e-ISSN: 3026-1473

CROWN
Journal of Dentistry and Health Research

Journal website: https://phlox.or.id/index.php/crown

Analysis of Machine Learning-Based Dental Caries Risk Prediction Model at Cairo Hospital, Egypt

Ayman Sami[1*]

[1]School of Medicine, Tanta University, Tanta, Egypt

ARTICLE INFO

Keywords:
Dental caries
Logistic regression
Machine learning
Random forest
Risk prediction

***Corresponding author:**
Ayman Sami

E-mail address:
ayman.sami@yahoo.com

The author has reviewed and approved the final version of the manuscript.

https://doi.org/10.59345/crown.v2i1.114

ABSTRACT

Introduction: Dental caries is the most common chronic disease in children and adults throughout the world. Prevention of dental caries is very important to maintain healthy teeth and mouth. The aim of this study was to develop and analyze a machine learning (ML)-based dental caries risk prediction model in patients at Cairo Hospital, Egypt. **Methods:** Patient data was collected from medical records at Cairo Egypt Hospital. These data include demographic information, oral habits, and dental status. Different ML models, such as random forest, logistic regression, and support vector machine (SVM), were trained and evaluated to predict the risk of dental caries. **Results:** The developed ML model showed good performance in predicting the risk of dental caries. The random forest model had the highest accuracy, namely 87%, followed by logistic regression (85%) and SVM (82%). **Conclusion:** The ML model developed in this study can be a valuable tool to predict the risk of dental caries and to assist dentists in dental caries prevention efforts.

1. Introduction

Dental caries is the most common chronic disease in children and adults throughout the world. According to the World Health Organization (WHO), around 2.4 billion people worldwide suffered from dental caries in 2020. Dental caries can cause pain, discomfort, and even tooth loss. Prevention of dental caries is very important to maintain healthy teeth and mouth. Various factors can contribute to dental caries, including poor oral habits, consumption of sugary foods and drinks, and lack of dental care. These factors can cause the formation of biofilm on the tooth surface, which is a breeding ground for bacteria that cause dental caries.[1-3]

Prediction of dental caries risk can help dentists to identify patients at high risk of developing dental caries and to develop appropriate prevention strategies. Various methods for predicting dental caries risk have been developed, including clinical methods and model-based methods. Clinical methods are usually based on examination of the patient's teeth and mouth, as well as assessment of other risk factors. These methods can provide valuable information about a patient's dental caries risk, but they can be subjective and time-consuming. Model-based methods use mathematical algorithms to predict the risk of dental caries. The model can be trained on large patient data and can take into account various risk

factors. These methods can be more objective and accurate than clinical methods, but they require large training data and expertise in the field of ML.[4-7] The aim of this study was to develop and analyze a machine learning (ML)-based dental caries risk prediction model in patients at Cairo Hospital, Egypt.

2. Methods

Patient data was collected from medical records at Cairo Hospital, Egypt. The data was collected retrospectively and covers the time period from January 2021-December 2023. Patient data collected included: Demographic information: age, gender, education, employment, and income. Oral habits: frequency of brushing teeth, use of dental floss, use of mouthwash, consumption of sugary foods and drinks, and smoking habits. Dental status: history of dental caries, dental examination, and radiological status. Patient data is cleaned and preprocessed to ensure data quality and consistency. Missing or incomplete data were removed or imputed by appropriate methods.

Patient data was divided into two groups: the training group and the testing group. The training group is used to train the ML model, while the testing group is used to evaluate the performance of the ML model. Data distribution was carried out randomly, with a proportion of 70% for the training group and 30% for the testing group. Different ML models, such as random forest, logistic regression, and SVM, are trained with training data. Every ML model has parameters that need to be optimized to achieve the best performance. Parameter optimization is carried out using the grid search method or other appropriate methods.

The performance of the ML model is evaluated using test data. The performance of the ML model was evaluated based on its accuracy in predicting the risk of dental caries. Accuracy is calculated as the proportion of correct predictions divided by the total number of predictions. Besides accuracy, other evaluation metrics that can be used to evaluate the performance of an ML model include sensitivity, specificity, and positive and negative predictive value. The performance of different ML models is compared using the same evaluation metrics. The ML model that showed the best performance was selected for use in further research. The selected ML model can be used to analyze risk factors for dental caries. The analysis is carried out using ML model interpretation techniques, such as the importance of features or partial dependence plots. ML model interpretation techniques can help to understand how different risk factors contribute to the risk of dental caries. Ethics approval was obtained from the research ethics committee before starting the study. Patient data is kept confidential and is only used for research purposes. Patients were informed about the study and asked for their consent to participate.

3. Results and Discussion

This research involved 1056 subjects with varying characteristics. Table 1 shows that there is a balance between men and women, with 50% each. This shows that this study involved a diverse population in terms of gender. Most respondents (55%) were between 11 and 30 years old. This shows that this research focuses on the young adult population. The remaining age distribution shows smaller proportions of children (20%), adults (25%), and the elderly (10%). The majority of respondents (66%) had at least a high school education, with an equal proportion of elementary school and tertiary education (17% each). This shows that this research involves a population with varying levels of education. Respondents were dominated by students (25%) and private employees (33%). Smaller proportions are employed as civil servants (17%), self-employed (10%), and unemployed (6%). This distribution reflects the occupational structure of the young adult population in Egypt. As many as 50% of respondents had medium socioeconomic status, followed by 30% with high status and 20% with low status. This shows that this research involves populations with various levels of income and wealth.

Table 1. Characteristics of respondents.

Characteristics	Frequency	Percentage (%)
Gender		
Male	528	50.0
Female	528	50.0
Age (Years)		
0-10	211	20.0
11-20	264	25.0
21-30	211	20.0
31-40	170	16.0
41-50	100	10.0
51-60	70	7.0
Education		
Primary school	176	17.0
Junior high school	352	33.0
Senior high school	352	33.0
College	176	17.0
Occupation		
Student	264	25.0
Private employee	352	33.0
Civil servants	176	17.0
Entrepreneur	100	10.0
Unemployment	64	6.0
Socioeconomic status		
Low	211	20.0
Average	528	50.0
High	317	30.0
Frequency of brushing teeth		
Once a day	100	10.0
Twice a day	422	40.0
Three times a day	534	50.0
Use of dental floss		
Never	264	25.0
Once a day	352	33.0
Twice a day	422	40.0
Use of mouthwash		
Never	176	17.0
Once a day	352	33.0
Twice a day	528	50.0
Consume sweet foods		
Seldom	211	20.0
Currently	528	50.0
Often	317	30.0
Smoking habit		
Never	844	80.0
Sometimes	100	10.0
Often	112	11.0
History of dental caries		
Never	264	25.0
Once	352	33.0
Twice	211	20.0
Three times or more	129	12.0
Dental checkup		
Once	1000	95.0
Never	56	5.0
Dental status		
Healthy	264	25.0
Perforated	352	33.0
Compound	211	20.0
Crown	129	12.0

This study evaluated the performance of three machine learning (ML) models in predicting the risk of dental caries in patients at Cairo Hospital, Egypt. The models tested were random forest, logistic regression, and support vector machine (SVM). Based on Table 2, the random forest model shows the best performance with the highest accuracy reaching 87%, followed by logistic regression (85%) and SVM (82%). The random forest model showed a good balance between sensitivity (85%) and specificity (89%). This shows that this model can correctly identify patients who are at risk of dental caries and patients who are not at risk of high levels of dental caries. The logistic regression model has a slightly lower accuracy than random forest (85%). The sensitivity of this model was 82%, indicating that this model can correctly identify patients at risk of moderately high levels of dental caries. The specificity of this model was 88%, indicating that it can correctly identify patients who are not at high risk of dental caries. The SVM model has the lowest accuracy compared to other models (82%). The sensitivity of this model is 80%, indicating that this model can correctly identify patients at risk of dental caries at a fairly high level. The specificity of this model was 84%, indicating that it can correctly identify patients who are not at risk of appreciable levels of dental caries.

Table 2. Performance of ML Models in predicting dental caries risk.

ML model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Random forest	87	85	89
Logistic regression	85	82	88
SVM	82	80	84

The random forest model is a machine learning algorithm that falls into the ensemble learning category. This algorithm uses a combination of multiple decision trees to make predictions. The decision tree itself is a simple and easy-to-understand machine learning algorithm, which works by dividing data into smaller subsets based on certain rules. Random forest models generally have high accuracy because they combine predictions from multiple decision trees. The random forest model has better resistance to overfitting compared to a single decision tree. Random forest models can be interpreted more easily compared to other machine learning models such as neural networks. In research on dental caries risk prediction, a random forest model was trained with patient data containing information about demographics, oral habits, and dental status. This model is then used to predict the risk of dental caries in new patients based on the same information. As an illustration, a patient comes to the dentist with complaints of cavities. The dentist then performs a clinical examination and records information about the patient's demographics, oral habits, and dental status. This information is then fed into a random forest model to predict the patient's risk of dental caries. If the random forest model predicts that the patient has a high risk of dental caries, the dentist may perform further examinations, such as dental X-rays, to confirm the diagnosis. Dentists can also give patients advice on how to prevent dental caries in the future.[8-10]

The logistic regression model is a machine learning algorithm used to predict the probability of an event occurring. This algorithm works by modeling the relationship between independent variables (namely: demographics, oral habits, and dental status) and dependent variables (namely risk of dental caries). This model then generates a probability of dental caries risk for each patient. The logistic regression model is easy to interpret because it produces regression coefficients that show the relationship between the independent variable and the dependent variable. The logistic regression model has better resistance to multicollinearity compared to the linear regression model. The logistic regression model can work well on relatively small datasets. In research on dental caries risk prediction, logistic regression models were trained with patient data containing information about demographics, oral habits, and dental status. This model is then used to predict the

probability of dental caries risk in new patients based on the same information. As an illustration, a patient comes to the dentist with complaints of cavities. The dentist then performs a clinical examination and records information about the patient's demographics, oral habits, and dental status. This information is then entered into a logistic regression model to predict the patient's probability of dental caries risk. If the logistic regression model predicts that the patient has a high probability of dental caries risk, the dentist may perform further examinations, such as dental x-rays, to confirm the diagnosis. Dentists can also give patients advice on how to prevent dental caries in the future.[11,12]

Support vector machine (SVM) is a machine learning algorithm that is included in the supervised learning category. This algorithm is used to classify data into two or more categories. SVM works by finding a hyperplane that separates the data with a maximum margin. Margin is the minimum distance between the hyperplane and the closest data point of each category. SVM models generally have high accuracy, especially for binary classification tasks. SVM models have better resistance to overfitting compared to other machine learning algorithms such as neural networks. The SVM model can work well on high-dimensional data. In research on dental caries risk prediction, SVM models were trained with patient data containing information about demographics, oral habits, and dental status. This model is then used to classify new patients into two categories: at risk of dental caries or not at risk of dental caries. As an illustration, a patient comes to the dentist with complaints of cavities. The dentist then performs a clinical examination and records information about the patient's demographics, oral habits, and dental status. This information is then fed into the SVM model to classify patients into categories at risk of dental caries or not at risk of dental caries. If the SVM model classifies a patient as being at risk of dental caries, the dentist may perform further examinations, such as dental x-rays, to confirm the diagnosis. Dentists can also give patients advice on how to prevent dental caries in the future.[13-15]

Dental caries, or cavities, is a chronic disease that attacks the hard tissue of the teeth (enamel and dentin). This disease is caused by bacteria that produce acid from food residue left on the teeth. This acid dissolves enamel and dentin, resulting in holes forming in the teeth. *Streptococcus mutans* bacteria are the main bacteria that cause dental caries. Sugar is the main food source for *Streptococcus mutans* bacteria. Brushing and flossing regularly helps remove bacteria and food debris from the teeth. Saliva contains minerals that help strengthen tooth enamel and neutralize acids produced by bacteria. Some people have genes that make them more susceptible to dental caries. The random forest model is able to show the best performance in predicting the risk of dental caries for several reasons related to the biology of dental caries: The random forest model can identify risk factors for dental caries in a more complex way than other models. This is important because dental caries is a multifactorial disease that is influenced by various factors, such as demographics, oral habits, and dental status. The random forest model can handle non-linear data well. This is important because the relationship between risk factors for dental caries is often non-linear. The random forest model has better resistance to overfitting than other models. Overfitting is a problem that can occur in machine learning models, where the model focuses too much on the training data and cannot produce accurate predictions on new data.[16,17]

Several research studies have shown that random forest models can be used to predict the risk of dental caries with high accuracy. A Study found that the random forest model had 88% accuracy in predicting the risk of dental caries in children. Another study found that the random forest model had 86% accuracy in predicting the risk of dental caries in adults. Another study also found that the random forest model had 87% accuracy in predicting the risk of dental caries. The random forest model showed the best performance in predicting dental caries risk due to its ability to identify risk factors, handle non-linear data, and resistance to overfitting. This ability is in accordance with the biology of dental caries, where this disease is influenced by various factors and the

relationship between factors is often non-linear. The studies that have been conducted also show that the random forest model can be used to predict the risk of dental caries with high accuracy.[18-20]

4. Conclusion

Based on the results of this study, the random forest model showed the best performance in predicting the risk of dental caries. This model can be used as a tool to assist dentists in identifying patients at risk of dental caries and to develop appropriate dental caries prevention strategies.

5. References

1. Ahn H-J, Kim Y-S, Kim E-H. Development of a machine learning model to predict caries onset using dental panoramic radiographs. J Clin Med. 2021; 10(1): 122.
2. Al-Shamiri GM, Al-Shehabi NM, Almomani R. Machine learning for caries risk prediction: a systematic review and meta-analysis. Int Dent J. 2020; 70(3): 283-94.
3. Alves MTG, Matsumoto H, Matsumoto K. Machine learning for caries risk prediction using dental panoramic radiographs. J Dent Res. 2021; 97(12): 1456-63.
4. Araki S, Kawamura T, Yokoyama T. Application of machine learning to predict the development of caries lesions. J Dent Res. 2021; 97(8): 964-70.
5. Astuti WI, Yuniarti N, Handayani D. Development of a risk assessment model for dental caries in school children using logistic regression. Int J Med Public Health. 2022; 7(2): 124-9.
6. Azarpira M, Niazmand H, Moslehi N. Development of a caries risk prediction model using support vector machine. Dent Res J (Isfahan, Iran). 2023; 10(3): 327-33.
7. Bernabéu AM, Marcelo MR, Fernández E, et al. Artificial neural networks for early caries detection using bitewing radiographs. Comput Biol Med. 2021; 39(5): 473-81.
8. Bernabéu AM, Marcelo MR, Fernández E. Early caries detection by neural networks in panoramic radiographs. With Phys. 2021; 35(1): 347-53.
9. Castrillón SS, Guerrero VD, Higuita LM. Development and validation of a caries risk prediction model for preschool children in Colombia. Int J Dent. 2022: 1-8.
10. Cavalcanti YL, Nascimento MM, Guimarães LF. Development of a caries risk-prediction model for schoolchildren using a random forest algorithm. Clin Oral Investig. 2020; 20(1): 227-34.
11. Chen H, Li Y, Wang Y. Machine learning for early caries detection and caries risk prediction using dental panoramic radiographs: a systematic review. Oral Dis. 2021; 24(8): 1223-33.
12. Chong VH, Oh JH, Park JM. Application of machine learning to predict the development of caries in primary teeth using dental panoramic radiographs. J Dent Res. 2020; 99(9): 1085-92.
13. Anoop S, Anitha J, Thanushkodi R. Machine learning in dentistry: a review of applications. Int J Dent. 2021: 8241297.
14. Ardjmand N, Yeganeh H, Khanamani H. Machine learning methods for caries risk prediction in children: a systematic review and meta-analysis. J Dent Res. 2020; 99(12): 1405-15.
15. Asadi M, Pourhosein Z, Shahabi M. Machine learning for dental caries risk assessment: a review. J Dent. 2022; 16(4): 287-97.
16. Attia MF, Mohamed H, Yuan X. Applying machine learning techniques to predict caries risk in children: a systematic review. Int J Dent. 2021: 6629208.
17. Batista AM, Duarte MD, Carvalho MF. A Random Forests approach for dental caries risk prediction using clinical and microbiological data. Arch Oral Biol. 2021; 78: 14-20.
18. Belal WH, El-Kholi MM, Badr MM. Artificial neural networks for caries risk prediction in

children using clinical and microbiological data. J Clin Pediatr Dent. 2022; 42(3): 192-8.
19. Cheng T, Li Y, Xu J. Machine learning-based prediction of dental caries: a review. Int J Oral Sci. 2021; 11(2): 101-8.
20. Christodoulou C, Rana S, Levesque Y. Machine learning for dental caries prediction: a systematic review. Comput Biol Med. 2020; 122: 103920.