0.64-1.10	*0.197*	0.93	0.60-1.43	*0.741*
Triglycerides						
< 200 mg/dL	1.00			1.00		
≥ 200 mg/dL	1.17	1.02-1.34	*0.024*	0.97	*0.77-1.21*	*0.779*
HDL cholesterol						
< 240 mg/dL	1.00			1.00		
≥ 240 mg/dL	1.10	0.95-1.28	*0.219*	1.01	0.77-1.32	*0.946*
Abdominal obesity						
No	1.00			1.00		
Yes	0.92	0.80-1.06	*0.254*	0.92	0.73-1.17	*0.501*
Metabolic syndrome						
No	1.00			1.00		
Yes	1.18	1.03-1.36	*0.016*	1.32	1.05-1.65	*0.017*
Blood pressure						
Normal	1.00			1.00		
Pre-hypertension	1.02	0.92-1.13	*0.658*	1.04	0.83-1.30	*0.751*
Hypertension	1.05	0.95-1.16	*0.305*	1.06	0.85-1.32	*0.610*
Hs-CRP (mg/L)						
< 1 mg/L	1.00			1.00		
1–3 mg/L	1.08	0.98-1.18	*0.110*	1.03	0.83-1.30	*0.747*
> 3 mg/L	1.03	0.90-1.17	*0.686*	1.06	0.85-1.32	*0.352*
Body mass index						
< 23 kg/m^2	1.00			1.00		
23-25 kg/m^2	1.15	1.02-1.32	*0.028*	1.00	0.80-1.26	*0.974*
≥ 25 kg/m^2	1.37	1.19-1.60	*< 0.001*	0.96	0.74-1.24	*0.754*
Smoking						
Never smoker	1.00			1.00		

Former smoker	0.63	0.54-0.73	*< 0.001*	0.96	0.76-1.21	*0.723*
Current smoker	0.52	0.44-0.62	*< 0.001*	0.89	0.68-1.18	*0.419*
Drinking quantity per occasion						
≤ 2 glasses	1.00			1.00		
3-4 glasses	0.93	0.86-1.11	*0.410*	0.86	0.64-1.16	*0.327*
5- 9 glasses	1.22	1.03-1.43	*0.068*	0.89	0.63-1.27	*0.528*
≥ 10 glasses	1.14	0.86-1.31	*0.332*	0.76	0.49-1.20	*0.245*
Frequency of binge drinking						
Never	1.00			1.00		
< 1 time/month	1.02	0.85-1.23	*0.813*	1.28	0.95-1.56	*0.133*
≥ 1 time/week	1.06	0.86-1.30	*0.578*	1.43	1.02-1.99	*0.037*
Self-rated stress level						
Very stressed	1.00			1.00		
Stressed	0.91	0.72-1.16	*0.435*	1.26	0.93-1.93	*0.339*
Slightly stressed	1.03	0.82-1.29	*0.782*	1.56	1.15-2.29	*0.050*
Stress free	1.08	0.84-1.37	*0.555*	1.45	1.05-2.18	*0.122*
Use of interdental cleaning aids						
Yes	1.00			1.00		
No	1.28	1.15-1.43	*< 0.001*	0.92	0.75-1.11	*0.380*
Frequency of tooth brushing per day						
1-2 times	1.00			1.00		
≥ 3 times	0.97	0.88-1.07	*0.548*	1.04	0.88-1.24	*0.621*
Dental visit within a year						
Yes	1.00			1.00		
No	1.04	0.94-1.15	*0.491*	0.89	0.74-1.07	*0.247*

HDL high-density lipoprotein, *Hs-CRP* high-sensitivity C-reactive protein, *OR* odds ratio, *CI* confidence interval.

Table 3. Relative importance of risk indicators in the prediction of total periodontitis with extreme gradient boost (XGBoost) and multilayer perceptron (MLP) models.

Rank	XGBoost (Gini impurity)	MLP (Permutation importance)	XGBoost (SHAP)	MLP (SHAP)
1	Age	Age	Age	Age
2	Sex	Smoking	Smoking	Smoking
3	Education	Education	Sex	Sex
4	Smoking	Sex	Education	Education
5	Blood Pressure	Use of interdental cleaning aids	Glycated hemoglobin	Use of interdental cleaning aids
6	Fasting blood glucose	Glycated hemoglobin	Blood Pressure	Household income
7	Diabetes	Frequency of binge drinking	Fasting blood glucose	Glycated hemoglobin
8	Use of interdental cleaning aids	Diabetes	Use of interdental cleaning aids	Blood Pressure
9	Glycated hemoglobin	Dental visit within a year	Diabetes	Dental visit within a year
10	Drinking quantity per occasion	Blood Pressure	Metabolic syndrome	Body mass index
11	Metabolic syndrome	Metabolic syndrome	Frequency of tooth brushing per day	Fasting blood glucose
12	Frequency of binge drinking	Body mass index	Drinking quantity per occasion	Frequency of tooth brushing per day
13	Abdominal obesity	Drinking quantity per occasion	Body mass index	Frequency of binge drinking

14	Frequency of tooth brushing per day	Frequency of tooth brushing per day	Frequency of binge drinking	Diabetes
15	Body mass index	Household income	Triglycerides	Abdominal obesity
16	Triglycerides	Fasting blood glucose	Household income	Metabolic syndrome
17	Dental visit within a year	Abdominal obesity	Abdominal obesity	Drinking quantity per occasion
18	Household income	Self-rated stress level	Hs-CRP	Hs-CRP
19	Hs-CRP	HDL cholesterol	Dental visit within a year	Triglycerides
20	Self-rated stress level	Hs-CRP	Self-rated stress level	Self-rated stress level
21	HDL cholesterol	Triglycerides	HDL cholesterol	HDL cholesterol

HDL high-density lipoprotein, *Hs-CRP* high-sensitivity C-reactive protein, *SHAP* Shapley additive explanations.

Table 4. Relative importance of risk indicators in the prediction of severe periodontitis with extreme gradient boost (XGBoost) and multilayer perceptron (MLP) models.

Rank	XGBoost (Gini impurity)	MLP (Permutation importance)	XGBoost (SHAP)	MLP (SHAP)
1	Glycated hemoglobin	Glycated hemoglobin	Glycated hemoglobin	Sex
2	Sex	Sex	Frequency of binge drinking	Education
3	Diabetes	Metabolic syndrome	Sex	Glycated hemoglobin
4	Drinking quantity per occasion	Education	Drinking quantity per occasion	Metabolic syndrome
5	Frequency of binge drinking	Self-rated stress level	Smoking	Frequency of binge drinking
6	Smoking	Frequency of binge drinking	Diabetes	Smoking
7	Education	Age	Age	Abdominal obesity
8	Hs-CRP	Dental visit within a year	Education	Diabetes
9	Blood Pressure	Smoking	Metabolic syndrome	Fasting blood glucose
10	Age	Fasting blood glucose	Self-rated stress level	Age
11	Body mass index	Triglycerides	Body mass index	Self-rated stress level
12	Fasting blood glucose	Use of interdental cleaning aids	Fasting blood glucose	Dental visit within a year
13	Triglycerides	Diabetes	Hs-CRP	Frequency of toothbrushing per day
14	Self-rated stress level	Household income	Use of interdental cleaning aids	Blood Pressure
15	Household income	Abdominal obesity	Blood Pressure	Body mass index
16	Metabolic syndrome	Frequency of toothbrushing per day	Triglycerides	Household income
17	Abdominal obesity	Drinking quantity per occasion	Abdominal obesity	Hs-CRP
18	Use of interdental cleaning aids	Body mass index	Frequency of toothbrushing per day	Drinking quantity per occasion
19	Frequency of toothbrushing per day	HDL cholesterol	Household income	Triglycerides
20	Dental visit within a year	Blood Pressure	Dental visit within a year	Use of interdental cleaning aids
21	HDL cholesterol	Hs-CRP	HDL cholesterol	HDL cholesterol

HDL high-density lipoprotein, *Hs-CRP* high-sensitivity C-reactive protein, *SHAP* Shapley additive explanations.

Supplementary Files