Received: 25 April 2022 | Revised: 1 June 2022 | Accepted: 26 June 2022

DOI: 10.1111/jcpe.13692

ORIGINAL ARTICLE

Journal of Clinical Periodontology WILEY

Systematic comparison of machine learning algorithms to develop and validate predictive models for periodontitis

Nasir Z. Bashir[1,2] | Zahid Rahman[3] | Sam Li-Sheng Chen[4]

[1]School of Oral and Dental Sciences, University of Bristol, Bristol, UK

[2]School of Mathematics and Statistics, The University of Sheffield, Sheffield, UK

[3]1859 Capital LLP, London, UK

[4]School of Oral Hygiene, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan

Correspondence
Nasir Z. Bashir, Bristol Dental School, Lower Maudlin Street, Bristol BS1 2LY, UK.
Email: nbashir562@gmail.com

Abstract

Aim: The aim of this study was to compare the validity of different machine learning algorithms to develop and validate predictive models for periodontitis.

Materials and Methods: Using national survey data from Taiwan ($n = 3453$) and the United States ($n = 3685$), predictors of periodontitis were extracted from the datasets and pre-processed, and then 10 machine learning algorithms were trained to develop predictive models. The models were validated both internally (bootstrap sampling) and externally (alternative country's dataset). The algorithms were compared across six performance metrics ([i] area under the curve for the receiver operating characteristic [AUC], [ii] accuracy, [iii] sensitivity, [iv] specificity, [v] positive predictive value, and [vi] negative predictive value) and two methods of data pre-processing ([i] machine-learning-based feature selection and [ii] dimensionality reduction into principal components).

Results: Many algorithms showed extremely strong performance during internal validation (AUC > 0.95, accuracy > 95%). However, this was not replicated in external validation, where predictive performance of all algorithms dropped off drastically. Furthermore, predictive performance differed according to data pre-processing methodology and the cohort on which they were trained.

Conclusions: Larger sample sizes and more complex predictors of periodontitis are required before machine learning can be leveraged to its full potential.

KEYWORDS

computing, machine learning, periodontitis, predictive modelling, statistics

Clinical Relevance

Scientific rationale for study: Machine learning is an exponentially growing field, with techniques being implemented across all disciplines, including periodontology. There is no current consensus on the best methods or algorithms to utilize for modelling periodontitis.

Principal findings: The performance of various algorithms differs depending upon which cohort of participants they are trained on, how the predictors are pre-processed, and whether they are validated internally or externally.

Practical implications: Until suitably large and complex datasets are readily available, the generalizability of more computationally intensive algorithms is not well established, and further research is required into exactly which predictors are required for accurate modelling.

1 | INTRODUCTION

Machine learning is an exponentially growing field, which is being implemented across a wide range of fields, including periodontology, due to the excitement surrounding its potential to make predictions that are not readily identifiable by humans (Sidey-Gibbons & Sidey-Gibbons, 2019; Farook et al., 2021). Algorithms of ever-increasing complexity are being developed and have been successfully leveraged across a huge range of disciplines, with seemingly "magical" predictive capabilities, in some cases (Mnih et al., 2015; He et al., 2016). Therefore, one might assume that machine learning may solve our lack of success in being able to predict those who may have periodontitis, with a high enough accuracy that such models can be utilized globally in day-to-day clinical practice.

However, the issue comes in identifying exactly how and where machine learning may offer benefits for the challenge of diagnosing periodontal disease. High-performance algorithms are typically designed with big datasets in mind (i.e., hundreds of thousands or millions of observations) and the predictors often have a highly complex and difficult-to-entangle relationship with the disease at hand (Vabalas et al., 2019; Thomas et al., 2020). However, in periodontology, it is extremely rare for clinical studies to reach even 1000 participants, let alone hundreds of thousands or millions. In addition, our predictors of periodontitis are often crude, such as socio-demographic status and self-reported oral health measures, which do not capture the true biological and inflammatory state of participants (Du et al., 2018). Furthermore, predictive models often lack external validation and, therefore, we do not know whether these algorithms generalize to other heterogeneous populations. Therefore, it is important for those working in the field to be aware that we may not yet be in a position to fully exploit machine learning methods, despite all of the advances being made.

The aim of this study is to utilize datasets that are "typical" of what may be expected in periodontology; moderate-sized, observational data, with predictors including socio-demographics, health behaviours, metabolic health, and self-reported oral health. We will develop and validate multivariate predictive models for periodontal diagnosis and assess whether more complex algorithms do appear to provide any substantial benefits under such conditions.

2 | MATERIALS AND METHODS

This study was designed according to Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (Collins et al., 2015).

2.1 | Data source: Taiwan

The Taiwanese data for this study were derived from a cross-sectional study conducted in Taiwan, which aimed to take a sample of participants from the Taiwanese population aged 18 years or older (Lai et al., 2015). The study adhered to the tenets of the Declaration of Helsinki and institutional review board approval was obtained (TMUJIRB No. 201207011). The sampling method for this study involved Taiwan being divided into four geographical areas and, due to population size varying by area, a probabilities-proportional-to-size method was used to randomly sample participants of different age groups from each area (Lohr, 2019). Each area was assessed by community-based integrated screening (CIS), incorporating periodontal examination into the examination protocol (Chen et al., 2004). In brief, the CIS is an adult disease screening programme implemented in several counties in Taiwan, which primarily aims to screen for cancers and metabolic syndrome. The periodontal status was assessed through the community periodontal index (CPI), and the examination consisted of CPI scores in five categories: healthy, gingival bleeding, calculus, shallow pocketing (4–5 mm), and deep pocketing (≥6 mm) (Ainamo et al., 1982; World Health, 1997). A total of 99 periodontal surveys were conducted from 2007 to 2008. Examiners used a WHO probe demarcated according to corresponding CPI scores. From the distal surfaces of the four canines, the entire dentition was divided into sextants, comprising two anterior and four posterior sextants. The first and second premolars were used as the index teeth for each posterior sextant, the right central incisor for the upper anterior sextant, and the left central incisor for the lower anterior sextant. Each index tooth was assessed at six sites (mesiobuccal, mid-buccal, distobuccal, mesiolingual, mid-lingual, and distolingual), and the CPI scores were recorded in a hierarchical manner, such that the highest score recorded for any site in a given sextant was taken to be representative for that sextant. The highest score across the six sextants was then taken to be representative of that participant. Participants were defined as having a positive diagnosis of periodontitis if they had a CPI ≥3.

2.2 | Data source: United States

The American data for this study were derived from the 2011 to 2012 National Health and Nutritional Examination Survey (NHANES), which is a cross-sectional study conducted by the National Center for Health Statistics, a division of the Centers for Disease Control and Prevention. The study adhered to the tents of the Declaration of Helsinki and institutional review board approval was obtained (NCHS Protocol #2011-17). The sampling method for this study was a multistage, clustered approach where participants were first interviewed at home and then invited to a mobile examination centre for further investigations, examination, and tests. These further tests included an oral examination conducted by state-licensed dental practitioners, during which the dentition was charted, and a six-point periodontal pocketing chart recorded. This involved all teeth except for third molars being examined, and the periodontal probing depth, recession, and clinical attachment loss were recorded at six sites (mesiobuccal, mid-buccal, distobuccal, mesiolingual, mid-lingual, and distolingual). For NHANES, participants were assessed according to the CDC-AAP case definitions for population-based surveillance of periodontitis, where individuals are classified as having no, mild, moderate, or severe periodontitis

(Eke et al., 2012). For this study, individuals who did not fall into the "no" category were diagnosed as having periodontitis.

2.3 | Predictors

Potential predictors of periodontitis, which were common between the two datasets, were extracted, which were broadly classified into the categories of demographics, health behaviours, metabolic health, and oral health. Demographic variables included age (years), sex (male/female), and education (less than high school/high school graduate or above). Health behaviours included smoking status (never/former or current) and alcohol consumption (never/former or current). Metabolic health variables included body mass index (in kilogram per square metre), waist circumference (in centimetre), systolic and diastolic blood pressure (in millimetre of mercury), fasting plasma glucose (in milligrams per decilitre), serum triglycerides (in milligrams per decilitre), and high-density lipoprotein (HDL) cholesterol (in milligrams per decilitre). Oral health variables included whether participants had visited a dentist within the past year, whether they had noticed any mobile teeth, and whether they used floss. For this study, we included participants who were at least 30 years of age without any missing periodontal data, and other missing data were imputed using multivariate imputation.

2.4 | Data pre-processing

Prior to model development, predictors are often transformed or specifically selected in such a way that aims to maximize the predictive validity of the fitted models (Mishra et al., 2011; Hira & Gillies, 2015). Here, we tested two different methods of data pre-processing:

1. Recursive feature elimination with cross-validation (RFECV): All of the predictors were min–max normalized on a scale of 0–1 and then the best predictors were selected using an RFECV algorithm. The RFECV algorithm was applied to a gradient boosting machine (GBM) in order to identify the most suitable predictors (Natekin & Knoll, 2013). In brief, the GBM algorithm iteratively selected 80% of the cohort to develop decision trees to assess which variables best predicted periodontitis and in what combination. This was averaged across 60 decision trees where each tree cast a vote on the optimal predictors. The RFECV algorithm implemented this GBM by iteratively adding additional predictors until the maximum was reached (maximum was equal to the square root of the total number of predictors). The combination of selected features was assessed by the GBM through 10-fold cross-validation.
2. Dimensionality reduction: All of the predictors were scaled and projected into a lower-dimensional space using principal components analysis (PCA). In brief, PCA generates synthetic variables (synthetic variables are linear combinations of the original variables) that maximize variance and minimize information loss, in essence trying to represent the data in the smallest number of dimensions possible (Jolliffe & Cadima, 2016). We selected the optimal number of principal components to represent the data by Minka's maximum likelihood estimation (MLE), which has been shown to be more optimal than cross-validation (Minka, 2001). Minka's MLE involves interpreting PCA as a density estimation problem and then uses Bayesian model selection to identify the optimal number of principal components.

2.5 | Model training

Ten machine learning classifiers were trained on the cohort data to predict the presence of periodontitis: (i) AdaBoost; (ii) artificial neural networks (ANNs); (iii) decision trees; (iv) a Gaussian process (GP); (v) K-nearest neighbours (KNN); (vi) linear support vector classification (SVC); (vii) linear discriminant analysis (LDA); (viii) logistic regression; (ix) random forests (RF); and (x) Naïve Bayes. A description of the algorithms and the underlying methodology is presented in Table 1. Where algorithms had tuning parameters, these were optimized using grid-search 10-fold cross-validation (Bergstra & Bengio, 2012). In brief, this meant that anytime an algorithm took tuning parameters, it was provided a set of feasible values. Any time this algorithm was then trained, it was trained using all permutations of the feasible tuning parameter values. The optimal value for each tuning parameter was then identified through 10-fold cross-validation.

2.6 | Model validation

We tested two methods for validating the fitted models

1. Internal validation: This was done by bootstrap sampling the training cohort with replacement, to create a validation sample equal in size to the training sample. The fitted models were then used to predict the periodontal diagnosis of the individuals in this bootstrap validation sample.
2. External validation: The model was developed by training the algorithm on the cohort from one country and then using the fitted model to predict the periodontal diagnosis for the individuals in the other country's cohort (i.e., an algorithm was trained on the data from the Taiwanese cohort and then this trained algorithm was used to predict the periodontal status of the individuals in the US cohort).

The predictive validity of each model was assessed using six metrics: (i) area under the curve for the receiver operating characteristic (AUC); (ii) accuracy; (iii) sensitivity; (iv) specificity; (v) positive predictive value; and (vi) negative predictive value. All analyses were programmed in Python version 3.9.7.

3 | RESULTS

3.1 | Population description

The full description of the included cohorts is presented in Table 2. In brief, the Taiwanese sample comprised 3453 participants, and the

TABLE 1 Summary of the tested algorithms

Algorithm name	Algorithm type	Algorithm function
AdaBoost	Ensemble method	Fits iterations of weak learners and assigns weights to these learners, in order to make a consensus-based classification of observations.
Artificial neural network	Neural network	Multi-layer perceptron with each neuron acting as a linear regression undergoing a non-linear transformation.
Decision tree	Classification and regression trees	Non-parametric tree-based classifier, which continues to split all internal nodes until all of the derived leaves are pure.
Gaussian process	Laplace approximation	Probabilistic classification with a logistic link function to approximate a non-Gaussian posterior.
K-nearest neighbours	Non-parametric classification	Non-parametric instance-based learning algorithm, where votes from neighbouring data points are used to classify each data point.
Linear support vector classification	Support vector machine	Fits a linear kernel, minimizing the squared hinge loss function, in order for efficient classification.
Linear discriminant analysis	Decision surface	Class conditional densities are fit using Bayes' rule, and a linear boundary between observations is identified.
Logistic regression	Generalized linear model	An extension of the linear model, which implements a logit link function to allow for binary classification.
Random forests	Ensemble method	Multiple decision trees are fit on subsamples of the data and predictions are averaged across the trees in a consensus-based manner to classify observations.
Naïve Bayes	Bayesian classification	Implements Bayes' theorem under the "naïve" assumption of conditional independence between all predictors, given the class to which they belong.

prevalence of periodontitis was 61.3% (no participant was excluded due to missing periodontal data). Participants with periodontitis were more likely to be older, male, have a lower educational attainment, be a former or current smoker, be non-drinkers, have a higher body mass index, waist circumference, systolic and diastolic blood pressure, fasting plasma glucose, serum triglycerides and HDL lipoproteins, have not visited the dentist in the last year, have noticed mobile teeth, and not use floss. The American sample comprised 3685 participants and the prevalence of periodontitis was 48.7% (881 out of 4566 participants aged 30 or older were excluded due to missing periodontal data). Participants with periodontitis were more likely to be older, male, have a lower educational attainment, be a former or current smoker, be non-drinkers, have a higher body mass index, waist circumference, systolic blood pressure, fasting plasma glucose, serum triglycerides, lower HDL lipoproteins, have not visited the dentist in the last year, have noticed mobile teeth, and not use floss. A heatmap for the Spearman rank correlation between each of the variables across the two cohorts is presented in Figure 1.

3.2 | Internal validation

The full results for the predictive validity of the algorithms trained and internally validated on the Taiwanese and American data are presented in Tables 3 and 4, respectively.

For the Taiwanese data, the strongest performing algorithms following RFECV feature selection were RF (AUC: 0.97, accuracy: 97.5%), followed by decision trees (AUC: 0.89, accuracy: 89.3%). All other algorithms were roughly similar, with AUC in the 0.55–0.65 range and accuracy approximately in the 60%–70% range. The strongest performing algorithms following PCA were RF (AUC: 0.99, accuracy: 99.3%), followed by decision trees (AUC: 0.97, accuracy: 96.8%) and GP (AUC: 0.79, accuracy: 80.7%). All other algorithms were roughly similar, with AUC in the 0.55–0.65 range and accuracy approximately in the 60%–70% range.

For the American data, the strongest performing algorithms following RFECV feature selection were KNN (AUC: 1.00, accuracy: 100.0%), followed by RF (AUC: 0.98, accuracy: 98.1%) and decision trees (AUC: 0.86, accuracy: 86.2%). All other algorithms were roughly similar, with AUC in the 0.60–0.70 range and accuracy approximately in the 60%–70% range. The strongest performing algorithms following PCA were KNN (AUC: 1.00, accuracy: 100.0%), followed by RF (AUC: 0.98, accuracy: 98.1%), decision trees (AUC: 0.94, accuracy: 94.1%), and GP (AUC: 0.95, accuracy: 92.0%). All other algorithms were roughly similar, with AUC in the 0.60–0.70 range and accuracy approximately in the 60%–70% range.

3.3 | External validation

The full results for the predictive validity of all algorithms trained on the Taiwanese data and then externally validated on the American data are presented in Table 5. The results for the algorithms trained on the American data and then externally validated on the Taiwanese data are presented in Table 6.

For the models trained on Taiwanese data and validated on American data, all algorithms saw a drastic drop-off in predictive performance, as compared with the results from the internal validation. Regardless of the method of data pre-processing, all algorithms

TABLE 2 Characteristics of the included cohorts

Characteristic	Taiwan ($n = 3453$)		NHANES ($n = 3685$)	
	Healthy ($n = 1338$ [38.7%])	Periodontitis ($n = 2115$ [61.3%])	Healthy ($n = 1892$ [51.3%])	Periodontitis ($n = 1793$ [48.7%])
Demographics				
Age (years), mean (SD)	45.6 (11.3)	50.4 (11.8)	51.8 (15.6)	55.3 (13.8)
Sex, *n* (%)				
Male	438 (32.7)	977 (46.2)	794 (42.0)	1034 (57.7)
Female	900 (67.3)	1138 (53.8)	1098 (58.0)	759 (42.3)
Education, *n* (%)				
Less than high school	481 (35.9)	857 (64.1)	402 (21.2)	528 (29.4)
High school or above	1056 (49.9)	1059 (50.1)	1490 (78.8)	1265 (70.6)
Health behaviours				
Smoking status, *n* (%)				
Never	1123 (83.9)	1553 (73.4)	1127 (59.6)	896 (50.0)
Former or current	215 (16.1)	562 (26.6)	765 (40.4)	897 (50.0)
Alcohol consumption, *n* (%)				
Never	1030 (77.0)	1648 (77.9)	243 (12.8)	261 (14.6)
Former or current	308 (23.0)	467 (22.1)	1649 (87.2)	1532 (85.4)
Metabolic health				
Body mass index (kg/m^2), mean (SD)	24.2 (5.0)	27.0 (73.0)	28.7 (6.5)	29.3 (6.4)
Waist circumference (cm), mean (SD)	78.1 (10.7)	80.1 (11.3)	98.4 (15.2)	100.8 (15.2)
Systolic blood pressure (mmHg), mean (SD)	122.9 (30.0)	127.6 (26.8)	123.2 (17.5)	126.8 (17.7)
Diastolic blood pressure (mmHg), mean (SD)	79.9 (27.7)	81.0 (12.1)	71.3 (11.7)	71.3 (12.4)
Fasting plasma glucose (mg/dl), mean (SD)	94.2 (24.5)	98.3 (29.0)	108.1 (22.6)	114.0 (29.3)
Serum triglycerides (mg/dl), mean (SD)	117.1 (108.0)	131.2 (130.1)	130.4 (62.8)	138.4 (75.7)
High-density lipoprotein (mg/dl), mean (SD)	55.5 (20.1)	56.2 (24.3)	53.3 (14.6)	52.0 (14.9)
Oral health				
Dental visit in last year, *n* (%)				
No	515 (38.5)	921 (43.5)	785 (41.5)	915 (51.0)
Yes	823 (61.5)	1194 (56.5)	1107 (58.5)	878 (49.0)
Mobile teeth, *n* (%)				
No	1209 (90.4)	1725 (81.6)	1653 (87.4)	1359 (75.8)
Yes	129 (9.6)	390 (18.4)	239 (12.6)	434 (24.2)
Uses floss, *n* (%)				
No	460 (34.4)	921 (43.5)	714 (37.7)	713 (39.8)
Yes	878 (65.6)	1194 (56.5)	1178 (62.3)	1080 (60.2)

Abbreviation: NHANES, National Health and Nutritional Examination Survey.

showed similar performance, with AUC typically in the 0.50–0.55 range and accuracy approximately in the 50%–60% range.

For the models trained on American data and validated on Taiwanese data, all algorithms saw a drastic drop-off in predictive performance, as compared with the results from the internal validation. Regardless of the method of data pre-processing, all algorithms showed similar performance, with AUC typically in the 0.50–0.60 range and accuracy approximately in the 50%–60% range.

4 | DISCUSSION

In this study, we develop and validate predictive models for periodontitis on national survey data from Taiwan and United States, using six different metrics to compare two methods of data pre-processing, 10 machine learning algorithms, and two methods of model validation. We find that the performance of the various algorithms differs depending upon which cohort of participants they are trained on, how

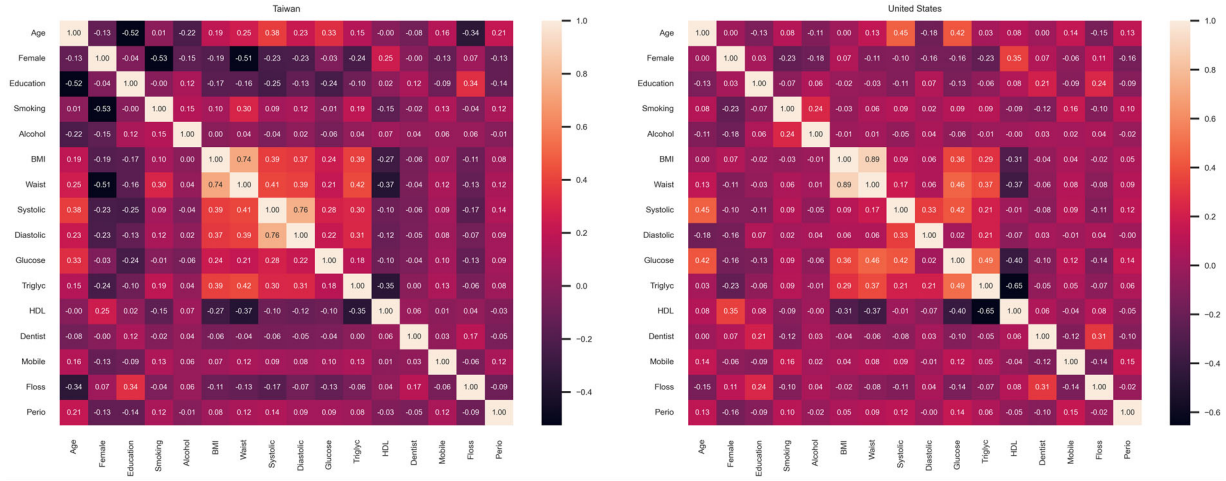

FIGURE 1 Heatmaps showing Spearman correlation coefficients between variables in the Taiwanese and American cohorts

the predictors are pre-processed, and whether they are validated internally or externally.

One of the most pertinent findings of this study is that the extremely strong predictive performance of many of the algorithms during internal validation (perfect predictive performance with KNN and near-perfect predictive performance with RF and decision trees) could not be replicated during external validation. This can be explained by a number of reasons, the first of which is model overfitting. This is the process by which the machine learning algorithms learn patterns in the training data that are specific to only that cohort and not replicated across other cohorts of individuals, and such patterns may arise solely by chance (i.e., noise), rather than due to a true causal relationship. This means that when the algorithms were validated in other cohorts, where these same relationships are not observed, the predictive performance drops off drastically; readers can refer to mathematical literature on the bias-variance trade-off to learn more about this phenomenon (Hastie et al., 2009). Secondly, our variables are crude and may only act as surrogate markers for the underlying inflammatory nature of the condition in individuals, meaning they are not powerful predictors of whether an individual does have periodontitis. Finally, the causes and risk factors of diseases may differ from population to population, meaning the features that are strong predictors of periodontitis in one population do not appear to be in another population. This is critical because it means that we cannot assume that just because a model works effectively in one group of individuals, it will then generalize to other cohorts who are heterogeneous to the cohort of individuals on which the model was originally trained (as we have seen in this present study).

There are limitations to this study that must be noted. Firstly, it is not possible to generalize the findings, which we have derived from just two datasets, to the broader problem of predicting periodontal disease across all populations. In these two specific cohorts, using a specific selection of predictors, we have identified shortcomings in the generalizability and predictive performance of certain algorithms. However, this is not to say that in other populations, using different predictors, the findings would be the same. In fact, this is a key motivator for why any machine learning model should be tested robustly, using multiple cohorts and a broad spectrum of predictors. It is very likely that models will need to be tailored in a way that is specific to the population on which they will be utilized. The second limitation comes from understanding that optimization of machine learning algorithms is an ongoing, unsolved challenge. The mathematical theory underlying optimization methods is still being developed, and, in many cases, it is impossible to prove that certain algorithms have identified a globally optimum solution, rather than a locally optimum solution (Zhang, 2004). In addition, there is an infinite combination of values that can be used for the tuning parameters; therefore, they must be reduced down to a subspace of plausible values, as is the case with grid-search cross-validation. The combination of these two factors means that, for any given algorithm, there may exist a way to adjust the parameters to produce a more optimal solution (i.e., an increased predictive validity).

Clearly, an issue we face in periodontology is that the number of observations typically collected is only in the hundreds or, occasionally, low thousands (Du et al., 2018). There are no studies published with sample sizes of clinically diagnosed participants in the millions. This instantly means that many of the more complex algorithms, such as ANNs or RF, are very unlikely to perform to their full potential in predicting periodontitis. Furthermore, the nature of the predictors used means that we are not yet in a position to exploit such powerful algorithms to their full extent. Conventional predictors such as sociodemographics, smoking, and self-reported oral health are known to display some weak-to-moderate correlation with periodontitis, but do not actually represent the true underlying biological state of the participants. Assimilating more complex and valid predictors, such as genetic or *-omics* data, which span hundreds of thousands or millions of participants, is incredibly expensive and resource-intensive. Despite reductions in the cost required to obtain such data, it is still not

TABLE 3 Performance of machine learning algorithms developed on Taiwanese data: Internal validation

	Pre-processing: RFECV feature selection						Pre-processing: Minka-MLE PCA					
	Model validation: Internal validation on bootstrap sample						Model validation: Internal validation on bootstrap sample					
Algorithm	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
AdaBoost	0.61 (0.59, 0.63)	63.8% (62.2, 65.4)	64.4% (62.8, 66.0)	60.0% (58.3, 61.6)	91.2% (90.3, 92.2)	20.7% (19.3, 22.0)	0.63 (0.61, 0.65)	64.4% (62.8, 66.0)	66.2% (64.6, 67.8)	57.6% (55.9, 59.2)	85.3% (84.1, 86.5)	31.4% (29.8, 32.9)
ANN	0.53 (0.51, 0.55)	64.4% (62.8, 66.0)	66.3% (64.7, 67.9)	57.5% (55.9, 59.2)	85.0% (83.8, 86.2)	32.0% (30.4, 33.5)	0.63 (0.61, 0.65)	65.5% (63.9, 67.0)	67.1% (65.5, 68.7)	59.6% (58.0, 61.3)	85.3% (84.1, 86.5)	34.2% (32.6, 35.7)
Decision tree	0.89 (0.88, 0.90)	89.3% (88.3, 90.4)	91.4% (90.4, 92.3)	86.2% (85.0, 87.3)	91.2% (90.2, 92.1)	86.4% (85.3, 87.6)	0.97 (0.96, 0.98)	96.8% (96.3, 97.4)	97.9% (97.5, 98.4)	95.2% (94.4, 95.9)	96.9% (96.3, 97.5)	96.8% (96.2, 97.4)
GP	0.63 (0.61, 0.65)	65.9% (64.3, 67.5)	67.6% (66.0, 69.2)	60.2% (58.6, 61.8)	84.9% (83.8, 86.1)	35.9% (34.3, 37.5)	0.79 (0.78, 0.80)	80.7% (79.4, 82.0)	80.7% (79.3, 82.0)	80.7% (79.4, 82.1)	90.0% (89.0, 91.0)	66.0% (64.4, 67.6)
KNN	0.64 (0.62, 0.66)	68.1% (66.5, 69.6)	69.9% (68.4, 71.4)	63.0% (61.4, 64.6)	83.9% (82.7, 85.2)	43.0% (41.4, 44.7)	0.62 (0.60, 0.64)	67.4% (65.9, 69.0)	69.0% (67.5, 70.6)	62.6% (60.9, 64.2)	84.8% (83.6, 86.0)	40.1% (38.5, 41.8)
SVC	0.60 (0.58, 0.62)	64.3% (62.7, 65.9)	66.1% (64.5, 67.7)	57.5% (55.8, 59.1)	85.5% (84.3, 86.6)	30.9% (29.4, 32.5)	0.59 (0.57, 0.61)	64.4% (62.8, 66.0)	66.3% (64.7, 67.9)	57.5% (55.9, 59.2)	85.0% (83.8, 86.2)	32.0% (30.4, 33.5)
LDA	0.59 (0.57, 0.61)	64.6% (63.0, 66.2)	66.6% (65.0, 68.1)	57.7% (56.0, 59.3)	84.6% (83.4, 85.8)	33.1% (31.5, 34.7)	0.58 (0.56, 0.60)	64.2% (62.6, 65.8)	66.3% (64.7, 67.9)	56.9% (55.2, 58.5)	84.4% (83.2, 85.6)	32.4% (30.8, 33.9)
Logistic regression	0.60 (0.59, 0.62)	64.9% (63.3, 66.5)	66.7% (65.1, 68.3)	58.5% (56.9, 60.2)	85.1% (83.9, 86.3)	33.0% (31.5, 34.6)	0.59 (0.57, 0.61)	64.4% (62.8, 66.0)	66.2% (64.6, 67.8)	57.7% (56.0, 59.3)	85.3% (84.1, 86.5)	31.5% (29.9, 33.0)
RF	0.97 (0.96, 0.98)	97.5% (97.0, 98.0)	96.4% (95.8, 97.0)	99.4% (99.1, 99.6)	99.6% (99.4, 99.8)	94.2% (93.4, 95.0)	0.99 (0.99, 0.99)	99.3% (99.0, 99.6)	99.1% (98.7, 99.4)	99.7% (99.5, 99.9)	99.8% (99.7, 100.0)	98.5% (98.1, 98.9)
Naïve Bayes	0.58 (0.56, 0.60)	51.1% (49.4, 52.7)	76.8% (75.4, 78.2)	43.5% (41.8, 45.1)	28.6% (27.1, 30.2)	86.4% (85.2, 87.5)	0.59 (0.57, 0.61)	54.7% (53.0, 56.3)	71.9% (70.4, 73.4)	44.9% (43.3, 46.6)	42.5% (40.8, 44.1)	73.9% (72.4, 75.4)

Note: All values are reported with 95% confidence intervals.

Abbreviations: ANN, artificial neural network; AUC, area under the curve for the receiver operating characteristic; GP, Gaussian process; KNN, K-nearest neighbours; LDA, linear discriminant analysis; MLE, maximum likelihood estimation; NPV, negative predictive value; PCA, principal components analysis; PPV, positive predictive value; RF, random forest; RFECV, recursive feature elimination with cross-validation; SVC, (linear) support vector classification.

TABLE 4 Performance of machine learning algorithms developed on National Health and Nutritional Examination Survey data: Internal validation

	Pre-processing: RFECV feature selection						Pre-processing: Minka-MLE PCA					
	Model validation: Internal validation on bootstrap sample						Model validation: Internal validation on bootstrap sample					
Algorithm	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
AdaBoost	0.67 (0.65, 0.69)	62.1% (60.6, 63.7)	60.3% (58.7, 61.9)	63.8% (62.2, 65.3)	59.6% (58.0, 61.2)	64.5% (62.9, 66.0)	0.68 (0.66, 0.70)	62.5% (61.0, 64.1)	60.3% (58.7, 61.9)	64.7% (63.1, 66.2)	62.0% (60.5, 63.6)	63.0% (61.4, 64.5)
ANN	0.68 (0.66, 0.70)	65.2% (63.7, 66.8)	62.6% (61.0, 64.2)	68.0% (66.4, 69.5)	66.8% (65.3, 68.3)	63.8% (62.3, 65.4)	0.69 (0.67, 0.71)	65.8% (64.3, 67.4)	62.9% (61.4, 64.5)	69.0% (67.5, 70.5)	68.6% (67.1, 70.0)	63.4% (61.8, 64.9)
Decision tree	0.86 (0.84, 0.88)	86.2% (85.1, 87.3)	88.2% (87.2, 89.2)	84.6% (83.4, 85.8)	81.9% (80.7, 83.1)	90.1% (89.1, 91.0)	0.94 (0.93, 0.95)	94.1% (93.3, 94.8)	96.3% (95.6, 96.9)	92.3% (91.4, 93.1)	91.0% (90.1, 92.0)	96.8% (96.2, 97.4)
GP	0.71 (0.69, 0.73)	68.3% (66.8, 69.8)	65.7% (64.2, 67.3)	70.9% (69.4, 72.4)	69.6% (68.1, 71.1)	67.1% (65.6, 68.7)	0.95 (0.93, 0.97)	92.0% (91.2, 92.9)	93.9% (93.1, 94.7)	90.5% (89.6, 91.5)	89.0% (88.0, 90.0)	94.8% (94.1, 95.5)
KNN	1.00 (1.00, 1.00)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)	1.00 (1.00, 1.00)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)	100.0% (100.0, 100.0)
SVC	0.64 (0.62, 0.66)	62.4% (60.9, 64.0)	60.8% (59.2, 62.4)	63.9% (62.3, 65.4)	59.1% (57.5, 60.7)	65.4% (63.9, 67.0)	0.63 (0.61, 0.65)	61.7% (60.1, 63.2)	60.2% (58.6, 61.7)	62.9% (61.3, 64.4)	57.2% (55.7, 58.8)	65.6% (64.1, 67.2)
LDA	0.63 (0.61, 0.65)	61.7% (60.2, 63.3)	60.1% (58.5, 61.6)	63.2% (61.6, 64.7)	58.3% (56.7, 59.9)	64.9% (63.3, 66.4)	0.62 (0.60, 0.64)	61.7% (60.1, 63.3)	60.2% (58.6, 61.8)	62.9% (61.3, 64.4)	57.2% (55.6, 58.8)	65.8% (64.2, 67.3)
Logistic regression	0.63 (0.61, 0.65)	62.4% (60.9, 64.0)	60.9% (59.3, 62.4)	63.8% (62.2, 65.3)	58.8% (57.3, 60.4)	65.7% (64.2, 67.2)	0.62 (0.60, 0.63)	62.4% (60.9, 64.0)	61.0% (59.5, 62.6)	63.6% (62.0, 65.1)	57.9% (56.3, 59.5)	66.5% (65.0, 68.0)
RF	0.98 (0.98, 0.98)	98.1% (97.7, 98.6)	98.8% (98.5, 99.2)	97.5% (97.0, 98.0)	97.2% (96.7, 97.7)	99.0% (98.6, 99.3)	0.98 (0.98, 0.98)	98.1% (97.7, 98.5)	98.4% (98.0, 98.8)	97.8% (97.3, 98.3)	97.5% (97.0, 98.0)	98.6% (98.2, 99.0)
Naïve Bayes	0.60 (0.58, 0.62)	59.3% (57.8, 60.9)	59.2% (57.6, 60.8)	59.4% (57.8, 61.0)	46.5% (44.9, 48.1)	71.0% (69.5, 72.4)	0.61 (0.59, 0.63)	60.1% (58.6, 61.7)	59.6% (58.0, 61.2)	60.5% (58.9, 62.1)	50.2% (48.6, 51.8)	69.1% (67.6, 70.6)

Note: All values are reported with 95% confidence intervals.

Abbreviations: ANN, artificial neural network; AUC, area under the curve for the receiver operating characteristic; GP, Gaussian process; KNN, K-nearest neighbours; LDA, linear discriminant analysis; MLE, maximum likelihood estimation; NPV, negative predictive value; PCA, principal components analysis; PPV, positive predictive value; RF, random forest; RFECV, recursive feature elimination with cross-validation; SVC, (linear) support vector classification.

TABLE 5 Performance of machine learning algorithms developed on Taiwanese data: External validation

	Pre-processing: RFECV feature selection						Pre-processing: Minka-MLE PCA					
	Model validation: External validation on NHANES cohort						Model validation: External validation on NHANES cohort					
Algorithm	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
AdaBoost	0.53 (0.51, 0.55)	51.5% (49.9, 53.1)	50.1% (48.5, 51.7)	79.4% (78.1, 80.7)	97.9% (97.5, 98.4)	7.6% (6.7, 8.4)	0.56 (0.54, 0.58)	54.7% (53.0, 56.3)	52.0% (50.3, 53.6)	69.8% (68.4, 71.3)	90.6% (89.7, 91.6)	20.6% (19.3, 21.9)
ANN	0.51 (0.49, 0.53)	49.6% (48.0, 51.2)	49.0% (47.3, 50.6)	52.7% (51.1, 54.3)	82.5% (81.3, 83.7)	18.5% (17.2, 19.8)	0.52 (0.50, 0.54)	51.1% (49.5, 52.7)	49.9% (48.2, 51.5)	56.6% (55.0, 58.2)	83.5% (82.3, 84.7)	20.4% (19.1, 21.7)
Decision tree	0.53 (0.51, 0.55)	52.5% (50.9, 54.1)	50.8% (49.2, 52.4)	56.8% (55.2, 58.4)	75.0% (73.6, 76.4)	31.2% (29.7, 32.7)	0.51 (0.49, 0.53)	49.7% (48.0, 51.3)	48.6% (47.0, 50.2)	51.3% (49.6, 52.9)	59.9% (58.3, 61.5)	40.0% (38.4, 41.5)
GP	0.55 (0.53, 0.57)	53.7% (52.1, 55.3)	51.3% (49.7, 52.9)	77.7% (76.4, 79.1)	95.8% (95.2, 96.5)	13.8% (12.7, 15.0)	0.50 (0.48, 0.52)	49.6% (48.0, 51.2)	49.0% (47.3, 50.6)	52.7% (51.1, 54.3)	82.5% (81.3, 83.7)	18.5% (17.2, 19.8)
KNN	0.56 (0.54, 0.58)	55.0% (53.3, 56.6)	52.5% (50.9, 54.1)	61.7% (60.2, 63.3)	78.9% (77.6, 80.2)	32.2% (30.7, 33.8)	0.52 (0.50, 0.54)	50.8% (49.2, 52.4)	49.7% (48.0, 51.3)	54.3% (52.7, 56.0)	76.6% (75.2, 77.9)	26.4% (25.0, 27.9)
SVC	0.54 (0.52, 0.56)	53.2% (51.6, 54.8)	51.0% (49.4, 52.6)	78.2% (76.9, 79.5)	96.4% (95.8, 97.0)	12.3% (11.3, 13.4)	0.53 (0.51, 0.55)	52.1% (50.5, 53.7)	50.5% (48.8, 52.1)	59.7% (58.1, 61.3)	85.3% (84.1, 86.4)	20.7% (19.4, 22.0)
LDA	0.50 (0.48, 0.52)	51.4% (49.8, 53.0)	50.5% (48.9, 52.1)	51.5% (49.9, 53.1)	8.3% (7.4, 9.1)	92.3% (91.5, 93.2)	0.53 (0.51, 0.55)	52.0% (50.4, 53.6)	50.4% (48.8, 52.0)	58.9% (57.3, 60.5)	84.2% (83.0, 85.4)	21.4% (20.1, 22.7)
Logistic regression	0.55 (0.53, 0.57)	53.7% (52.1, 55.3)	51.3% (49.7, 52.9)	75.5% (74.1, 76.9)	95.0% (94.3, 95.7)	14.6% (13.5, 15.8)	0.53 (0.51, 0.55)	52.6% (51.0, 54.2)	50.7% (49.1, 52.4)	61.3% (59.7, 62.8)	86.2% (85.1, 87.3)	20.7% (19.4, 22.0)
RF	0.57 (0.55, 0.59)	56.4% (54.8, 58.0)	53.1% (51.5, 54.7)	71.3% (69.8, 72.7)	89.2% (88.2, 90.2)	25.4% (24.0, 26.8)	0.53 (0.51, 0.55)	50.6% (49.0, 52.2)	49.6% (48.0, 51.2)	57.9% (56.3, 59.5)	89.2% (88.2, 90.2)	14.0% (12.9, 15.1)
Naïve Bayes	0.50 (0.48, 0.52)	48.8% (47.2, 50.5)	48.7% (47.1, 50.3)	55.9% (54.3, 57.5)	98.5% (98.2, 98.9)	1.7% (1.3, 2.2)	0.55 (0.53, 0.57)	54.5% (52.9, 56.1)	52.6% (51.0, 54.2)	57.5% (55.9, 59.1)	65.8% (64.3, 67.3)	43.8% (42.2, 45.4)

Note: All values are reported with 95% confidence intervals.

Abbreviations: ANN, artificial neural network; AUC, area under the curve for the receiver operating characteristic; GP, Gaussian process; KNN, K-nearest neighbours; LDA, linear discriminant analysis; MLE, maximum likelihood estimation; NHANES, National Health and Nutritional Examination Survey; NPV, negative predictive value; PCA, principal components analysis; PPV, positive predictive value; RF, random forest; RFECV, recursive feature elimination with cross-validation; SVC, (linear) support vector classification.

TABLE 6 Performance of machine learning algorithms developed on National Health and Nutritional Examination Survey data: External validation

	Pre-processing: RFECV feature selection						Pre-processing: Minka-MLE PCA					
	Model validation: External validation on Taiwanese cohort						Model validation: External validation on Taiwanese cohort					
Algorithm	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
AdaBoost	0.58 (0.56, 0.60)	52.0% (50.3, 53.7)	75.1% (73.7, 76.6)	43.7% (42.1, 45.4)	32.4% (30.8, 33.9)	83.0% (81.8, 84.3)	0.53 (0.51, 0.55)	51.1% (49.4, 52.8)	64.5% (62.9, 66.1)	41.2% (39.5, 42.8)	44.7% (43.1, 46.4)	61.1% (59.5, 62.8)
ANN	0.59 (0.57, 0.61)	57.8% (56.2, 59.5)	69.8% (68.3, 71.4)	46.7% (45.0, 48.3)	54.8% (53.1, 56.5)	62.6% (60.9, 64.2)	0.51 (0.49, 0.53)	52.9% (51.2, 54.5)	62.0% (60.4, 63.7)	39.9% (38.2, 41.5)	59.3% (57.7, 60.9)	42.7% (41.0, 44.3)
Decision tree	0.54 (0.52, 0.56)	52.4% (50.8, 54.1)	65.2% (63.6, 66.8)	42.0% (40.3, 43.6)	48.0% (46.3, 49.7)	59.5% (57.9, 61.1)	0.51 (0.49, 0.53)	49.2% (47.6, 50.9)	61.6% (60.0, 63.2)	39.0% (37.4, 40.7)	45.4% (43.8, 47.1)	55.2% (53.6, 56.9)
GP	0.58 (0.56, 0.60)	59.5% (57.8, 61.1)	67.4% (65.9, 69.0)	47.8% (46.1, 49.5)	65.4% (63.9, 67.0)	50.0% (48.3, 51.7)	0.54 (0.52, 0.56)	52.9% (51.2, 54.5)	66.1% (64.5, 67.6)	42.5% (40.9, 44.2)	47.4% (45.7, 49.0)	61.5% (59.9, 63.1)
KNN	0.57 (0.55, 0.59)	58.0% (56.4, 59.7)	67.0% (65.5, 68.6)	46.3% (44.6, 48.0)	62.0% (60.4, 63.6)	51.8% (50.1, 53.5)	0.53 (0.51, 0.55)	50.5% (48.8, 52.2)	64.9% (63.3, 66.5)	41.1% (39.5, 42.8)	41.7% (40.1, 43.4)	64.3% (62.8, 65.9)
SVC	0.59 (0.57, 0.61)	55.2% (53.5, 56.8)	73.5% (72.1, 75.0)	45.3% (43.7, 47.0)	41.9% (40.2, 43.5)	76.2% (74.7, 77.6)	0.54 (0.52, 0.56)	52.4% (50.8, 54.1)	66.2% (64.6, 67.7)	42.4% (40.7, 44.0)	45.8% (44.1, 47.4)	63.0% (61.4, 64.6)
LDA	0.54 (0.52, 0.56)	45.4% (43.7, 47.0)	79.1% (77.7, 80.4)	41.0% (39.4, 42.7)	14.7% (13.5, 15.8)	93.9% (93.1, 94.7)	0.54 (0.52, 0.56)	52.4% (50.8, 54.1)	66.1% (64.6, 67.7)	42.3% (40.7, 44.0)	45.7% (44.1, 47.4)	63.0% (61.4, 64.6)
Logistic regression	0.59 (0.57, 0.61)	56.8% (55.1, 58.4)	72.1% (70.6, 73.6)	46.2% (44.6, 47.9)	48.0% (46.4, 49.7)	70.6% (69.1, 72.1)	0.55 (0.53, 0.57)	52.6% (50.9, 54.3)	66.3% (64.8, 67.9)	42.5% (40.8, 44.1)	45.9% (44.2, 47.5)	63.2% (61.6, 64.8)
RF	0.59 (0.57, 0.61)	54.4% (52.7, 56.0)	74.4% (72.9, 75.8)	44.9% (43.3, 46.6)	38.9% (37.2, 40.5)	78.8% (77.5, 80.2)	0.55 (0.53, 0.57)	52.4% (50.7, 54.0)	64.2% (62.6, 65.8)	41.4% (39.8, 43.1)	50.3% (48.6, 52.0)	55.6% (53.9, 57.3)
Naïve Bayes	0.53 (0.51, 0.55)	62.4% (60.8, 64.0)	62.7% (61.0, 64.3)	58.4% (56.8, 60.1)	95.5% (94.8, 96.2)	10.1% (9.1, 11.1)	0.54 (0.52, 0.56)	51.0% (49.3, 52.6)	68.3% (66.7, 69.8)	42.3% (40.6, 43.9)	37.3% (35.6, 38.9)	72.6% (71.2, 74.1)

Note: All values are reported with 95% confidence intervals.

Abbreviations: ANN, artificial neural network; AUC, area under the curve for the receiver operating characteristic; GP, Gaussian process; KNN, K-nearest neighbours; LDA, linear discriminant analysis; MLE, maximum likelihood estimation; NPV, negative predictive value; PCA, principal components analysis; PPV, positive predictive value; RF, random forest; RFECV, recursive feature elimination with cross-validation; SVC, (linear) support vector classification.

feasible to do so for the vast number of participants, which we require to truly maximize the effectiveness of the more computationally intensive algorithms. The existing evidence indicates that, for clinical machine learning problems, more complex algorithms do not appear to provide any significant benefits above and beyond conventional logistic regression (Christodoulou et al., 2019). Our findings are largely in agreement with this, and until suitably large and complex datasets are available, models such as logistic regression may be the preferred option due to the greater ease with which they can be explained and interpreted by clinicians.

For future researchers who wish to exploit machine learning in order to maximize our effectiveness in predicting periodontitis, the findings of this study essentially have three implications. Firstly, we require that vastly more data are collected, which is quite easily facilitated by the swathes of computing power available at most research institutes and the ability to disseminate big data very easily. The challenge comes in how we ensure the participants have received an accurate clinical diagnosis. It is very unlikely that one institute alone could produce this quantity of data, and we may take some inspiration from the field of genetic epidemiology, where biobanks from various different countries share data and collaborate with one another in order to harmonize data so that researchers can work on much larger sample sizes (Palmer et al., 2011). Secondly, assimilating data on variables that are more likely to represent the true biological state of individuals will be beneficial. Again, it is unlikely that this can be implemented on a large scale by a single institute, so collaboration may be beneficial. Furthermore, if we can identify which molecular markers we believe to be clinically relevant in some smaller studies, we can then specifically focus on collecting large-scale data for these markers only, which could further help mitigate the costs associated with such efforts. Third, models that have a strong predictive performance but have only been validated internally should not be taken at face value. Robust external validation is required before we can speak to the generalizability of such algorithms and the real-world impact they may have on identifying high-risk individuals.

In conclusion, larger sample sizes and more complex predictors of periodontitis are required before machine learning can be leveraged to its full potential for the prediction of periodontitis. In addition, any models that are developed should undergo robust external validation before claims are made regarding their potential impact on the wider population.

AUTHOR CONTRIBUTIONS

Conception and design of work: Nasir Z. Bashir. *Acquisition and analysis of data*: Nasir Z. Bashir and Sam Li-Sheng Chen. *Interpretation of data*: Nasir Z. Bashir, Zahid Rahman, and Sam Li-Sheng Chen. *Drafting, revising, and final approval*: Nasir Z. Bashir, Zahid Rahman, and Sam Li-Sheng Chen.

ACKNOWLEDGEMENTS

The authors thank Dr. Hongmin Lain for his assistance in data collection.

FUNDING INFORMATION

No funding was received for this research.

CONFLICT OF INTEREST

All authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data for NHANES are publicly available from the Centers for Disease Control and Prevention at https://www.cdc.gov/nchs/nhanes/index.htm. The data for the Taiwanese cohort are available on request from the College of Oral Medicine, Taipei Medical University, but not publicly available due to privacy and ethical reasons.

ETHICS STATEMENT

The protocols for the Taiwanese (TMUJIRB No. 201207011) and NHANES (NCHS Protocol #2011-17) studies both received ethical board approval.

ORCID

Nasir Z. Bashir https://orcid.org/0000-0001-7416-7610
Sam Li-Sheng Chen https://orcid.org/0000-0001-9750-3015

REFERENCES

Ainamo, J., Barmes, D., Beagrie, G., Cutress, T., Martin, J., & Sardo-Infirri, J. (1982). Development of the World Health Organization (WHO) community periodontal index of treatment needs (CPITN). *International Dental Journal*, *32*, 281–291.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Chen, T. H., Chiu, Y. H., Luh, D. L., Yen, M. F., Wu, H. M., Chen, L. S., Tung, T. H., Huang, C. C., Chan, C. C., Shiu, M. N., Yeh, Y. P., Liou, H. H., Liao, C. S., Lai, H. C., Chiang, C. P., Peng, H. L., Tseng, C. D., Yen, M. S., Hsu, W. C., & Chen, C. H. (2004). Community-based multiple screening model: Design, implementation, and analysis of 42,387 participants. *Cancer*, *100*, 1734–1743. https://doi.org/10.1002/cncr.20171

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, *13*, 1. https://doi.org/10.1186/s12916-014-0241-z

Du, M., Bo, T., Kapellas, K., & Peres, M. A. (2018). Prediction models for the incidence and progression of periodontitis: A systematic review. *Journal of Clinical Periodontology*, *45*, 1408–1420. https://doi.org/10.1111/jcpe.13037

Eke, P. I., Page, R. C., Wei, L., Thornton-Evans, G., & Genco, R. J. (2012). Update of the case definitions for population-based surveillance of periodontitis. *Journal of Periodontology*, *83*, 1449–1454. https://doi.org/10.1902/jop.2012.110664

Farook, T. H., Jamayet, N. B., Abdullah, J. Y., & Alam, M. K. (2021). Machine learning and intelligent diagnostics in dental and orofacial pain management: A systematic review. *Pain Research & Management*, *2021*, 6659133. https://doi.org/10.1155/2021/6659133

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, *2015*, 198363. https://doi.org/10.1155/2015/198363

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*, 20150202. https://doi.org/10.1098/rsta.2015.0202

Lai, H., Su, C. W., Yen, A. M., Chiu, S. Y., Fann, J. C., Wu, W. Y., Chuang, S. L., Liu, H. C., Chen, H. H., & Chen, L. S. (2015). A prediction model for periodontal disease: Modelling and validation from a National Survey of 4061 Taiwanese adults. *Journal of Clinical Periodontology*, *42*, 413–421. https://doi.org/10.1111/jcpe.12389

Lohr, S. L. (2019). *Sampling: Design and analysis*. Chapman and Hall/CRC.

Minka, T. (2001). Automatic choice of dimensionality for PCA (Technical report 514). MIT Media Lab Vision and Modeling Group.

Mishra, D., Dash, R., Rath, A. K., & Acharya, M. (2011). Feature selection in gene expression data using principal component analysis and rough set theory. *Advances in Experimental Medicine and Biology*, *696*, 91–100. https://doi.org/10.1007/978-1-4419-7046-6_10

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533. https://doi.org/10.1038/nature14236

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 21. https://doi.org/10.3389/fnbot.2013.00021

Palmer, L. J., Burton, P. R., & Smith, G. D. (2011). *An introduction to genetic epidemiology*. Policy Press.

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, *19*, 64. https://doi.org/10.1186/s12874-019-0681-4

Thomas, R., Bruin, W., Zhutovsky, P., & van Wingen, G. (2020). *Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders* (pp. 249–266). Elsevier Science.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS One*, *14*, e0224365. https://doi.org/10.1371/journal.pone.0224365

World Health Organization (Ed.). (1997). *Oral health surveys: Basic methods* (4th ed.). World Health Organization.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the twenty-first international conference on Machine learning* (pp. 116). Association for Computing Machinery, Banff, Alberta, Canada.

How to cite this article: Bashir, N. Z., Rahman, Z., & Chen, S. L.-S. (2022). Systematic comparison of machine learning algorithms to develop and validate predictive models for periodontitis. *Journal of Clinical Periodontology*, *49*(10), 958–969. https://doi.org/10.1111/jcpe.13692