International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predictive Analysis of Periodontal Disease Progression Using Machine Learning: Enhancing Oral Health Assessment and Treatment Planning

T. K. Lakshmi[1], Dheeba J.[*2]

Submitted: 26/05/2023 **Revised**: 09/07/2023 **Accepted**: 28/07/2023

Abstract: Artificial Intelligence (AI) has revolutionized various aspects of our lives, offering solutions to numerous problems and bridging gaps between reality and business. Within the realm of AI, emerging technologies such as machine learning and deep learning have become prominent in transforming the way we analyze data, make decisions, and address challenges. With the exponential growth of data usage and storage, these technologies have assumed a vital role in data analytics, storage management, and decision-making processes. As digital transformation continues to reshape industries and services worldwide, the healthcare sector, including oral health services, necessitates complete digitalization. Oral diseases, prevalent across all age groups, often go neglected until they reach a painful and severe stage, leading to potential tooth damage. To counteract such consequences, the field of dentistry requires digitalization for timely diagnosis, effective decision-making, patient management, and predictive capabilities. This research paper focuses on leveraging these emerging technologies to predict the progression of Periodontitis, a common oral disease. In this study, machine learning classifiers are employed to analyze and predict the disease. Additionally, cross-validation methods, feature extraction techniques, and ensemble learning strategies are implemented and evaluated. The performance metrics are compared for various classifiers including Naïve Bayes, Support Vector Machine, Random Forest, Logistic Regression, K Nearest Neighbors, and Decision Tree classifiers. These classifiers are applied to a dataset of 1000 periodontitis patients, resulting in impressive accuracies of 95.5%, 100%, 100%, 100%, 99.5%, and 99% respectively, for the classification of chronic localized and chronic generalized periodontitis. Through this research, we highlight the potential of machine learning and other AI techniques in revolutionizing the field of dentistry. By harnessing the power of predictive analysis, accurate diagnosis, timely interventions, and improved patient management can be achieved, ultimately enhancing oral health outcomes. This study serves as a significant step towards integrating advanced technologies into dentistry, contributing to the overall digital transformation of the healthcare industry.

Keywords: *Periodontitis, dental disease prediction, machine learning, feature selection, feature engineering, accuracy, oral health, decision making, COVID 19, supervised learning, classifiers, digitalization, artificial intelligence, data analysis, gum disease, cross validation*

1. Introduction

Taking care of your teeth is important for your general health [1]. Studies have shown a strong link between oral infections like gum disease and systemic diseases like coronary heart disease (atherosclerosis and myocardial infarction), stroke, infective endocarditis, bacterial pneumonia, and even pregnancy complications like low birth weight babies, pre term babies, and a higher risk of miscarriage. Gum disease is the most prevalent persistent inflammatory illness, and it affects many other regions of the body. If you don't brush and floss regularly, the bacteria in your mouth will form a thin film called plaque, which, if left untreated, can harden into tartar that can't be removed by brushing. Since most people only visit the dentist when they're in pain, neglecting regular checkups can lead to serious issues like periodontitis and tooth loss. Gingivitis is an inflammation of the gums that, if left untreated, can progress to periodontitis. Periodontitis destroys the ligaments and alveolar bone that hold teeth in place, which can be extremely painful. Some of the most frequent risk factors for periodontitis are age, sex, behaviours including smoking and drinking alcohol, socioeconomic status, and certain systemic disorders [4]. Recent investigations have shown a strong correlation between COVID - 19 patients' drug usage and oral issues [5]. New research suggests that the black, yellow, and white fungus that has become common in covid recovered patients' needs the attention of an oral surgeon or it may result in new dental problems; similarly, the entire world is suffering from different COVID - 19 variants, many cases ended with deaths, and those who got recovered are still facing other health problems. According to a 2004 study conducted by the Dental Council of India, 57.7% of Indians aged 12-15 have periodontal disease, 67.7% of Indians aged 35-44 have periodontal disease, and 89.6% of Indians aged 65-74 have periodontal disease. Periodontal disease affects between 20 and 50 percent of people worldwide. Heart issues are more common in

[1] Research scholar
SCOPE, Vellore Institute of Technology, Vellore
tklakshmiphd@gmail.com
[1] Associate Professor
SCOPE, Vellore Institute of Technology, Vellore
** Corresponding Author Email: dheeba.j@vit.ac.in*

adults over the age of 65, and people with gum disease have a 19% higher chance of developing one. Patients with diabetes who also suffer from gum disease have a 3.2-fold increased chance of dying [6].

While periodontitis cannot be healed but may be managed with frequent dental visits and follow-up care, the early stage of gingivitis can be cured with correct prompt treatment and keeping excellent oral hygiene. Therefore, maintaining a high standard of oral hygiene is crucial to one's overall health. Due to the severity of the consequences of ignoring periodontitis, early detection of the condition is of the utmost importance. Therefore, the present research incorporates the use of machine learning techniques to anticipate the onset of periodontal disease. The field of artificial intelligence (AI) encompasses a broad variety of new technologies that have found useful applications in healthcare and related fields. Data science without exception for pre-processing and cleaning huge data, feature engineering and extraction for appropriate data analysis, machine learning classifiers and regression algorithms, deep learning methodologies through which an image data set can be used by convolutional neural networks, recurrent neural networks, deep reinforcement learning, self-organizing maps, etc., are all examples of how AI and its subsets can be put to use in the healthcare industry. By identifying the most important characteristics in target forecasts, feature engineering is the most effective approach for reducing high-dimensional data sets using current algorithms [8]. In this research, we apply a variety of feature selection techniques to determine the best combination of ranking, threshold, and other factors for accurate target prediction. It's possible that the accessible data in real time is of a high dimension, taking up a lot of room and processing time. Additionally, not all criteria are equally important when making predictions about targets. As a result, selectkbest, information gain, feature importance method, Pearson correlation, recursive feature selection methods, gradient boosting methods, and many others are required for real-time applications to reduce high-dimensional data to lower dimensions. Classifiers are also used to predict periodontal disease, with performance measures including as accuracy, precision, recall, and f score being used to evaluate and compare the various models' efficacy. Thus, the current study aids the dentist in the digital method of case prediction and acts as a decision-making supporting tool, and this method can also be used for demonstration purposes to help aspiring dentists, newcomers to the field, and house surgeons understand the relationship between different features that are responsible in predicting the target variable.

2. Recent Studies in This Area

Automated image-based periodontitis screening was the focus of 2017 research by Asghar Tabatabaei Balaei, Philip de Chazal, and colleagues. Twenty patients in Sydney and twenty-four patients in Berlin had their intra-oral photographs taken. They employed a recursive feature elimination approach of feature selection, developed a logistic regression classifier, and got an accuracy of 66.7% and a precision of 60% after adjusting brightness and contrast during pre-processing. Also, they have used picture pairs of patients before and after therapy to test the classifier's ability to distinguish between healthy and ill individuals [9].

In 2019, photos were employed as input into convolutional neural networks (CNN) for periodontal detection by Jaehan Joo, Sinjin Jeong, et al. Images were resized, cropped, and trimmed as part of the pre-processing phase. The 5 convolution and pooling layers, 1 fully linked layer, and 1 output layer make up the CNN. Accuracy of 81% was achieved [10] using the ReLu and softmax activation routines.

In 2020, Maryam Farhadian, Parisa Shokouhi, and coworkers used Support vector machines using a variety of kernel functions to identify 300 patients with periodontitis. To prevent model overfitting, we performed 10-fold cross validation and found that the linear kernel had an accuracy of 81.7%, the polynomial kernel of 81.1%, the radial kernel of 88.7%, and the sigmoid kernel of 87.5% [11].

Machine learning classifiers have been effectively applied and deployed for illness prediction with excellent acceptable accuracies [12], as examined and tabulated by Gopi Battineni, Getu Gamo Sagaro etal in 2020.

Mouthwash samples were used to predict periodontitis by Eun-Hye Kim, Seunghoon Kim, et al. utilising machine learning methods such as random forest, SVM, logistic regression, and neural network with 5 fold cross validation in the R package in 2020 [13].

Caries, gingivitis, periodontitis, root fractures, cysts, orthodontic extractions, cephalometric analysis, and many more were among the dental conditions that Sanjeev B. Khanagar, Ali Al-ehaideb, et al. studied and tabulated in 2021 [14].

Using machine learning classifiers to predict a target variable, S. B. Kotsiantis in 2011 emphasised the importance of relevant, irrelevant, and redundant information. Extensive examination of filter, wrapper, and other hybrid approaches, as well as a plethora of search algorithms, were all laid out in detail in this study [15].

Forward and backward selection approaches, as well as the Pearson coefficient, information gain, feature ranking, wrapper, filter, and embedding methods, were all discussed in detail by Isabelle Guyon and Andre Elisseeff in a 2003 work [16].

Using two datasets and 10 fold cross validation, Jie Cai, Jiawei Luo, et al. (2018) studied and implemented a variety of feature selection techniques, including supervised filter and wrapper approaches, ReliefF, Information Gain, Ensemble feature selections, and SVM-Recursive feature selection techniques.

In 2017, Chuan Liua, Wenyong Wang, and colleagues implemented a novel feature selection method called LW Sequential forward selection. Unlike traditional wrapper filter methods, LW Sequential forward selection is inexpensive and yields superior results due to its focus on measuring degree of separation with 2 linearly dividable classes.

All of the known feature selection approaches were published and theoretically described in great detail by Jovic, K. Brkic et al. in 2015. There was a comprehensive tabular summary of all the current techniques and their many uses in the article. All of the techniques were summarised in a concise table [19].

In 2010, Luka Cehovin and Zoran Bosnic implemented and compared five feature selection methods with various classifiers across a total of 20 datasets spanning a variety of domains, with each step involving the removal of the feature with the lowest quality index, the process being repeated, and the classification accuracy being recorded.

3. Proposed Workflow

Step 1. Data Collection and Entry:
Step 2. Gather a dataset containing relevant information for the prediction of periodontitis.
Step 3. Ensure the data is entered accurately and completely.
Step 4. Data Pre-processing, Cleaning, and Outlier Removal:
Step 5. Handle missing data by imputation or removal based on the nature and quantity of missing values.
Step 6. Clean the data by addressing inconsistencies, errors, and anomalies.
Step 7. Detect and remove outliers that may adversely affect the analysis.
Step 8. Implementation of Machine Learning Supervised Classifiers and Performance Evaluation:
Step 9. Choose appropriate machine learning supervised classifiers, such as Naïve Bayes, Support Vector Machine, Random Forest, Logistic Regression, K Nearest Neighbors, and Decision Tree.
Step 10. Train the classifiers using the pre-processed data.
Step 11. Evaluate the performance of each classifier using appropriate metrics like accuracy, precision, recall, and F1 score.
Step 12. Implementation of Cross-validation Techniques and Performance Evaluation:
Step 13. Implement four types of cross-validation techniques: K-fold Cross-Validation, Leave-One-Out Cross-Validation, Stratified Cross-Validation, and Shuffle-Split Cross-Validation.
Step 14. Apply each cross-validation technique to the trained classifiers.
Step 15. Evaluate the performance of each classifier using the chosen cross-validation techniques and compare the results.
Step 16. Application of Feature Selection Methods and Performance Evaluation:
Step 17. Employ feature selection methods such as Recursive Feature Elimination, Principal Component Analysis, or SelectKBest.
Step 18. Select the most relevant features from the dataset based on the chosen method.
Step 19. Re-train the classifiers using the selected features.
Step 20. Evaluate the performance of each classifier using the reduced feature set and compare the results.
Step 21. Application of Ensemble Methods and Performance Evaluation:
Step 22. Utilize ensemble methods like Bagging, Boosting, or Stacking.
Step 23. Combine the predictions of multiple classifiers using ensemble techniques.
Step 24. Train the ensemble models using the pre-processed data.
Step 25. Evaluate the performance of each ensemble method and compare the results.
Step 26. Comparison of Results:
Step 27. Compare and analyze the performance of all the different techniques employed in previous steps.
Step 28. Consider metrics such as accuracy, precision, recall, and F1 score to assess the effectiveness of each approach.
Step 29. Identify the most accurate and reliable method for predicting periodontitis based on the results obtained.

By following this algorithm, the study aims to optimize the prediction of periodontitis using machine learning, cross-validation, feature selection, and ensemble techniques. The ultimate goal is to determine the most effective approach for accurate diagnosis and timely intervention in the field of dentistry.

4. Proposed Work

With the participants' permission, we obtained data on 1,000 periodontic patients from K T Super speciality Dental hospital Tirupati using 26 characteristics. Both diabetic and non-diabetic patients are represented in the dataset used in this study. Tonetti and Greenwell's publication isn't the only one to reference the American Academy of Periodontology's 2017 international workshop on periodontitis for information on how to recognise the various stages and grades of gum disease. Grading is acquired from risk factors and evidence during testing, and the disease's severity is classified into four phases based on clinical attachment loss, bone loss, probing depth, and tooth loss, which may be either systemic or confined [21]. Those who smoke, have diabetes, practise poor hygiene, have high stress levels, have immune deficiencies, have crooked teeth, or are experiencing hormonal imbalances or changes are at a much higher risk for the disease and its complications [22]. Other research [23] suggests that clinical attachment losses, pocket depth, and haemorrhage on probing are the most important variables in disease grading and staging. Since prior studies shown a considerable correlation between diabetes and periodontitis, our present analysis included data from around 50% diabetic individuals. Diabetics have a higher risk of developing periodontitis. Age, stress, genetics, hormone imbalances, osteoporosis, smoking, HIV, hematologic illnesses, and drugs are all linked to periodontitis [24]. As a result, the present work employs a wide range of machine learning technologies to analyse and predict periodontitis, including feature selection methods to enhance the model, all four cross validation methods to prevent over fitting, and various ensemble methods to acquire the best performance in terms of accuracy. Ultimately, six machine learning models are constructed, and the results are compared.

4.1 Information Gathering

Patients' demographics (age, gender, number of teeth, whether they were diabetic or not, smoking and alcohol use, pan and betel nut consumption), clinical findings (gingival index, periodontal index, tooth mobility grades 1-3, furcation grades 1-3, radiological findings (alveolar bone loss grades 1-3, black triangle grades 1-3, severity impac), and radiological findings were collected and analysed. The goal value is the patient's periodontal disease severity, which may be either chronic generalised periodontitis (CGP), chronic localised periodontitis (CLP), or chronic generalised gingivitis (CGG). Patients' records including CGG were scrubbed from the dataset because of the present study's narrow emphasis on periodontitis. Thus, the present dataset has 26 characteristics, including the target variable, and 1000 records including data from 493 CGP and 507 CLP patients. As stated in Table 1, the aforementioned characteristics are determined using industry-standard techniques [25] [26] [27] [28] [29] [30]. All of the computed and measured data is input by hand into an Excel spreadsheet.

S.NO	METHOD	FEATURE OBTAINED
1	Loe and silness	Gingival index
2	Russell’s index	Periodontal index
3	Goldman classification	Furcation
4	Miller’s classification	Tooth mobility
5	Goldman and Cohen	Alveolar bone loss
6	Nordland and Tarnow	Black Triangle

Table 1: Standard methods used for measuring few parameters considered in the dataset

4.2 Data Preprocessing

Data in the actual world may be inconsistent and noisy, and there may be human or machine mistake during manual input. Preprocessing the data before beginning to construct the model is essential for achieving reliable outcomes. There are many phases involved in prepping the data, including cleaning, integrating, transforming, and reducing [31]. Treatment of missing values, inconsistent data, normalisation, standardisation, aggregation, feature selection and extraction, dimensionality/data reduction, treatment of outliers, scaling, augmentation, treatment of imbalanced data, and conversion of categorical to numerical are all examples of pre-processing steps [32]. We use standard scaler to normalise the data, we convert the categorical values to numeric using functions we write, we display a boxplot to identify outliers, and we handle them using the "winsorizing" approach [33].

4.3 Classifiers derived from machine learning

Six different machine learning classifiers—Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), Logistic Regression

(LR), and Decision Tree (DT)—are chosen after preprocessing techniques and outliers treatment are completed in order to test and predict gum disease using all input features without applying any feature selections. Python 3.8.5 in Jupyter notebook on Windows 10 Pro Lenovo G50 64-bit with Intel(R) Core(TM) i3-5010U CPU @ 2.10GHz for 1000 records with an 80%/20% split between train and test, and the Scikit and data science libraries and packages.

4.4 Cross-Validation Methods

Overfitting and underfitting are more likely to occur in the model. In the case of overfitting, the model does well on the training data but not on the test data. Under fitting has poor results on both the training set and the test set. The efficiency of the classifier or model will suffer as a direct result of these issues. Regularisation techniques such as L1 (Lasso regression), L2 (Ridge regression), and drop outs [34] may be used to prevent overfitting, as can early halting or a slower learning rate, cross validations, pre and post pruning procedures, feature selections, enlarging training data, and feature choices. Underfitting may be avoided if enough data is supplied to boost learning. Our data set is utilised in conjunction with cross-validation methods to produce a machine learning model that is both accurate and robust when applied to fresh data (avoiding overfitting) [35]. K(10) Fold cross validation, Stratified K(10) fold, hold out, and leave one out approaches are used in this study to compare the performance metrics of each chosen machine learning classifier and to prevent over fitting. For this investigation, we use a cross-validation scheme with a 10-fold safety margin. In this phase, we compare the findings acquired in the previous step with the results obtained after using all four cross validation approaches.

4.5 Methods for selecting features to implement

Reducing the dimensionality of inputs, irrelevant and redundant features, and improving the performance of machine learning models all need feature selection. Filters, wrappers, and embedding methods are the three means by which features are chosen. Filtering techniques make use of classifiers to evaluate features and then choose features based on their relative ranking. Wrapper approaches implement the classifier using a reduced set of characteristics. Better feature subset selection leads to improved performance, therefore the cycle continues until that point is reached. In the case of embedded approaches, feature selection occurs during the learning process itself [36]. After developing a machine learning model, numerous different feature selection approaches were applied and compared for their accuracy. Fifteen, eleven, and eleven features are chosen after being evaluated using selectkbest, information gain, sequential feature selection - step forward and step backward feature selection techniques. The outcomes of feature selection are then examined across a variety of machine learning classifiers, including Nave Bayes, Support vector machine, Random forest, K closest neighbour, logistic regression, and decision tree models.

4.6 Ensemble Techniques

In this section machine learning, combining the judgements made by many Meta algorithms may reduce volatility and bias and improve model performance as a whole. The term "Ensemble Methods" describes these procedures, which may operate in either a sequential or parallel fashion. Common strategies include bagging, boosting, and stacking [37]. Decisions based on the review of numerous models are more likely to be correct than those based on the evaluation of a single model. Ensembles are based on this principle. Ensemble stands out from other approaches that employ max voting, averages, or weighted averages due to its distinctive voting procedure. Bagging techniques, such as random forest boosting techniques, are also widely used [38]. The present research uses ensemble techniques and compares the outcomes. Using the obtained performance metrics, we conclude that the methods led to more accurate and better predictions of periodontal disease, with 5 ensemble methods giving 100% accuracy predictions of periodontal disease. These methods include bagging, extremely randomised tree ensemble, ensemble by random forest, ada boost (DT), gradient boost ensemble, histogram based gradient boosting, voting classifier multi model ensemble, and XG boost ensemble.

4.7 Analysing the Data

In this analysis, we evaluate information from 1,000 patient records of periodontal patients, including those of diabetics who exhibited 26 characteristics. In step 4, we preprocess this dataset as described above.Two: six machine learning classifiers are chosen after preprocessing, and successful implementation and prediction results are achieved with these classifiers. For NB, SVM, RF, KNN, LR, and DT, the corresponding levels of accuracy are 95.5%, 100%, 100%, 99.5%, 100%, and 99%. After that, we implement four different cross-validation approaches to make sure our classifiers aren't overfitting, and then we use feature-selection strategies to get more accurate results. When compared to the boosting findings, the final predictions achieved by using ensemble techniques. Therefore, several machine learning algorithms are used to the dataset in an effort to provide the best prediction outcomes; after the goal is reached, the granular outputs acquired are addressed in Section 5 below.

5. Results & Discussion

The current study's dataset includes radiological and clinical results from periodontal patients, as well as demographic information such as smoking status, number of teeth, and diabetes status. The research has employed many machine learning methods, yielding respectable prediction accuracies. Accuracy, f1 score, recall, precision, receiver operating characteristics, correlation coefficients, and many more are only some of the metrics that may be used to evaluate a model's efficacy [39][40]. Python in Jupyter notebook is used to implement Scikit learn and other data science tools and packages.

S.No	Algorithm 80-20% split	Output label	Precision	Recall	F1 score	Accuracy (%) (After outliers treatment)
1	Naïve Bayes	0	0.96	0.95	0.96	95.5
		1	0.95	0.96	0.95	
2	SVM	0	1.00	1.00	1.00	100
		1	1.00	1.00	1.00	
3	Random forest	0	1.00	1.00	1.00	100
		1	1.00	1.00	1.00	
4	KNN	0	0.99	1.00	0.99	99.5
		1	1.00	0.99	1.00	
5	Logistic regression	0	1.00	1.00	1.00	100
		1	1.00	1.00	1.00	
6	Decision tree	0	0.98	1.00	0.99	99
		1	1.00	0.98	0.99	

Table 2: Performance Evaluation Results

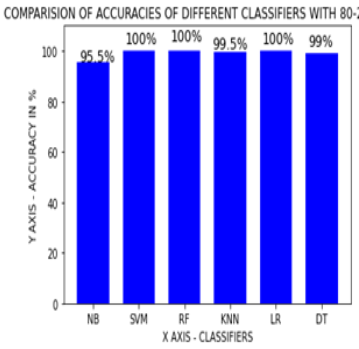

Fig 1 (a)

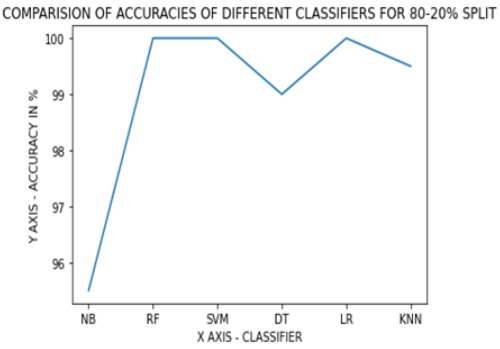

Fig 1 (b)

Fig 1 (a) & 1 (b): Comparison of accuracy of NB, SVM, RF, KNN, LR, DT classifiers

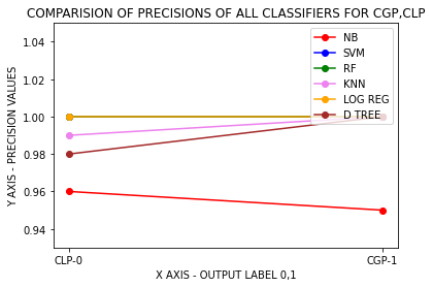

Fig 2 (a)

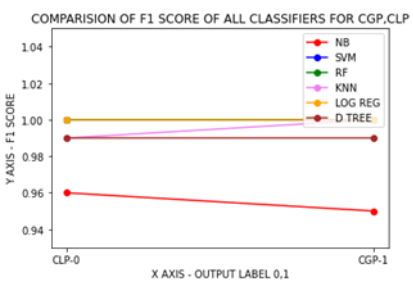

Fig 2 (b)

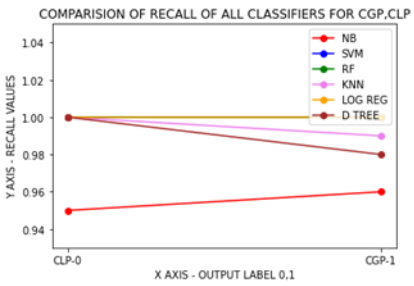

Fig 2 (c)

Fig 2 (a) Precision **2 (b)** recall **2 (c):** F1 scores plot of NB, SVM, RF, KNN, LR, and DT

Table 2 shows the results of 6 machine learning classifiers on the dataset without applying feature selections, with accuracies of 95.5%, 100%, 100%, 99.5%, 100%, and 99% using an 80-20% split of training-testing data for NB, SVM, RF, KNN, LR, and DT, respectively. A value of 0 indicates chronic localised periodontitis, whereas a value of 1 indicates chronic periodontitis throughout. The accuracy, recall, and F1 scores of the aforementioned 6 classifiers are compared in Fig. 2. The following phase involves the use of cross-validation procedures, and the results are summarised in table 3 below.

CLASSIFIER / CROSS VALIDATION	NAÏVE BAYES ACCURACY (%)	SVM ACCURACY (%)	RF ACCURACY (%)	KNN ACCURACY (%)	LR ACCURACY (%)	DT ACCURACY (%)
HOLD OUT	95.5	100	100	99.5	100	99
K(10) FOLD	95.8	98.1	100	98.6	99.8	100
STRATIFIED K(10) FOLD	95.8	98.1	100	98.6	100	100
LEAVE 1 OUT	95.8	98.2	100	98.6	100	100

Table 3: cross validation methods vs. accuracies obtained by 6 machine learning classifiers

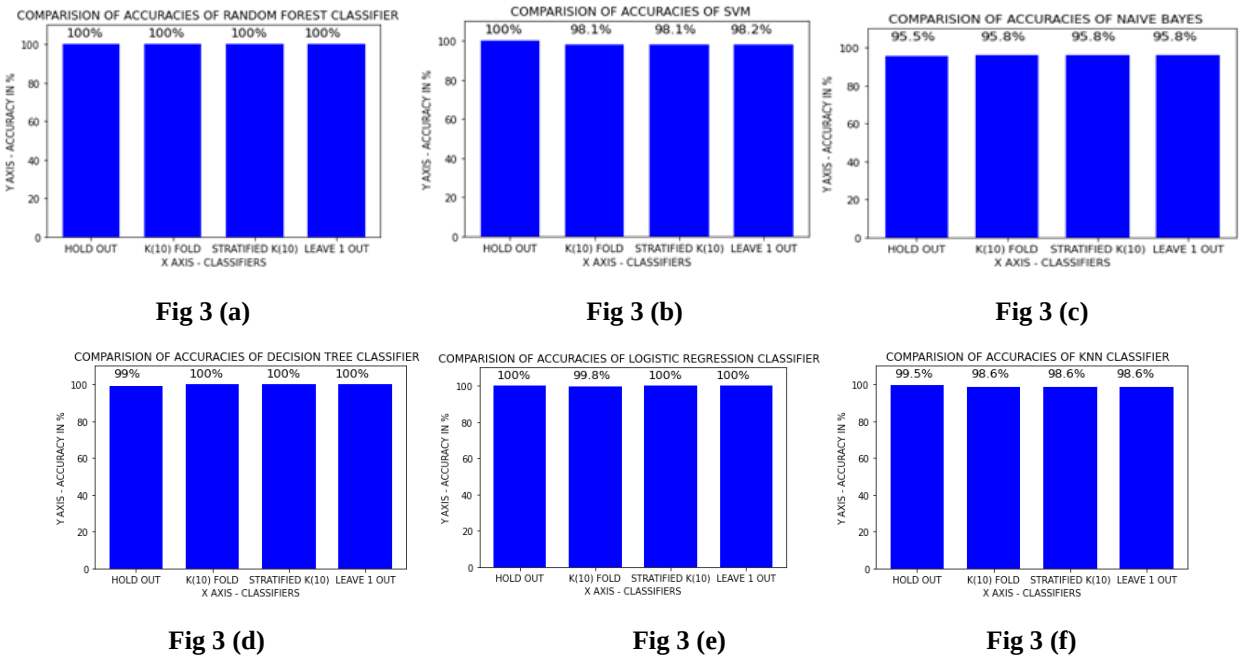

Fig 3 (a) **Fig 3 (b)** **Fig 3 (c)**

Fig 3 (d) **Fig 3 (e)** **Fig 3 (f)**

Fig 3 (a) (b) (c) (d) (e) (f): cross validation methods vs. accuracies obtained by 6 machine learning classifiers NB, SVM, RF, KNN, LR & DT respectively

Cross-validation techniques are used to ensure that a model is not overfit and to determine how well it performs on generic data that is unknown to the classifier. K (10) Fold cross validation, stratified K (10) fold, hold out, and leave one out methods are implemented in this paper, and the results obtained are not much different from those obtained without cross validation methods, as shown in Table 3 and fig 3 above. Other such methods include leave one out, leave p out, k fold, stratified k fold, repeated random sub sampling, time series [41]. Predictions made on the direct train test split size 80-20% of the data set of 1000 records taken show very little and minute difference when compared with predictions made on the various cross validation techniques and the accuracies obtained by each classifier (see Figure 3 (a) Naive Bayes, (b) Support vector, (c) random forest, (d) K nearest neighbour, (e) logistic regression, (f) decision tree). In this work, we analysed machine learning classifiers in depth, paying special attention to the areas of investigation indicated in the methodology. Feature selection approaches are then

applied and accuracy comparisons are made, as shown in fig. 4 (a) pick k best method, during the in-depth analysis of the illness by means of a machine learning classifier. FIGURE 4 (B) INFORMATION GAIN METHOD 97.2, 99.2, 100%, 98.6, 100%, and 100% FIGURE 4 (C) STEP FORWARD SEQUENTIAL SELECTION 97.2, 98.6, 100%, 98.6, 100%, and 100% Figure 4(d): Reverse step-by-step sequential selection for probabilities of 99.2%, 99.8%, 100%, 99.1%, 99.7%, and 100% The following graph shows the accuracy of six different classifiers: NB, SVM, RF, KNN, LR, and DT, with values of 99.6%, 98.8%, 100%, 99.1%, 99.5%, and 99.8%, respectively.

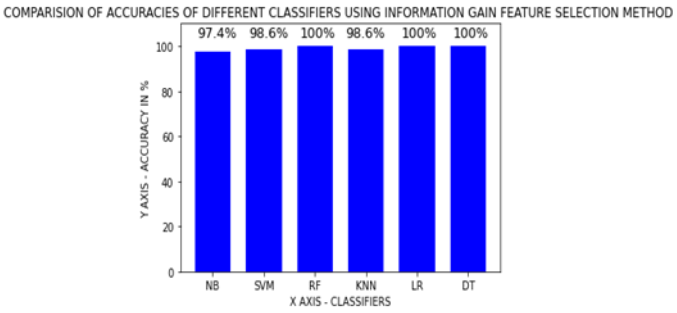

Fig 4 (a): Accuracy of classifiers obtained by Select k best feature selection method

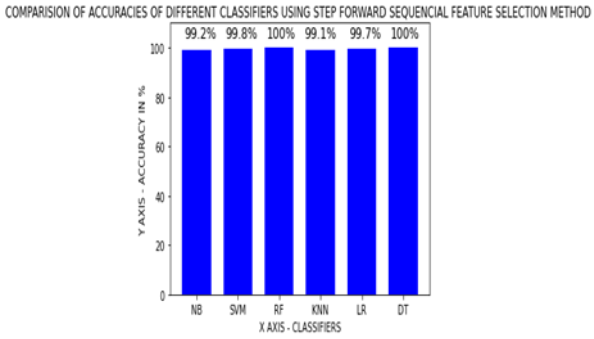

Fig 4 (c): Accuracy of classifiers by step forward sequential feature selection method

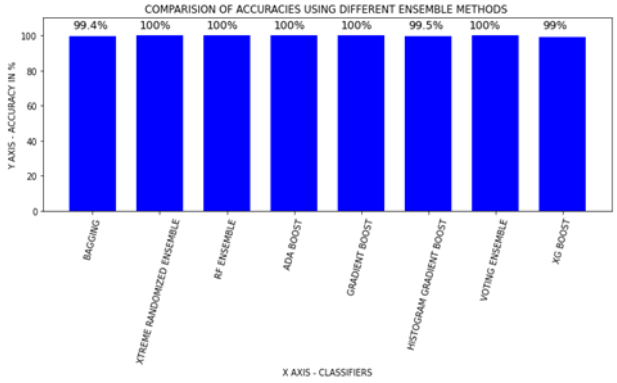

Fig 4 (d): Accuracy of classifiers by step backward sequential feature selection method

To double-check if the classifiers' results are in line with the ensemble techniques' results, a variety of ensemble methods are employed in the final stage. In conclusion, the research yielded very excellent beneficial findings that may be utilised by dentists as a decision-making help, and the accuracies for all techniques employed are almost equal. Table 4 and fig. 5 below summarise the results of putting eight different ensemble algorithms to use on a dataset of 1000 input records and showing the accuracies gained on train and test data.

Table 4: Accuracy obtained by Ensemble methods

S.NO	ENSEMBLE METHOD	ACCURACY ON TEST DATA (%)	ACCURACY ON TRAIN DATA (%)
1	Bagging classifier	99.4	99.5
2	Extremely Randomised Tree - Ensemble	100	100
3	Ensemble By Random forest	100	100
4	AdaBoost by Decision Tree	100	100
5	Gradient Boost Ensemble	100	100
6	Histogram Based Gradient Boosting	99.5	99.75
7	Voting Classifier Multi Model Ensemble	100	100
8	XG Boost	99	100

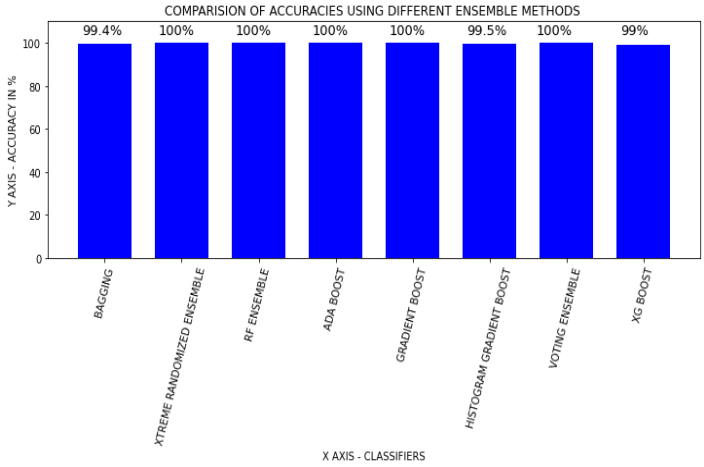

Fig 5: Accuracy obtained by Ensemble methods

Results acquired after executing all ML techniques for the supplied dataset are therefore commendable, since the accuracies gained in all the approaches are almost similar, and the research has yielded productive outcomes.

6. Conclusion & Future enhancement

Artificial intelligence, and its subfields machine learning and deep learning, play a crucial role in the health care and ancillary service sectors, enabling earlier disease prediction, more accurate diagnosis, and faster treatment. Oral illnesses are one example of a more general issue space where machine learning has been employed to find answers. Since periodontitis is the most common form of gum disease, this study employs a variety of machine learning technologies to diagnose and predict the condition in a dataset of 1,000 periodontics patients, both diabetic and non-diabetic. Outliers are removed from the dataset using the winsorizing technique, and then six different machine learning classifiers—Naive Bayes, support vector, random forest, K nearest neighbour, logistic regression, and decision tree—are modelled with accuracies of 95.5%, 100%, 100%, 99.5%, 100%, and 99%, respectively. In order to prevent overfitting and accomplish dimensionality reduction, the dataset is then subjected to a variety of cross validation techniques, including the K fold, stratified k fold, hold out, and leave one out methods, and feature selection methods, including the select Kbest, information gain, sequential feature selection methods. Finally, eight ensemble methods are put into practise to compare and contrast the results. The findings are encouraging, and the accuracies gained are indicative of the reliability of the forecasts. Therefore, this kind of research may be useful for dentists and fellow students in diagnosing and predicting the incidence of periodontitis, and can be utilised as a supportive aid for correct decision making in time to prevent further damage of alveolar bone loss and other supporting structures. The technique may also be used as a teaching tool to show aspiring periodontists how to execute their jobs. Given that prior research has linked periodontitis to systemic diseases and that diabetes was included as a feature in the current study, there is room for further investigation into

the correlation between the two conditions. Specifically, future research could investigate the pattern by which other systemic diseases impact the gums and their supporting structures, and deep learning models could be developed to investigate the pattern and its association.

References

[1] RM Baiju, Elbe Peter etal, "Oral Health and Quality of Life: Current Concepts", Journal of Clinical and Diagnostic Research, Jun 2017, 11(6): ZE21-ZE26

[2] Xiaojing Li, Kristin M. Kolltveit etal, "Systemic Diseases Caused by Oral diseases", Clinical Microbiology Reviews, Oct 2000, 13(4): 547-558

[3] Eija Könönen, Mervi Gursoy etal, "Periodontitis: A Multifaceted Disease of Tooth-Supporting Tissues", Journal of Clinical Medicine", Jul 2019, 8(8):1135

[4] TK Madiba and A Bhayat, "Periodontal disease - risk factors and treatment options", South African dental journal, Oct 2018, 73(9): 571 - 575

[5] Arkadiusz Dziedzic and Robert Wojtyczka, "The impact of coronavirus infectious disease 19 (COVID-19) on oral health", Oral diseases, Apr 2020,00:1-4

[6] Muhammad Ashraf Nazir, "Prevalence of periodontal disease, its association with systemic diseases and prevention", International Journal of Health Sciences, Jun 2017, 1(2):72-80

[7] Divya Tandona and Jyotika Rajawata, "Present and future of artificial intelligence in dentistry", Journal of Oral Biology and Craniofacial Research, Dec 2020, 10(4):391-396

[8] Neel Shimpi , Susan McRoy etal, "Development of a periodontitis risk assessment model for primary care providers in an interdisciplinary setting", Technology and Health Care, May 2019,1: 1–12

[9] Asghar Tabatabaei Balaei, Philip de Chazal etal, "Automatic Detection of Periodontitis Using Intra-Oral Images", IEEE 2017,3906-3909

[10] Jaehan Joo, Sinjin Jeong etal, "Periodontal Disease Detection Using Convolutional Neural Networks", IEEE ICAIIC 2019, 360-362

[11] Maryam Farhadian, Parisa Shokouhi etal, "A decision support system based on support vector machine for diagnosis of periodontal disease", BMC Res Notes, Jul 2020, 13:337

[12] Gopi Battineni, Getu Gamo Sagaro etal, "Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis", Journal of personalized medicine, Jun 2020, 10(2):21

[13] Eun-Hye Kim, Seunghoon Kim etal, "Prediction of Chronic Periodontitis Severity Using Machine Learning Models Based On Salivary Bacterial Copy Number", Frontiers in Cellular and Infection Microbiology, Nov 2020, 10:698

[14] Sanjeev B. Khanagar, Ali Al-ehaideb etal, "Developments, application, and performance of artificial intelligence in dentistry A systematic review", Journal of dental sciences, Jan 2021, 16(1):508-522

[15] S. B. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview", Artificial Intelligence Review, May 2011

[16] Isabelle Guyon and Andre Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, Mar 2003, 3:1157-1182

[17] Jie Cai, Jiawei Luo etal, "Feature selection in machine learning: A new perspective", Neurocomputing, Jul 2018, 300: 70-79

[18] Chuan Liua, Wenyong Wang etal, "A new feature selection method based on a validity index of feature subset", Pattern recognition letters, Jun 2017, 92:1-8

[19] A. Jovic, K. Brkic etal, "A review of feature selection methods with applications", IEEE 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Jul 2015, 1200-1205

[20] Luka Cehovin and Zoran Bosnic, "Empirical evaluation of feature selection methods in classification", Intelligent data analysis, Jan 2010, 14(3):265-281

[21] Maurizio S. Tonetti, Henry Greenwell etal, "Staging and grading of periodontitis: Framework and proposal of a new classification and case definition", Journal of clinical periodontology, Feb 2018, 45(20): 149-161

[22] Centers for Disease control and prevention website

https://www.cdc.gov/oralhealth/conditions/periodontal-disease.html/Causes

[23] Meliha Germen, Ulku Baser etal, "Periodontitis Prevalence, Severity, and Risk Factors: A Comparison of the AAP/CDC Case Definition and the EFP/AAP Classification", International Journal

of Environmental Research and Public Health, Mar 2021, 18:3459-3466

[24] Denis F. Kinane, Melanie Peterson etal, "Environmental and other modifying factors of the periodontal diseases", Periodontology 2000, Feb 2006, 40:107-119

[25] Stanley P. Hazen, "Indices for the measurement of gingival inflammation in clinical studies of oral hygiene and periodontal disease", Journal of Periodontal research, 1974, 9(14): 61-77

[26] Anitha Subbappa, Aruna Ganganna etal, "Russell's Periodontal Index: To Score or not to score", International Journal of Information Research and Review, December, 2019, 6(12): 6647-6649

[27] Dr. Amit Mani, Dr. Rosiline James etal, "Classifications for Furcation Involvement", Galore International Journal of Health Sciences and Research, Mar 2018, 3(1):15-17

[28] Larry Laster, Kenneth W. Laudenbach, "An Evaluation of Clinical Tooth Mobility Measurement", Journal of Periodontal, Oct 1975, 46(10):603-607

[29] Nigel G. Clarke, Robert S. Hirsch etal, "Periodontitis and angular alveolar lesions: A critical distinction", Oral surgery oral medicine oral pathology, May 1990, 69(5):564-571

[30] W. Peter Nordland, and Dennis P. Tarnow, "A Classification System for Loss of Papillary Height", Journal of Periodontol, Oct 1998, 69(10):1124-1126

[31] A. Sivakumar and R.Gunasundari, "A Survey on Data Pre-processing Techniques for Bioinformatics and Web Usage Mining", International Journal of Pure and Applied Mathematics, 2017, 117(20):785-794

[32] Cheng Fan, Meiling Chen etal, "A Review on Data Pre-processing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data", Frontiers in energy research, Mar 2021, 9

[33] Salkind, Neil J, "Winsorize", Encyclopaedia of Research Design: SAGE Publications, 2010

[34] Xue Ying, "An Overview of Over fitting and its Solutions", IOP Conf. Series: Journal of Physics 1168 022022, 2019

[35] Daniel Berrar, "Cross-validation", Encyclopaedia of Bioinformatics and Computational Biology, Apr 2019, 1: 542-545

[36] Esra Mahsereci Karabuluta , Selma Ayşe Özel etal, "A comparative study on the effect of feature selection on classification accuracy", Procedia Technology, May 2012, 1:323-327

[37] David Opitz and Richard Maclin, "Popular Ensemble Methods: An Empirical Study", Journal of Artificial Intelligence Research, Aug 1999, 11: 169-198

[38] Kristina Machova, Miroslav Puszta etal, "A comparison of the bagging and the boosting methods using the decision tree classifiers", Computer science and information systems, Dec 2006, 3(2):57-72

[39] Yangguang Liu, Yangming Zhou etal, "A Strategy on Selecting Performance Metrics for Classifier Evaluation", International Journal of Mobile Computing and Multimedia Communications, Dec 2014, 6(4): 20-35

[40] Lakshmi TK, Dheeba J, "Digitalization in Dental problem diagnosis, Prediction and Analysis: A Machine Learning Perspective of Periodontitis", International Journal of Recent Technology and Engineering, Jan 2020, 8(5): 67-74

[41] Michael W Browne, "Cross-Validation Methods", Journal of Mathematical Psychology, Mar 2000, 44(1):108-132

[42] Lakshmi, T. K., and J. Dheeba. "Digitalization in dental problem diagnosis, prediction and analysis: A machine learning perspective of periodontitis." *Int. J. Recent Technol. Eng* 8.5 (2020): 67-74.

[43] T.K., Lakshmi; DHEEBA, J.. Classification and Segmentation of Periodontal Cyst for Digital Dental Diagnosis Using Deep Learning. Computer Assisted Methods in Engineering and Science, [S.l.], v. 30, n. 2, p. 131–149, oct. 2022.

[44] Lakshmi, T. K., and J. Dheeba. "Digital Decision Making in Dentistry: Analysis and Prediction of Periodontitis Using Machine Learning Approach." *International Journal of Next-Generation Computing* 13.3 (2022).

[45] Mr. B. Naga Rajesh. (2019). Effective Morphological Transformation and Sub-pixel Classification of Clustered Images. International Journal of New Practices in Management and Engineering, 8(01), 08 - 14. https://doi.org/10.17762/ijnpme.v8i01.74

[46] Tanaka, A., Min-ji, K., Silva, C., Cohen, D., & Mwangi, J. Predictive Analytics for Healthcare Resource Allocation. Kuwait Journal of Machine Learning, 1(4). Retrieved from

http://kuwaitjournals.com/index.php/kjml/article/view/150

[47] Keerthi, R. S., Dhabliya, D., Elangovan, P., Borodin, K., Parmar, J., & Patel, S. K. (2021). Tunable high-gain and multiband microstrip antenna based on liquid/copper split-ring resonator superstrates for C/X band communication. Physica B: Condensed Matter, 618 doi:10.1016/j.physb.2021.413203

Authors

Mrs T.K.LAKSHMI, B.Tech CSE, M.Tech Bioinformatics, Gold Medallist, M.Tech CSE and currently Pursuing PhD in CSE at Vellore institute of Technology, Vellore in the department of School of Computer Science and Engineering. She has been awarded as **Best Teacher**, **Young professor of the year**, **Dr. Sarvepalli RadhaKrishnan Life time Achievement award**, **Utthama Acharya, Young professor of the year in women category** and at present working in **Malla Reddy University, Hyderabad in the School of Engineering.** Area of research interests are Artificial intelligence, Machine & deep learning applications, Data science and predictive analytics, Medical image processing & Bioinformatics, Mobile applications & Security

DR. DHEEBA. J, B.E CSE, M.E CSE, MBA HRM and PhD in I & C, working as Associate professor in SCOPE, Vellore institute of Technology, Vellore. She has been awarded with Sir **C. V. Raman research award** from IET (UK) Chennai Network. She is an IBM **Certified Database Associate** also having 365 **Citations** for publications in SCI, SCOPUS, SPRINGER, dblp and other highly reputed journals. She worked on a funded project titled "Energy Efficient Automated Power Management System" funded by the ICT Academy, Tamil Nadu. Her areas of interests are Algorithms, AI, Medical Imaging, Network security, Mobile communications and CN