ORIGINAL ARTICLE / *ARTICLE ORIGINAL*

Pedodontics / *Pédodontie*

CARIES RISK PREDICTION IN PEDIATRIC PATIENTS USING MACHINE LEARNING TECHNIQUES: A RETROSPECTIVE STUDY

Nagarathna P J[1] | Vishnu Priya Veeraraghavan[2] | Aysha Jebin A[3] | Shikar Daniel[4] | Kaladhar Reddy Aileni[5] | Dr Santosh R Patil[6]

Introduction: Dental caries is a multifactorial disease prevalent in children and is often influenced by behavioral, environmental, and genetic factors. Traditional caries risk assessment methods have a limited accuracy and objectivity. Machine learning (ML) models provide a data-driven approach for predicting caries risk, thereby enabling targeted interventions.

Objectives: The main objective of this study is to develop and validate ML models for predicting caries risk in pediatric patients aged 6–12 years by incorporating clinical, behavioral, dietary, and socioeconomic factors.

Methods: This retrospective observational study included 148 children aged 6–12 years. Data, including demographic details, Decayed Missing Filling Treatment (DMFT) scores, dietary habits, fluoride exposure, and socioeconomic factors, were collected from clinical records and structured interviews. The dataset was preprocessed using imputation, normalization, and feature-selection techniques. Five ML models (Logistic Regression, Random Forest, SVM, XGBoost, and Neural Networks) were trained and evaluated using metrics, such as accuracy, sensitivity, specificity, and AUC-ROC. Patients were stratified into low, moderate, and high-risk categories based on predictions.

Results: The XGBoost model achieved the highest AUC-ROC (0.94), followed by the Neural Networks (0.92). DMFT score (35.2%), sugary food consumption (28.7%), and fluoride exposure (18.3%) were the most significant predictors of caries risk. Risk stratification classified 21.0% of patients as high-risk, emphasizing the need for targeted preventive measures. Significant associations were observed between caries risk and fluoride exposure ($P < 0.01$) and sugary food consumption ($P < 0.05$).

Conclusions: ML models, particularly XGBoost, provide accurate and actionable caries risk predictions in children and outperform traditional assessment methods. The integration of ML tools in clinical practice can enhance personalized prevention and resource allocation, ultimately improving pediatric oral health outcomes.

Keywords: Dental Caries. Risk Assessment, Machine Learning, Artificial Intelligence, Prediction

Corresponding author:
Aysha Jebin A, E-mail: ayshpathu3@gmail.com , Phone: 9895205190 , Facsimile number: 080-28416199.
Department of Periodontics, Krishnadevaraya college of dental sciences and Hospital, Rajiv Gandhi University of Health Sciences, Bangalore, Karnataka, India, Pin Code-562157

Conflicts of interest:
The authors declare no conflicts of interest.

1. Professor, Department of Pediatric Dentistry,Chhattisgarh Dental College and Research Institute, India.
2. Professor, Centre of Molecular Medicine and Diagnostics, Saveetha Dental College and Hospitals, Saveetha University, India.
3. Senior Lecturer, Department of Periodontics, Krishnadevaraya college of dental sciences and Hospital, India.
4. Senior Lecturer, Department of Oral Medicine and Radiology, Chhattisgarh Dental College and Research Institute, India.
5. Professor, Department of Preventive Dentistry, College of Dentistry, Jouf University, Kingdom of Saudi Arabia,
6. Professor, Department of Oral Medicine and Radiology, Chhattisgarh Dental College and Research Institute, India.

PRÉDICTION DU RISQUE DE CARIE CHEZ LES PATIENTS PÉDIATRIQUES À L'AIDE DE TECHNIQUES D'APPRENTISSAGE AUTOMATIQUE: UNE ÉTUDE RÉTROSPECTIVE

Introduction: La carie dentaire est une maladie multifactorielle fréquente chez l'enfant souvent influencée par des facteurs comportementaux, environnementaux et génétiques. Les méthodes traditionnelles d'évaluation du risque carieux présentent une précision et une objectivité limitées. Les modèles d'apprentissage automatique (AA) offrent une approche basée sur les données pour prédire le risque carieux, permettant ainsi des interventions ciblées.

Objectifs: L'objectif principal de cette étude est de développer et de valider des modèles d'AA pour prédire le risque carieux chez les patients pédiatriques âgés de 6 à 12 ans en intégrant des facteurs cliniques, comportementaux, alimentaires et socio-économiques.

Méthodes: Cette étude observationnelle rétrospective a porté sur 148 enfants âgés de 6 à 12 ans. Les données, notamment les données démographiques, les scores de traitement des obturations manquantes (TCMM), les habitudes alimentaires, l'exposition au fluorure et les facteurs socio-économiques, ont été recueillis à partir de dossiers cliniques et d'entretiens structurés. L'ensemble de données est prétraité par des techniques d'imputation, de normalisation et de sélection de caractéristiques. Cinq modèles ML (régression logistique, forêt aléatoire, SVM, XGBoost et réseaux neuronaux) ont été entraînés et évalués à l'aide de mesures telles que la précision, la sensibilité, la spécificité et l'ASC-ROC. Les patients ont été classés en catégories de risque faible, modéré et élevé en fonction des prédictions.

Résultats: Le modèle XGBoost a obtenu l'ASC-ROC la plus élevée (0,94), suivi des réseaux neuronaux (0,92). Le score DMFT (35,2 %), la consommation d'aliments sucrés (28,7 %) et l'exposition au fluorure (18,3 %) étaient les prédicteurs les plus significatifs du risque de carie. La stratification du risque a classé 21,0 % des patients comme étant à haut risque, soulignant la nécessité de mesures préventives ciblées. Des associations significatives ont été observées entre le risque de carie et l'exposition au fluorure ($p < 0,01$) et la consommation d'aliments sucrés ($p < 0,05$).

Conclusions: Les modèles d'apprentissage automatique, en particulier XGBoost, fournissent des prédictions précises et exploitables du risque de carie chez les enfants et surpassent les méthodes d'évaluation traditionnelles. L'intégration des outils d'apprentissage automatique dans la pratique clinique peut améliorer la prévention personnalisée et l'allocation des ressources, améliorant ainsi les résultats en matière de santé bucco-dentaire pédiatrique.

Mots clés: Caries dentaires. Évaluation des risques, apprentissage automatique, intelligence artificielle, prédiction

Introduction

Dental caries are one of the most prevalent chronic diseases affecting children worldwide and pose significant challenges to oral and overall health [1]. Although largely preventable, more than 530 million children worldwide have dental caries in the primary dentition, and most of the decayed teeth are untreated [2]. The multi-factorial nature of caries, involving a complex interplay between host factors, diet, microbial activity, and environmental influences, makes its prevention and management particularly challenging [3].

Early identification of children at risk of caries is essential for timely intervention and prevention [4]. Traditional caries risk assessment methods rely on clinical examination and subjective evaluation of risk factors, such as diet, oral hygiene practices, fluoride exposure, and socioeconomic status [5]. However, these approaches often lack accuracy and objectivity, leading to underestimation or overestimation of the caries risk. The integration of advanced technologies such as machine learning (ML) offers a promising avenue for improving risk prediction by analyzing large datasets and identifying complex patterns that may be overlooked by traditional methods [6].

Machine learning, a subset of artificial intelligence (AI), uses algorithms to learn from data and make predictions or decisions without explicit programming. In dentistry, ML has shown potential for diagnostic imaging, treatment planning, and risk assessment [7]. Specifically, ML models can synthesize a wide range of caries risk factors, such as the Decayed, Missing, and Filled Teeth (DMFT) index, dietary habits, fluoride exposure, family history, and socioeconomic status to generate accurate and personalized risk predictions. These models not only enhance precision but also provide actionable insights for tailoring preventive strategies [8].

Despite promising applications of ML in dentistry, its use in pediatric caries risk prediction remains relatively under explored. Pediatric populations present unique challenges, including rapidly changing dietary and oral hygiene behaviors, variable fluoride exposure, and differing susceptibility to caries owing to developmental factors [9]. Furthermore, the inclusion of diverse risk factors, such as parental education, access to preventive care, and family history, highlights the complexity of risk modeling in children [10].

This study aimed to develop and validate ML models for predicting caries risk in children aged 6–12 years by leveraging a comprehensive dataset that included clinical, behavioral, dietary, and socioeconomic factors. By integrating advanced ML algorithms, this study sought to overcome the limitations of traditional risk assessment methods and provide a robust, data-driven approach for the early identification of high-risk individuals. The findings of this study have the potential to inform personalized preventive strategies, improve resource allocation, and ultimately reduce the burden of caries in pediatric populations.

Materials and Methods

Study Design and Ethical Approval

This retrospective observational study was designed to develop and validate machine learning (ML) models for predicting caries risk in pediatric patients aged 6–12 years. A total of 148 children were included and their data were collected from clinical records and structured interviews. The study protocol was approved by the institutional ethics committee with approval number Ref#GGSDC/Dean/Res/21/14 and the study adhered to the ethical principles outlined in the Declaration of Helsinki. Prior to data collection, written informed consent was obtained from the parents or legal guardians of all the participants.

This study was designed to leverage existing clinical and behavioral data, making it both resource-efficient and feasible within a given time frame. Using a retrospective data set, we ensured that the study utilized real-world, historical data to test the applicability of machine learning models in predicting caries risk. Additionally, the study design aimed to integrate multiple domains of risk factors, including clinical, dietary, oral hygiene, fluoride exposure, and socioeconomic indicators to create a robust predictive model.

Participant Selection

The participants were selected based on a set of predefined inclusion and exclusion criteria to ensure that the study cohort was representative and suitable for the study's objectives. Children aged 6–12 years were eligible for inclusion if they had complete dental records including detailed information on dietary habits, oral hygiene practices, fluoride exposure, socioeconomic status, and clinical oral health findings. These parameters were critical for ensuring that the necessary data points were available for the development and validation of the machine learning models.

Inclusion Criteria

Children aged 6–12 years with complete dental records from the past 12 months were included to provide recent and relevant clinical data. Additionally, the availability of behavioral and lifestyle information, such as dietary habits and oral hygiene practices, was mandatory to incorporate critical modifiable risk factors into predictive models. Documented exposure to fluoride through toothpaste, mouth rinses, or professional treatments was another essential inclusion criterion, given the significant role of fluoride in caries prevention. Finally, socioeconomic information, including parental education and household income, was required to examine the potential influence of social determinants on caries risk.

Exclusion Criteria

Children with systemic diseases known to affect oral health, such as diabetes or immunodeficiency, were excluded to eliminate confounding variables that could skew the assessment of caries risk. Similarly, participants with incomplete or missing data for critical variables, such as dietary habits, fluoride exposure, or socioeconomic factors, were excluded to maintain the integrity of the data set and ensure robust predictive modeling. Additionally, patients undergoing orthodontic treatment were not included because such interventions could influence oral health outcomes and potentially confound the analysis.

A total of 148 participants met these criteria and were included in this study. This sample size was considered sufficient for developing and validating machine-learning models while balancing computational feasibility.Patient confidentiality was ensured by anonymizing the data. Access to the data set was restricted to authorized personnel, and all procedures complied with the ethical guidelines for retrospective studies.

Data Collection

Demographic Data

Demographic data were collected to examine the potential influence of social and environmental factors on the risk of caries. Information included the child's age and gender, along with the residential area (urban or rural), which provided insight into access to oral healthcare resources. Socioeconomic status (SES) indicators such as parental education level and household income were also recorded to evaluate their association with caries risk. These variables helped establish a comprehensive context for analyzing disparities in oral health outcomes [11].

Clinical Data

Clinical data served as a cornerstone for assessing the current and past caries experiences of the participants. The DMFT score was recorded for permanent teeth, while decayed, missing, and filled teeth in primary dentition were assessed separately. Cavitated teeth were identified based on ICDAS criteria, which include visual-tactile inspection, avoiding unnecessary probing to prevent damage. Additionally, plaque accumulation and overall oral hygiene status were evaluated to understand the relationship between oral hygiene practices and the risk of caries. These clinical parameters are essential for validating the predictions of the machine-learning model.

Behavioral and Lifestyle Factors

Behavioral and lifestyle factors were included to account for modifiable risks associated with caries. The frequency of sugary food and beverage consumption was categorized as daily, weekly, or rarely, providing a detailed picture of dietary habits. Sugary foods were defined as items containing added sugars or free sugars, such as sucrose, fructose, and glucose. Examples include candies, sugary beverages, pastries, and processed snacks. The study focused on dietary sugars contributing to cariogenic risks. Oral hygiene practices, such as frequency of tooth-brushing, use of fluoride toothpaste, flossing, and mouth rinses, were recorded to evaluate their preventive effects. Additionally, data on the history of dental visits and preventive care, including professional fluoride application and sealant placement, were collected to determine their impact on reducing caries risk. These variables provide actionable insights into personalized prevention strategies.

Fluoride Exposure

Fluoride exposure, which is a critical determinant of caries prevention, was assessed using multiple sources. Information on the use of fluoridated toothpaste or mouth rinses was documented along with access to community water fluoridation. The frequency and type of professional fluoride treatment were also recorded to understand the cumulative protective effects of fluoride on dental health. These data were crucial for assessing the role of fluoride in caries prevention and its integration into the predictive model.

Family History and Genetic Predisposition

Family history and genetic predisposition were included to explore the inherited and shared environmental factors influencing caries risk. The presence of a family history of dental caries among immediate family members was documented to identify any potential genetic links. Any known hereditary conditions affecting enamel quality, salivary composition, or immune responses were also recorded. These factors contribute to the interplay between genetic and environmental influences in caries development.

Data Anonymization and Organization

To ensure patient confidentiality, all the data were anonymized by assigning unique patient codes. An electronic database was created to store and organize the collected data, thereby facilitating seamless integration into the machine learning analysis. This structured approach to data collection and management ensured the reliability and integrity of the findings.

Data Pre-processing and Feature Engineering

To ensure that the data set was suitable for machine-learning analysis, several pre-processing steps were implemented.

Handling Missing Data: Missing values were addressed using the k-nearest neighbor (KNN) algorithm for numerical variables and mode imputation for categorical variables. This approach ensures that no significant information is lost while maintaining data integrity.

Encoding Categorical Variables: Categorical variables such as fluo-

ride use and SES were converted into numerical formats using one-hot encoding. Binary variables, such as the presence or absence of active caries, were labeled as 0 or 1.

Normalization and Standardization: Continuous variables, such as DMFT scores and frequency of sugary food consumption, were standardized using z-scores. This step reduces the impact of outliers and ensures uniformity across different measurement scales.

Feature Selection Process and Results

To enhance the interpret-ability and performance of the machine learning models, recursive feature elimination (RFE) was used for feature selection. RFE is a wrapper-based feature-selection technique that imperatively evaluates a subset of features by training the model and removing the least significant feature at each iteration until the optimal subset is obtained. In this study, RFE was implemented using a Random Forest classifier as the base model owing to its ability to handle nonlinear relationships and feature interactions.

Process Details

1. Initial Feature Set: The initial data set included 25 features spanning the demographic, clinical, behavioral, dietary, and socioeconomic domains.
2. Evaluation Criteria: Features were ranked based on their contribution to model performance using the mean decrease in impurity (Gini importance).
3. Iterative Removal: Features with the lowest importance scores were sequentially removed, and the model's performance was re-evaluated at each step using cross-validation.
4. Optimal Subset: The process identified an optimal subset of 15 features that maximized the predictive accuracy of the model while reducing computational complexity.

Features Retained

The final selected features included:

- Clinical Factors: DMFT score, presence of active caries, and plaque index.
- Behavioral Factors: Frequency of sugary food consumption, brushing frequency, and use of fluoride toothpaste.
- Dietary Factors: Fluoride exposure (toothpaste, mouth rinse, professional application) and community water fluoridation.
- Socioeconomic Indicators: Parental education and household income.
- Family History: Family history of dental caries and genetic predisposition.

Features Removed: Non-contributory features such as residential area (urban/rural) and history of dental sealant placement were excluded, as they demonstrated minimal impact on the model's performance metrics.

Outcome of RFE: The feature selection process improved the computational efficiency of the machine-learning models by reducing the dimensionality of the dataset. Models trained on the selected features demonstrated enhanced accuracy with no significant loss in sensitivity or specificity compared to the full data set.

Machine Learning Model Development

Machine learning models have been developed using supervised learning techniques to classify patients into different caries risk categories. The data set was randomly split into training (80%) and testing (20%) subsets using stratified sampling to ensure an even distribution of the risk levels. The following models were implemented:

Logistic Regression: Used as a baseline model, logistic regression provided a simple framework for binary classification and allowed comparison with more advanced algorithms.

Random Forest Classifier: This ensemble method constructs multiple decision trees and aggregates their predictions. Random forest was chosen for its ability to handle nonlinear relationships and interactions between variables.

Support Vector Machines (SVM): An SVM with a radial basis function (RBF) kernel was employed to capture complex patterns in the data.

Gradient Boosting (XGBoost): This boosting algorithm optimizes classification accuracy by imperatively reducing errors in predictions.

Neural Networks: A multi-layer perceptron (MLP) was tested to evaluate the applicability of deep learning in predicting caries risk.

Model Training and Validation

The training subset was used to train each model. Hyper parameter tuning was performed using a grid search with 10-fold cross-validation to optimize parameters such as the number of trees in the random forest and the learning rate in XGBoost. Over-fitting was mitigated using regularization techniques and cross-validation.

Model Evaluation

The performance of each model was evaluated on the testing subset using the following metrics.

- Accuracy, sensitivity (recall), specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).
- The AUC-ROC is the primary metric for model comparison, as it provides a comprehensive measure of model performance.

Patients were stratified into three risk categories based on predicted probabilities: low (<30%), moderate (30–70%), and high (>70%) risk.

Statistical Analysis

Descriptive statistics were calculated for the demographic, clinical, and behavioral variables. Associations between categorical variables and caries risk were analyzed using Chi-square tests, while independent

t-tests were used to compare continuous variables, such as DMFT scores, across risk groups. Multivariate logistic regression was used to identify independent predictors of caries risk.Data analysis and modeling were conducted using Python (version 3.9) with libraries, such as Scikit-learn for ML, Pandas for data manipulation, and Matplotlib for visualization.

Results

The demographic and clinical characteristics of the participants showed disparities in caries risk. Urban participants (60.8%) had significantly higher DMFT scores compared to rural participants (p = 0.045). Fluoride exposure was reported in 74.3% of the children, and those with fluoride exposure exhibited significantly lower DMFT scores (p = 0.012). No significant differences in caries risk were observed between sexes, indicating that other factors, such as behavioral and socioeconomic variables, may play a more substantial role (Table 1). The mean DMFT score of 3.2 ± 1.5 reflects a moderate prevalence of caries within this cohort.

Table 2 presents key behavioral and socioeconomic factors associated with caries risk. A majority of children (59.5%) consumed sugary foods three or more times per week, while 70.3% reported brushing their teeth at least twice daily. Additionally, 64.2% of participants had a family history of caries. Notably, only 43.9% of parents had attained higher education.

The performance metrics in Table 3 confirm that XGBoost is the most effective model for predicting caries risk in pediatric patients, achieving the highest accuracy (91.5% Â ± 1.3), sensitivity (89.7% Â ± 1.2), specificity (93.2% Â ± 1.0), AUC-ROC (0.94; 95% CI: 0.92–

Table 1. Demographic and Clinical Characteristics of Participants

Characteristic	N (%)	Statistical Significance (p-value)
Age (mean ± SD)	8.7 ± 2.3	-
Gender		-
Male	78 (52.7%)	
Female	70 (47.3%)	
Residential Area		**0.045*** (Urban vs. Rural)
Urban	90 (60.8%)	
Rural	58 (39.2%)	
Mean DMFT Score	3.2 ± 1.5	**<0.001*** (High vs. Low Risk Groups)
Fluoride Exposure		**0.012*** (Yes vs. No)
Yes	110 (74.3%)	
No	38 (25.7%)	

*significant if p<0.05

Table 2. Behavioral and Socioeconomic Factors

Factor	N (%)
Frequency of Sugary Food (≥3 times/week)	88 (59.5%)
Brushing Frequency (≥2 times/day)	104 (70.3%)
Family History of Caries	95 (64.2%)
Parent Education (≥College)	65 (43.9%)

Table 3. Model Performance Metrics with F1-Score

Model	Accuracy (% ± SD)	Sensitivity (% ± SD)	Specificity (% ± SD)	AUC-ROC (95% CI)	F1-Score (% ± SD)
Logistic Regression	82.4 ± 2.1	80.1 ± 2.3	84.5 ± 1.8	0.85 (0.82–0.89)	79.3 ± 2.0
Random Forest	89.2 ± 1.5	87.3 ± 1.7	90.8 ± 1.4	0.91 (0.88–0.93)	87.8 ± 1.8
SVM	85.7 ± 1.9	84.1 ± 2.0	87.5 ± 1.6	0.88 (0.85–0.90)	84.9 ± 1.7
XGBoost	91.5 ± 1.3	89.7 ± 1.2	93.2 ± 1.0	0.94 (0.92–0.96)	90.6 ± 1.4
Neural Network	90.2 ± 1.6	88.4 ± 1.5	91.6 ± 1.3	0.92 (0.90–0.94)	89.2 ± 1.5

0.96), and F1-score (90.6% Â ± 1.4). Neural Networks followed closely with an accuracy of 90.2% Â ± 1.6, AUC-ROC of 0.92 (95% CI: 0.90–0.94), and F1-score of 89.2% Â ± 1.5. Random Forest also performed well, with an AUC-ROC of 0.91 (95% CI: 0.88–0.93) and F1-score of 87.8% ± 1.8, but it was slightly less precise than XGBoost. Logistic Regression showed the lowest performance, with an F1-score of 79.3% ± 2.0, and an AUC-ROC of 0.85 (95% CI: 0.82–0.89).

The AUC-ROC values, as shown in Figure 1, demonstrate that XGBoost achieved the highest discriminatory ability among the models evaluated.

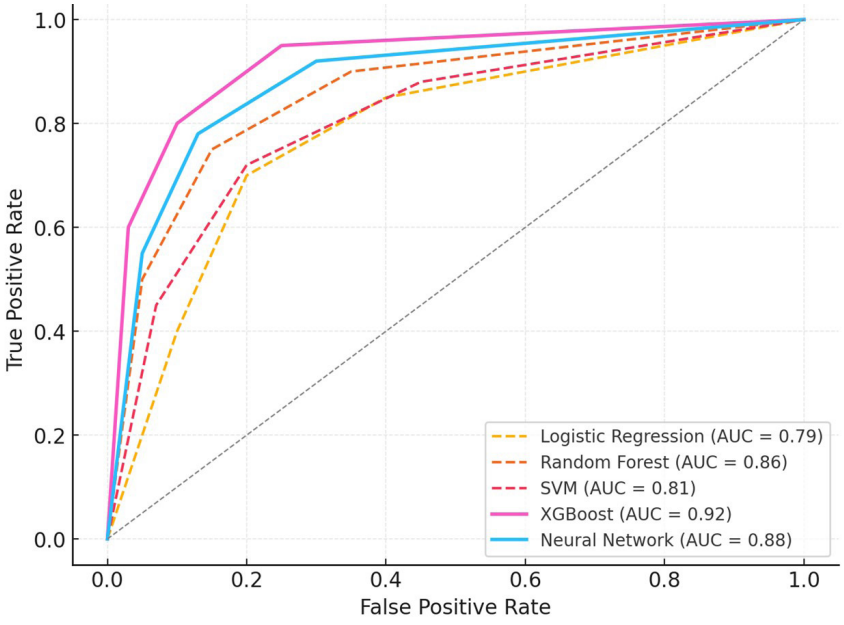

Figure 1. AUC-ROC Curve for ML Models Caries Risk Categories Predicted by XGBoost

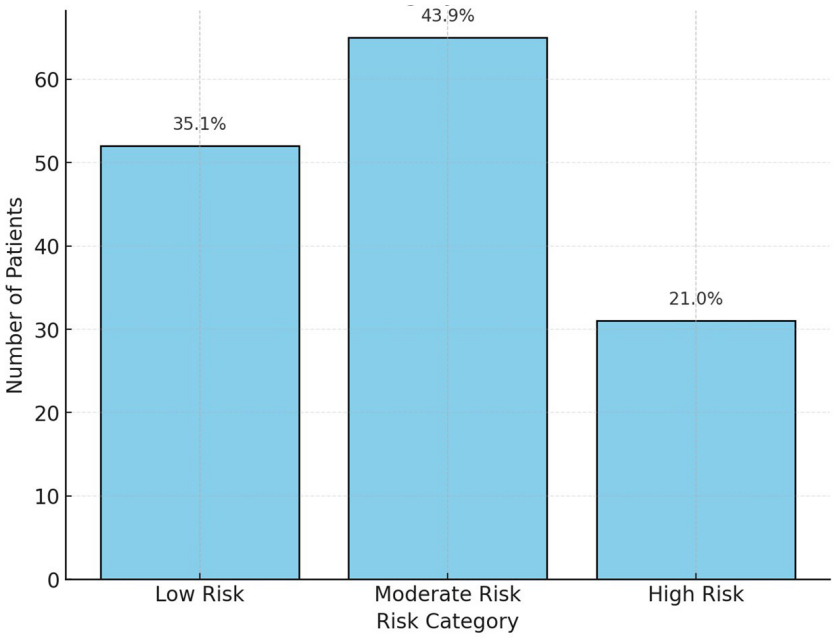

Figure 2. Risk Category Distribution

Table 4 summarizes the distribution of participants across the three caries risk categories predicted using the XGBoost model. The largest group (43.9%) fell under the moderate-risk category, while 21.0% were classified as high-risk.

Table 4. Caries Risk Categories Predicted by XGBoost

Risk Category	N (%)
Low Risk	52 (35.1%)
Moderate Risk	65 (43.9%)
High Risk	31 (21.0%)

Figure 2. visually represents the distribution of the risk categories among the study participants. The bar graph illustrates that while a significant proportion of children fall under the low- and moderate-risk categories, the high-risk group still represents a substantial 21%.

Feature Importance (XGBoost)

Table 5 details the feature importance rankings from the XGBoost model. DMFT scores contributed 35.2% to the overall importance, followed by the frequency of sugary food consumption (28.7%) and fluoride exposure (18.3%). Parental education accounted for 6.4% of the feature importance.

Table 5. Feature Importance (XGBoost)

Feature	Importance (%)
DMFT Score	35.2
Frequency of Sugary Food	28.7
Fluoride Exposure	18.3
Family History of Caries	11.4
Parental Education	6.4

Discussion

This study aimed to develop and validate machine learning (ML) models to predict caries risk in pediatric patients by leveraging a comprehensive dataset comprising clinical, behavioral, dietary, and socioeconomic variables. The findings demonstrated that ML models, particularly XGBoost, can achieve high predictive accuracy, sensitivity, and specificity, offering a robust approach to the early identification of children at risk for dental caries.

The findings of this study demonstrated that the XGBoost model achieved superior performance compared to traditional methods, such as logistic regression, with an AUC-ROC of 0.94. This high level of performance underscores the effectiveness of gradient boosting algorithms in capturing complex, nonlinear relationships and interactions between variables commonly seen in multifactorial conditions such as dental caries. Gradient boosting models such as XGBoost are used to analyze diverse datasets with interdependent variables, providing accurate and actionable risk predictions. These results align with those of prior research, emphasizing the utility of machine learning (ML) in improving dental risk assessments.

For instance, Çiftçi and Aşantoğrol used ML models to predict caries risk groups and oral health-related risk factors in adults. Their study demonstrated a high predictive accuracy, highlighting the potential of ML in caries risk assessment. Similarly, their findings emphasized the importance of key predictors, such as dietary habits, fluoride exposure, and past caries experience, which align closely with the predictors identified in the present study [8]. This convergence underscores the ability of ML to integrate diverse variables into reliable, clinically relevant predictive models.

Such studies reinforce the observations of this study, in which XGBoost excelled by integrating multifactorial data and producing actionable predictions with high reliability. The identification of DMFT scores, sugary food consumption, and fluoride exposure as the most significant predictors further aligns with prior findings that emphasize the importance of both clinical and behavioral factors in caries risk assessment. These consistent results across studies highlight the robustness of ML tools in addressing dental caries, paving the way for more widespread adoption of such technologies in both clinical and public health settings [12-14]

Moreover, this study underscores the importance of incorporating diverse predictors into ML models. DMFT scores emerged as the most important variable, consistent with the existing evidence that past caries experience is a strong indicator of future risk. The significant contributions of dietary habits and fluoride exposure to the predictive model further emphasized the multifactorial nature of caries. Previous research by Pang et al. also highlighted the importance of integrating behavioral and environmental factors into AI-based caries risk assessments [15].

Stratification of patients into low-, moderate-, and high-risk categories is a key strength of this study, providing actionable insights for personalized prevention. The finding that 21.0% of the participants were classified as high-risk underscores the need for targeted interventions in this subgroup. High-risk patients can benefit from preventive measures such as fluoride varnishes, dietary counseling, and regular monitoring, as recommended by the European Academy of Pediatric Dentistry (EAPD) [16]. Similarly, moderate-risk patients may require tailored interventions to prevent progression to high-risk status.

The utility of risk stratification extends from individual patient care to public health planning. By identifying high-risk populations, health care providers and policymakers can allocate resources more effectively. For example, community-based fluoride programs or school-based oral health education initiatives could be prioritized in regions with a higher prevalence of high-risk individuals. Such targeted approaches have been shown to reduce caries incidence and improve oral health outcomes [17].

Traditional caries risk assessment tools, such as the Caries Management by Risk Assessment (CAMBRA) protocol, rely on clinician judgment and subjective evaluation of risk factors [18]. While effective, these methods are often limited by variability in clinician expertise and the inability to analyze complex interactions between variables. By contrast, ML models can process large datasets with multiple interdependent variables, providing objective and reproducible predictions. This advantage was evident in the current study, where ML models achieved higher predictive accuracy than the conventional approaches reported in the literature [19, 20].

However, the integration of ML models into clinical practice requires careful consideration of interpretability. Clinicians may find it challenging to understand the decision-making processes of complex algorithms, such as XGBoost or neural networks. Explainable AI (XAI) techniques, such as Shapley Additive Explanations (SHAP), can be used to elucidate the contributions of individual variables to model predictions, enhancing clinician trust and adoption [21].

The significant role of sugary food consumption and fluoride exposure in caries risk prediction highlights the importance of addressing the modifiable risk factors. Frequent consumption of sugary foods and beverages was the second most important predictor, contributing 28.7% of the XGBoost model. This finding aligns with the well-established relationship between dietary sugars and caries development documented in numerous epidemiological studies. For instance, the WHO recommends limiting free sugar intake to less than 10% of the total en-

ergy consumption to reduce caries risk [22].

Fluoride exposure, the third most important predictor in this study, accounted for 18.3% of the model's predictive power. The protective effect of fluoride against caries is well documented, with evidence supporting its role in enhancing enamel remineralization and inhibiting demineralization [23]. Community water fluoridation, fluoridated toothpaste, and professional fluoride applications have been shown to significantly reduce caries prevalence in children [23]. The findings of this study reinforce the need for the continued promotion of fluoride use as a cornerstone for caries prevention.

A family history of caries, a proxy for genetic predisposition and shared environmental factors, contributed 11.4% to the predictive model. Genetic susceptibility to caries has been linked to variations in enamel formation, salivary composition, and immune responses, underscoring the complex interplay between genetics and caries risk [24]. Additionally, parental education and socioeconomic status (SES) were associated with caries risk, albeit with lower importance scores. These findings are consistent with research indicating that lower SES is linked to limited access to preventive care, unhealthy dietary habits, and reduced fluoride use [25].

The implementation of ML-based caries risk prediction models has several implications in clinical practice. First, these models can facilitate early identification of high-risk patients, enabling timely and targeted interventions. Second, ML tools can support evidence-based decision making, reduce reliance on subjective assessments, and improve diagnostic accuracy. Third, by stratifying patients based on risk, clinicians can prioritize resources for those who need them the most, thereby enhancing the efficiency of dental care delivery.

From a public health perspective, ML models can inform population-level strategies for the prevention of caries. For example, predictive models can identify geographic areas with a higher prevalence of high-risk children, thereby guiding the allocation of resources for community-based interventions. Additionally, ML tools can be integrated into school-based oral health programs to provide educators and health workers with data-driven insights to tailor preventive measures.

Despite its strengths, this study had several limitations. First, the sample size may have limited the generalization of the findings. Larger multi center datasets are needed to validate the models and ensure their applicability across diverse populations. Second, the retrospective design of the study relied on the availability and quality of existing records, which may have introduced selection bias or data inaccuracies. Prospective studies with standardized data collection protocols are warranted to address these issues.

Another limitation is the potential under representation of certain risk factors, such as salivary biomarkers or oral micro-biome composition, which were not included in this study. Future research should explore the integration of biological markers into ML models to enhance their predictive accuracy. Longitudinal studies are needed to assess the long-term performance of these models in predicting caries risk over time.

Finally, the successful implementation of ML models in clinical practice requires addressing barriers, such as cost, technical expertise, and clinician acceptance. User-friendly interfaces and clinician training programs are critical to ensure the widespread adoption of these tools.

Conclusion

This study demonstrated the potential of machine learning models, particularly XGBoost, to provide accurate and clinically relevant predictions of caries risk in pediatric patients. By incorporating diverse clinical, behavioral, and environmental factors, these models offer a data-driven approach to risk assessment, thereby addressing the limitations of traditional methods. These findings highlight the importance of modifiable risk factors, such as diet and fluoride use, in caries prevention and underscore the need for targeted interventions for high-risk populations.

Future efforts should focus on validating these models in larger, more diverse cohorts, integrating novel risk factors, and developing user-friendly tools for clinical applications. By leveraging the power of AI, dentistry can advance toward a more predictive, personalized, and preventive approach to caries management, ultimately improving oral health outcomes in children.

Acknowledgement

We would like to extend our sincere gratitude to doctors and supporting staff of all the associated dental institutions for this study protocol formulation and patient data analysis etc.We would like to extend our gratitude to the technical team for their assistance in creating Machine learning model software for this particular study and its application.

Ethical approval

The study was approved by Institutional Ethical Committee Board of Chattisgarh Dental College and Research Institute -Ref#GGSDC/Dean/Res/21/14 .

Statement of informed consent

Written informed consent was obtained from the patients' guardians for publication of this article. A copy of the written consent is available for review by the Editor-in Chief of this journal on request.

References

1. Dye BA. The Global Burden of Oral Disease: Research and Public Health Significance. J Dent Res. 2017 Apr;96(4):361-363. doi: 10.1177/0022034517693567.
2. Chen J, Chen W, Lin L, Ma H, Huang F. The prevalence of dental caries and its associated factors among preschool children in Huizhou, China: a cross-sectional study. Front Oral Health. 2024 Aug 30;5:1461959. doi: 10.3389/froh.2024.1461959.
3. Spatafora G, Li Y, He X, Cowan A, Tanner ACR. The Evolving Microbiome of Dental Caries. Microorganisms. 2024 Jan 7;12(1):121. doi: 10.3390/microorganisms12010121.
4. Zou J, Du Q, Ge L, et al. Expert consensus on early childhood caries management. Int J Oral Sci. 2022 Jul 14;14(1):35. doi: 10.1038/s41368-022-00186-0.
5. Nie E, Jiang R, Islam R, Li X, Yu J. Evaluation of caries risk assessment practices among dental practitioners in Guangzhou, China: a cross-sectional study. Front Oral Health. 2024 Oct 16;5:1458188. doi: 10.3389/froh.2024.1458188.
6. Sadegh-Zadeh SA, Bagheri M, Saadat M. Decoding children dental health risks: a machine learning approach to identifying key influencing factors. Front Artif Intell. 2024;7:1392597. Published 2024 Jun 17. doi:10.3389/frai.2024.1392597
7. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. J Dent Res. 2020;99(7):769-774. doi:10.1177/0022034520915714
8. Çiftçi BT, Aşantoğrol F. Utilization of machine learning models in predicting caries risk groups and oral health-related risk factors in adults. BMC Oral Health. 2024;24(1):430. Published 2024 Apr 8. doi:10.1186/s12903-024-04210-z
9. Butera A, Maiorani C, Morandini A, et al. Evaluation of Children Caries Risk Factors: A Narrative Review of Nutritional Aspects, Oral Hygiene Habits, and Bacterial Alterations. Children (Basel). 2022 Feb 15;9(2):262. doi: 10.3390/children9020262.
10. Alexander KE, Brijnath B, Mazza D. Parents' decision making and access to preventive healthcare for young children: applying Andersen's Model. Health Expect. 2015 Oct;18(5):1256-69. doi: 10.1111/hex.12100.
11. Featherstone JDB, Crystal YO, Alston P, et al. Evidence-Based Caries Management for All Ages-Practical Guidelines. Front Oral Health. 2021 Apr 27;2:657518. doi: 10.3389/froh.2021.657518. PMID: 35048005; PMCID: PMC8757692.
12. Ogwo C, Brown G, Warren J, Caplan D, Levy S. Predicting dental caries outcomes in young adults using machine learning approach. BMC Oral Health. 2024 May 3;24(1):529. doi: 10.1186/s12903-024-04294-7.
13. Park YH, Kim SH, Choi YY. Prediction Models of Early Childhood Caries Based on Machine Learning Algorithms. Int J Environ Res Public Health. 2021 Aug 15;18(16):8613. doi: 10.3390/ijerph18168613.
14. Ramos-Gomez F, Marcus M, Maida CA, et al. Using a Machine Learning Algorithm to Predict the Likelihood of Presence of Dental Caries among Children Aged 2 to 7. Dent J (Basel). 2021;9(12):141. Published 2021 Dec 1. doi:10.3390/dj9120141
15. Pang L, Wang K, Tao Y, Zhi Q, Zhang J, Lin H. A New Model for Caries Risk Prediction in Teenagers Using a Machine Learning Algorithm Based on Environmental and Genetic Factors. Front Genet. 2021;12:636867. Published 2021 Mar 11. doi:10.3389/fgene.2021.636867
16. Toumba KJ, Twetman S, Splieth C, Parnell C, van Loveren C, Lygidakis N. Guidelines on the use of fluoride for caries prevention in children: an updated EAPD policy document. Eur Arch Paediatr Dent. 2019;20(6):507-516. doi:10.1007/s40368-019-00464-2
17. Nghayo HA, Palanyandi CE, Ramphoma KJ, Maart R. Oral health community engagement programs for rural communities: A scoping review. PLoS One. 2024 Feb 6;19(2):e0297546. doi: 10.1371/journal.pone.0297546.
18. Featherstone JDB, Crystal YO, Alston P, Chaffee BW, Doméjean S, Rechmann P, Zhan L, Ramos-Gomez F. Evidence-Based Caries Management for All Ages-Practical Guidelines. Front Oral Health. 2021 Apr 27;2:657518. doi: 10.3389/froh.2021.657518.
19. Alzaid N, Ghulam O, Albani M, Alharbi R, Othman M, Taher H, Albaradie S, Ahmed S. Revolutionizing Dental Care: A Comprehensive Review of Artificial Intelligence Applications Among Various Dental Specialties. Cureus. 2023 Oct 14;15(10):e47033. doi: 10.7759/cureus.47033.
20. Al-Khalifa KS, Ahmed WM, Azhari AA, et al. The Use of Artificial Intelligence in Caries Detection: A Review. Bioengineering (Basel). 2024;11(9):936. Published 2024 Sep 18. doi:10.3390/bioengineering11090936
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;30:4765–74.

22. World Health Organization. Sugars and dental caries. Geneva: WHO; 2017. https://www.who.int/publications-detail/9789241549028/

23. Mankar N, Kumbhare S, Nikhade P, Mahapatra J, Agrawal P. Role of Fluoride in Dentistry: A Narrative Review. Cureus. 2023 Dec 21;15(12):e50884. doi: 10.7759/cureus.50884.

24. Abbasoğlu Z, Tanboğa İ, Küchler EC, et al. Early childhood caries is associated with genetic variants in enamel formation and immune response genes. Caries Res. 2015;49(1):70-7. doi: 10.1159/000362825. Epub 2014 Dec 18. PMID: 25531160; PMCID: PMC4376372.

25. Yousaf M, Aslam T, Saeed S, Sarfraz A, Sarfraz Z, Cherrez-Ojeda I. Individual, Family, and Socioeconomic Contributors to Dental Caries in Children from Low- and Middle-Income Countries. Int J Environ Res Public Health. 2022 Jun 10;19(12):7114. doi: 10.3390/ijerph19127114.