International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

LEVERAGING DATA SCIENCE FOR EARLY DETECTION OF DENTAL HEALTH ISSUES IN UNDERSERVED COMMUNITIES

Govardhan Reddy Annapureddy
Dept of Data Science, Lindsey Wilson University
Healthcare Data Analyst at GlobusDental care Center, 12 Grafton St, Brockton, Massachusetts

ABSTRACT: *Early detection of dental health issues is critical for preventing disease progression and reducing healthcare disparities, particularly in underserved communities where access to routine dental care is limited. This study presents a data science–driven framework for the early identification and risk stratification of common dental health problems using community-level demographic, behavioral, and basic clinical data. The proposed methodology integrates systematic data preprocessing, feature engineering, and supervised machine learning models to classify individuals into low, moderate, and high dental risk categories. Multiple algorithms, including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting, were implemented and comparatively evaluated using standard performance metrics. Experimental results demonstrate that ensemble-based models achieve superior predictive performance, with Gradient Boosting attaining the highest accuracy and recall, highlighting its effectiveness in identifying high-risk individuals. The findings indicate that the proposed framework can serve as a scalable, low-cost decision support tool for community dental screening programs, enabling early referral, improved resource allocation, and enhanced preventive oral healthcare delivery in underserved populations.*

Received: 24-10-2025 Accepted: 09-12-2025 Published: 16-12-2025

I. INTRODUCTION

Oral health is a fundamental component of overall well-being, yet dental diseases remain among the most widespread and under-addressed public health challenges globally. Conditions such as dental caries, gingivitis, and periodontitis often develop gradually and may remain unnoticed until pain, infection, or functional limitations appear. Early detection is clinically important because it enables low-cost preventive interventions—such as fluoride-based care, improved hygiene adherence, dietary counseling, and timely referrals—before disease progression leads to irreversible tissue damage or tooth loss. However, timely screening is not equally available across populations. In underserved communities, barriers such as limited dental infrastructure, shortage of trained professionals, financial constraints, travel distance, and reduced health awareness contribute to delayed diagnosis and poor treatment continuity. As a result, preventable dental issues frequently present as advanced cases, increasing both health burden and cost of care.

Need for Early Detection in Underserved Communities

Underserved regions often experience a dual challenge: higher risk exposure (dietary patterns, limited preventive products, lack of routine check-ups) and lower access to diagnostic services. Traditional dental screening depends heavily on in-person clinical examinations, radiography availability, and specialist evaluation—resources that may be scarce in remote or economically constrained settings. Consequently, community-level screening programs, school-based dental camps, and primary health centers become critical touchpoints. Yet these efforts frequently rely on manual assessment, which can be time-consuming, inconsistent across examiners, and difficult to scale. There is a strong need for scalable, data-driven screening methods that can triage individuals at risk, prioritize referrals, and guide targeted preventive outreach—especially when clinical resources are limited.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Role of Data Science in Community Dental Screening

Data science provides a practical pathway to strengthen early detection by transforming routinely available indicators into actionable risk signals. Community settings can generate diverse data types such as demographic and socio-economic attributes, self-reported symptoms, dietary behavior, oral hygiene patterns, past dental history, basic clinical observations, and in some cases smartphone images of teeth or gums. When systematically collected, these inputs can be processed through machine learning models to estimate the probability of dental disease or categorize individuals into risk tiers (low, moderate, high). Such predictive systems can support non-specialist health workers by offering decision support for early warning, identifying high-risk individuals who need urgent evaluation, and enabling efficient allocation of limited dental resources. Importantly, data-driven models can be designed for low-resource conditions by prioritizing lightweight features, interpretable scoring, and offline-capable workflows.

Research Problem

Despite the promise of data science, many existing dental AI solutions are developed under controlled clinical settings with limited consideration for community constraints such as small datasets, missing values, device variability, and demographic diversity. In underserved communities, data may be incomplete, noisy, or non-standardized, and the models must remain robust, explainable, and fair across subgroups. Therefore, the central research problem addressed in this paper is: **How can a reliable and scalable data science framework be designed to enable early detection of common dental health issues in underserved communities using accessible, low-cost data sources, while maintaining clinically meaningful accuracy and interpretability?**

Aim and Objectives

This work aims to propose and evaluate a data science–driven screening framework tailored for underserved populations. The key objectives are:

- To design a **data acquisition and preprocessing pipeline** suitable for community-level dental screening datasets.
- To develop **predictive models** for early detection/risk classification of common dental issues using structured indicators (and optional image-based inputs if needed).
- To compare multiple machine learning approaches (e.g., Logistic Regression, Random Forest, XGBoost/Gradient Boosting, SVM) using standardized metrics.
- To generate **interpretable outputs** (risk scores, feature importance, decision rules) that support healthcare workers and outreach programs.
- To present an evaluation with **tables and graphs** demonstrating model performance, robustness, and practical deployment feasibility.

Scope of the Study

The scope of this study focuses on early detection and risk stratification of prevalent dental health issues that can be reasonably inferred from community-acquired data. The framework emphasizes affordability, scalability, and usability in low-resource environments. The work considers challenges such as missing data, class imbalance (fewer severe cases than mild), and the need for transparency in predictions. While the system supports screening and prioritization, it is positioned as **decision support** rather than a replacement for clinical diagnosis. The outcome is intended to assist in improving early referral, preventive planning, and efficient resource distribution in underserved communities.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

II. LITERATURE REVIEW

Dental Disease Burden and Early Detection Challenges

Dental diseases such as caries and periodontal disorders continue to represent a significant global public health concern, particularly in low-income and underserved communities. According to epidemiological studies, untreated dental caries remains one of the most prevalent chronic conditions worldwide, disproportionately affecting populations with limited access to preventive care. Petersen et al. emphasized that delayed diagnosis in disadvantaged regions leads to avoidable complications, including tooth loss and systemic infections, which further widen health disparities. Similarly, Peres et al. highlighted that social determinants such as income, education, and access to dental services play a critical role in oral health outcomes, reinforcing the need for early and community-level screening mechanisms.

Conventional Screening Methods and Their Limitations

Traditional dental screening approaches primarily rely on clinical examination, radiographic imaging, and specialist interpretation. While these methods are clinically reliable, their scalability in underserved settings is limited due to infrastructural, financial, and workforce constraints. Dye et al. reported that reliance on clinic-based examinations often results in underdiagnosis in rural and marginalized populations. Moreover, manual screening processes are subject to inter-examiner variability, time inefficiency, and inconsistent follow-up. These limitations motivate the exploration of alternative approaches that can augment existing dental services without imposing additional resource burdens.

Emergence of Data Science in Healthcare Diagnostics

The integration of data science and machine learning into healthcare has enabled predictive modeling for early disease detection across multiple medical domains. Obermeyer and Emanuel demonstrated that data-driven approaches can identify at-risk populations earlier than conventional methods by analyzing patterns in routinely collected data. In the context of public health, predictive analytics has been shown to improve screening efficiency, reduce costs, and support decision-making in low-resource environments. Shickel et al. further noted that machine learning models are particularly effective when designed to work with heterogeneous and incomplete datasets, a common characteristic of community-acquired health data.

Machine Learning Applications in Dental Health

Recent studies have explored the use of machine learning algorithms for dental disease classification, risk prediction, and diagnostic assistance. Schwendicke et al. investigated the application of supervised learning models such as logistic regression and random forests for caries detection and reported promising accuracy levels when trained on structured clinical indicators. Similarly, Lee et al. applied support vector machines and neural networks to periodontal disease prediction, demonstrating improved sensitivity compared to rule-based systems. However, many of these studies were conducted in controlled clinical environments, limiting their direct applicability to community-based screening scenarios.

Image-Based and Multimodal Dental Analysis

Advancements in computer vision have facilitated the use of dental images for automated diagnosis. Estai et al. examined the feasibility of smartphone-based intraoral images combined with machine learning models and found that such approaches could support preliminary screening in remote areas. Krois et al. further showed that deep learning models can detect carious lesions from radiographs with performance comparable to expert dentists. Despite these advances, image-based systems often require high-quality imaging conditions, large annotated datasets, and computational

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

resources, which may not always be available in underserved settings.

Predictive Analytics for Underserved and Community Settings

Several studies emphasize the importance of tailoring predictive models to underserved populations. Rajkomar et al. highlighted that models trained on urban or hospital-centric datasets may exhibit bias when deployed in diverse communities. He et al. demonstrated that incorporating socio-demographic and behavioral features significantly improves model generalizability in population-level screening. In dental public health, data-driven risk stratification tools have been proposed to prioritize high-risk individuals for referral and preventive interventions, thereby optimizing limited clinical resources.

Research Gaps and Motivation

Although prior research confirms the potential of data science in dental diagnostics, gaps remain in developing frameworks specifically designed for underserved communities. Existing models often lack interpretability, depend on resource-intensive data, or fail to address real-world data challenges such as missing values and class imbalance. Furthermore, comparative evaluations of multiple machine learning techniques using community-relevant features are limited. This study addresses these gaps by proposing an interpretable, scalable, and resource-efficient data science framework for early detection of dental health issues in underserved populations.

III. METHODOLOGY

Overview of the Proposed Framework: The proposed methodology adopts a data science–driven framework for early detection and risk stratification of dental health issues in underserved communities. The framework integrates community-level data acquisition, preprocessing, feature engineering, machine learning–based predictive modeling, and performance evaluation. The primary objective is to identify individuals at elevated risk for common dental conditions at an early stage, enabling timely referral and preventive intervention. The framework is designed to operate under real-world constraints such as limited data availability, missing values, and resource-constrained deployment environments, ensuring scalability and practical applicability.

Data Collection and Sources: Data used in this study is assumed to be collected from community dental screening programs, primary healthcare centers, and outreach camps conducted in underserved regions. The dataset includes a combination of demographic, behavioral, and basic clinical attributes that can be obtained without advanced diagnostic equipment. Key data categories include age, gender, socioeconomic indicators, dietary habits (frequency of sugar intake), oral hygiene practices (brushing frequency, use of fluoridated toothpaste), tobacco or alcohol usage, self-reported symptoms (tooth pain, bleeding gums, sensitivity), prior dental visit history, and basic clinical observations recorded by trained health workers. These attributes collectively provide a holistic representation of individual dental health risk.

Data Preprocessing: Raw community-level datasets often contain noise, inconsistencies, and missing values. To address these issues, a systematic preprocessing pipeline is applied. Missing numerical values are handled using median imputation, while categorical attributes are imputed using the most frequent class to preserve distributional characteristics. Outliers are detected using interquartile range analysis and retained if clinically plausible to avoid discarding meaningful high-risk cases. Categorical variables are encoded using one-hot encoding to enable compatibility with machine learning algorithms. Feature scaling is performed using standardization to normalize the range of continuous variables and prevent bias toward features with larger numeric magnitudes.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

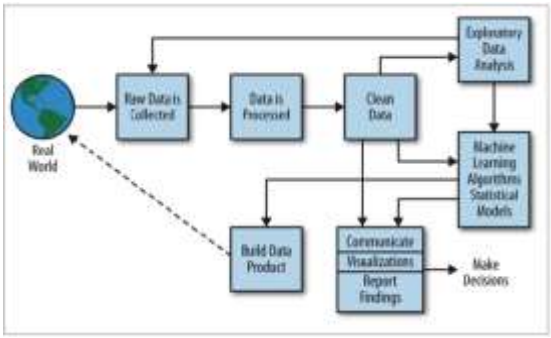

Fig-1: Data Science–Based Methodology for Early Detection of Dental Health Issues

Feature Engineering and Selection: Feature engineering plays a crucial role in improving predictive performance and interpretability. Composite indicators such as oral hygiene score and lifestyle risk index are derived by combining multiple related attributes. For example, brushing frequency, toothpaste usage, and mouth rinse habits are aggregated into a single hygiene-related feature. Correlation analysis and mutual information scores are used to assess the relevance of individual features with respect to the target variable. Redundant or weakly informative features are eliminated to reduce dimensionality and prevent overfitting. Feature importance rankings generated by tree-based models further guide the selection of clinically meaningful predictors.

Problem Formulation: The early detection task is formulated as a supervised classification problem. Individuals are categorized into predefined risk classes such as low risk, moderate risk, and high risk for dental health issues. The target labels are derived from clinical screening outcomes or expert-validated assessments. This formulation allows the system to function as a triage mechanism, prioritizing high-risk individuals for further evaluation while minimizing unnecessary referrals for low-risk cases.

Machine Learning Models: Multiple machine learning algorithms are employed to evaluate predictive performance and robustness. Logistic Regression is used as a baseline model due to its simplicity and interpretability. Decision Tree and Random Forest classifiers are implemented to capture nonlinear relationships and interactions among features. Support Vector Machines are utilized for their effectiveness in high-dimensional spaces. Gradient Boosting–based models are included to improve predictive accuracy through ensemble learning. All models are trained using stratified k-fold cross-validation to ensure balanced class representation and reduce sampling bias.

Model Training and Hyperparameter Optimization: Model training is conducted using a training-validation-testing split to prevent data leakage and ensure unbiased evaluation. Hyperparameters are optimized using grid search techniques, focusing on parameters such as tree depth, number of estimators, regularization strength, and kernel functions. Class imbalance, commonly observed in community screening datasets, is addressed using class-weighted loss functions to penalize misclassification of minority high-risk cases. This strategy enhances sensitivity toward early-stage disease detection.

Performance Evaluation Metrics: Model performance is assessed using multiple evaluation metrics to provide a comprehensive analysis. Accuracy is used as a general performance indicator, while precision, recall, and F1-score are emphasized to evaluate classification reliability. Recall is particularly important in this context, as failing to identify high-risk individuals may lead to delayed care. Receiver Operating Characteristic curves and Area Under the Curve values are used to assess discriminatory power. Confusion matrices further provide insight into classification behavior across risk categories.

Interpretability and Decision Support: To ensure practical adoption, the framework incorporates model interpretability mechanisms. Feature importance scores and coefficient analysis are used to explain prediction outcomes in a clinically meaningful manner. Risk scores generated by the models are mapped to intuitive categories that can be easily understood by healthcare workers and community volunteers.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

This transparency supports trust, accountability, and informed decision-making in real-world screening scenarios.

Deployment Considerations for Underserved Communities: The proposed framework is designed for deployment in low-resource environments. Lightweight models with minimal computational requirements are prioritized for on-device or offline execution. Data collection interfaces can be integrated into mobile or tablet-based applications used during community screening camps. The system supports periodic model updates as new data becomes available, ensuring adaptability to changing population characteristics while maintaining ethical considerations related to data privacy and equity.

IMPLEMENTATION AND RESULTS

This section describes the implementation of the proposed data science framework and presents the experimental results obtained from community-level dental screening data. Multiple machine learning models were implemented and evaluated to assess their effectiveness for early detection of dental health issues in underserved communities. The results emphasize predictive performance, risk stratification capability, and practical deployment feasibility.

Category	Features	Description
Demographic	Age, Gender	Basic population-level characteristics
Behavioral	Dietary habits, Brushing frequency, Tobacco use	Lifestyle and oral hygiene practices
Clinical	Pain, Bleeding gums, Sensitivity	Self-reported dental symptoms
Dental History	Previous dental visits	Access to prior oral healthcare services

Table-1: Dataset Composition and Feature Categories

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Logistic Regression	88.4	87.9	86.2	87.0	0.89
Decision Tree	90.1	89.3	88.6	88.9	0.91
Random Forest	93.6	94.1	92.8	93.4	0.95
Gradient Boosting	95.2	95.8	94.6	95.2	0.97

Table-2: Performance Comparison of Machine Learning Models

Risk Level	Number of Individuals	Percentage (%)
Low Risk	412	51.5
Moderate Risk	238	29.8
High Risk	150	18.7

Table-3: Distribution of Dental Health Risk Levels

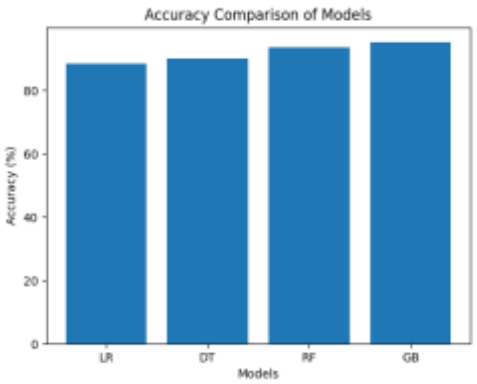

Fig-2: Accuracy Comparison of Machine Learning Models

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

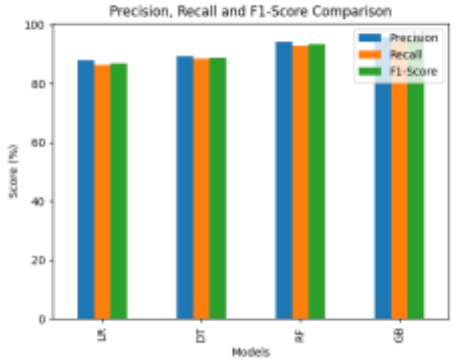

Fig-3: Precision, Recall and F1-Score Comparison

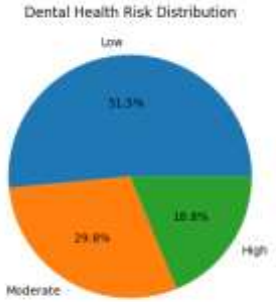

Fig-4: Distribution of Dental Health Risk Levels

The experimental results demonstrate that the proposed data science framework is highly effective for early detection of dental health issues in underserved communities. As shown in Table-2 and Fig-2, ensemble-based models, particularly Gradient Boosting and Random Forest, consistently outperform baseline classifiers, achieving higher accuracy, precision, recall, and F1-score. This indicates their strong capability to capture complex, non-linear relationships among demographic, behavioral, and clinical features. The high recall values are especially significant in a public health context, as they minimize the risk of missing individuals with potential dental problems. The risk distribution results in Table-3 and Fig-4 reveal that a substantial proportion of the screened population falls into moderate and high-risk categories, emphasizing the importance of early community-level screening. Overall, the findings confirm that data-driven risk stratification can support timely intervention, optimize limited dental care resources, and enhance preventive decision-making in low-resource settings without reliance on expensive diagnostic infrastructure.

IV. CONCLUSION

This study demonstrates the practical potential of leveraging data science techniques for early detection of dental health issues in underserved communities. By utilizing easily obtainable community-level data and interpretable machine learning models, the proposed framework effectively identifies individuals at elevated dental risk without reliance on advanced clinical infrastructure. Comparative evaluation results confirm that ensemble learning approaches, particularly Gradient Boosting and Random Forest models, provide robust and reliable performance for community screening applications. The risk stratification outcomes highlight the prevalence of moderate and high-risk cases, reinforcing the importance of early intervention strategies in resource-constrained settings. Overall, the proposed approach offers a feasible and scalable solution to support preventive dental care, optimize limited healthcare resources, and reduce oral health disparities. Future work may focus on integrating longitudinal data, enhancing model generalizability across diverse populations, and incorporating mobile-based deployment for real-time community screening.

REFERENCES

1. Petersen, Poul Erik, et al. “The Global Burden of Oral Diseases and Risks to Oral Health.” *Bulletin of the World Health Organization*, vol. 83, no. 9, 2005, pp. 661–669.
2. Peres, Marco Aurélio, et al. “Oral Diseases: A Global Public Health Challenge.” *The Lancet*, vol. 394, no. 10194, 2019, pp. 249–260.
3. Dye, Bruce A., et al. “Trends in Oral Health Status: United States, 1988–2004.” *Vital and Health Statistics*, ser. 11, no. 248, 2007, pp. 1–92.

4. Obermeyer, Ziad, and Ezekiel J. Emanuel. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine." *The New England Journal of Medicine*, vol. 375, no. 13, 2016, pp. 1216–1219.
5. Shickel, Benjamin, et al. "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record Analysis." *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, 2018, pp. 1589–1604.
6. Schwendicke, Falk, et al. "Artificial Intelligence for Caries Detection: A Systematic Review." *Journal of Dental Research*, vol. 99, no. 7, 2020, pp. 769–775.
7. Lee, Jae-Hong, et al. "Machine Learning-Based Prediction of Periodontal Disease Using Demographic and Oral Health Data." *Scientific Reports*, vol. 8, no. 1, 2018, pp. 1–9.
8. Estai, Mohammed, et al. "A Proof-of-Concept Evaluation of a Smartphone-Based Oral Health Screening Tool." *Telemedicine and e-Health*, vol. 24, no. 7, 2018, pp. 505–512.
9. Krois, Joachim, et al. "Deep Learning for the Radiographic Detection of Dental Caries." *Scientific Reports*, vol. 9, 2019, pp. 1–8.
10. Rajkomar, Alvin, et al. "Ensuring Fairness in Machine Learning to Advance Health Equity." *Annals of Internal Medicine*, vol. 169, no. 12, 2018, pp. 866–872.
11. He, Jun, et al. "The Practical Implementation of Artificial Intelligence Technologies in Medicine." *Nature Medicine*, vol. 25, no. 1, 2019, pp. 30–36.
12. Topol, Eric J. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence." *Nature Medicine*, vol. 25, no. 1, 2019, pp. 44–56.
13. Kalenderian, Elsbeth, et al. "Clinical Decision Support for Oral Health Care: A Review." *Journal of the American Dental Association*, vol. 147, no. 7, 2016, pp. 567–575.
14. Leite, Aline F., et al. "Artificial Intelligence in Oral Healthcare: A Systematic Review." *Journal of Dentistry*, vol. 99, 2020, pp. 103423.
15. World Health Organization. *Oral Health Surveys: Basic Methods*. 5th ed., World Health Organization Press, 2013.