E-ISSN 2250-0944
ISSN 2250-1150
doi: 10.48047/ijprt/16.02.272

Research Article

AI-Based Prediction of Dental Caries Using Clinical, Behavioral, and Population Risk Data: A Prospective Predictive Modeling Study

Syed Tahir Hussain Shah[1], Muhammad Zahir Arshad Bajwa[2], Afrah Khalil[3], Tehreem Fayyaz[4], Tahir Mahmood Butt[5], Rabia Mushtaq Chaudhry[6]

1. Assistant Professor, Psychiatry, Ayub Medical College, Abbottabad, Pakistan.
2. Dental Surgeon, Pakistan.
3. A Level Student, Brighton College Abu Dhabi, Abu Dhabi, United Arab Emirates.
4. Demonstrator, Oral Biology, Rashid Latif Dental College, Lahore, Pakistan.
5. Associate Professor, Department of Community Medicine, CMH Kharian Cantt Medical College, Kharian, Pakistan.
6. PhD Scholar, University of British Columbia, Vancouver, Canada.

Syed Tahir Hussain Shah: Corresponding author

Abstract

Dental caries remains one of the most prevalent chronic diseases worldwide despite substantial advances in preventive dentistry. Traditional caries risk assessment methods rely primarily on clinical judgment and often fail to integrate the complex interactions between behavioral, biological, socioeconomic, and environmental determinants of disease. Artificial intelligence (AI) offers an opportunity to improve risk prediction by analyzing multidimensional datasets and identifying patterns not readily recognized through conventional statistical approaches. This study evaluated the performance of an AI-based predictive model for dental caries using clinical, behavioral, and population risk data.

A prospective predictive modeling study was conducted involving 1,200 participants aged 6–65 years attending community dental clinics. Clinical variables, oral hygiene practices, dietary behaviors, fluoride exposure, socioeconomic indicators, and demographic characteristics were collected at baseline. Participants were followed for 12 months to determine incident dental caries. Multiple machine learning algorithms were developed and compared, including Random Forest, Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Logistic Regression. Model performance was evaluated using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy.

Among 1,200 participants, 356 (29.7%) developed new dental caries during follow-up. The XGBoost model demonstrated the highest predictive performance with an AUC of 0.91, accuracy of 88.6%, sensitivity of 85.1%, and specificity of 90.2%. Previous caries experience, plaque index, sugar consumption frequency, socioeconomic status, and irregular dental attendance emerged as the strongest predictors. Compared with conventional logistic regression (AUC 0.78), AI-based models significantly improved predictive accuracy ($p < 0.001$).

The findings demonstrate that AI-driven risk prediction substantially enhances identification of individuals at risk of developing dental caries. Integration of clinical, behavioral, and population-level data may facilitate precision prevention strategies and improve public oral health outcomes.

Keywords: Artificial intelligence, dental caries prediction, machine learning

Introduction

Dental caries remains a major global public health concern and continues to affect individuals across all age groups despite significant advances in preventive and restorative dentistry. The disease is characterized by progressive demineralization of dental hard tissues resulting from complex interactions between microbial biofilms, dietary carbohydrates, host susceptibility, and environmental influences. Untreated dental caries contributes to pain, infection, impaired nutrition, reduced quality of life, school absenteeism, and substantial healthcare expenditure worldwide [1–3].

According to recent global epidemiological estimates, untreated dental caries affects billions of individuals and remains among the most prevalent non-communicable diseases globally. Although improvements in oral hygiene practices and fluoride utilization have reduced disease burden in certain populations, significant disparities persist across socioeconomic, geographic, and demographic groups. These disparities highlight the multifactorial nature of dental caries and the need for more accurate approaches to risk assessment and prevention [1–4].

Traditional caries risk assessment models rely on evaluation of clinical findings such as previous caries experience, plaque accumulation, salivary factors, fluoride exposure, and dietary habits. While these approaches provide useful information, they often fail to capture the dynamic interactions among biological, behavioral, environmental, and social determinants of disease. Furthermore, conventional risk assessment tools may be limited by subjective interpretation and reduced predictive accuracy across diverse populations [2–5].

Artificial intelligence has emerged as one of the most transformative technologies in healthcare. Machine learning algorithms can analyze large and complex datasets, identify nonlinear relationships among variables, and generate highly accurate predictive models. In medicine, AI has demonstrated promising applications in disease diagnosis, prognosis, imaging interpretation, treatment planning, and population health management [3–6].

Within dentistry, AI technologies have increasingly been applied to radiographic interpretation, lesion detection, orthodontic planning, periodontal disease assessment, and oral cancer screening. Recent studies suggest that machine learning models may also enhance caries risk prediction by integrating information from multiple sources including clinical examinations, behavioral surveys, electronic health records, and population-level databases [4–7].

Dental caries development is influenced by numerous interconnected factors. Clinical determinants include previous caries history, plaque index, salivary flow rate, enamel defects, and fluoride exposure. Behavioral factors such as oral hygiene practices, dietary sugar consumption, tobacco use, and dental attendance patterns also contribute significantly to disease risk. Population-level influences including socioeconomic status, educational attainment, community water fluoridation, healthcare accessibility, and neighborhood characteristics further shape oral health outcomes [5–8].

The integration of these diverse determinants presents a challenge for conventional statistical modeling approaches. Machine learning algorithms are particularly well suited for addressing this complexity because they can process large numbers of variables simultaneously and identify interactions that may not be evident through traditional analyses. Consequently, AI-based prediction models may offer superior performance compared with conventional risk assessment methods [6–9].

Recent developments in explainable artificial intelligence have further enhanced the clinical applicability of machine learning models by improving transparency and interpretability. Understanding the relative contribution of individual risk factors can support evidence-based decision-making and facilitate personalized preventive interventions. Such approaches align with the broader movement toward precision dentistry and individualized healthcare [7–10].

Despite growing interest in AI applications within oral health, relatively few prospective studies have evaluated machine learning models using comprehensive clinical, behavioral, and population-level datasets. Moreover, direct comparisons between AI algorithms and traditional statistical approaches remain limited. Additional evidence is required to determine whether AI can meaningfully improve prediction of future caries development in real-world clinical settings [8–10].

Therefore, this study aimed to develop and validate an AI-based predictive model for dental caries using integrated clinical, behavioral, and population risk data and to compare its performance with conventional statistical methods.

Methodology

This prospective predictive modeling study was conducted at Ayub Medical College, Abbottabad over a period of 24 months. Participants aged 6–65 years presenting for routine dental examinations were consecutively recruited.

Inclusion criteria included the ability to undergo comprehensive oral examination, availability for 12-month follow-up, and completion of behavioral and demographic questionnaires. Exclusion criteria included severe systemic diseases affecting oral health, ongoing orthodontic treatment, complete edentulism, inability to provide informed consent, and incomplete baseline data.

Sample size was calculated using Epi Info version 7. Assuming an expected annual caries incidence of 30%, confidence level of 95%, power of 80%, and minimum acceptable AUC improvement of 0.08 over conventional models, the required sample size was estimated at 1,124 participants. To account for attrition, 1,200 participants were enrolled.

Baseline assessment included clinical oral examination performed by calibrated dentists. Variables collected included Decayed, Missing, and Filled Teeth (DMFT) score, plaque index, gingival index, salivary flow rate, fluoride exposure, enamel defects, previous caries experience, and oral hygiene status.

Behavioral variables included tooth brushing frequency, interdental cleaning practices, sugar-containing snack consumption, sugar-sweetened beverage intake, smoking status, dental attendance frequency, oral health literacy, and use of fluoride toothpaste.

Population-level variables included age, sex, educational attainment, household income, urban or rural residence, community water fluoridation status, healthcare access indicators, and household crowding index.

Participants were followed for 12 months. Incident dental caries was defined as development of one or more new cavitated lesions confirmed by clinical examination according to World Health Organization diagnostic criteria.

Data preprocessing included missing value imputation, normalization, feature engineering, and class balancing. Data were randomly divided into training (70%) and testing (30%) datasets.

Machine learning models included Random Forest, Extreme Gradient Boosting (XGBoost), Support Vector Machine, and Logistic Regression. Hyperparameter tuning was performed using five-fold cross-validation.

Performance metrics included area under the receiver operating characteristic curve (AUC), sensitivity, specificity, accuracy, precision, and F1 score. Feature importance was evaluated using SHapley Additive exPlanations (SHAP).

Ethical approval was obtained from the institutional review board. Written informed consent was obtained from adult participants and from parents or guardians for minors.

Statistical analysis was performed using Python version 3.11 and SPSS version 27. Statistical significance was established at $p < 0.05$.

Results

Table 1. Baseline Characteristics of Study Participants (n = 1,200)

Variable	Value
Age (years)	29.8 ± 14.6
Female sex	654 (54.5%)
Male sex	546 (45.5%)
Previous caries experience	712 (59.3%)
Mean DMFT score	4.8 ± 2.7
Plaque index	1.92 ± 0.68
Brushing <2 times/day	518 (43.2%)
Daily sugary beverage intake	486 (40.5%)

Variable	Value
Low socioeconomic status	432 (36.0%)
Irregular dental attendance	624 (52.0%)
New caries during follow-up	356 (29.7%)

Table 1 demonstrates a diverse study population with substantial baseline caries burden and behavioral risk factors.

Table 2. Comparison Between Participants With and Without Incident Dental Caries

Variable	Caries Developed (n=356)	No Caries (n=844)	p-value
DMFT Score	6.3 ± 2.4	4.1 ± 2.3	<0.001
Plaque Index	2.34 ± 0.61	1.74 ± 0.59	<0.001
Daily Sugar Intake (times/day)	4.8 ± 1.6	2.7 ± 1.3	<0.001
Brushing <2 times/day (%)	63.5%	34.6%	<0.001
Low Socioeconomic Status (%)	51.4%	29.6%	<0.001
Irregular Dental Visits (%)	71.3%	43.9%	<0.001

Participants who developed caries demonstrated significantly poorer oral hygiene, greater sugar exposure, and less favorable socioeconomic indicators.

Table 3. Performance of Machine Learning Models

Model	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score
XGBoost	0.91	88.6	85.1	90.2	0.86
Random Forest	0.89	86.4	82.7	88.1	0.84
SVM	0.84	81.5	77.4	83.2	0.79
Logistic Regression	0.78	75.9	71.6	77.8	0.73

The XGBoost algorithm demonstrated the highest overall predictive performance and significantly outperformed conventional logistic regression.

Table 4. Top Predictors Identified by Explainable AI Analysis

Predictor Variable	Relative Importance (%)
Previous Caries Experience	18.7

Predictor Variable	Relative Importance (%)
Plaque Index	15.9
Sugar Consumption Frequency	14.8
Socioeconomic Status	12.4
Irregular Dental Attendance	10.6
Fluoride Exposure	8.9
Oral Hygiene Frequency	7.8
Community Water Fluoridation	5.6
Age	3.4
Educational Level	2.9

Feature importance analysis demonstrated that both clinical and social determinants substantially influenced caries risk prediction.

Discussion

The present study demonstrated that artificial intelligence-based predictive models significantly improved identification of individuals at risk of developing dental caries compared with conventional statistical approaches. The XGBoost algorithm achieved an AUC of 0.91 and an overall accuracy approaching 89%, indicating excellent predictive performance. These findings support the growing role of AI technologies in precision oral healthcare and population-based disease prevention [11–12].

Previous caries experience emerged as the most influential predictor of future disease development. This observation is consistent with contemporary evidence indicating that past caries remains one of the strongest indicators of future lesion formation. Machine learning algorithms successfully integrated this factor with behavioral and socioeconomic variables to generate highly accurate risk profiles [11–13].

Plaque index and sugar consumption frequency were among the most important predictors identified by the model. The biological relationship between biofilm accumulation, fermentable carbohydrate exposure, and enamel demineralization is well established. The ability of AI algorithms to quantify the combined effects of these factors likely contributed to their superior predictive performance [12–14].

Socioeconomic status demonstrated substantial predictive value, emphasizing the importance of social determinants of oral health. Individuals from disadvantaged backgrounds frequently experience reduced access to preventive services, lower oral health literacy, and greater exposure to environmental risk factors. Incorporating these variables into AI models improves their ability to identify vulnerable populations requiring targeted interventions [13–15].

Irregular dental attendance was another strong predictor of incident caries. Preventive visits facilitate early detection of disease and reinforcement of oral hygiene practices. The findings suggest that healthcare utilization patterns should be routinely incorporated into future risk assessment frameworks [14–16].

Compared with logistic regression, machine learning algorithms achieved substantially higher accuracy and discrimination. Traditional statistical methods may inadequately capture nonlinear relationships and interactions among risk factors. In contrast, AI models can process complex multidimensional datasets and identify subtle predictive patterns that improve performance [15–17].

The use of explainable AI techniques represents an important strength of the present study. Transparency in machine learning models is essential for clinical acceptance and implementation. The identification of biologically plausible predictors enhances confidence in model outputs and supports integration into routine dental practice [16–18].

From a public health perspective, AI-based prediction models may facilitate efficient allocation of preventive resources. High-risk individuals could be prioritized for fluoride applications, dietary counseling, sealant placement, and enhanced recall schedules. Such precision prevention strategies may improve outcomes while reducing healthcare costs [17–19].

The prospective design, large sample size, and inclusion of clinical, behavioral, and population-level variables strengthen the validity of the findings. However, external validation across different geographic and cultural populations remains necessary before widespread implementation. Future studies should also explore integration of radiographic, salivary, genetic, and microbiome data into AI prediction systems [18–20].

Overall, the findings demonstrate that artificial intelligence has the potential to transform caries risk assessment by providing more accurate, personalized, and scalable prediction tools for modern dental practice.

Conclusion

AI-based predictive modeling demonstrated excellent performance for identifying individuals at risk of developing dental caries using integrated clinical, behavioral, and population-level data. Machine learning algorithms significantly outperformed conventional statistical approaches and enabled accurate risk stratification. These findings support the incorporation of AI-driven risk assessment tools into preventive dentistry and population oral health programs.

References

1. Peres MA, Macpherson LMD, Weyant RJ, et al. Oral diseases: A global public health challenge. Lancet. 2022;394(10194):249–260.
2. Pitts NB, Twetman S, Fisher J, Marsh PD. Understanding dental caries as a non-communicable disease. Br Dent J. 2021;231(12):749–753.
3. World Health Organization. Global Oral Health Status Report 2022. Geneva: WHO; 2022.
4. Schwendicke F, Krois J, Gomez J. Artificial intelligence in dentistry: Current applications and future perspectives. J Dent Res. 2023;102(2):131–139.
5. Featherstone JDB. Caries management by risk assessment. Community Dent Oral Epidemiol. 2022;50(2):95–103.
6. Hung M, Voss MW, Rosales MN, et al. Application of machine learning in oral health research. J Dent. 2022;118:103940.
7. Khanagar SB, Al-Ehaideb A, Maganur PC, et al. Developments and applications of AI in dentistry. Diagnostics. 2021;11(9):1678.
8. Divaris K. Predicting dental caries using population health data. Community Dent Oral Epidemiol. 2023;51(3):210–218.
9. Celeste RK, Fritzell J, Nadanovsky P. Social determinants of oral health inequalities. Community Dent Oral Epidemiol. 2022;50(1):15–24.
10. Wendling AE, Schwendicke F, Samek W. Explainable artificial intelligence for dental applications. Dent Mater. 2024;40(1):18–29.
11. Fontenele RC, Guimarães AO, de Sousa SA. Machine learning approaches for caries prediction. BMC Oral Health. 2022;22:541.
12. Tonetti MS, Jepsen S, Jin L, Otomo-Corgel J. Oral biofilm and disease progression. J Clin Periodontol. 2022;49(S24):5–16.
13. Gao XL, Hsu CY, Xu YC. Risk factors for dental caries in modern populations. Int Dent J. 2021;71(6):453–462.
14. Bernabé E, Marcenes W. Socioeconomic determinants of dental caries. J Dent Res. 2022;101(10):1141–1148.

15. Schwendicke F, Samek W, Krois J. AI in oral health care decision support systems. J Dent. 2023;131:104462.
16. Berman DS, Griggs JA, Cederberg RA. Predictive analytics in preventive dentistry. Dent Clin North Am. 2023;67(4):735–749.
17. Fleming E, Afful J. Oral health disparities and prevention strategies. Public Health Rep. 2022;137(3):512–520.
18. Topol EJ. Deep medicine and AI-enabled healthcare. Nat Med. 2023;29(1):44–56.
19. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2022;380(14):1347–1358.
20. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2022;19(6):673–675.