

# A MACHINE LEARNING SCREENING MODEL FOR PREDICTING THE DEVELOPMENT OF CERVICAL DENTAL LESIONS

Iryna I. Zabolotna<sup>1</sup>, Tatiana L. Bogdanova<sup>2</sup>, Volodymyr I. Azarenkov<sup>3</sup>, Olena S. Genzytska<sup>1</sup>, Andrii A. Komlev<sup>1</sup>

<sup>1</sup>Donetsk National Medical University, Kropyvnytskyi, Ukraine

<sup>2</sup>National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine

<sup>3</sup>National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine

## Abstract

**Introduction.** Predicting the risk of development of cervical dental pathology is a difficult task due to the multifactorial nature of its etiopathogenesis and limited knowledge of risk factors.

**Aim.** To develop and test a computer model for predicting the development of cervical dental lesions in young patients.

**Materials and methods.** The survey consisted of 272 patients (mean age  $24.3 \pm 6.9$  years), in whom risk factors for the development of a wedge-shaped defect, cervical caries and enamel erosion were determined, which became the input data for the computer model. The Extreme Gradient Boosting (XGBoost) tree-based machine learning method implemented in the Python programming language using the scikit-learn and XGBoost libraries was used. Synthetic Minority Over-sampling Technique (SMOTE) was additionally applied to increase the efficiency of predicting less common enamel erosion among the examined individuals.

**Results.** When developing the models, the priority was given to recall over accuracy and specificity. This contributed to reducing the number of missed cases for each pathology. The highest discriminatory ability (ROC-AUC) = 0.84 (Receiver Operating Characteristic curve – Area under the curve) in combination with a high level of recall (recall = 0.82) corresponded to the model for predicting cervical caries of teeth. This confirmed the feasibility of using the XGBoost algorithm to identify complex relationships in nonlinear combinations of the indicators. The model for predicting a wedge-shaped defect of teeth also had high recall (recall = 0.83) but the moderate value of ROC-AUC (0.64) that emphasizes the presence of nonlinear dependent predictors. Particular scientific interest has the model for predicting erosion of tooth enamel which was created under conditions of a low prevalence of pathology among the examined. However, the results showed an acceptable level of recall (recall = 0.47) and moderate discriminatory ability (ROC-AUC = 0.72). This allowed us to determine that the problem of small sample was successfully solved.

**Conclusions.** The presented machine learning screening model helps identify patients with increased risk of developing cervical dental lesions. Its use will make it possible to justify the prescription of preventive measures to young patients.

**Keywords:** dental caries, dental erosion, ensemble learning, prevention, risk factors

## INTRODUCTION

Timely diagnostic and preventive measures contribute to the improvement of the prediction of the prevalence of cervical dental pathology and determine the conservative or surgical tactics of their treatment [1, 2]. And if the intervention is started with younger age groups, this will allow for a reduction in the incidence at an older age [2]. Therefore, the relevance of early diagnosis is connected with preserving the integrity of hard tissues

and, accordingly, prolonging tooth functioning and life quality [3]. One of the ways to solve this issue is to perform patients' preventive examination while screening. In this regard, it is necessary to have a clear idea of the risk factors (RFs) of the development of cervical dental lesions and the share of each of them [4]. But today most of the information is of a generalized nature that is not enough for their practical implementation. So, the diagnosis of each type of lesion is relevant, namely cervical caries (CC), wedge-shaped defect (WSD) and

erosion (E) of tooth enamel in order to distinguish RFs in populations. Further influence on them will contribute to increasing the effectiveness of preventive measures at the individual level [5, 6].

As a result of the studies conducted by the authors, the RFs of the development of WSD, CC, and E of tooth enamel in young patients were determined [7]. Some of the predictors were modifiable, allowing changes in them to be compared during subsequent examinations [8]. This makes it possible to observe their dynamics, assess the effectiveness of the prescribed measures, and, if necessary, make corrections.

Nowadays, when using mathematical modeling, the assessment of the risks of the development of various pathologies, including dental ones, becomes more objective and scientifically justified [1, 2]. Computer technologies are increasingly used for this purpose [9, 10]. Such an approach can significantly simplify and accelerate the process of prediction, and can become the basis for the development of individual prevention plans. However, existing methods need to be improved [8]. In addition, most of the known mathematical models relate to the prediction of dental caries in children [8, 10]. They can accurately identify patients with a high risk of caries but they have shown insufficient effectiveness for patients with a low risk of developing dental pathology [11].

Logistic regression analysis is more often used to predict non-carious cervical dental lesions [12, 13]. Alternatively, the ensemble model allows for an increase in the diagnostic accuracy based on medical imaging [3, 9, 14]. However, to date, there are no mathematical models for predicting the development of cervical dental lesions in young patients.

### AIM

To develop and test a computer model for predicting the development of cervical dental lesions (WSD, CC, and E of enamel) in young patients in order to promptly identify patients with a high risk and increase the effectiveness of primary preventive measures.

### MATERIALS AND METHODS

The survey consisted of 272 patients (174 women and 98 men) aged 18-44 years (mean age  $24.3 \pm 6.9$  years) of different social groups and professions. The selection criteria included young age according to the WHO classification (2016), the absence of alcohol and drug addiction, neoplasms, tuberculosis, HIV/AIDS, hepatitis C, mental disorders, pregnancy, lactation period, occupational hazards. Among the examined patients, 60 were diagnosed with WSD, 50 with CC and 15 with E of tooth enamel. More than 120 predictors of cervical dental lesions were identified in the patients [7]. Later, twenty-seven RFs that had a statistically proven significant effect

on the occurrence of at least one of the diagnosed cervical dental pathologies were used to develop a computer prediction model. Extreme Gradient Boosting (XGBoost) tree-based machine learning method was chosen as the main approach. It was implemented in the Python programming language (version 3.9.12, [www.python.org](http://www.python.org)) using the scikit-learn and XGBoost libraries [12].

For E of tooth enamel, which is less common among the examined patients, the balanced sampling method Synthetic Minority Over-sampling Technique (SMOTE) was applied to compensate class imbalance and increase the prediction efficiency. It consists of an artificial increase in the number of cases. Together with XGBoost, this increases the generalization ability due to the gradual correction of errors of early trees and optimization of loss functions [15]. More detailed characteristics of the developed statistical computer model were given in a previous publication [12].

In the process of training the model, the probabilities of developing cervical dental lesions were calculated for each factor and threshold values were determined for classification into risk groups (high, medium, low). The Youden criterion and Receiver Operating Characteristic curve (ROC) analysis were used for that purpose. This was due to the fact that when creating a screening model, the priority belongs to the threshold value at which the difference between recall (True Positive Rate) and the proportion of false positive predictions (False Positive Rate) is the largest. While predicting that approach allowed the detection the maximum possible number of patients even with a low risk of dental pathology and determined the choice of the classification threshold. The latter for all three types of cervical dental lesions was 0.010 and it was significantly lower than the standard value of 0.5. It is such a low threshold value that can make the proposed model an effective screening tool and minimize the omission of patients with suspicion of the risk of developing pathology even with an increase in the number of false positive predictions.

The effectiveness of the model was assessed by Stratified K-Fold Cross-Validation ( $n=5$ ) with the calculation of ROC-AUC indicators (Area Under the Curve), accuracy, and recall, as well as the construction of confusion matrices and the analysis of feature importance.

### RESULTS

#### *Modeling the prediction of the development of CC of teeth*

In order to predict the development of CC of teeth, a tree-based machine learning model was created based on the XGBoost algorithm without using class balancing. The assessment of the predictive characteristics of the model was carried out using Stratified K-Fold Cross-Validation ( $n = 5$ ) with the obtaining of out-of-fold

predicted probabilities. Based on ROC-analysis using the criterion of maximizing the difference between recall and the proportion of false positive predictions (the Youden criterion), the optimal classification threshold was determined. That threshold was necessary to balance the

ratio between patients with a risk of development of CC of teeth and to limit the number of false predictions. The effectiveness of the model was evaluated by the ROC-AUC and recall indicators that were obtained on a training sample of 272 patients (Table 1).

Table 1

Quality Metrics of the XGBoost Model for Predicting CC of Teeth

Metric name	Metric value
Threshold	0.010
Accuracy	0.651
Precision	0.323
Recall (Sensitivity)	0.820
F1-score	0.463
ROC-AUC	0.8412

The determined recall of the model (recall = 0.82) showed its high ability to detect patients with an existing risk of CC of teeth. This confirms the ability of the model to correctly identify the vast majority of patients with an existing minimal risk of developing CC of teeth which is fundamentally important for screening models of primary diagnosis.

The obtained relatively low precision (precision = 0.32) is expected within the screening approach. This is a consequence of minimizing the number of missed cases (false negatives) by increasing the frequency of false positive predictions (false positives). The chosen strategy is clinically appropriate at the stage of assessing RFs of CC of teeth because it prevents missing patients

who require further observation or the implementation of therapeutic and preventive measures.

The balanced nature of the model is confirmed by the value of F1-score = 0.46 which reflects a compromise between recall and classification accuracy. At the same time, the high value of ROC-AUC = 0.84 emphasizes its good discriminative ability in distinguishing patients with a minimal risk of developing CC of teeth from those without it.

Fig. 1A presents a confusion matrix showing the ratio between true positive, false positive, true negative and false negative predictions. It also confirms the effectiveness of the proposed prediction model as a primary screening tool.

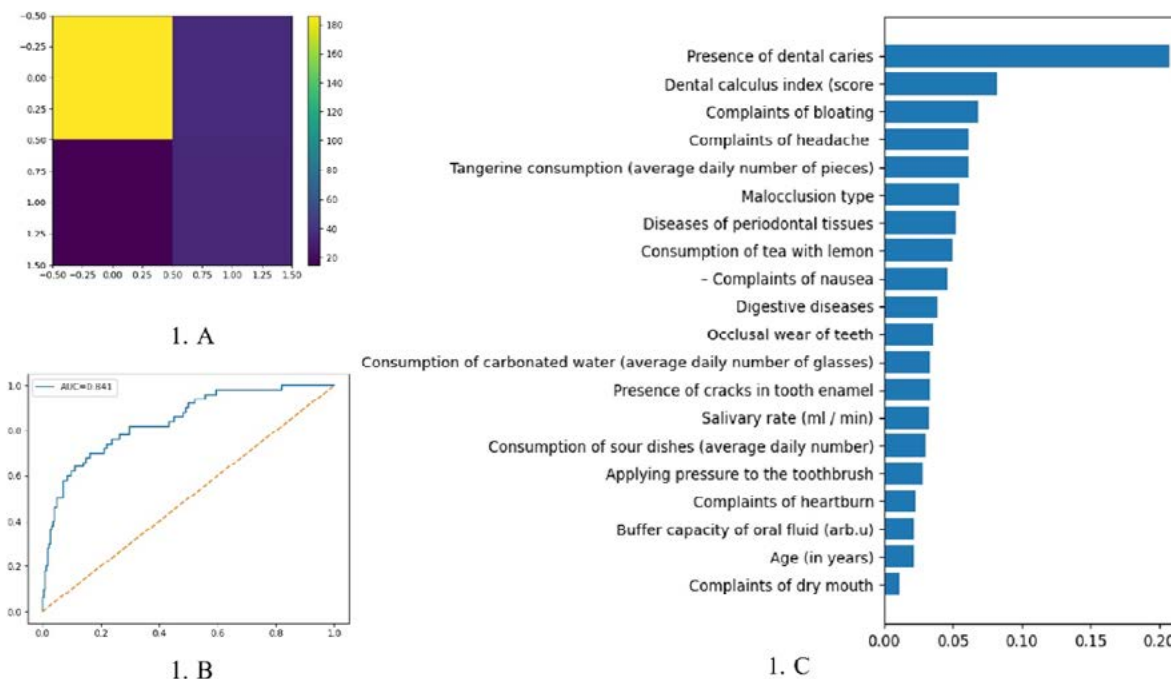


Figure 1. The characteristics of the model for CC of teeth: A. Confusion matrix; B. ROC curve; C. Gradation of predictors.

The ROC curve was made in order to evaluate the ability of the model to distribute patients with a risk of CC of teeth. As can be seen from Fig. 1B, the model demonstrated a fairly good discriminatory ability. The

area under the ROC curve is AUC = 0.841, which confirms the high probability of correct distribution of patients into risk groups of the development of CC of teeth and patients who do not belong to any risk group,

i.e., without any signs of pathology. The ROC curve by its shape significantly deviates from the diagonal line. This indicates the advantage of the model over random prediction in a fairly wide range of threshold values and confirms its ability to maintain high recall even when the classification threshold is changed. Thus, the screening orientation of the model of CC of teeth is emphasized.

The calculated AUC value also allows classifying the model as a qualitative one for the initial detection of patients with a risk of the development of CC of teeth. Such patients require further in-depth examination, an increase in the number of preventive examinations and/or a decrease in the interval between them. The presented algorithm also allows for assessing the significance of all considered RFs. Thus, a history of caries, a dental calculus index value greater than 0.6 points, and complaints of bloating and frequent headaches were among the most significant predictors of the development of CC of teeth (Fig. 1C). In general, the model provides reliable identification of patients with an existing risk of developing this dental pathology, it is characterized

by sufficient discriminatory ability and allows effective distinction patients of the risk group from those without it.

#### *Modeling the prediction of the development of WSD of teeth*

A screening tree-based classification model was also created to predict the development of WSD of teeth. As for CC of teeth, the XGBoost algorithm was used for this purpose. A single approach to training and evaluating the models was proposed for both cervical dental pathologies. The emphasis was placed on minimizing the omission of patients with a risk of their development. The use of a classification threshold at the level of 0.010 made it possible to increase recall of the models for identifying patients of the risk group. At the same time, the precision indicators were reduced expectedly. But in general, this is a characteristic feature of screening systems. According to the results of testing, the model showed high recall = 0.83 in identifying patients with a risk of developing WSD of teeth (Table 2). This confirms its ability to correctly identify the majority of patients with an increased risk of developing this cervical dental pathology.

Table 2

**Quality Metrics of the XGBoost Model for Predicting WSD of Teeth**

Metric name	Metric value
Threshold	0.010
Accuracy	0.437
Precision	0.259
Recall (Sensitivity)	0.833
F1-score	0.395
ROC-AUC	0.636

The obtained relatively low values of precision (precision = 0.26) and overall classification accuracy (accuracy = 0.44) indicate the presence of a significant number of false positive predictions. But, this is an expected consequence of using a screening strategy where recall is the priority over specificity. Therefore, such indicators do not reduce the clinical significance of the model since false positive results can be clarified during further additional examination of a patient.

The value of F1-score = 0.40 indicates a moderately balanced efficiency of the model. At the same time, ROC-AUC = 0.64 indicates a moderate discriminatory ability to distribute patients into groups with a minimal risk of WSD of teeth and patients without any signs of pathology. All the obtained indicators are acceptable for primary screening models.

Confusion matrix (Fig. 2A) illustrates the distribution of predictions and real values and reflects the relationship between clinical data and prediction results for the presence of WSD of teeth. The model correctly classified 67 patients without cervical dental lesions, but 145 patients without WSD of teeth were classified as a risk group. This indicates low specificity of the model.

The indicated number of false positive results is also expected and acceptable within the framework of the screening approach.

As for patients with clinically diagnosed WSD of teeth, the model correctly identified 51 cases, missing only 9 patients. This confirms the high recall of the algorithm, and it is consistent with the obtained recall value. Thus, the confusion matrix demonstrated a clearly expressed priority of the model in favor of identifying the maximum number of patients with probable pathology at the expense of reducing the classification accuracy of patients without cervical dental lesions.

Fig. 2B presents the ROC curve which clearly demonstrates the ability of the model to distinguish patients with a minimal risk of developing WSD and without cervical dental pathology. The area under the ROC curve is equal to AUC = 0.636 that fully corresponds to the moderate discriminatory ability of the model. At the same time, the AUC value exceeds 0.5. This indicates that the model provides better classification quality compared to random guessing, but the result is insufficient and further optimization of the set of RFs or expansion of the amount of the training sample seems advisable.

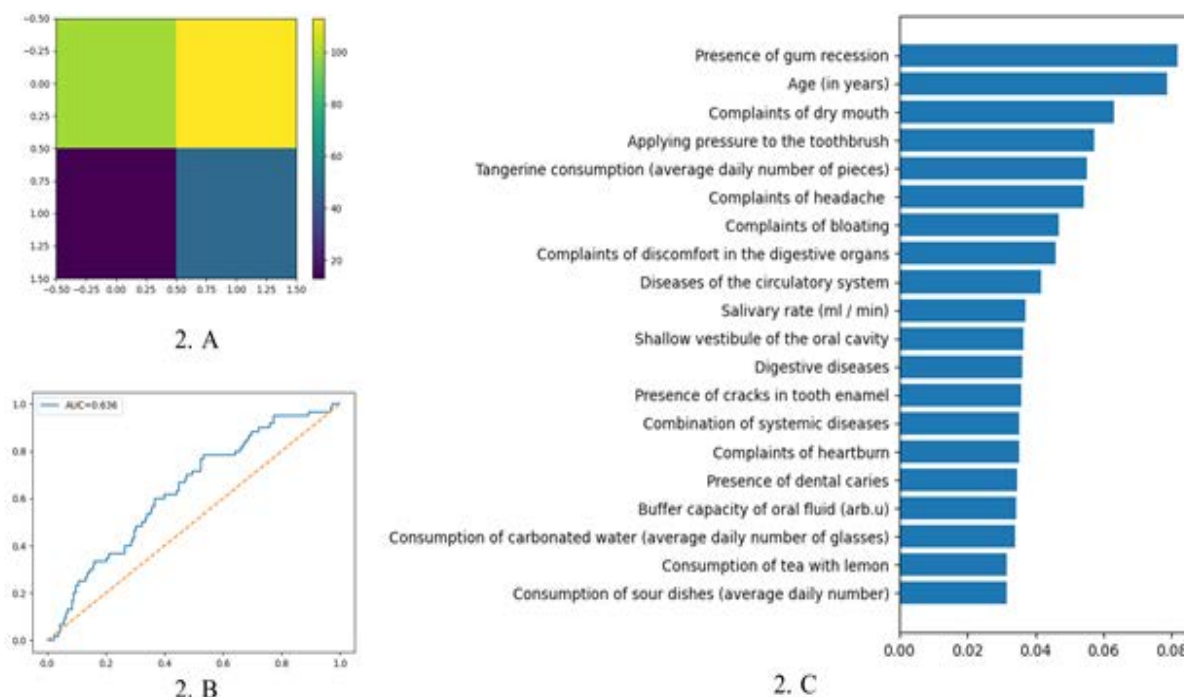


Figure 2. Model characteristics for WSD of teeth: A. Confusion matrix; B. ROC curve; C. Gradation of predictors.

The ROC curve does not approach the ideal upper limit of the graph either. This emphasizes the complexity of the prediction task, possible influence of class imbalance in the sample, the presence of linear and nonlinear relationships between the predictors. But high recall of the model at low threshold values is consistent with its screening purpose.

Fig. 2C shows the distribution of the significance of RFs in the model for predicting WSD of teeth. Among them, the most significant variables were the patient's diagnosed gum recession, age over 25 years, and the presence of complaints of dry mouth. The calculated results are generally consistent with existing ideas about the cumulative nature of the effect of etiopathogenetic factors on hard dental tissues and the role of a decrease in the salivary rate in the development of cervical dental pathology. According to the modeling results, the patient's dietary preference, such as eating more than two tangerines per day was attributed to the most significant predictor as well. Thus, the proposed machine learning model confirmed the multifactorial etiopathogenesis of WSD of teeth and clearly presented a quantitative assessment of the contribution of specific RFs to its development. The obtained calculated limited classification accuracy and overall predictive ability of the

model do not reduce the clinical significance, and they are sufficient for practical application.

#### *Modeling the prediction of the development of E of tooth enamel*

To predict the development of E of tooth enamel, a screening tree-based classification model was also created using the XGBoost algorithm. However, it methodologically differs from the prediction models of WSD and CC of teeth. This is due to the existence of a pronounced imbalance of classes and the specifics of the prevalence of this pathology in the presented sample (15 cases of E of tooth enamel among 272 examined patients). Therefore, the SMOTE method was additionally applied, which allowed for reducing the impact of data imbalance and ensured training of the classifier in conditions of a limited sample. The priority of maximizing recall and minimizing omissions of patients with initial manifestations of cervical dental pathology was chosen as the main criterion for evaluating the quality of the model. The working classification threshold was determined by the Youden criterion based on ROC analysis, and it was set at 0.01. The use of such a threshold, together with sample balancing, contributed to an increase in recall of the model, but at the same time, as expected, it reduced its accuracy and balance of the classification (Table 3).

Table 3

#### Quality Metrics of the XGBoost Model for Predicting E of Tooth Enamel

Metric name	Metric value
Accuracy	0.673
Precision	0.080
Recall (Sensitivity)	0.467
F1-score	0.136
ROC-AUC	0.716

The calculated quality indicators of the model confirm its high recall in identifying patients with a minimal risk of developing E of tooth enamel. Thus, the value of recall = 0.467 indicates the ability to correctly identify almost half of the patients with existing manifestations of pathology in the training sample. As with previous models, this is a key requirement for screening tools for early detection of dental pathology.

At the same time, the values of precision (precision = 0.08) and overall classification accuracy (accuracy = 0.673) emphasize the existence of a certain number of false positive predictions. This is an expected consequence of the simultaneous use of the SMOTE method and the chosen priority of recall over specificity. However, this approach does not reduce the clinical significance of the model, as false positive results can be clarified during the patient's further examinations.

The obtained F1-score value = 0.136 confirms the low balance between recall and accuracy of the model. When it was developed, that was an expected result in the conditions of a low prevalence of E of tooth enamel in the training sample and an absolute priority of recall of the model over other characteristics. At the same time, ROC-AUC = 0.716 shows a moderate discriminatory ability of the developed algorithm, which allows for distinguishing patients with a risk of E of tooth enamel from those without any signs of damage.

The discrepancy matrix (Fig. 3A) shows that the model correctly identified 6 patients but missed 9 cases among 15 patients with E of tooth enamel. Among 257 patients without any signs of E of tooth enamel, 176 patients were correctly

attributed to the group without any risk while 81 patients were falsely attributed to the risk group.

The ROC curve (Fig. 3B) demonstrates that the model for E of tooth enamel significantly exceeds the efficiency of random prediction. ROC-AUC = 0.716 indicates a moderate discriminatory ability of the algorithm. This phenomenon can be explained by the small number of positive cases in the sample. The curve also confirms high recall of the model at low decision-making thresholds that corresponds to the screening purpose and the focus on minimizing the omission of patients with initial manifestations of dental pathology.

The distribution of the predictors by informativeness showed that the most significant risk signs of E of tooth enamel include the presence of a combination of systemic diseases in the patient, a shallow vestibule of the oral cavity, occlusal wear of the teeth, complaints of frequent headaches (Fig. 3C). This additionally confirms the multifactorial etiopathogenesis of E of tooth enamel in young patients and justifies a differential approach to the prescription of therapeutic and preventive measures.

Thus, as a result of class balancing with the help of the SMOTE method and reducing the classification threshold, the developed model provided high recall. This allowed patients to be distributed into risk groups in conditions of low prevalence of dental pathology. The determined moderate discriminatory ability (ROC-AUC = 0.716) and the nature of the distribution of false positive predictions generally correspond to the screening purpose of the model and allow it to be recommended for use in practical dentistry.

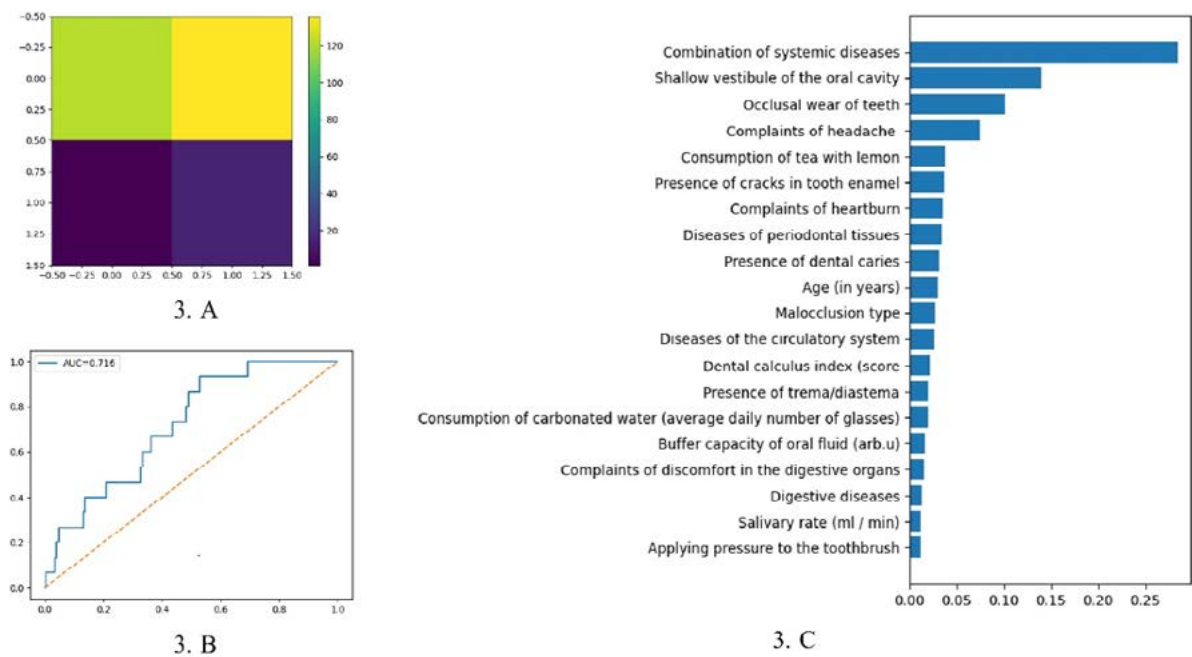


Figure 3. Model characteristics for E of tooth enamel: A. Confusion matrix; B. ROC curve; C. Gradation of predictors.

### Practical application in new patients

To check the capability of the model to work with new patients, the study provided for the introduction of new data into the Excel template, which is processed by the previously obtained code. That is, a single Python program code is used which combines the training of machine models for predicting cervical dental lesions in a sample of 272 patients and their application to assess the risk of pathology

development in new patients. After preparing databases and constructing the models, the code automatically saves the algorithm and threshold values. This allows the prediction process to be performed in the future without retraining. As can be seen in Fig. 4, the program calculates the probability of development of each type of cervical dental lesion separately, generates a percentage risk indicator and its categorical interpretation («low», «medium», «high»).

Probability of WSD development	Modeled WSD risk	Interpretation of WSD risk	Probability of CC development	Modeled CC risk	Interpretation of CC risk	Probability of E development	Modeled E risk	Interpretation of E risk
0,13726235		100 High	0,044899691	14,7	Low	0,242870644		100 High

Figure 4. An example of predicting the risk of pathologies based on machine-learned models with output to an Excel spreadsheet.

Gradation by risk levels is based on the calculated percentage risk, i.e. the patient is automatically attributed to one of three categories, namely low (0-33%), medium (34-66%) or high (67-100%) risk. Such columns as «Modeled risk» and «Interpretation of risk» are created in Excel for the convenience of the doctor's work. The modeling results are additionally highlighted in intuitively understandable colors. Thus, green corresponds to a low level, yellow to a medium level, and red to a high level. In our opinion, this approach contributes to clear visualization of patients with an increased risk of development of cervical lesions and helps to correct the clinical situation at the screening stage.

Before performing risk prediction for a new patient, the entered data is pre-processed. All predictors whose values are missing in the file, are automatically marked as missing (NaN). Then, the missing values are replaced by mean values in the corresponding columns using the SimpleImputer with the «mean» strategy. This approach ensures the correctness of the prediction even with incomplete data and prevents model errors due to missing values.

A prediction of the probability of developing dental pathology is calculated for each patient with the help of the XGBoost model saved after training. The obtained probability values represent a continuous indicator from 0 to 1. This corresponds to the quantitative characteristic of the individual risk of the development of a specific cervical dental lesion.

It is proposed to use a combined approach for WSD and CC of teeth. In this regard, the predicted probability is combined with the number of risk factors in a specific patient. The weights for integrating both the predicted probability and the number of factors are automatically selected based on correlation matching with the training sample. This approach allows for balancing the contribution of two different risk components, namely the prediction of the XGBoost model and the patient's RFs. One component may dominate over the other without weight correction and this leads to distortions in risk modeling. For example, a patient with a great number of

RFs but a low value of the predicted probability from the model may receive a high or low integral risk that does not correspond to reality.

Thus, correlation matching allows for «adjusting» the weights to the structure of the training sample, finding the optimal ratio between the predicted probability and the factor scale. With this approach, the integral risk is maximally consistent with the real presence of dental pathology in patients in the training sample. Therefore, two main advantages are provided, namely increasing the accuracy of the prediction for specific patients and improving the interpretability of the results for a doctor. The integral risk is more relevant to clinical data reflecting both the statistically predicted probability and the actual available RFs. A doctor can assess the contribution of each individual predictor to a general prediction that makes clinical decision-making mathematically justified and eliminates subjective assessment.

For E of tooth enamel, the predicted risk is calculated by normalizing the posterior estimate between the minimum probability of a positive class and the optimal threshold determined while training. This allows obtaining a comparable risk for pathology with a low prevalence, where standard approaches may show an unstable result due to class imbalance.

Thus, machine selection of optimal weights through correlation matching is not only a technical solution. It is a key mechanism for increasing the clinical relevance of the model in terms of the reliability of the prediction for a new patient.

## DISCUSSION

The presented study is the first of a proposed comprehensive system for predicting cervical dental lesions in young patients. Its results demonstrated a new potential of innovative technologies in clinical dentistry. The leading predictors of cervical dental lesions identified by the authors confirm their multifactorial nature and they are consistent with recent research data [5-6, 13].

The proposed model allowed for systematizing various RFs that would contribute to reducing the subjectivity of medical assessment and, thus, would increase the reproducibility of prediction results, simplify the diagnostic stage by reducing the number and duration of procedures [3, 14]. The developed model includes the factors that were partially used in other tools [8]. This approach corresponds to modern ideas about the introduction of digital technologies into preventive dentistry, the task of which is the early identification of patients with an increased probability of developing pathology.

In order to assess the risk of cervical dental lesions within the framework of the screening approach, it was appropriate to apply tree-based machine learning models using the XGBoost algorithm. For all created models, recall was a priority over accuracy and specificity. This approach allowed for minimizing the number of missed cases of CC, WSD and E of tooth enamel and it seems methodologically justified. This is due to the fact that the main task of primary prevention is the timely identification of patients from the risk group with subsequent correction of existing individual RFs, the prescription of therapeutic and preventive measures and dynamic monitoring of the condition of the cervical region of the teeth.

The model for predicting E of tooth enamel is of particular importance in which the SMOTE method was used to create it. The results obtained showed positive experience in using tree-based machine learning models for dental pathology of low prevalence among the examined patients. This became possible with the correct adaptation of the training methodology and the selection of classification threshold values.

The presented screening model had limitations in specificity, which are manifested in an increase in the number of false positive predictions. This drawback does not reduce its clinical value but it provides for further improvement of the predictive system and justifies the feasibility of searching for methods to increase the classification accuracy. The conducted analysis of available scientific publications showed the absence of studies that would comprehensively compare machine learning models for predicting cervical dental lesions with the indication of recall, specificity and overall accuracy indicators separately for CC, WSD and E of tooth enamel. Therefore, direct comparison of the obtained results with the data of other scientists is limited. This indirectly indicates the novelty of the presented approach and determines the need for further research at the same time.

## CONCLUSIONS

1. The developed machine learning screening model allows for identifying young patients with an increased risk of cervical dental lesions. Therefore, it can be used as a primary screening tool to justify the feasibility of prescribing preventive measures and conducting dynamic monitoring of patients.

2. The predictors that were used to create a computer prediction model can be determined directly during the patient's dental examination. This increases its practical significance and prospects for use in everyday medical practice.

3. The application of the model for a new patient is seen as relevant. The availability of convenient visual tools for a dentist, the integration of quantitative risk assessment and predictive probability almost completely eliminate subjectivity in the process of predicting dental pathology.

**Prospects for further research.** Further improvement of the developed mathematical models is planned in order to increase their predictive accuracy and clinical relevance. In this regard, the work will be carried out to optimize the specificity of the model while maintaining its high recall. Combining models for more accurate identification of patients with intact cervical areas of teeth and reducing the number of false positive predictions is seen as promising.

## COMPLIANCE WITH ETHICAL REQUIREMENTS

The study was conducted in accordance with the principles of World Medical Association's Declaration of Helsinki «Ethical Principles for Medical Research Involving Human Subjects» (1964-2008) and completely excluded restrictions on the patient's interests and harm to his health (conclusion of the Bioethics Commission of Donetsk National Medical University No. 43 dated January 21, 2021). All patients gave written informed consent to participate in the study.

## FUNDING AND CONFLICT OF INTEREST

The study has no external funding sources. There is no conflict of interest.

## AUTHOR CONTRIBUTIONS

Zabolotna I. I.<sup>A, B, D, E, F</sup>

Bogdanova T. L.<sup>B, C, D, F</sup>

Azarenkov V. I.<sup>E</sup>

Genzytska O. S.<sup>B</sup>

Komlev A. A.<sup>B</sup>

## REFERENCES

1. Fernández-Barrera, M. F., Lara-Carrillo, T., Scougall-Vilchis, R. J., Pontigo-Loyola, A. P., Mora-Acosta, M., Acuña-Gonzalez, G. R., Casanova-Sarmiento, J. A., Escoffié-Ramírez, M., Medina-Solis, C. E., & Maupomé, G. (2024). Effect of sealant versus fluoride varnish on dental caries incidence in Mexican children: A randomized controlled clinical trial. *J Stoma*, 77(3), 161-167. <https://doi.org/10.5114/jos.2024.143583>
2. Olley, R. C., & Sehmi, H. (2017). The rise of dentine hypersensitivity and tooth wear in an ageing population. *Br Dent J*, 223(4), 293-297. <https://doi.org/10.1038/sj.bdj.2017.715>
3. Supriyadi, M. R., Samah, A. B. A., Muliadi, J., Awang, R. A. R., Ismail, N. H., Majid, H. A., Othman, M. S. B., & Hashim, S. Z. B. M. (2025). A systematic literature review: exploring the challenges of ensemble model for medical imaging. *BMC Med Imaging*, 25(1), 128. <https://doi.org/10.1186/s12880-025-01667-4>
4. Proshchenko, A. M. (2024). Prohnozuvannya vynyknennya bol'ovoho syndromu dysfunktsiyi SNSHCHS u patsiyentiv z oklyuziyno-artykulyatsiynymy rozladamy [Prediction of TMJ dysfunction pain syndrome in patients with occlusive articulation disorders]. *Bulletin of Dentistry*, 3(53), 75-82. <https://doi.org/10.35220/2078-8916-2024-53-3.13>
5. Levri, L., Di Benedetto, G., & Raspanti, M. (2014). Dental wear: A scanning electron microscope study. *BioMed Research International*, 2014, 340425. <https://doi.org/10.1155/2014/340425>
6. Ramsay, D.S., Marilyn Rothen, M., Scott, J., & Cunha-Cruz, J. (2015). Tooth wear and the role of salivary measures in general practice patients. *Clin Oral Investig*, 19(1), 85-95. <https://doi.org/10.1007/s00784-014-1223-4>
7. Zabolotna, I. I., & Bohdanova, T.L. (2025). Analiz faktoriv ryzyku vynyknennya i prohresuvannya nekarioznykh pryshykovykh urazhen' zubiv [Analysis of risk factors of the development and progression of non-carious cervical lesions of teeth]. *Innovation in stomatology*, 2, 75-83. <https://doi.org/10.35220/2523-420X/2025.2.13>
8. Vodorig, Y. Y., Brailko, N. M., Dvornyk, A. V., & Tkachenko, I. M. (2024). Ohlyad suchasnykh metodyk otsinky ryzyku poyavy kariyesu [A review of contemporary methods for caries risk assessment]. *Actual problems of modern medicine: Bulletin of Ukrainian Medical Stomatological Academy*, 24(4), 277-283. <https://doi.org/10.31718/2077-1096.24.4.277>
9. Alsubai, S. (2023). Enhancing prediction of tooth caries using significant features and multi-model classifier. *PeerJ Comput Sci*, 9, e1631. <https://doi.org/10.7717/peerj-cs.1631>
10. Liubarets S. F. (2018). Prohnozuvannya rozvytku kariyesu yak uskladnennya porushen' formuvannya zubiv u ditey [Predicting of the development of caries as the complications of the disturbances of teeth formation in children]. *Bulletin of problems in biology*, 1(1), 367-370. <https://doi.org/10.29254/2077-4214-2018-1-1-142-367-370>
11. Kryvenchuk, Yu., & Oleskevych, S. (2023). Informatsiyna systema dlya stomatolohichnoyi kliniky z mozhlyvistyuu vyyavlennya kariyesu na panoramnykh znimkakh zubiv [Creation of caries detection system]. *Herald of Khmelnytskyi National University*, 1(1), 271-275. <https://doi.org/10.31891/2307-5732-2023-317-1-271-275>
12. Zabolotna, I. I., & Bohdanova, T.L. (2025). Vybir optymal'noho variantu statystychnoyi komp'yuternoyi modeli prohnozuvannya vynyknennya pryshykovykh urazhen' zubiv [Selection of the optimal variant of a statistical computer model for predicting the development of cervical lesions of teeth]. *Colloquium-journal*, 66(259), 14-17. <https://doi.org/10.5281/zenodo.17520847>
13. Kong, W., Ma, H., Qiao, F., Xiao, M., Wang, L., Zhou, L., Chen, Y., Liu, J., Wang, Y., & Wu, L. (2024). Risk factors for noncarious cervical lesions: A case-control study. *J Oral Rehabil*, 51(9), 1684-1691. <https://doi.org/10.1111/joor.13772>
14. Bui, T. H., Hamamoto, K., & Paing, M. P. (2022). Automated Caries Screening Using Ensemble Deep Learning on Panoramic Radiographs. *Entropy (Basel)*, 24(10), 1358. <https://doi.org/10.3390/e24101358>
15. Inacio, V., Rodriguez Alvarez, M. X., & Gayoso Diz, P. (2021). Statistical Evaluation of Medical Tests. *Annual Review Statistics and Its Application*, 8, 41-67. <https://doi.org/10.1146/annurev-statistics-040720-022432>

## Резюме

### СКРИНІНГОВА МОДЕЛЬ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВИНИКНЕННЯ ПРИШИЙКОВИХ УРАЖЕНЬ ЗУБІВ

Ірина І. Заболотна<sup>1</sup>, Тетяна Л. Богданова<sup>2</sup>, Володимир І. Азаренков<sup>3</sup>, Олена С. Гензицька<sup>1</sup>, Андрій А. Комлев<sup>1</sup>

<sup>1</sup>Донецький національний медичний університет, м. Кропивницький, Україна

<sup>2</sup>Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна

<sup>3</sup>Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна

**Вступ.** Прогнозування ризику виникнення пришийкової патології зубів є складним завданням через багатофакторність її етіопатогенезу та обмежені знання про фактори ризику.

**Мета.** Розробити та апробувати комп'ютерну модель прогнозування розвитку пришийкових уражень зубів у пацієнтів молодого віку.

**Матеріали та методи.** Вибірку склали 272 пацієнти (середній вік  $24,3 \pm 6,9$  роки), у яких було визначено фактори ризику виникнення клиновидного дефекту, пришийкового карієсу, ерозії емалі, що стали вхідними даними до комп'ютерної моделі. Було використано метод деревоподібного машинного навчання Extreme Gradient Boosting (XGBoost), реалізований на мові програмування Python із використанням бібліотек scikit-learn і XGBoost. Для підвищення ефективності прогнозування менш поширеної серед обстежених ерозії емалі додатково було застосовано Synthetic Minority Over-sampling Technique (SMOTE).

**Результати.** При розробці моделей було надано пріоритет чутливості (recall) над показниками точності та специфічності. Це сприяло зменшенню кількості пропущених випадків для кожної з патологій. Найвища дискримінаційна здатність (ROC-AUC) = 0,84 (Receiver Operating Characteristic curve – Area under the curve) у поєднанні з високим рівнем чутливості (recall = 0,82) відповідає моделі прогнозування пришийкового карієсу зубів. Це підтвердило доцільність використання алгоритму XGBoost для виявлення складних взаємозв'язків у нелінійних сполученнях показників. Модель прогнозування клиновидного дефекту зубів також мала високу чутливість (recall = 0,83), але помірне значення ROC-AUC (0,64), що підкреслює наявність нелінійно залежних між собою предикторів. Особливий науковий інтерес представляє модель прогнозування ерозії емалі зубів, яка була створена в умовах незначної поширеності патології серед обстежених. Проте отримані результати показали прийнятний рівень чутливості (recall = 0,47) та помірну дискримінаційну здатність (ROC-AUC = 0,72). Це дозволило визначити, що проблема малої вибірки була успішно вирішена.

**Висновки.** Представлена скринінгова модель машинного навчання сприяє ідентифікації осіб з підвищеним ризиком виникнення пришийкових уражень зубів. Її використання дозволить обґрунтувати призначення профілактичних заходів пацієнтам молодого віку.

**Ключові слова:** карієс зуба, ерозія зуба, ансамблеве навчання, профілактика, фактори ризику

Received: 11.12.2025

Accepted: 5.02.2026