

Development and Validation of Machine Learning Algorithm with Oversampling Technique in Limited Data Scenarios for Prediction of Present and Future Restorative Treatment Need

Elina Väyrynen, Otso Tirkkonen, Henna Tiensuu, Jaakko Suutala, Vuokko Anttonen, Marja-Liisa Laitala, Katri Kukkola, Saujanya Karki

Submitted to: JMIR Medical Informatics
on: April 04, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 21

 Figures 22

 Figure 1..... 23

 Figure 2..... 24

 Figure 3..... 25

Multimedia Appendixes 26

 Multimedia Appendix 1..... 27

 Multimedia Appendix 2..... 27

Development and Validation of Machine Learning Algorithm with Oversampling Technique in Limited Data Scenarios for Prediction of Present and Future Restorative Treatment Need

Elina Väyrynen¹ DDS; Otso Tirkkonen¹ BDS; Henna Tiensuu² PhD; Jaakko Suutala² Prof Dr; Vuokko Anttonen¹ Prof Dr, DDS; Marja-Liisa Laitala¹ Prof Dr, DDS; Katri Kukkola³ MSci; Saujanya Karki¹ PhD, BDS, MPH

¹Research Unit of Population Health Faculty of Medicine University of Oulu Oulu FI

²Biomimetics and Intelligent Systems Group Faculty of Information Technology and Electrical Engineering University of Oulu Oulu FI

³Optoelectronics and Measurement techniques Faculty of Information Technology and Electrical Engineering University of Oulu Oulu FI

Corresponding Author:

Saujanya Karki PhD, BDS, MPH
Research Unit of Population Health
Faculty of Medicine
University of Oulu
Aapistie 3
Oulu
FI

Abstract

Background: Untreated dental caries is the most common health condition globally. Because of this, new strategies need to be developed to reduce the manifestations of dental caries.

Objective: The aim of this study was to develop and test a machine learning (ML) algorithm for detecting present and predicting future carious lesions in the adolescent population, utilizing a set of easy- to-collect predictive variables. In addition, another aim was to deal with an imbalanced and small dataset with an oversampling method.

Methods: This population-based study was conducted among secondary schoolchildren, between 13–17 years of age, from northern parts of Finland in 2022. After the inclusion criteria was met, a total of n=218 participants were included in this study. The inclusion criteria consisted of participants having completed a web-based risk assessment questionnaire and undergone a clinical examination at Public Healthcare Services. Dental caries (ICDAS4-6) and active initial caries (ICDAS2+,3+) were considered as outcomes. Several predictors, such as behavioral and dietary habits, were included. An eXtreme Gradient Boosting (XGBoost) model was developed, tested and assessed for its predictive performance. A 4-fold cross-validation (CV) was performed using the nested resampling technique. The random over-sampling examples (ROSE) method and the k-nearest neighbors (KNN) classifiers were utilized for all four folds. The mean (SD) performance of all folds was computed.

Results: The prevalence of dental caries was 65.56% (ICDAS2+,3+,4-6). The mean (SD) area under the curve (AUC) was 0.769 (0.042) and the mean (SD) F1-score was 0.816 (0.055) for the XGBoost model. Similarly, the mean (SD) AUC and mean (SD) F1-scores after oversampling were 0.744 (0.045) and 0.787 (0.035), respectively. The SHapley Additive exPlanations (SHAP) values were calculated for all four folds to assess feature importance, revealing that previous dental fillings were the feature most strongly associated with the need for restorative treatment.

Conclusions: Based on performance metrics, the ML algorithm developed and tested in this study can be considered good. The ML algorithm could serve as a cost-effective screening tool for dental professionals in identifying the risk the future restorative treatment needs. However, future studies with longitudinal cohorts and longitudinal data, along with external validation for generalizability, are needed to validate our model.

(JMIR Preprints 04/04/2025:75117)

DOI: <https://doi.org/10.2196/preprints.75117>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR publication, my title and abstract will remain visible to all users.

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Application of Machine Learning Algorithm with Oversampling Technique in Limited Data Scenarios for Prediction of Present and Future Restorative Treatment Need

Väyrynen Elina ¹, Tirkkonen Otso ¹, Tiensuu Henna ², Suutala Jaakko ², Anttonen Vuokko ¹, Laitala Marja-Liisa ^{1,3}, Kukkola Katri ⁴, Karki Saujanya ¹

¹*Research Unit of Population Health, Faculty of Medicine, University of Oulu, Oulu, Finland*

²*Biomimetics and Intelligent Systems Group, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland*

³*Medical Research Center Oulu and Oulu University Hospital, Oulu, Finland*

⁴*Optoelectronics and Measurement techniques, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland*

Corresponding Author

Saujanya Karki, (BDS, MPH, PhD)
Research Unit of Population Health
Faculty of Medicine University of Oulu
Aapistie 3 Oulu, 90220 Finland
Email : saujanya.karki@oulu.fi

Abstract

Background: Untreated dental caries is the most common health condition globally. Because of this, new strategies need to be developed to reduce the manifestations of dental caries. The aim of this study was to develop and test a machine learning (ML) algorithm for detecting present and predicting future carious lesions in the adolescent population, utilizing a set of easy- to-collect

predictive variables. In addition, another aim was to deal with an imbalanced and small dataset with an oversampling method.

Methods: This population-based study was conducted among secondary schoolchildren, between 13–17 years of age, from northern parts of Finland in 2022. After the inclusion criteria was met, a total of $n=218$ participants were included in this study. The inclusion criteria consisted of participants having completed a web-based risk assessment questionnaire and undergone a clinical examination at Public Healthcare Services. Dental caries (ICDAS₄₋₆) and active initial caries (ICDAS_{2+,3+}) were considered as outcomes. Several predictors, such as behavioural and dietary habits, were included. An eXtreme Gradient Boosting (XGBoost) model was developed, tested and assessed for its predictive performance. A 4-fold cross-validation (CV) was performed using the nested resampling technique. The random over-sampling examples (ROSE) method and the k-nearest neighbors (KNN) classifiers were utilized for all four folds. The mean (SD) performance of all folds was computed.

Results: The prevalence of dental caries was 65.56% (ICDAS_{2+,3+,4-6}). The mean (SD) area under the curve (AUC) was 0.769 (0.042) and the mean (SD) F1-score was 0.816 (0.055) for the XGBoost model. Similarly, the mean (SD) AUC and mean (SD) F1-scores after oversampling were 0.744 (0.045) and 0.787 (0.035), respectively. The SHapley Additive exPlanations (SHAP) values were calculated for all four folds to assess feature importance, revealing that previous dental fillings were the feature most strongly associated with the need for restorative treatment.

Conclusions: Based on performance metrics, the ML algorithm developed and tested in this study can be considered good. The ML algorithm could serve as a cost-effective screening tool for dental professionals in identifying the risk the future restorative treatment needs. However, future studies with longitudinal cohorts and longitudinal data, along with external validation for generalizability, are needed to validate our model.

KEYWORDS oral health; machine learning; dentistry; dental caries, caries management

Introduction

Dental caries is the most common dietary-microbial disease, requiring regular exposure to fermentable carbohydrates in the diet. Enrichment of acid-producing and acid-tolerating microorganism in dental plaque leads to demineralized tooth structure, which, in turn, can lead to loss of tooth structure, ultimately resulting in cavities (dental caries) [1]. Risks for dental caries include physical, biological, environmental, behavioural, and lifestyle-related factors [2]. The

balance between pathological and protective factors, such as insufficient exposure to fluoride or irregular brushing of teeth, influences the initiation and progression of dental caries [3].

Individual-level risk recognition is of utmost importance, since recent dental caries management protocol prioritizes early prevention and minimal intervention at an individual level [4]. The availability of Caries Risk Assessment Tools (CRAT) has assisted clinicians in risk identification, as well as in risk minimization. However, a majority of the existing CRAT require either dental visits or measurements of saliva parameters at dental clinics (face-to-face methods). A systematic review considered the possibility of using reduced Cariogram (without saliva parameters), one of the CRATs, due to its better performance when compared to a full Cariogram [5]. The use of artificial intelligence (AI) and machine learning (ML) in the medical field is gaining attention, globally. The possibilities of an automated dental caries risk prediction method, utilizing machine learning algorithms, needs to be explored. A recent study highlighted the possibility of using ML algorithms and easy-to-collect predictors in screening active dental caries and urgent treatment needs in school-age children. However, the questionnaire mainly consisted of predictors covering physical, mental and social aspects [6], missing out on important predictors such as dietary habits and oral health-related behaviours. As the development and progression of carious lesions is multifactorial, such information is important to consider.

AI can be defined as the non-biological ability of a computer trying to imitate human intelligence to accomplish complex tasks, such as problem-solving and decision-making [7]. ML is a subset of AI designed to identify patterns or make predictions based on used data. ML algorithms can model non-linear and high dimensional characteristics, such as health data [8]. In addition to being the latest and often the most popular technology, ML algorithms have the ability to learn itself and improve over time when exposed to more data [9]. The use of ML models can improve patient care by providing individualized outcome predictions and by reducing standardized processes, allowing clinicians to spend more time with patients [10]. In dentistry, the literature showed that various ML algorithms such as logistic regression, decisions trees, random forest and extreme gradient boosting are used in predicting dental caries. However, these studies are in their early stages and more research needs to be conducted to validate these methods [11]. Furthermore, the ground truth in the above mentioned studies is based only on clinical examinations (visual-tactile), even though it is recommended to perform both clinical and radiographic examinations to minimize the risk of misdiagnosis in caries evaluation [12]. In this regard, the severity and activity of carious lesions are crucial while deciding on the treatment path [13,14]. It is important to include real-life clinical environment observations when training and testing ML algorithms.

There are however some challenges such as high costs, data security and legal restrictions in dentistry, making the acquisition of individual-level comprehensive data more difficult [15,16]. Studies have utilized electronic health records (EHR) and national registries for training ML algorithms [17]. However, in dentistry, availability of extensive EHR data or national registries is scarce [18]. Therefore, to overcome the challenges associated with small datasets, oversampling techniques can be explored. Oversampling is a data augmentation technique that aims to rebalance the training data distribution by amplifying the volume of instances that belong to the under-represented class, helping to correct the imbalance between minority and majority examples [19], [20]. Another challenge in dentistry is also related to class imbalance. When the number of patients with a target disease differs from the healthy population, the situation is referred to as the imbalanced data problem. The accuracy of ML models can be affected by these imbalances.

Therefore, the aim of this study was to develop and test a ML algorithm in detecting present and predicting future carious lesions among adolescents utilizing a set of easy- to-collect predictive variables. Another aim was to deal with challenges due to the imbalanced and small dataset in cariology with the use of the oversampling method.

Methods

Study population and data sources

This cross-sectional study utilized data collected for the Digileap project, conducted among secondary school children in 2022 aged 13–17-years from the northern parts of Finland. Prior to the study, the sample size ($n=246$) was calculated based on the prevalence of dental caries from a previous study [21] with 95% confidence intervals and a precision set at 0.05, assuming that the total population of 13–15-year-olds is 100 000 [22]. Participants completed a web-based risk assessment questionnaire within their school premises, and their oral health records were registered at the Public Healthcare Services during 2022-2023, from where they were later requested through Findata services [23]. Findata is the Finnish Social and Health Data Permit authority, which grants permits for the secondary use of social and health care data, improving data protection for individuals. The inclusion criteria for this study included the following: (1) being 13–17-years old with signed informed consent, (2) completion of a web-based risk assessment questionnaire, and (3) completion of a dental examination performed at Public Dental Clinics during 2022-2023. After meeting the inclusion criteria, a total of $n=218$ participants were included in this study.

Ethical issues

The entire study was conducted in full accordance with the World Medical Association Declaration

of Helsinki. The ethical committee of Northern Ostrobothnia Hospital District approved the study protocol (EETTMK 62/2021) and Finnish Medicines Agency (FIMEA) [24] also issued the Medical Device Permit (2022/007715). In addition, study permissions were also obtained from the Public Healthcare Services in Kuusamo, Ylivieska, Oulu and Liminka. Oral health records (dental caries registered at the public dental services) were obtained from the Finnish Social and Health Data Permit Authority Findata [25] with a data permit (THL/6268/14.02.00/2021). All the schools were contacted prior to the study via an official email requesting the participation of schoolchildren and their parents. Over 15 years old participants signed the informed consent, and the informed consent was asked beforehand from parents of participants under 15 years of age. Participation was completely voluntary, and the participants had a right to withdraw their participation at any given phase of the study.

Study variables

Outcome variables

Initial active carious lesion (enamel and/or dentin caries) and all cavitated lesions were considered as a main outcome variable for this study. Carious lesion were diagnosed using the International Caries Detection and Assessment System (ICDAS) criteria with the aim to differentiate between different stages of dental caries. ICDAS stands for the assessment of the caries process by stage (noncavitated or cavitated) and activity (active or arrested/inactive). The “+” symbol indicates a caries lesion that is active and progressing. The “-” symbol indicates an inactive lesion with no active progression and the tooth surface is considered as sound [26].

To describe the transition of caries lesions in this study, the ICDAS score of 0 or ICDAS scores of 2- and ICDAS 3- were considered as sound in contrast to the ICDAS scores of 2+, 3+, 4, 5, 6 being considered as diseased. The ICDAS 2+ and 3+ codes were merged into one category (ICDAS_{2+, 3+}) to represent non-cavitated lesions or microcavitated active lesions and the ICDAS 4, 5 and 6 codes were used as one category (ICDAS₄₋₆) to represent cavitated lesions. In a nutshell, any initial active carious lesion (enamel and/or dentin caries) and all cavitated lesions were considered as a main outcome variable for this study (ICDAS_{2+, 3+, 4-6}) [13]. ICDAS 1 was not found from our study population due to challenges in diagnostics: these lesions are characterized by the first visual changes in enamel, often appearing as white spots or lines that are only visible when the tooth is carefully air-dried for 5 seconds.

Caries assessment with surface-by-surface evaluation was conducted by a licensed dentist following the Finnish Current Care Guidelines [27]. All teeth of the participants were examined with halogen light with a surface reflecting mirror and explorer and fiber-optic transilluminator followed by radiographic examination if needed. A radiographic examination was suggested if (1) one localized enamel breakdown lesion was found or (2) the patient had several initial caries lesions or (3) the patient had dental caries risk factors or suspicion that the patient might have hidden dental caries lesions or (4) radiographs had not been taken in the past few years [28]. Previously, data collected from the Finnish public healthcare records were shown to not be inferior to the calibrated examiners [29].

Predictors

For this study, age, gender, oral health-related behaviours such as frequency of toothbrushing, toothbrush type, toothpaste type, frequency of fluoride toothpaste use, interdental cleaning frequency, frequency of xylitol use, additive sugar consumption and smoking habits were considered as independent variables. The questionnaire consisted of information about food and drink consumption of the. Participants were asked to report the amounts and frequencies. The average daily consumption was calculated for each product and multiplied by the quantity consumed. The additive sugar consumption was calculated using the FINELI database, a Finnish national food consumption database maintained by the Finnish Institute for Health and Welfare (THL)[46]. Using the FINELI database, total sugar content was matched to the food items and the amount of sugar (g) in each item was calculated. Finally, the total daily sugar intake was calculated for each food item consumed per day (daily added sugar intake). Likewise, local factors such as recent restorations, extracted teeth, bleeding when brushing, and the dry mouth index were considered as predictors. To complement the self-reported survey, clinical data on missing teeth and dental fillings were also included as predictors. The dry mouth index included questions such as ‘Does your mouth feel dry when you eat?’, ‘Do you have difficulties swallowing some food?’, ‘Do you have to drink in order to make it easier to swallow dry food?’. Participants responded yes/no, and answers were combined as one continuous variable achieving values from 0-3.

Model development and training

In this study, the eXtreme Gradient Boosting (XGboost) algorithm was applied to predict the outcome variable. The model was trained using the R software version 4.3.1 (R Foundation for

Statistical Computing Vienna, Austria, [30]).

Model fitting (training and testing)

The training and testing of the ML models were performed using the nested resampling technique with 4-fold cross-validation (CV). In a typical K-fold CV, the dataset is randomly and evenly split into K parts. The model is built using K – 1 parts of the dataset, called the training set, and evaluated based on the remaining part, known as the test set. This process is repeated K times so that each part is used as a test set once [31]. Four splits were created, each containing 75% of data as training set and 25% of unseen data as a test set.

Model training protocol

For each four folds, the ML models were built by use of a MLR-package [32]. During the training hyperparameter, tuning was conducted for all four folds. The tuned model was evaluated based on a validation set, which was independent of the nested sampling testing set. Finally, the optimized model's performances were evaluated on each fold's test set. For all four folds, the model training protocol was then repeated using the new oversampled dataset. The model training protocol is shown in Figure 1.

Oversampling

As the total sample was $n = 218$, the oversampling technique was utilized as suggested [33]. The random over-sampling examples (ROSE) -method and the k-nearest neighbors (KNN) classifier were utilized for all folds [34][35]. Oversampling simulated $n = 2000$ new synthetic participants to the data set ($P = 0.05$). In the context of oversampling techniques, "P" refers to the proportion or percentage of the minority class instances that are to be oversampled. In the complete dataset, the number of participants with carious lesion was $n = 143$. The training set consisted of 75% of the data and the number of participants with carious lesion in the training set was $0.75 \times 143 = 107$. Using the oversampling method, $2000 \times 0.95 = 1900$ new participants with carious lesions were created. As a result, the total number of participants with carious lesions increased to $107 + 1900 = 2007$. Likewise, in the complete dataset, the number of participants with sound teeth was $n = 75$. The training set included of 75% of the complete data, and the number of participants with sound teeth in the training set was $0.75 \times 75 = 56$. Using the oversampling method, a total of $n = 100$ ($2000 \times 0.05 = 100$) new synthetic participants with sound teeth were created. This resulted the total number of participants with sound teeth to $n = 156$.

Model evaluation

To assess the predictive performance of the ML models, the area under the curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), no information rate (NIR), precision, recall and F1-score for each predictive model were calculated. The mean (SD) performance of all folds was computed.

SHapley Additive exPlanations (SHAP) values were computed for all folds and also after oversampling. The SHAP values were computed to determine the importance of each variable in predicting the dental caries outcomes of this study. The SHAP values are an additive feature importance measure which represent the responsibility of each feature in pushing the model output away from its base value [36]. This study was reported in accordance with the TRIPOD+AI statement for developing or evaluating the performance of prediction models [37].

Results

The prevalence of carious lesions (enamel and dental caries) in this study group was 65.56% (ICDAS₂₊, ₃₊, ₄₋₆). More than half of the study participants were female and had a habit of tooth brushing twice daily. Almost three-fourths of the participants reported using fluoride toothpaste while 48.6 % reported using an electronic toothbrush. A majority of the participants (89.0 %) did not smoke, and the mean daily added sugar intake was 51.6 grams. Additionally, more than half of the participants had dental restorations. The demographic characteristics and descriptive analysis of the categorical independent variables are shown in Table 1.

The performance metrics of all folds of the XGBoost model before and after oversampling, are shown in Table 2. Additionally, the mean performance across all folds is presented. The mean (SD) AUC value, which evaluates the model's ability to discriminate between carious and sound teeth was good AUC 0.769 (0.042); after oversampling 0.744 (0.045). The mean (SD) accuracy, that evaluates performance of models was also high, 0.752 (0.059) and after oversampling 0.725 (0.034). The AUC and accuracy values were complemented by the F1-scores. The F1-score is the harmonic mean of precision and recall, and it provides a comprehensive evaluation of a model with an imbalanced dataset. The mean (SD) F1-score was 0.816 (0.055); after oversampling 0.787 (0.035). The ability of the model in predicting carious lesions (true positive cases), expressed as a mean (SD) sensitivity of 0.854 (0.115); after oversampling 0.783 (0.089). They were considered high. The ability to predict sound teeth (true negative cases) expressed as mean (SD) specificity of 0.559 (0.129), and the value

slightly increased after oversampling 0.607 (0.145). These, in turn, were considered low.

After evaluating the performance of each model, the SHAP values were computed for all four folds and also after oversampling. The feature that most strongly predicted with the need for present and future restorative treatment was previous fillings in all folds, followed by the total added sugar intake, frequency of smoking, toothpaste type, frequency of toothbrushing, varying between four folds. Interestingly, the importance of minor predictors slightly increased after the oversampling method was applied in all folds. Fillings and total added sugar intake were in the top four most important features in every fold before oversampling. There was more variation in the folds after oversampling, but clearly, fillings remained the most important feature. SHAP values for each four folds of the XGBoost model before and after oversampling are shown in Figure 2.

Discussion

The main aim of this work was to develop and test a ML algorithm in predicting carious lesions among the adolescent population, utilizing a set of easy- to-collect predictors and evaluate the importance of each predictor. Another aim was to utilize a novel approach of oversampling in cariology research to deal with and improve the imbalance of a small dataset. The ML models developed and tested in this study performed well in predicting present and future restorative treatment needs among adolescents. Despite the drop in performance metrics after oversampling, the parameters were within the acceptable range, supporting the positive performance of the ML algorithm in the present study.

The XGBoost algorithm used in the present study performed well. This is comparable with previous studies from Brazil [38], [39]. In both studies, the XGBoost outperformed compared to other ML algorithms such as the logistic regression and decision tree. Both studies utilized the socioeconomic variables such as income and parents' employment as predictors for carious lesions. However, those variables were not considered in this study, since in Finland, all individuals under the age of 18 are entitled to free dental care. The aim was to utilize easy-to-collect predictors. A recent study also utilized ML algorithms and easy-to-collect information when screening active dental caries and urgent treatment needs in adolescents and concluded that the Naïve Bayes outperformed other models [6]. However, that study particularly considered physical, mental and social factors rather than behavioural factors. Furthermore, both clinical and radiographic examinations were performed in the present study to minimize the risk of over- or under-diagnosis. In the real-life clinical

environment, the use of radiographic methods is considered advantageous when deciding the need of operative care especially when a patient presents with a ICDAS score of 3 [28]. The drop in performance after oversampling in the present study is comparable with the previous study from the US [6]. Xiong et al. considered the Synthetic Minority Over-Sampling Technique (SMOTE) for oversampling in their studies. However, the Random Over-Sampling Examples (ROSE) was utilized in this study for oversampling. The ROSE oversampling technique creates more balanced data, drawing new artificial samples from the feature space neighbourhood around the minority class [40]. In contrast, SMOTE works effectively with continuous variables rather than categorical or binary variables as in the present study. Additionally, the k-Nearest Neighbour Classifier (KNN) data cleaning method was utilized in this study.

Dental caries is a multifactorial disease influenced by individual, biological, behavioural, and environmental factors [41]. In literature, a past caries experience was found to be the most powerful caries predictor [42]. High consumption of carbohydrates increased the chance of developing dental caries [43] and xylitol-containing products significantly prevented caries, when compared with other non-xylitol products [44]. These findings from previous literature are in line with the results from this study, as delivered from the feature importance SHAP values. The past caries experience, non-use of fluoridated toothpaste, socio-economic level and a higher frequency of sugar consumption were predictors that influenced caries progression the most in a previous longitudinal study that aimed to predict dental caries in primary and permanent teeth among children aged 1–5 years old [38]. Likewise, the use of dental floss, unhealthy food consumption, self-declared race and exposure to fluoridated water were the most predictive variables in another previous study [39]. The previous fillings (explaining past caries experience) and total added sugar intake (explaining high consumption of carbohydrates) were the most predictive variables in this study.

The application of SHAP values in this study enhances the interpretability of the ML model, providing a transparent understanding of how each feature contributes to predictions. Interpretability is a key element of explainable AI (XAI), which plays a critical role in ensuring that the ML models are not only accurate, but also transparent and reliable for real-world application. In contrast to black-box models, XAI make the model's decisions more comprehensible and trustworthy. Understanding the rationale behind predictions helps in validating the model's clinical relevance.

Strengths and Limitations

One of the strengths of this study is that the outcome variable, carious lesions, was recorded by a licensed dentist based on both clinical and radiographic evaluations, following the Finnish Current

Care Guidelines [28]. Another strength was the absence of missing values, ensured by the strict inclusion criteria of this study. Selection bias due to voluntary participation in this study can be considered as a limitation. Another limitation is the potential for response bias due to use of a self-reported questionnaire. However, the aim was to keep the questionnaire short and simple in order to minimize response bias. Finally, generalizability of ML algorithm may be questionable as the model did not undergo external validation. Future studies with longitudinal cohorts are needed both to validate our models and to perform external validation. In ML, validating provides evidence that a model is reliable and performs sufficiently with new data. External validation also requires testing the model on independent populations to assess its applicability [45].

Clinical implication

Rising healthcare costs associated with restorative treatment requires justification in early prevention and control of dental caries. New strategies need to be developed to reduce social impacts, such as aesthetic and functional disturbances, on both the individual and societal level. A potential application of ML algorithms in dental caries prognostic studies enables evidence-based personalized dental care that could assist in decreasing dental caries prevalence globally. The ML model developed and tested in this study has the potential to identify possible risk factors of dental caries before the onset of actual dental caries lesions. In this study, each SHAP indicated the importance of each feature in dental caries progression and the information gained can be transformed into a deeper dental caries risk assessment. Algorithm-based risk assessment tools can be integrated into electronic health records and used in electronic pre-assessment forms. Information about dental caries risk is valuable for both patients and dental professionals, influencing treatment and prevention plans, follow-up and patient education [46]. Linking the machine learning algorithms to intraoral images using deep learning algorithms is expected to increase dental screening potential.

Individuals encounter unique challenges in adhering to behavioural changes. To overcome these obstacles, behavioural change interventions need to be both multifaceted and personalized. Health behaviours factors, such as unhealthy food consumption, can be modified by health promotion policies and strategies. This study is unique and innovative because it is the first study to use machine learning models in dental caries prediction in adolescents using easy-to-collect predictors. In the future, after further development and external validation, this ML-model could be used as a risk assessment tool and even be integrated into health record systems, which would be beneficial for the patient and the healthcare professionals in saving time and resources [47]. For the caries risk

assessment tool to be successful, it needs to be inexpensive, user-friendly and open for everyone, even in low-income countries.

Conclusions

Despite the small and imbalanced dataset, XGBoost performed well in predicting restorative treatment among adolescents with and without oversampling method in this study. The results from this study suggest the potential feasibility of the ML models in caries risk assessment enabling easier, cost-effective, less time-consuming and more effective decision-making. However, future studies with longitudinal data as well as external validation are needed to validate our models.

Acknowledgements

We would like to acknowledge all the participants and project partners of this study. We would also like to acknowledge Ville Kaikkonen, Eero Molkoselka, and Laura Pentti for their contributions to the data collection.

Author contributions

EV: Conceptualization, Data Curation, Visualization, Writing - Original Draft. OT: Conceptualization, Data Curation, Formal analysis, Visualization, Writing - Original Draft. HT: Conceptualization, Data Curation, Formal analysis, Visualization, Writing - Review & Editing, Supervision. JS: Writing - Review & Editing, VA: Writing - Review & Editing, MLL: Writing - Review & Editing, KK: Writing - Review & Editing, SK: Conceptualization, Data Curation, Formal analysis, Visualization, Writing - Review & Editing, Supervision.

Funding information

This project was financially supported by The European Regional Development Fund (project: *Digileap of oral health toward virtual reception*; A76934), Minerva Foundation, Finnish Dental Society Apollonia, and Finnish Medical foundation. Elina Väyrynen also received a personal research grant from Finnish cultural foundation, Terttu-foundation and Minerva-foundation.

Conflict of interest statement

All authors declared no conflicts of interest with respect to this article.

Data availability statement

Due to national legislation, restrictions apply to the availability of clinical data at individual level, which were used with the Finnish Social and Health Data Permit Authority (FINDATA). Findata was responsible for the pseudonymization of the data and ensuring the anonymity of the results (THL/6268/14.02.00/2021).

References

1. A. Tanner, C. Kressirer, and L. Faller, "Understanding Caries From the Oral Microbiome Perspective," *J Calif Dent Assoc*, vol. 44, pp. 437–446, Jul. 2016, doi: 10.1080/19424396.2016.12221036.
2. R. H. Selwitz, A. I. Ismail, and N. B. Pitts, "Dental caries," *The Lancet*, vol. 369, no. 9555, pp. 51–59, 2007, doi: [https://doi.org/10.1016/S0140-6736\(07\)60031-2](https://doi.org/10.1016/S0140-6736(07)60031-2).
3. N. B. Pitts *et al.*, "Dental caries," *Nat Rev Dis Primers*, vol. 3, no. 1, p. 17030, 2017, doi: 10.1038/nrdp.2017.30.
4. N. P. T. Innes *et al.*, "A Century of Change towards Prevention and Minimal Intervention in Cariology," *J Dent Res*, vol. 98, no. 6, pp. 611–617, May 2019, doi: 10.1177/0022034519837252.
5. N. Su, M. D. Lagerweij, and G. J. M. G. van der Heijden, "Assessment of predictive performance of caries risk assessment models based on a systematic review and meta-analysis," *J Dent*, vol. 110, p. 103664, 2021, doi: <https://doi.org/10.1016/j.jdent.2021.103664>.
6. D. Xiong *et al.*, "Development of short forms for screening children's dental caries and urgent treatment needs using item response theory and machine learning methods," *PLoS One*, vol. 19, no. 3, p. e0299947, Mar. 2024, doi: 10.1371/journal.pone.0299947.
7. A. S. Ahuja, "The impact of artificial intelligence in medicine on the future role of the physician," *PeerJ*, vol. 7, p. e7702, Oct. 2019, doi: 10.7717/peerj.7702.
8. J. Krois, C. Graetz, B. Holtfreter, P. Brinkmann, T. Kocher, and F. Schwendicke, "Evaluating Modeling and Validation Strategies for Tooth Loss," *J Dent Res*, vol. 98, no. 10, pp. 1088–1095, Jul. 2019, doi: 10.1177/0022034519864889.
9. I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.
10. A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," *BMC Med Res Methodol*, vol. 22, no. 1, p. 282, Nov. 2022, doi: 10.1186/s12874-022-01758-8.
11. L. T. Reyes, J. K. Knorst, F. R. Ortiz, and T. M. Ardenghi, "Scope and challenges of machine learning-based diagnosis and prognosis in clinical dentistry: A literature review," *J Clin Transl Res*, vol. 7, no. 4, pp. 523–539, Aug. 2021.
12. L. T. Reyes, J. K. Knorst, F. R. Ortiz, and T. M. Ardenghi, "Machine Learning in the Diagnosis and Prognostic Prediction of Dental Caries: A Systematic Review," *Caries Res*, vol. 56, no. 3, pp. 161–170, May 2022, doi: 10.1159/000524167.
13. K. R. Ekstrand, T. Gimenez, F. R. Ferreira, F. M. Mendes, and M. M. Braga, "The International Caries Detection and Assessment System – ICDAS: A Systematic Review," *Caries Res*, vol. 52, no. 5, pp. 406–419, Mar. 2018, doi: 10.1159/000486429.
14. A. I. Ismail, N. B. Pitts, and M. Tellez, "The International Caries Classification and Management System (ICCMS™) An Example of a Caries Management Pathway," *BMC Oral Health*, vol. 15, no. S1, p. S9, Dec. 2015, doi: 10.1186/1472-6831-15-S1-S9.
15. F. Schwendicke and J. Krois, "Data Dentistry: How Data Are Changing Clinical Care and Research,"

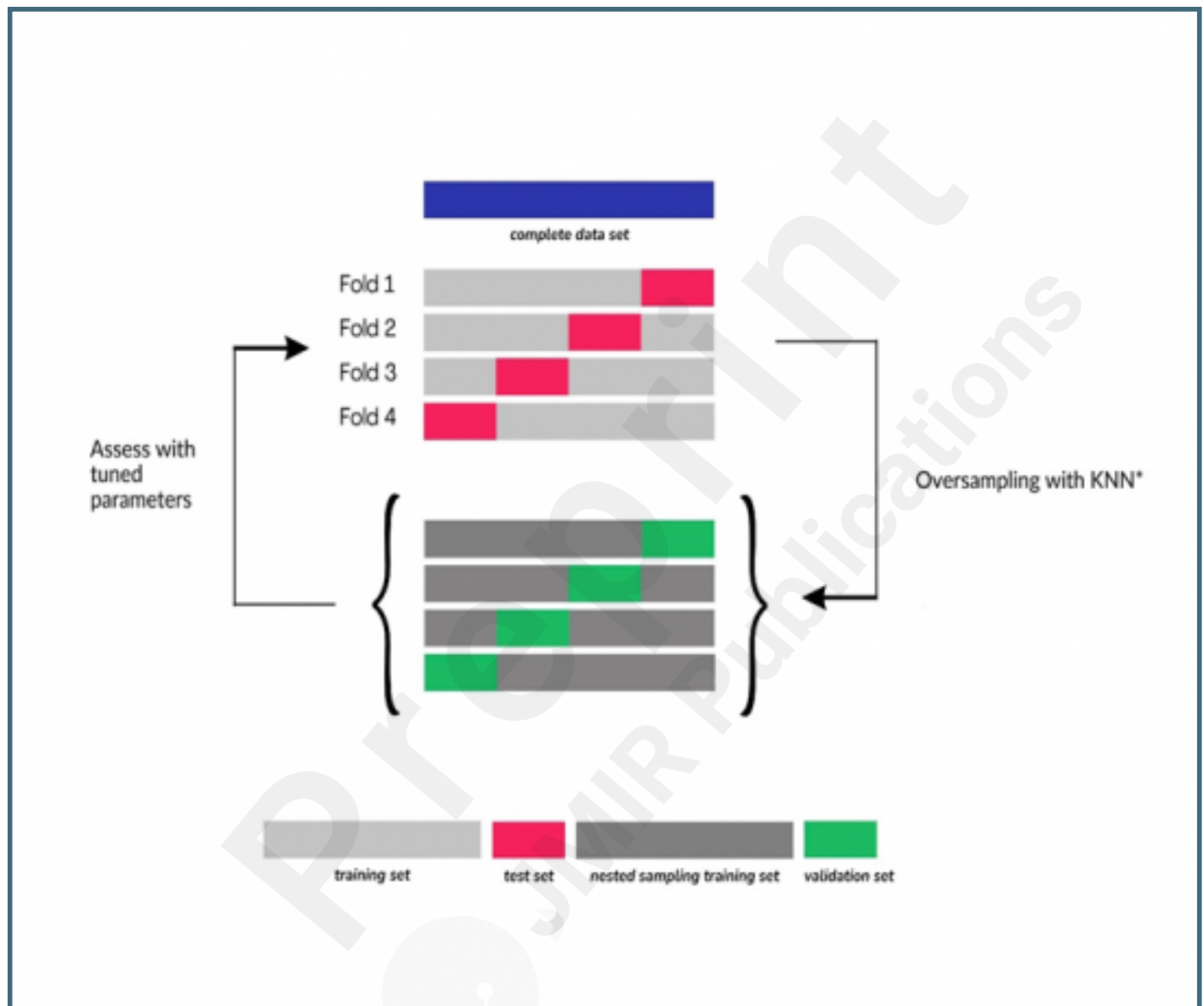
- J Dent Res*, vol. 101, no. 1, pp. 21–29, Jan. 2022, doi: 10.1177/00220345211020265.
16. R. Ducato, “Data protection, scientific research, and the role of information,” *Computer Law & Security Review*, vol. 37, p. 105412, 2020, doi: <https://doi.org/10.1016/j.clsr.2020.105412>.
 17. L. Swinckels *et al.*, “The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review,” *J Med Internet Res*, vol. 26, p. e48320, Aug. 2024, doi: 10.2196/48320.
 18. T. Tiwari, J. S. Patel, and G. G. Nascimento, “Big Data and Oral Health Disparities: A Critical Appraisal,” *J Dent Res*, vol. 104, no. 2, pp. 119–130, Feb. 2025, doi: 10.1177/00220345241285847.
 19. K. Fujiwara *et al.*, “Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis,” *Front Public Health*, vol. 8, p. 178, May 2020, doi: 10.3389/fpubh.2020.00178.
 20. I. Alkhawaldeh, I. Albalkhi, and A. Nashwan, “Challenges and limitations of synthetic minority oversampling techniques in machine learning,” *World J Methodol*, vol. 13, Dec. 2023, doi: 10.5662/wjm.v13.i5.373.
 21. A. L. Suominen-Taipale, E. Widström, and R. Sund, “Association of Examination Rates with Children’s National Caries Indices in Finland,” *Open Dent J*, vol. 3, pp. 59–67, Apr. 2009, doi: 10.2174/1874210600903010059.
 22. Statistics of Finland, “Statistics of Finland.” Accessed: Sep. 03, 2024. [Online]. Available: <https://stat.fi/til/vaerak/index.html>
 23. “FINDATA.” Accessed: Sep. 03, 2024. [Online]. Available: <https://findata.fi/>
 24. “FIMEA - Finnish Medicines Agency.” Accessed: Mar. 28, 2025. [Online]. Available: <https://fimea.fi/en/frontpage>
 25. “Findata – Finnish Social and Health Data Permit Authority.” Accessed: Mar. 28, 2025. [Online]. Available: <https://findata.fi/>
 26. A. I. Ismail *et al.*, “The International Caries Detection and Assessment System (ICDAS): an integrated system for measuring dental caries,” *Community Dent Oral Epidemiol*, vol. 35, no. 3, pp. 170–178, Jun. 2007, doi: <https://doi.org/10.1111/j.1600-0528.2007.00347.x>.
 27. “Tooth Restoration Treatment: Current Care Guidelines.” Accessed: Dec. 03, 2024. [Online]. Available: <https://www.kaypahoito.fi/hoi50117>
 28. “Dental Caries Management: Current Care Guideline.” Accessed: Dec. 03, 2024. [Online]. Available: <https://www.kaypahoito.fi/hoi50127>
 29. H. Hausen, S. Kärkkäinen, and L. Seppä, “Caries Data Collected from Public Health Records Compared with Data Based on Examinations by Trained Examiners,” *Caries Res*, vol. 35, no. 5, pp. 360–365, 2001, doi: 10.1159/000047475.
 30. R Core Team, “R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,” Vienna, Austria. Accessed: Feb. 07, 2025. [Online]. Available: <https://www.r-project.org>
 31. S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, “Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases,” *Comput Biol Chem*, vol. 97, p. 107619, 2022, doi: <https://doi.org/10.1016/j.compbiolchem.2021.107619>.
 32. B. Bischl *et al.*, “mlr: Machine Learning in R,” Aug. 30, 2013. doi: 10.32614/CRAN.package.mlr.
 33. H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and · Tang, “SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling,” *Applied Intelligence*, vol. 51, Mar. 2021, doi: 10.1007/s10489-020-01852-8.
 34. Nicola Lunardon, Giovanna Menardi, and Nicola Torelli, “ROSE: Random Over-Sampling Examples,” Mar. 01, 2013. doi: 10.32614/CRAN.package.ROSE.
 35. Y. Wang, Z. Pan, and Y. Pan, “A Training Data Set Cleaning Method by Classification Ability Ranking for the k-Nearest Neighbor Classifier,” *IEEE Trans Neural Netw Learn Syst*, vol. PP, pp. 1–13, Jun. 2019, doi: 10.1109/TNNLS.2019.2920864.

36. S. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*. 2017.
37. G. S. Collins *et al.*, “Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence,” *BMJ Open*, vol. 11, no. 7, p. e048008, Jul. 2021, doi: 10.1136/bmjopen-2020-048008.
38. L. Toledo Reyes *et al.*, “Early Childhood Predictors for Dental Caries: A Machine Learning Approach,” *J Dent Res*, vol. 102, no. 9, pp. 999–1006, Aug. 2023, doi: 10.1177/00220345231170535.
39. R. A. Bomfim, “Machine learning to predict untreated dental caries in adolescents,” *BMC Oral Health*, vol. 24, no. 1, p. 316, 2024, doi: 10.1186/s12903-024-04073-4.
40. N. A. Selamat, A. Abdullah, and N. Mat Diah, “Association features of smote and rose for drug addiction relapse risk,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 9, pp. 7710–7719, Oct. 2022, doi: 10.1016/j.jksuci.2022.06.012.
41. N. B. Pitts, S. Twetman, J. Fisher, and P. D. Marsh, “Understanding dental caries as a non-communicable disease,” *Br Dent J*, vol. 231, no. 12, pp. 749–753, 2021, doi: 10.1038/s41415-021-3775-4.
42. M. G. Cagetti, G. Bontà, F. Cocco, P. Lingstrom, L. Strohmer, and G. Campus, “Are standardized caries risk assessment models effective in assessing actual caries status and future caries increment? A systematic review,” *BMC Oral Health*, vol. 18, no. 1, p. 123, 2018, doi: 10.1186/s12903-018-0585-4.
43. A. Teshome, A. Muche, and B. Girma, “Prevalence of Dental Caries and Associated Factors in East Africa, 2000–2020: Systematic Review and Meta-Analysis,” *Front Public Health*, vol. 9, 2021, [Online]. Available: <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.645091>
44. J. ALHumaid and M. Bamashmous, “Meta-analysis on the Effectiveness of Xylitol in Caries Prevention,” *J Int Soc Prev Community Dent*, vol. 12, no. 2, 2022, [Online]. Available: https://journals.lww.com/jpcd/fulltext/2022/12020/meta_analysis_on_the_effectiveness_of_xylitol_in.1.aspx
45. F. Cabitza *et al.*, “The importance of being external. methodological insights for the external validation of machine learning models in medicine,” *Comput Methods Programs Biomed*, vol. 208, p. 106288, 2021, doi: <https://doi.org/10.1016/j.cmpb.2021.106288>.
46. T. C.-H. Ng, B. W. Luo, W. Y.-H. Lam, A. Baysan, C.-H. Chu, and O. Y. Yu, “Updates on Caries Risk Assessment—A Literature Review,” *Dent J (Basel)*, vol. 12, no. 10, 2024, doi: 10.3390/dj12100312.
47. S. Chen, X. Guo, and X. Ju, “The Design of Personalized Artificial Intelligence Diagnosis and the Treatment of Health Management Systems Simulating the Role of General Practitioners: International Conference, ICSH 2018, Wuhan, China, July 1–3, 2018, Proceedings,” 2018, pp. 26–40. doi: 10.1007/978-3-030-03649-2_3.

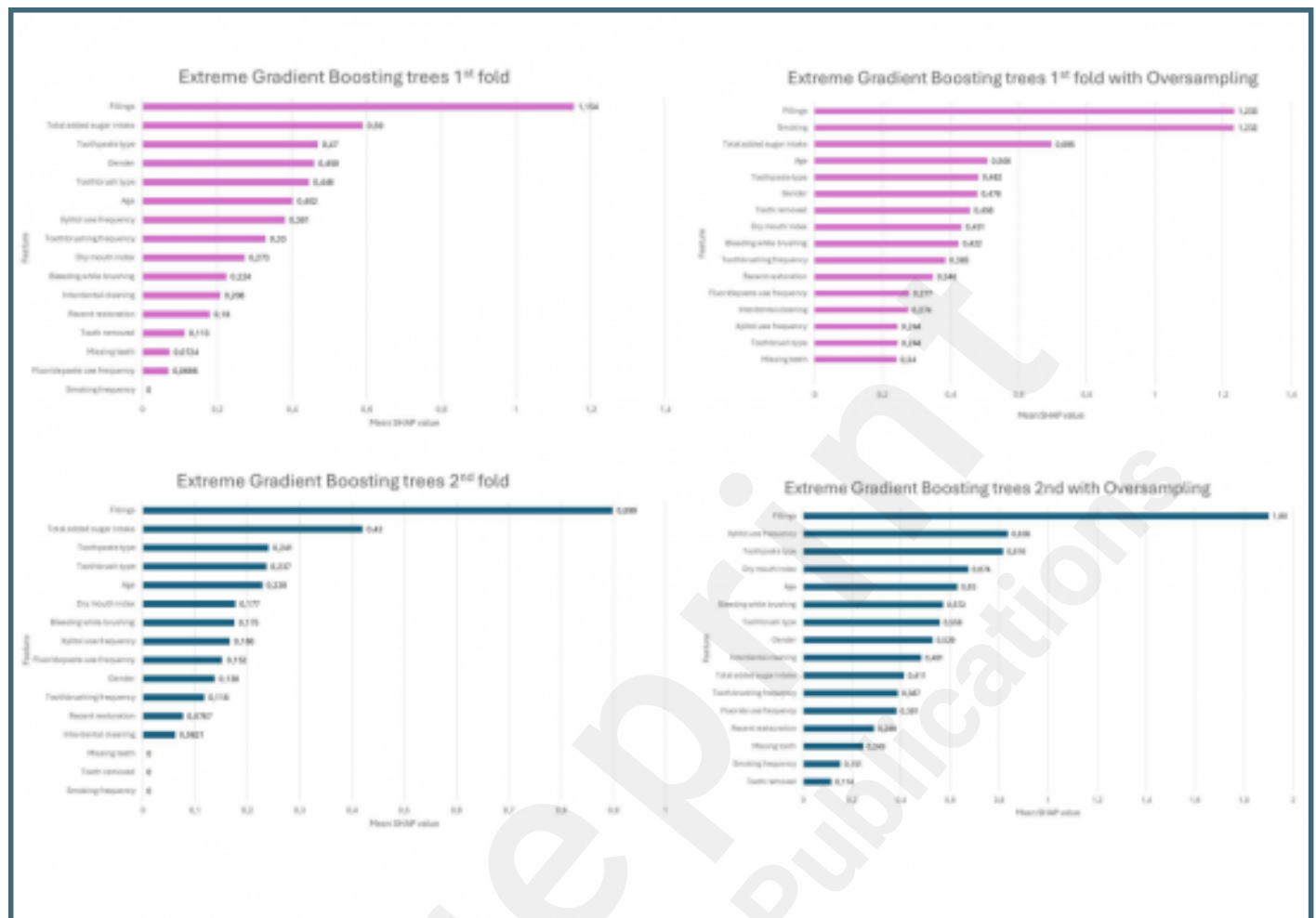
Supplementary Files

Figures

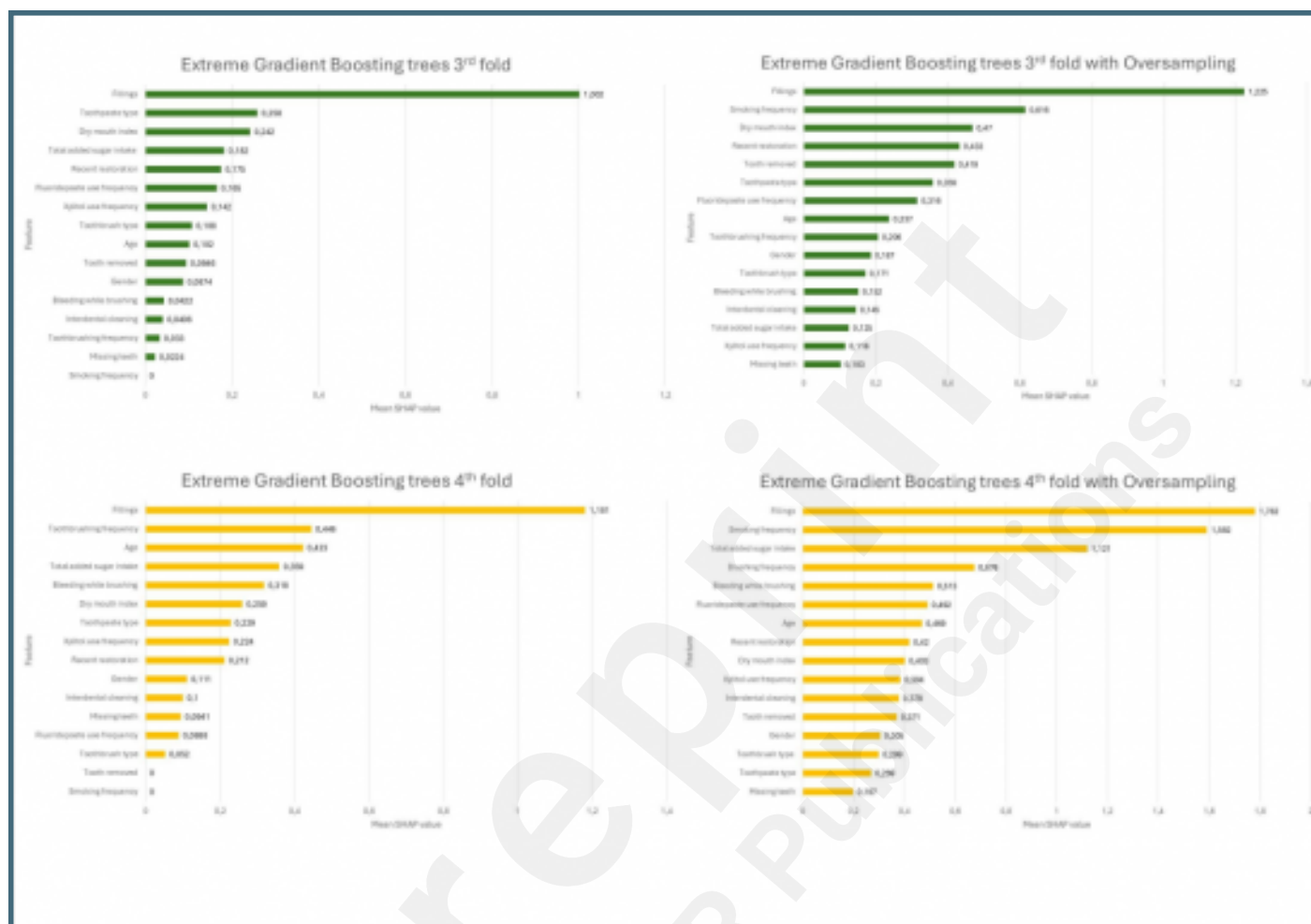
The model training protocol: The training and testing of ML algorithms were performed using 4-fold cross-validation, the nested resampling technique and hyperparameter tuning. The tuned model was then evaluated by a separate validation set, which was distinct from the nested sampling testing set. Finally, the performance of the optimized model was assessed on each fold's test set. This entire training protocol was then repeated for all four folds. The oversampling fold included both the original training set and the nested sampling training set.



SHAP values from folds 1 and 2, before and after oversampling.



SHAP values from folds 3 and 4, before and after oversampling. The absolute SHAP value shows how much a single feature affected the prediction of dental caries. The higher the SHAP value of a feature, the more likely it is to influence the prediction.



Multimedia Appendixes

Table 1. Demographic characteristics and descriptive analysis of the categorical independent variables.

URL: <http://asset.jmir.pub/assets/24b6cda41b8b11eeb4befd6ab8295a89.docx>

Table 2. Performance metrics of ML models before and after oversampling (OS), including the mean (SD) across all four folds.

URL: <http://asset.jmir.pub/assets/bd39973555f830f7ca6d6c7dcfe9d6d4.docx>

