Submitted: 11 September, 2024 Accepted: 07 February, 2025 Published: 03 September, 2025 DOI:10.22514/jocpd.2025.110

Journal of Clinical Pediatric Dentistry

ORIGINAL RESEARCH

Prediction of dental caries in children through machine learning

Sarah Ahmed Bahammam[1],*

[1]Preventive Dental Sciences Department, Dental College and Hospital, Taibah University, 4147 Medina, Kingdom of Saudi Arabia

***Correspondence**
sbahammam@taibahu.edu.sa
(Sarah Ahmed Bahammam)

Abstract

Background: Dental caries is a contagious disease that decays the structure of the tooth, with cavities of the tooth as the one of most common outcomes. It is categorized as one of the prevailing issues related to oral health. The study on dental caries has been examined for early prediction due to treatment cost and pain. Medical study in oral healthcare has depicted restrictions such as the requirement of time and funds consideration. Therefore, artificial intelligence (AI) has been employed lately to produce models that can predict the dental caries risk. **Methods**: This study employed machine learning techniques to predict dental caries among children. The questionnaire, containing both clinical and socio-demographic information, was divided into training and testing sets to train and evaluate four distinct predictive models: Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost) and Light Gradient-Boosting Machine (LightGBM). Our investigation aimed to assess the performance of each model performance in identifying cases of dental caries. Through key metrics such as precision, recall and specificity, we gauged effectiveness in distinguishing positive and negative cases. Additionally, the study *involved* generating Receiver Operating Characteristic curves and calculating the Area Under the Curve for each model to evaluate predictive capability. The models' predictions were then visualized and compared through graphical representations, including ROC curves. **Results**: LightGBM exhibited the highest overall performance, achieving an accuracy of 85% and demonstrating superior sensitivity and specificity in distinguishing between children with and without dental caries. Random Forest and XGBoost also yielded commendable results with accuracy rates of 83% and 84%, respectively. **Conclusions**: It is concluded that the results revealed promising predictive capabilities across all four models. Our findings provide valuable insights into the feasibility of utilizing machine learning algorithms for dental caries prediction, contributing to the advancement of preventive dental healthcare practices among children.

Keywords

Algorithms; Artificial intelligence; Dental caries; Machine learning; ROC curve; Sensitivity and specificity

1. Introduction

Dental caries is a global health issue, and its prevalence in children is generally high [1]. It is considered one of the most common infectious diseases of childhood, which is evident from the fact that it is seven times more common than hay fever or allergic rhinitis and five times more common than asthma in children [2]. Childhood caries is a disease that is a rising health burden globally. Later it can cause new caries in both permanent and primary dentition, impacting oral health lifelong [3]. In developing countries, almost 60% of the student population suffers from dental caries [4]. Under severe conditions, the pain and discomfort associated with dental caries can affect the quality of life of children and may lead to acute and chronic infections, changed eating and sleeping habits, educational loss, increased risk of hospitalization and acute and chronic conditions [5, 6].

There are four major risk factors for dental caries: plaque, dietary habits, time, saliva and susceptible tooth [7] Several factors that contribute to dental caries are not observable during clinical examination [8] but substantially lead to the disease development [9]. These factors are determined by clinical evaluation of different aspects such as attitudes, knowledge and oral and socioeconomic health behavior [10].

Risk factors are the signs that might lead to the occurrence of pathological conditions. It is essential to have an exact epidemiological examination tool for predicting the risk of caries in children [11]. A questionnaire is considered the best method to accomplish this objective.

This is an open access article under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

©2025 The Author(s). Published by MRE Press.

Recently, artificial intelligence (AI) advent has led to a significant amount of advancement in diagnostics related to medicine inclusive of the dentistry field, specifically its application in periodontal disease and cariology [12]. AI and deep learning (DL) are playing an important role in transforming dentistry practices by improving diagnostic accuracy, treatment planning and patient care. This is done by enhancing radiograph analysis, supporting periodontal disease detection and automating orthodontic assessments. For instance, AI optimizes restoration design using advanced CAD (Computer-Aided Design) and CAM (Computer-Aided Manufacturing) systems to enhance its overall efficiency and desired customization [13]. A recent study by Zhu *et al.* [14] utilized an AI framework with BDU-Net an nnU-Net to show excellent accuracy in diagnosing conditions like missing teeth, full crowns, residual roots and caries. Therefore, it is believed that AI can enhance diagnostic accuracy and speed resulting improved patient outcomes and reduced healthcare costs.

Machine learning (ML) has become an effective technique for understanding and examining large data and is being utilized in several ways in the medical field. The technique of ML is employed for future prediction results by learning active patterns between targeted data elements. The following algorithm exhibits an increased level of specificity, sensitivity and accuracy, surpassing the conventional techniques, hence, suggesting AI potential as a power tool for diagnostic purposes. Lately, notable achievements have been made in translation, speech and image recognition with the help of ML. In the area of dentistry in Korea the deep learning (DL) study utilizing X-ray, dental computed tomography, panoramic and intraoral images. Nonetheless, as compared to the significance and effectiveness of this technique, it is not being employed steadily in the field of dentistry, and utilization of this in many diverse areas is required. Apart from traditional methods, ML can detect caries lesions, and teeth with the help of basic information or surveys before the diagnosis made by a specialist. The cost, time and human resources needed for an oral health examination will be lessened. It can also identify those individuals that are at high risk. The majority of reviewed works are on prediction and diagnosis utilizing deep learning through medical images and radiography. We do not discuss the method based on images, we have utilized data on demographics, and behavior as input to detect caries in children.

The traditional diagnostic methods often fail to address non-clinical factors such as behavioral and socioeconomic contributors despite of significant burden of dental caries in children and its implications for lifelong oral health. ML gives a novel solution to predict caries risk efficiently and accurately by analyzing diverse datasets. Therefore, this study aims to develop and evaluate machine learning models, including Logistic Regression, Random Forest, XGBoost and LightGBM, for the prediction of dental caries among children aged 2 to 6 years. The study seeks to assess the predictive performance of these models in identifying children at risk of dental caries. The present study aims to enhance early diagnosis of dental caries through ML models like Logistic Regression, Random Forest, XGBoost and LightGBM to enable preventive interventions. The investigation aims to provide insights into the effectiveness of different machine learning algorithms for early detection of dental caries, contributing to the advancement of preventive strategies and interventions in pediatric oral health.

2. Materials and methods

This prospective observational study used a survey to collect data and incorporate the data related to the presence or absence of dental caries obtained from a clinical examination to predict the risk of occurrence of caries in a child. The prospective observational study was conducted from January 2024–March 2024. The population comprised Arabic and non-Arabic mother-child pairs with children ages ranging from 2 to 6 years having complaints of dental caries. Children were excluded if they had a particular need for maintaining a level of oral hygiene and/or their parents refused to take part in the survey.

Each child underwent a clinical examination performed by a trained pediatric dentist to ensure the presence or absence of dental caries. This clinical examination helped to identify visible dental caries, including cavities, white spot lesions, residual roots, general assessment of plaque buildup and tooth surface health. To select individuals from the target population, the study utilized a non-probability approach called convenience sampling. The needed sample size was determined using Epi-Info^{TM} (CDC. Atlanta, GA, USA. 2019) with the default assumptions of 40% outcome probability, 95% confidence level and 5% error margin (r = 0.4; α = 5%; β = 10%; n = 200) [15].

The following section depicts the dataset, preprocessing of data and the prediction model utilized for predicting dental caries in children. Four prediction methods were trained to utilize the reduced subsets of features gained with the help of algorithms of feature selection, and the ML method's performances were compared. A flow diagram of the predicting model of dental caries in children is shown in Fig. 1.

2.1 Machine learning algorithms

2.1.1 Logistic regression

Logistic regression (LR) is a type of model of data based on Categories in which dependent variables contain sequence and nominal scale and is utilized when the dependent variable value is 1 and 0, of which verification is performed of odd ratio. The following equation is of LR, where; y is the dependent variable (prediction of absence or presence of dental caries), β is the coefficient, which will present the association between the independent variable and dependent variable and x is the explanatory variable (predict the impact of factor variables) [15].

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

$$\frac{y}{1-y}; y = 1\ for\ infinity\ and\ 0\ for\ y = 0$$

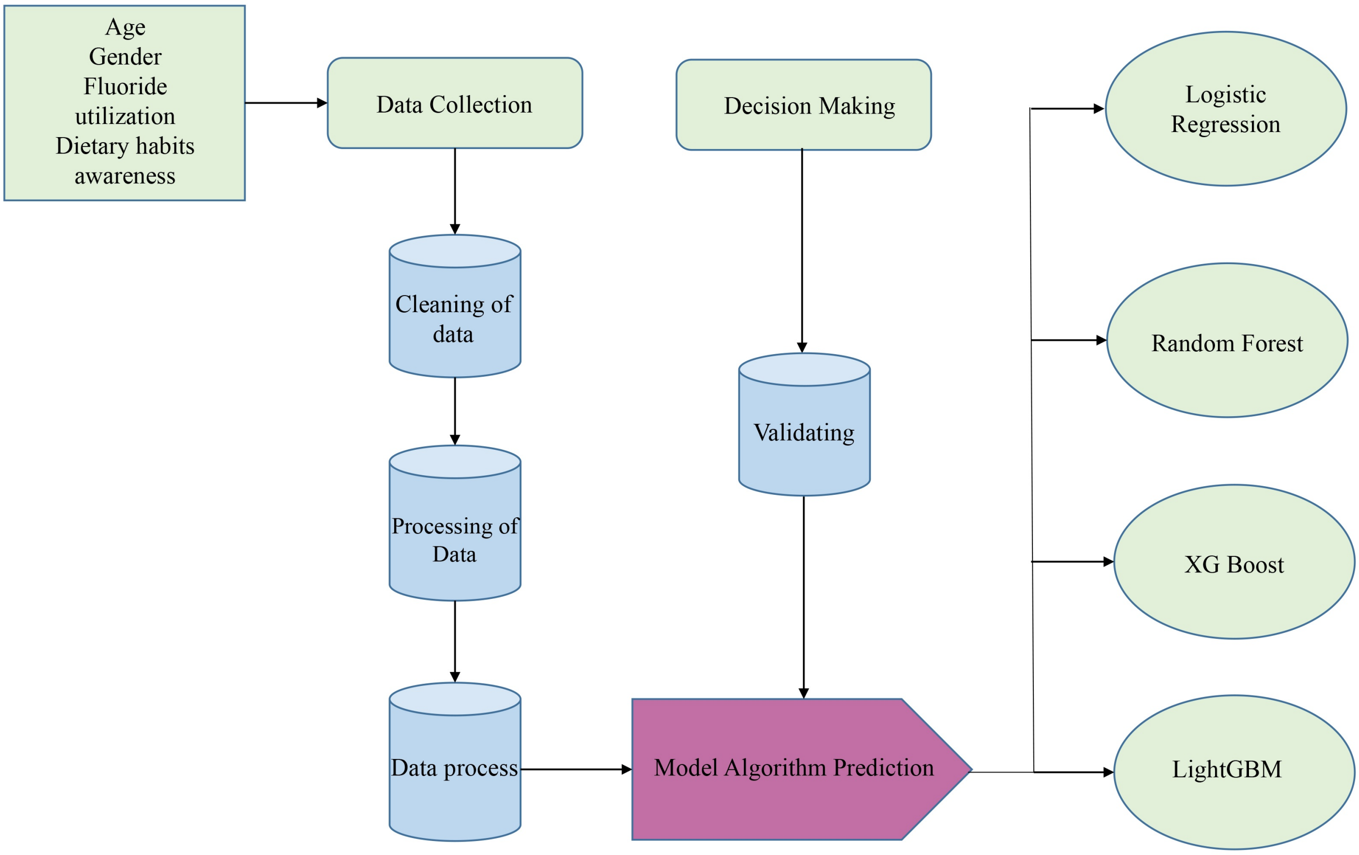

FIGURE 1. Algorithm utilization for prediction.

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

2.1.2 Random forest (RF)

Random forest (RF) is a method of decision-making in tree analysis, and it is a method of machine learning that examines datasets by spreading them into multiple trees. As the selection of variables is free, the dataset's overfitting can be prevented, and a high level of predictivity is gained. Multiple layers of the same size of data are collected with random sampling that enables duplication. The outcome is predicted by merging the results taken from every decision tree with average weight. The RF method has many benefits as compared to other techniques of mining data and can be utilized to enhance the research model understanding and performance for prediction [16].

2.1.3 XGBoost

XGBoost is a machine learning algorithm that uses tree-based analysis. It is also known as the gradient boosting technique. Boosting is a method for accuracy prediction. Gradient descent is performed by weighting the errors of those prediction models that are weak and reflecting them in the learning model which is next to produce a minimum amount of loss and strong prediction. This type of algorithm drastically lessens the time of model performance by employing a gradient boost. It has a strong level of durability that prevents the overfitting of a function [17].

2.1.4 LightGBM algorithm

Microsoft Research Asia designed another algorithm known as "LightGBM" employing the Gradient Boosting Decision Tree (GBDT) framework. It has also the objective of enhancing the efficiency of computers, so the problems of prediction of big data can be solved and work smoothly. It also gives better accuracy, low usage of memory and faster training speed [18].

2.2 Performance evaluation

For the evaluation of the machine learning algorithms' performance the recall, precision, F1 score and accuracy have been broadly utilized. Between them, recall indicates the capability of the model to predict samples correctly, precision is the model's ability to predict samples accurately, F1-score measures the recall and precision performance and accuracy shows the sample proportion that is predicted correctly.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\text{-}Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

$$Accuracy = \frac{TP + FP}{(TN + TP + FN + FP)}$$

2.3 Data analysis

The machine learning was performed using Python 3.12.1. For the reliability test, the Pingouin library was utilized along with pandas, scikit-learn, LightGBM, XGBoost, Matplotlib and PIL for dataset training.

3. Results

The survey items related to dental health, represented by questions Q1 to Q24, exhibit a commendable level of internal consistency with a calculated Cronbach's Alpha of 0.926 shown in Table 1. This high alpha value indicates robust reliability among the survey items, suggesting that they effectively measure the same underlying construct. In practical terms, respondents consistently provided coherent and reliable responses across the entire set of questions. A Cronbach's Alpha exceeding the conventional threshold of 0.70 indicates strong internal consistency. The result of 0.926 reflects a high degree of correlation among the survey items, reinforcing the reliability of our measurement instrument. This finding enhances the credibility and validity of our survey, affirming its efficacy in gauging the targeted aspects of dental health among the surveyed population.

TABLE 1. Cronbach's Alpha.

Cronbach's Alpha	Total number of Questions
0.926	24

The data Tables 2 and 3 reveal the demographics and multifaceted panorama of factors influencing children's oral health, with notable findings across distinct domains. There were a total of 229 male children and 271 female were present. A significant proportion of families (54.2%) exhibit a commendable socioeconomic status, underlining a potential link between economic well-being and oral health. However, concerning dental practices, a substantial number of children (60.6%) do not undergo annual check-ups, signaling a gap in preventive care. Alarmingly, only 19.8% adhere to the recommended practice of brushing their teeth twice daily. While fluoridated toothpaste enjoys relatively high usage (63.8%), supervision during tooth brushing is lacking in the majority (69.4%) of cases. The prevalence of sugary food consumption (77%) further underscores dietary habits contributing to oral health challenges. These findings underscore the need for targeted interventions, particularly in promoting regular dental check-ups, reinforcing oral hygiene practices and addressing dietary patterns. Enhancing awareness and access to preventive services could be instrumental in improving overall pediatric oral health outcomes.

TABLE 2. Demographic.

Categories	Frequency (n)
Gender	
Male	229
Female	271
Age (yr)	
2	105
3	92
4	90
5	122
6	91
Parent's education level	
Diploma	347
Graduate	121
Postgraduate	32

A substantial majority of parents (54.8%) demonstrate awareness of the importance of oral health, indicating a foundational understanding within the community. Positive parental attitudes toward dental care are prevalent (50.8%), contributing positively to children's oral health practices. However, there is a notable gap in knowledge of preventive measures, with only 39% of parents reporting sufficient awareness. Parental involvement in their child's oral hygiene practices is relatively balanced (51%), showcasing an opportunity for increased engagement. The data suggests there was significant peer influence (61.6%) on children's oral health behaviors. Cultural practices related to oral care are prominent among parents (54.8%), while traditional remedies affecting oral health are utilized by a substantial proportion (65.2%). Access to dental care poses challenges, with 38% residing near dental clinics, and 37.6% finding dental care affordable. Moreover, 28% of families have dental care covered by insurance. Additionally, children's engagement in regular physical activity is reported by 31%, highlighting a health behavior that may contribute to overall well-being. These findings collectively underscore the complex interplay of socio-cultural, economic and behavioral factors influencing pediatric oral health within the studied population.

In Table 4 the Logistic Regression model, the confusion matrix indicates that out of 98 instances, 77 were correctly predicted, resulting in an accuracy of approximately 77%. The model correctly identified 42 cases with dental caries (true positives) and 35 cases without dental caries (true negatives). However, there were 14 cases wrongly predicted as not having dental caries (false negatives) and 9 cases wrongly predicted as having dental caries (false positives). The precision is approximately 82.4%, representing the accuracy of positive predictions. The sensitivity (recall) is about 75%, demonstrating the model's ability to capture the majority of actual positive cases. The F1-Score, a balanced metric considering both precision and recall, is approximately 78.5%. Fig. 2 gives

TABLE 3. Questionnaire responses.

Questions	Yes	No
Is the family's socioeconomic status good?	271 (54.2%)	229 (45.8%)
Do the parents/guardians have high educational levels?	200 (40.0%)	300 (60.0%)
Has the child experienced dental caries before?	271 (54.2%)	229 (45.8%)
Does the child visit the dentist for check-ups at least once a year?	197 (39.4%)	303 (60.6%)
Does the child regularly utilize dental services?	154 (30.8%)	346 (69.2%)
Does the child brush their teeth twice a day?	99 (19.8%)	401 (80.2%)
Does the child use fluoridated toothpaste?	319 (63.8%)	181 (36.2%)
Is the child supervised during tooth brushing?	153 (30.6%)	347 (69.4%)
Does the child consume sugary foods and beverages?	385 (77.0%)	115 (23.0%)
Does the child get fluoride from toothpaste?	274 (54.8%)	226 (45.2%)
Is the fluoride content in toothpaste and water adequate?	305 (61.0%)	195 (39.0%)
Does the child have any chronic illnesses or conditions affecting oral health?	405 (81.0%)	95 (19.0%)
Does the child take any medications that may impact oral health?	85 (17.0%)	415 (83.0%)
Are parents aware of the importance of oral health?	274 (54.8%)	226 (45.2%)
Do parents have positive attitudes toward dental care?	254 (50.8%)	246 (49.2%)
Do parents have sufficient knowledge of preventive measures?	195 (39.0%)	305 (61.0%)
Are parents *involved* in the child's oral hygiene practices?	255 (51.0%)	245 (49.0%)
Is there peer influence on the child's oral health behaviors?	308 (61.6%)	192 (38.4%)
Are there cultural practices related to oral care?	274 (54.8%)	226 (45.2%)
Are traditional remedies or practices affecting oral health used?	326 (65.2%)	174 (34.8%)
Is the dental clinic located close to the child's residence?	190 (38.0%)	310 (62.0%)
Is dental care affordable for the family?	188 (37.6%)	312 (62.4%)
Is dental care covered by insurance?	140 (28.0%)	360 (72.0%)
Does the child engage in regular physical activity?	155 (31.0%)	345 (69.0%)

TABLE 4. Evaluation of models.

Models	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.77	0.82	0.75	0.79
Random Forest	0.83	0.84	0.86	0.85
XGBoost	0.84	0.84	0.88	0.86
LightGBM	0.85	0.87	0.86	0.86

XGBoost: Extreme Gradient Boosting; LightGBM: Light Gradient Boosting Machine.

an overview of the performance evaluation of models.

Moving to the Random Forest model, the confusion matrix illustrates its performance in dental caries prediction. Out of 98 instances, the model achieved an accuracy of around 82.7%, correctly identifying 48 cases with dental caries and 35 cases without dental caries. The false negatives and false positives were reduced compared to the Logistic Regression model, with 8 cases wrongly predicted as not having dental caries and 9 cases wrongly predicted as having dental caries. The precision is approximately 84.2%, and the sensitivity is about 85.7%. The F1-Score for the Random Forest model is approximately 84.9%. These metrics collectively suggest that the Random Forest model outperforms the Logistic Regression model in terms of accuracy and balanced prediction of dental caries cases.

For the XGBoost model, the confusion matrix reveals its performance in dental caries prediction. Out of 98 instances, the model achieved an accuracy of approximately 83.7%, correctly identifying 49 cases with dental caries and 35 cases without dental caries. The false negatives were 7, indicating instances wrongly predicted as not having dental caries, while 9 false positives represent instances wrongly predicted as having dental caries. The precision for XGBoost is around 84.5%, and the sensitivity is approximately 87.5%. The F1-Score for XGBoost is approximately 86.0%. These metrics suggest that the XGBoost model performs well in accurately predicting dental caries cases, with a balanced precision-recall trade-off.

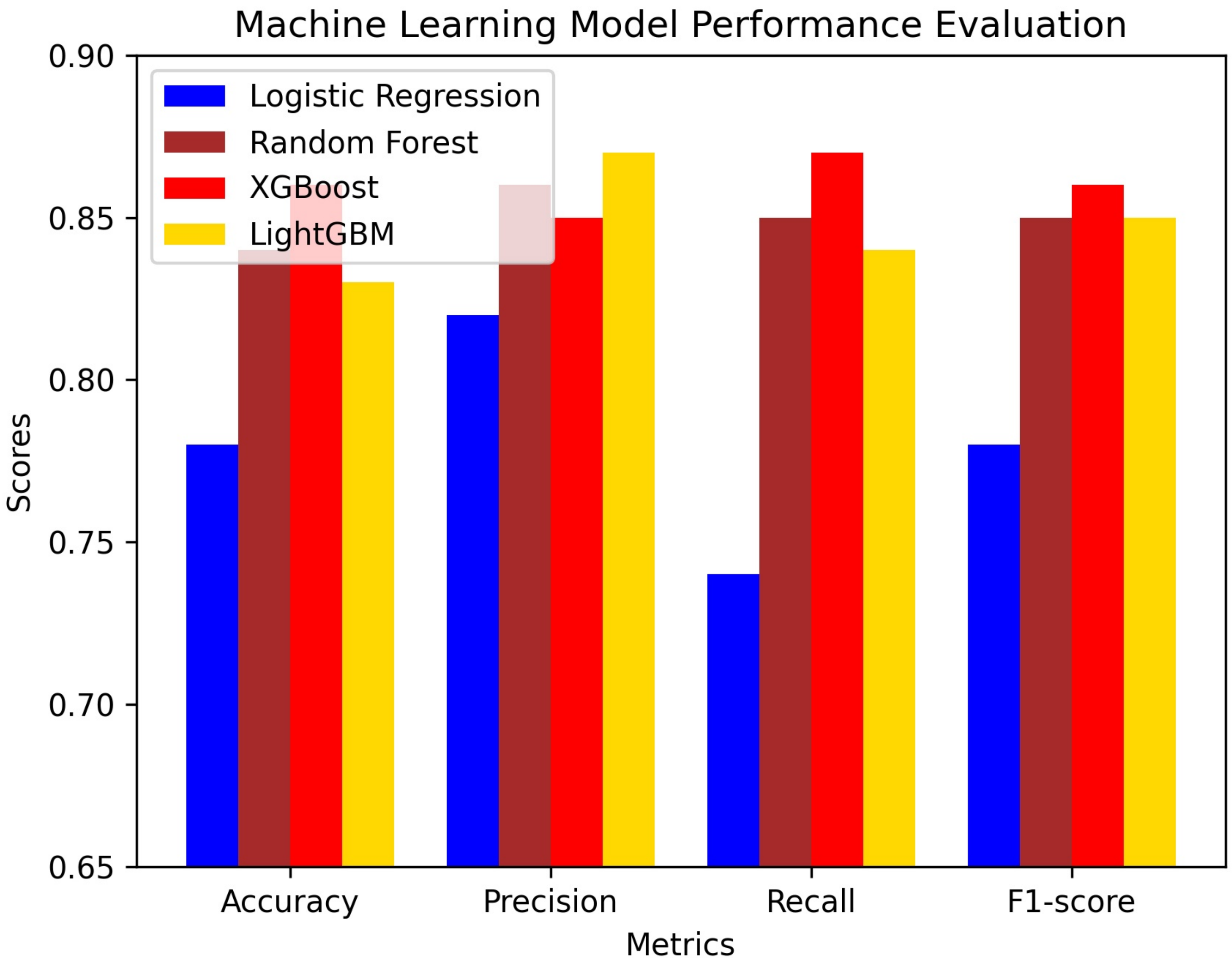

FIGURE 2. Performance evaluation Model. XGBoost: Extreme Gradient Boosting; LightGBM: Light Gradient Boosting Machine.

Moving to the LightGBM model, the confusion matrix illustrates its performance in dental caries prediction. Out of 98 instances, the model achieved an accuracy of around 84.7%, correctly identifying 48 cases with dental caries and 37 cases without dental caries. The false negatives were 8, and the false positives were 7. The precision for LightGBM is approximately 87.3%, and the sensitivity is around 85.7%. The F1-Score for LightGBM is approximately 86.5%. These metrics collectively suggest that the LightGBM model performs well, demonstrating a high level of accuracy and balanced prediction of dental caries cases.

Comparing these two advanced models, XGBoost and LightGBM, both exhibit strong performance, with similar accuracy and F1 scores. The choice between them may depend on other considerations such as computational efficiency, interpretability or specific requirements of the application.

Table 5 shows the specificity and sensitivity of the models. The lightGBM had the the highest specificity (0.84) whereas all the models had 0.80 specificity which measures the proportion of true negative cases of dental caries in children. The highest amount of sensitivity was reported at 0.88 in the XGBoost model and the lowest was found in logistic regression (0.75).

TABLE 5. Specificity and sensitivity of models.

Models	Specificity	Sensitivity
Logistic Regression	0.80	0.75
Random Forest	0.80	0.86
XGBoost	0.80	0.88
LightGBM	0.84	0.86

XGBoost: Extreme Gradient Boosting; LightGBM: Light Gradient Boosting Machine.

The ROC curve value range between 0.8 to 0.9 is considered excellent. Presently our result shows the values of the ROC curve were highest in LightGBM *i.e.*, 0.94, and the lowest in LR 0.83 which depicts that these models can diagnose dental caries in children with and without based on the test (Fig. 3).

4. Discussion

The current study found out that among 4 models LightGBM shows the highest accuracy and Logistic regression reported

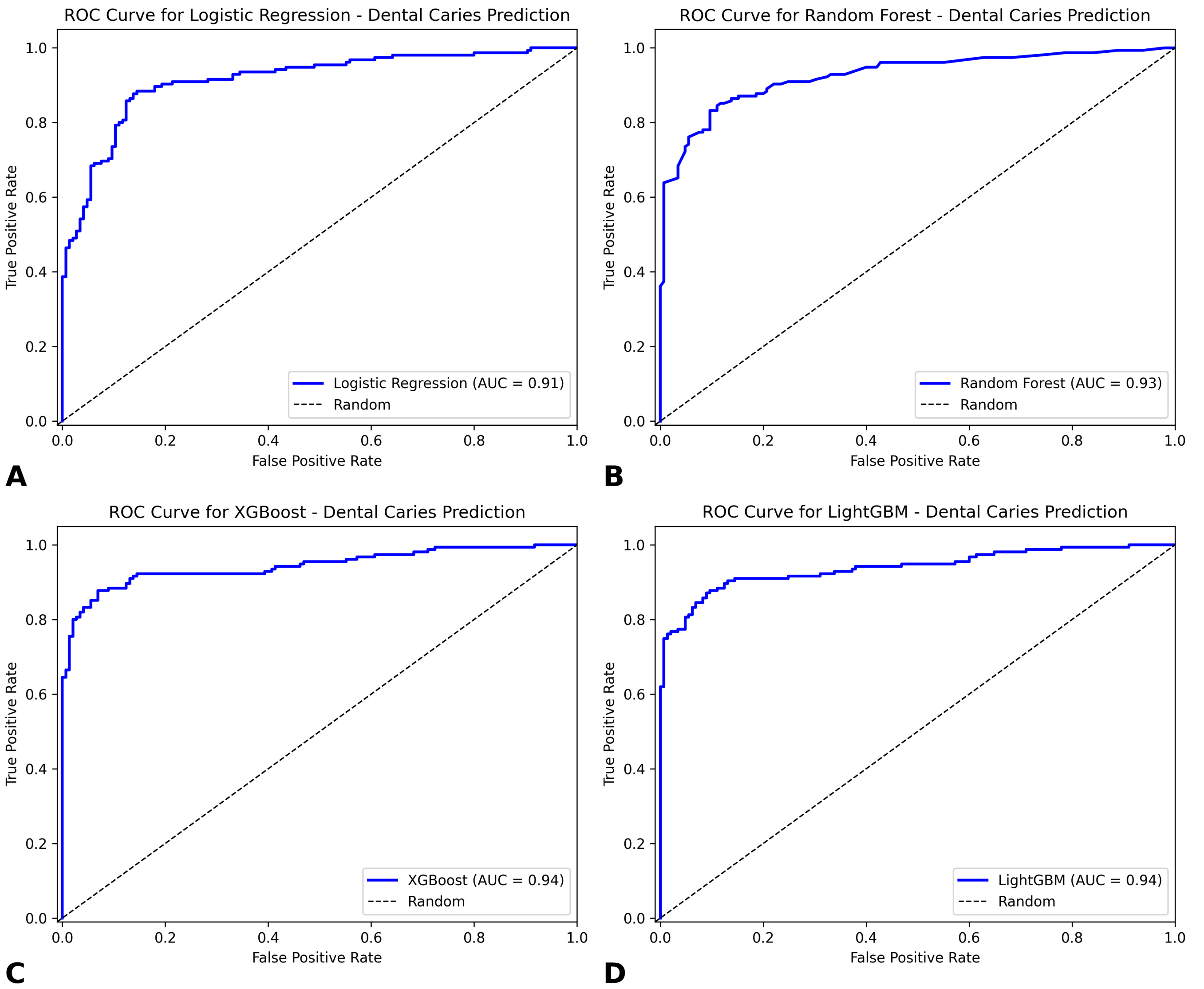

FIGURE 3. Receiver Operating Characteristics Curve and Area Under the Curve for Dental Caries Prediction. (A) ROC Curve for Logistic Regression. (B) ROC Curve for Random Forest. (C) ROC Curve for XGBoost. (D) ROC Curve for Light GBM. ROC: Receiver Operating Characteristic; AUC: Area Under the Curve; XGBoost: Extreme Gradient Boosting; LightGBM: Light Gradient Boosting Machine.

the lowest accuracy. Similarly, Karhade *et al*. [19] utilized 6404 individuals' data to research dental caries in children. The model gives a value of 0.74 ROC curve, 0.67 sensitivity and 0.64 positive value for prediction. Moreover, research was carried out on the prediction of caries with a 97% accuracy, the research found few existing associations between variables like income, age, watching television hours, dental visit last time and caries presence. Based on their study, age was one of the significant predictive values due to the root surface exposure of aged people. One obstacle that was found during the study was the low income of the families which prevented them from visiting the dentist for regular checkups. It was also reported that parameters like demographics, living style and variables of oral health were crucial features for their status of oral health [20].

Another study was conducted by the Center for Disease Control and Prevention Korea. Various algorithms of machine learning were applied and the performance of data was examined utilizing recall, precision, F1-score and accuracy. The algorithm of Random Forest outperformed other models with a precision of 94%, recall of 87%, F1-score of 90% and accuracy of 92%. The outcomes of the study stated that Machine learning is Highly recommended for dental professionals in helping them make early diagnoses and dental caries [21].

Wu *et al*. [22] examine the utilization of sequencing of 16S rRNA and machine learning for the prediction of tooth decay by detecting bacterial communities present in the oral cavity of the individual. The study model comprehends strong predictive ability, reporting that taking environmental and demographic factors could enhance the accuracy of prediction treatment. Ngnamsie Njimbouom *et al*. [23] generated a decision support system based on an algorithm of machine learning to help in planting treatment for dental caries. Similarly, research employs AI for detecting the concentration of fluoride in drinking water in Turkey, the outcomes reported that the utilization of AI was faster, more feasible and cheaper than the utilization of several methods of chemical analysis available [24].

ML helps to identify risk factors resulting in the development of caries among children, along with the generation of computer algorithms for considering different combinations of variables. The performance of the classifier demonstrates minimal bias to ensure applicability in diverse scenarios, with enhanced reliability as a tool for subject-specific insights to predict caries risk [25]. Additionally, it facilitates the early identification of dental caries. The present study is one of the recent studies focusing on creating oral health algorithms and tools. These resources could serve as valuable tools for dentists, healthcare professionals, and policymakers for screening, program evaluation, oral health assessments and planning public health strategies [26–31]. The study by Sadegh-Zadeh *et al.* [25] highlighted the value of routinely assessing children's risk for dental caries through expert evaluations and determining individual caries risk scores. As a result, there is an increased focus on personalized prevention strategies to reduce the likelihood of developing dental caries by leveraging ML models. A recent study by Toledo *et al.* [32] created and validated models to predict the progression of dental caries in primary and permanent teeth after 2 and 10 years of monitoring using the ML approach with early childhood data. The study concluded that using ML techniques proved to be promising for predicting caries development in both types of teeth by relying on simple, early childhood predictors.

ML models highlight the contributing factors resulting in the development of dental caries in primary teeth, such as the frequent consumption of sugary foods and a negative parental perception of their child's oral health [33]. The XGBoost algorithm, an advanced machine learning technique, uses an ensemble of weak prediction trees combined in an additive manner, which addresses the challenges in fitting previous models [34]. However, complex ML models like XGBoost are limited as they can be difficult to interpret. On the other hand, logistic regression (LR) models offer straightforward effect estimates and model outputs that are easy for most practitioners to comprehend, which makes them more commonly used in everyday practice [35]. According to Lundberg and Lee [36], XGBoost with a Shapley Additive exPlanations (SHAP) based framework enhances the interpretability of machine learning models, simplifies the process of identifying key predictors and facilitates more informed decision-making.

Application based on AI will lower the workload of dentists from daily tasks, raising health care at less cost for every person, and ultimately assist predictive, personalized and preventive dentistry. Despite AI advancement, it has not been fully entered into dental practice, majorly due to (1) inadequate standard and methods development; (2) restricted accessibility, comprehensiveness and availability of data; and (3) less awareness of the usefulness of AI. Due to their powerful capabilities concerning the analysis of data, these algorithms are likely to enhance the efficacy and accuracy of dental detection, assisting with treatment with the help of anatomic visualization, acting and examining outcomes and projecting the development and prediction of oral diseases. In various areas of dentistry, solely the right diagnosis ensures the plan of treatment correctly, and that is the only method to restore the health of the patient. The identification and plan for treatment depend on the knowledge of specialists but there are chances of high and multifactorial errors. Hence, methods such as AI, ML and statistics are a great hope for patients and doctors.

However, there are certain limitations of this study. The study does not explore underlying behavioral or psychological factors influencing dental hygiene practices, such as infrequent brushing and lack of supervision. Moreover, socioeconomic status is highlighted as a factor, however, the specific barriers faced by lower-income families (*i.e.* access to education or financial constraints) are not explained. Additionally, the findings lack generalizability due to the geographic specificity of the population studied and the cross-sectional design limits the ability to establish causal relationships between identified factors and oral health outcomes. The machine learning models that showed high predictive accuracy were not validated on external datasets, which overestimates their real-world applicability. Furthermore, the study does not examine the complex relationship between demographic, behavioral and cultural factors or the direct impact of health insurance coverage on preventive care.

5. Study implications

The study results provide some clinical implementations to improve pediatric oral health; such as promoting regular dental check-ups and educating families about the importance of preventive dental care. Dental professionals can develop interventions to address gaps in oral hygiene practices, particularly among children from low socioeconomic backgrounds. The study's predictive models, like LightGBM and XGBoost, can assist in the early identification of children at risk for dental caries to enable targeted preventive measures. Collaboration with schools and community organizations could further facilitate health education and access to affordable dental care, ensuring broader reach and long-term oral health improvements.

Several measures can be implemented to improve the condition of patients such as regular dental check-ups, promoting proper oral hygiene practices and supervision during brushing, especially for younger children. Moreover, encouraging a balanced diet rich in nutrients and limiting sugary food and beverage consumption can promote oral health significantly and reduce the risk of dental caries. Parents should be advised to actively participate and foster healthy habits from a young age only. Educating parents about the importance of preventive care that includes explaining role of fluoride and regular dental visits is very important. Proper communication between parents and dental care providers can help address specific concerns and tailor interventions to individual needs.

6. Conclusions

In conclusion, this study delved into the realm of dental caries prediction among children using advanced machine-learning models. Through the utilization of Logistic Regression, Random Forest, XGBoost and LightGBM, we aimed to assess the efficacy of these algorithms in identifying early signs of dental caries. Based on the sample size of 500, the machine learning models (Logistic Regression, Random Forest, XGBoost and

LightGBM) appear to be effective in predicting dental caries among children. However, the choice of the "best" model may depend on specific priorities (*e.g.*, maximizing precision or recall). Further validation on external datasets is recommended to assess the generalizability of the models. The results revealed promising predictive capabilities across all four models. Notably, LightGBM exhibited the highest overall performance, achieving an accuracy of 85% and demonstrating superior sensitivity and specificity in distinguishing between children with and without dental caries. Random Forest and XGBoost also yielded commendable results with accuracy rates of 83% and 84%, respectively.

Furthermore, the study highlighted the importance of features such as socio-economic status, oral hygiene practices and parental involvement in oral care, as they significantly contributed to the predictive power of the models. In essence, this study contributes valuable insights into the application of machine learning for early detection of dental caries among children, emphasizing the need for ongoing research and collaborative efforts to enhance the precision and reliability of predictive models in pediatric oral health.

7. Future studies

Further studies in the field of dentistry and ML need to focus on advanced predictive models for various dental conditions, such as periodontal diseases, oral cancers and malocclusions. It is possible to enhance the diagnostic accuracy by investigating ML integration with imaging technologies, like AI-driven analysis of radiographs or Cone-beam computed tomography systems (CBCT) scans. Future studies also need to evaluate ML-based personalized treatment planning to leverage patient data for tailored dental care. Additionally, ML should be integrated in dentistry platforms for remote diagnostics and patient monitoring in underprivileged regions.

AVAILABILITY OF DATA AND MATERIALS

The datasets used and analyzed during the current study are available from the corresponding author (SAB) upon reasonable request.

AUTHOR CONTRIBUTIONS

SAB—conceptualization, methodology, data collection, analysis, writing original draft and review & editing.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study has been approved by Taibah University, College of Dentistry Research Ethics Committee (TUCDREC), reference number TUCDREC/27112020/SABahamam. This study was conducted in accordance with ethical guidelines, and informed consent was obtained from all participants prior to their participation. For minors or individuals unable to provide consent, consent was obtained from their parent or legal guardian.

ACKNOWLEDGMENT

The author is very thankful to all the associated personnel in any reference that contributed to/for this research.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] Chi DL, Scott JM. Added sugar and dental caries in children: a scientific update and future steps. Dental Clinics of North America. 2019; 63: 17–33.

[2] He S, Wang J. Validation of the Chinese version of the caries impacts and experiences questionnaire for children (CARIES-QC). International Journal of Paediatric Dentistry. 2020; 30: 50–56.

[3] Zou J, Du Q, Ge L, Wang J, Wang X, Li Y, *et al*. Expert consensus on early childhood caries management. International Journal of Oral Science. 2022; 14: 35.

[4] Laksmiastuti SR, Budiardjo SB, Sutadi H. Validated questionnaire of maternal attitude and knowledge for predicting caries risk in children: epidemiological study in North Jakarta, Indonesia. Journal of International Society of Preventive and Community Dentistry. 2017; 7: S42–S47.

[5] Bud ES, Bica CI, Stoica OE, Vlasa A, Eșian D, Bucur SM, *et al*. Observational study regarding the relationship between nutritional status, dental caries, mutans streptococci, and lactobacillus bacterial colonies. International Journal of Environmental Research and Public Health. 2021; 18: 3551.

[6] Pakkhesal M, Riyahi E, Naghavi Alhosseini A, Amdjadi P, Behnampour N. Impact of dental caries on oral health related quality of life among preschool children: perceptions of parents. BMC Oral Health. 2021; 21: 68.

[7] Reddy P, Krithikadatta J, Srinivasan V, Raghu S, Velumurugan N. Dental caries profile and associated risk factors among adolescent school children in an urban South-Indian city. Oral Health and Preventive Dentistry. 2020; 18: 379–386.

[8] Zabokova Bilbilova E. Dietary factors, salivary parameters, and dental caries. In Zabokova Bilbilova E (ed.) Dental caries (pp. 1–20). IntechOpen: London, UK. 2020.

[9] Weatherspoon DJ, Horowitz AM, Kleinman DV. Maryland physicians' knowledge, opinions, and practices related to dental caries etiology and prevention in children. Pediatric Dentistry Journal. 2016; 38: 61–67.

[10] Tadakamadla SK, Tadakamadla J, Kroon J, Lalloo R, Johnson NW. Effect of family characteristics on periodontal diseases in children and adolescents—a systematic review. International Journal of Dental Hygiene. 2020; 18: 3–16.

[11] Chaffee BW, Featherstone JD, Gansky SA, Cheng J, Zhan L. Caries risk assessment item importance: risk designation and caries status in children under age 6. JDR Clinical & Translational Research. 2016; 1: 131–142.

[12] Ghaffari M, Zhu Y, Shrestha A. A review of advancements of artificial intelligence in dentistry. Dentistry Review. 2024; 4: 100081.

[13] Nambiar R, Nanjundegowda R. A comprehensive review of ai and deep learning applications in dentistry: from image segmentation to treatment planning. Journal of Robotics and Control. 2024; 5: 1744–1752.

[14] Zhu J, Chen Z, Zhao J, Yu Y, Li X, Shi K, *et al*. Artificial intelligence in the diagnosis of dental diseases on panoramic radiographs: a preliminary study. BMC Oral Health. 2023; 23: 358.

[15] Schober P, Vetter TR. Logistic regression in medical research. Anesthesia & Analgesia. 2021; 132: 365–366.

[16] Abhishek Sharma. Decision tree vs random forest | which is right for you? 2020. Available at: https://www.analyticsvidhya.com/

blog/2020/05/decision-tree-vs-random-forest-algorithm/ (Accessed: 05 March 2025).

[17] Dong W, Huang Y, Lehane B, Ma G. XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. Automation in Construction. 2020; 114: 103155.

[18] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. Lightgbm: a highly efficient gradient-boosting decision tree. 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA, 04–09 December 2017. Curran Associates Inc.: Red Hook, NY, USA. 2017.

[19] Karhade DS, Roach J, Shrestha P, Simancas-Pallares MA, Ginnis J, Burk ZJS, *et al*. An automated machine learning classifier for early childhood caries. Pediatric Dentistry Journal. 2021; 43: 191–197.

[20] Hung M, Voss MW, Rosales MN, Li W, Su W, Xu J, *et al*. Application of machine learning for diagnostic prediction of root caries. Gerodontology. 2019; 36: 395–404.

[21] Kang IA, Ngnamsie Njimbouom S, Lee KO, Kim JD. DCP: prediction of dental caries using machine learning in personalized medicine. Applied Sciences. 2022; 12: 3043.

[22] Wu TT, Xiao J, Sohn MB, Fiscella KA, Gilbert C, Grier A, *et al*. Machine learning approach identified multi-platform factors for caries prediction in child-mother dyads. Frontiers in Cellular and Infection Microbiology. 2021; 11: 727630.

[23] Ngnamsie Njimbouom S, Lee K, Kim JD. MMDCP: multi-modal dental caries prediction for decision support system using deep learning. International Journal of Environmental Research and Public Health. 2022; 19: 10928.

[24] Ataş M, Yeşilnacar Mİ, Demir Yetiş A. Novel machine learning techniques based hybrid models (LR-KNN-ANN and SVM) in prediction of dental fluorosis in groundwater. Environmental Geochemistry and Health. 2022; 44: 3891–3905.

[25] Sadegh-Zadeh SA, Rahmani Qeranqayeh A, Benkhalifa E, Dyke D, Taylor L, Bagheri M. Dental caries risk assessment in children 5 years old and under via machine learning. Dentistry Journal. 2022; 10: 164.

[26] Marcus M, Maida CA, Wang Y, Xiong D, Hays RD, Coulter ID, *et al*. Child and parent demographic characteristics and oral health perceptions associated with clinically measured oral health. JDR Clinical & Translational Research. 2018; 3: 302–313.

[27] Liu H, Hays RD, Marcus M, Coulter I, Maida C, Ramos-Gomez F, *et al*. Patient-Reported oral health outcome measurement for children and adolescents. BMC Oral Health. 2016; 16: 95.

[28] Wang Y, Hays RD, Marcus M, Maida CA, Shen J, Xiong D, *et al*. Developing children's oral health assessment toolkits using machine learning algorithm. JDR Clinical & Translational Research. 2020; 5: 233–243.

[29] Maida CA, Marcus M, Hays RD, Coulter ID, Ramos-Gomez F, Lee SY, *et al*. Child and adolescent perceptions of oral health over the life course. Quality of Life Research. 2015; 24: 2739–2751.

[30] Maida CA, Marcus M, Hays RD, Coulter ID, Ramos-Gomez F, Lee SY, *et al*. Qualitative methods in the development of a parent survey of children's oral health status. Journal of Patient-Reported Outcomes. 2017; 2: 7.

[31] Marcus M, Xiong D, Wang Y, Maida CA, Hays RD, Coulter ID, *et al*. Development of toolkits for detecting dental caries and caries experience among children using self-report and parent report. Community Dentistry and Oral Epidemiology. 2019; 47: 520–527.

[32] Toledo Reyes L, Knorst JK, Ortiz FR, Brondani B, Emmanuelli B, Saraiva Guedes R, *et al*. Early childhood predictors for dental caries: a machine learning approach. Journal of Dental Research. 2023; 102: 999–1006.

[33] Feldens CA, Dos Santos IF, Kramer PF, Vítolo MR, Braga VS, Chaffee BW. Early-life patterns of sugar consumption and dental caries in the permanent teeth: a birth cohort study. Caries Research. 2021; 55: 505–514.

[34] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). 2016; 785–794.

[35] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. The New England Journal of Medicine. 2019; 380: 1347–1358.

[36] Lundberg S, Lee S. A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA, 04–09 December. Long Beach, CA, USA. 2017.

How to cite this article: Sarah Ahmed Bahammam. Prediction of dental caries in children through machine learning. Journal of Clinical Pediatric Dentistry. 2025; 49(5): 158-167. doi: 10.22514/jocpd.2025.110.