*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 10 Issue 4 April 2022

Statistical Analysis and Machine Learning Amalgamated Convolution Neural Network (CNN) Approach on Correlating the Dental Plaque with Cardiac Illness

Ishita Singh[1]
Delhi Public School, Kaifi Azmi Marg, KD Colony, Sector 12, Rama Krishna Puram, New Delhi, Delhi - 110022, India
Email-id: v09431ishita@dpsrkp.net

Reetu Jain[2]
Chief Mentor
On My Own Technology Pvt Ltd, Mumbai, India
reetu.jain@onmyowntechnology.com

Abstract:

Dental plaque (DP) is a thin sticky film that coats the teeth and contains bacteria. The DP is caused due to the consumption of food with high percentage of carbohydrates, sugary foods and drinks, and fatty foods. The bacteria that feeds on the sugar of the foods produces acid which forms DP. However these bacteria reach the bloodstream, digestive and respiratory tracts can cause some serious diseases such as heart diseases, cancer, tumors etc. The present paper is focussed on finding the correlation of DP with other seismic factors and developing machine learning models to predict cardiovascular diseases (CVD). In achieving the aim of the paper, the analysis is carried out in three phases. First phase involves the development of a convolution neural network (CNN) model to identify and categorize DP on the basis of the thickness of plaque deposition. The second phase involves correlating DP with other seismic factors. Those seismic factors that showed positive and significant relation with DP are chosen as parameters for developing the ML models. Finally, ML models are developed using the six different algorithms namely XGBoost, Logistic Regression, Support Vector Machines, Random Forest, Decision Tree and k-Nearest Neighbor. The ML model built by XGBoost showed the best performance on the basis of training and testing accuracy, false negative and true positive values. The strength of the proposed approach is its practical applicability.

Keyword:- *Convolution Neural Network (CNN), Statistical examination, Data analysis, Dental health, Cardiac illness*

1. Introduction

Dental plaque (DP) is a thin sticky film that coats the teeth and contains bacteria. A layer of saliva called dental pellicle mostly composed of glycoproteins is formed on the surface of the teeth shortly after cleaning. The bacteria is then attached to these dental pellicles forming micro-colonies which mature and result in oral-diseases [1]. Many oral diseases such as gingivitis, caries and periodontitis are caused as a result of the DP [2]. Periodontal disease is the major cause of tooth loss among people of all ages [20]. Figure (1) shows a picture of DP.

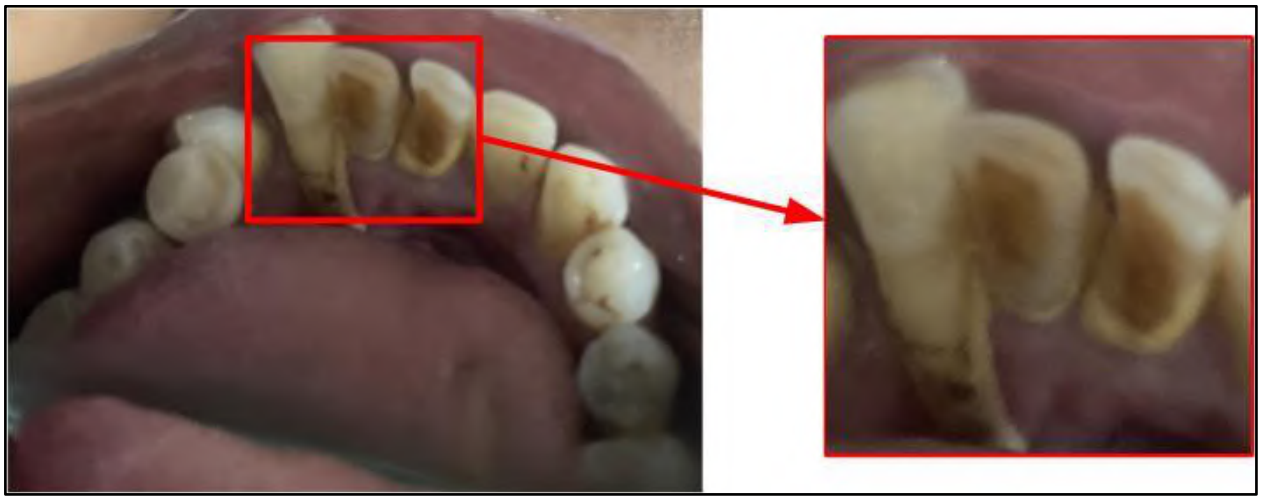


Figure (1): Picture of teeth with severe plaque

Mouth like other parts of the body is swarming with bacteria. However, most of these bacteria are harmless but once these bacteria reach the bloodstream, digestive and respiratory tracts can cause some serious diseases [4]. In general, good oral health care reduces the chances of flow of the bacteria into the bloodstream [5]. Nevertheless, some medications such as decongestants, antihistamines, painkillers, diuretics, antidepressants etc. reduces the flow of saliva which is responsible for washing of food and neutralizing the acids produced by the bacteria [6]. Above that these bacteria are responsible for forming DP.

In the present technologically advanced world where researchers are tirelessly working on developing new state-of-art sophisticated scientific technology to detect early signs of cancer, tumors, heart diseases, blood vessel diseases etc. Deep learning is one such type of modern technology which is tasked to perform complex and intricate jobs. The advantage of deep learning algorithms is its robustness and its ability to predict and classify different labeled and unlabeled data which makes it the most preferred tool for early diagnosis.

1.1. Review of the contemporary literatures

This section of the paper presents a review of contemporary literature that involves application of machine learning (ML) in detecting DP and correlates it with cardiac ailments. Cardiovascular disease is caused as a result of the blood pressure, body mass index (BMI) measurements and lipid profile which are inter-related to periodontal diseases [25]. Convolution neural network (CNN) is a deep learning algorithm that is widely used for detecting DP and predicting oral health. CNN is an unsupervised deep-learning tool which is mostly applied to analyze visual imagery [7]. CNN mimics the biological process of connectivity organization of the animal visual cortex [8]. The popularity of CNN application in the field of dental, oral and craniofacial imaging is heightening, as it has been continually applied to a broader spectrum of scientific studies [23]. Due to this reason, CNN is applied to detect DP from images.

In the paper [9], developed an image analysis technique to detect dental caries with an accuracy of 93%. However, the method failed to detect broken teeth from dental caries and also failed in detecting the depth of the caries. This drawback was rectified in [10] where the authors applied CNN to detect dental diseases on Quantitative Light-induced Fluorescence images. Dental caries are one of the chronic diseases caused by organic acids made from oral microbes [24]. In [11] authors extended the CNN to build a deep CNN model with an accuracy of 88% which was used to label a small dataset composed of 251 Radio visiography X-ray images of three distinct classes. The model developed was used to classify the tooth into 7 different types that can be used for automatically filling dental charts for forensic identification [12]. The model developed in [12] was trained with 3000 periapical radiographic images with an accuracy of 91%. A pre-trained Google Net Inception v3 CNN was used for it [13]. In the literature [14], pictures of microscopic DP are fed in a CNN model to classify healthy from unhealthy teeth.

The output obtained from the CNN model is compared with that of the AlexNet architecture. The potentiality of artificial intelligence (AI) in detecting dental diseases is reviewed in the literature [15].

In [3], researchers analyzed about 65,000 cardiovascular events which are taken from nearly a million people to conclude that there is a moderate correlation between oral health and cardiac illness. The study concluded that poor oral health is not directly linked with cardiovascular health but it is the bacteria that are associated with DP when travel to the blood vessels may cause coronary cardiac diseases. The plaque formed on the teeth is chemically the same as formed in the coronary artery that causes cardiovascular diseases (CVD) [16]. The oral health provides a complete image of the seismic health of the body [17]. In the present research scenario, scientists are integrating different AI and ML algorithms to detect and monitor the seismic condition of humans by correlating it with the oral health. [18] is one of the earliest known literatures that applied various ML algorithms to improve and rationalize the diagnostic procedures of Ischaemic heart disease. In [19] is a review based literature that comprises all the development made in the field of the contemporary state of ML based algorithms applied to cardiac CT till 2018. In contemporary literature [21], a ML model is proposed that computes a based risk score which has greater prognostic accuracy than the existing coronary computed tomography angiography integrated risk scores. In [22] an improved ML integrated ischaemia risk score is proposed to predict lesion-specific ischaemia by invasive fractional flow reserve, over stenosis, plaque measures and pre-test likelihood of coronary artery diseases (CAD). An image-based classification technique is developed in [26] for early prediction of CAD. For compiling the present study, many research papers have been reviewed but limiting the literature review part to the most significant and recent literatures.

1.2. Motivation and Novelties

From the literature reviewed for the study some of the gaps that are identified are as follows:

I. Although there are many literatures that involve the application of CNN in detecting DP, there are only few literatures that are capable of categorizing DP on the basis of their thickness.
II. The literature that correlates CVD with DP do not take into account the significance of the factors such as body mass index (BMI), systolic and diastolic pressure, lipid profile, glucose level in causing the cardiac illness.
III. Moreover, the existing literature either correlates CVD with DP or develops ML models to identify the DP. There is very little paper that simultaneously performs the two tasks.

In the present study, the first gap identified in the literature will be addressed by developing a tensor flow based CNN model. The model is capable of not only detecting DP but also capable of classifying it on the basis of plaque film thickness. The second gap identified will be addressed in two phases. In the first phase, different factors that are significant and potential measures of CVD will be identified by one way Analysis of Variance (ANOVA). In the second second phase those factors that showed positive and significant relation with CVD will be tested for correlation with DP. The third gap will be addressed by developing different ML algorithms that would predict the risk of CVD by taking into account the output from the CNN model and the values of the factors that showed positive and significant relationship with CVD.

The remainder of the paper is drafted in the following format. Section 2 briefly describes the preliminary concept of the methodology used for analyzing the dataset. Section 3 of the paper discusses the case study and the assumptions considered for solving the problem. Section 4 of the paper describes the result obtained after applying the proposed methodology in the case study and validates the proposed model. Finally the paper is concluded in section 5. The framework of the study is shown in figure (2).

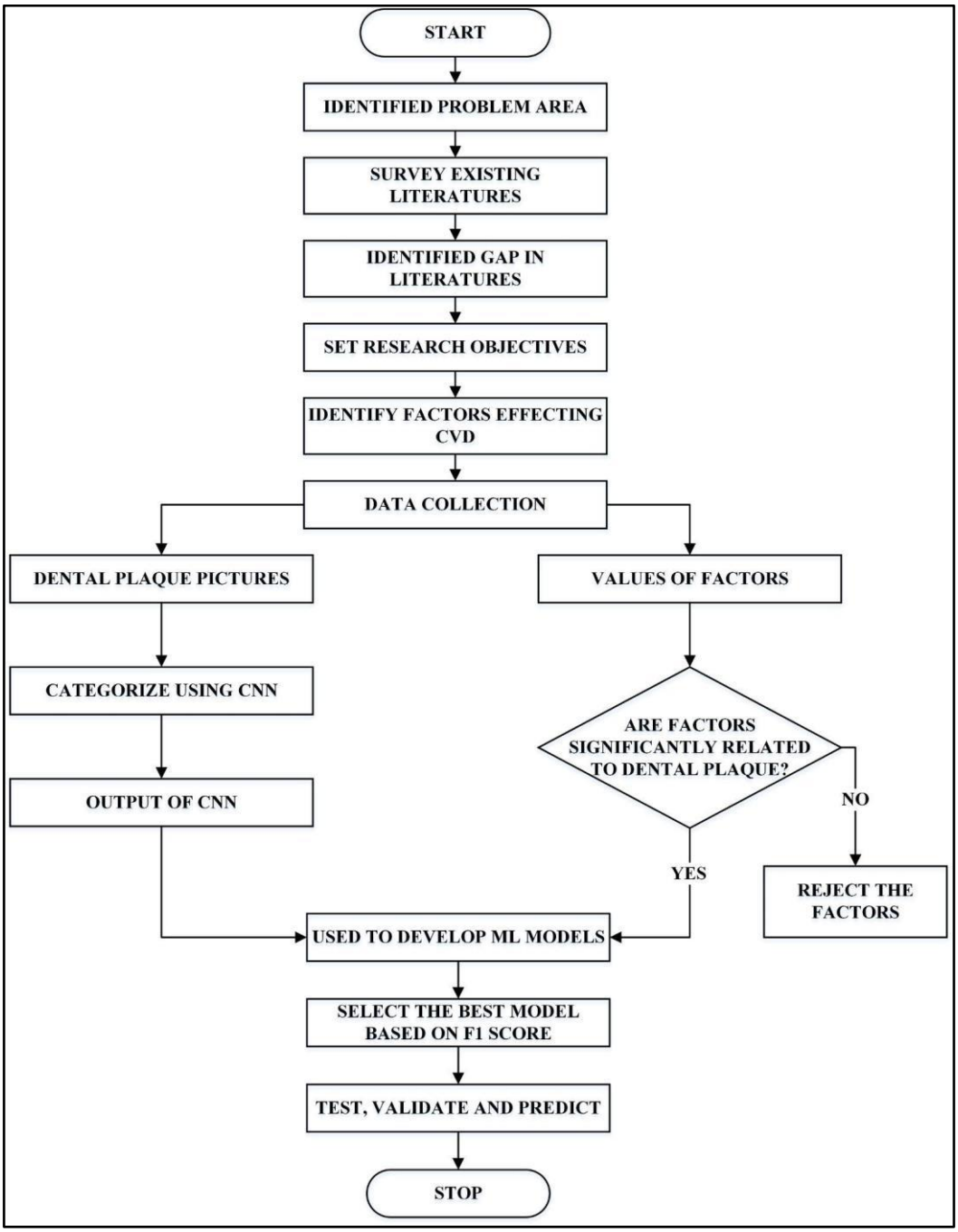


Figure (2) Framework of the present study

2. Preliminaries and methodology

This section of the paper briefly describes the preliminary concept required to analyze the data and subsequent development of the ML models. Three major concepts are used to create the ML models and analyze the data in the paper viz. Deep learning (DL), statistical analysis and machine learning. Initially, recalling the definition and concept of DL.

2.1. Deep Learning (DL)

Deep learning (DL) allows multi-layer processing computational models to recognise the pattern in the data with multiple levels of abstraction [27]. The DL methods have greatly improved AI utilization and are extremely

applied in the field of speech recognition, object recognition, facial recognition and many other intricate domains. The DL algorithm recognises the pattern among the different parameters of the dataset by using the backpropagation algorithm. CNN is a DL model that has brought about breakthroughs in processing images, video, speech and audio [28].

In the study, the CNN model is applied to detect and categorize the DP into five categories namely nil, trace, slight, moderate and heavy. The categorization is done on the basis of the thickness of the DP. Table 1 shows the relation of the DP thickness with the categorization.

Table 1: Relation of the DP thickness with the categorization

Serial no.	Categorization	Film thickness
1	Nil	No DP is formed
2	Trace	DP formation started
3	Slight	$\leq 1\ mm$
4	Moderate	$\leq 3\ mm$
5	Heavy	$\leq\ 5\ mm$

CNN has two parts. The first part is training and the second part is testing. In the training part, the model is determined for all good values of weights and the bias from labeled examples. In the second part, the model formed during the training is tested for a sample dataset. The predicted label is then compared with the actual label to determine the accuracy and preciseness of the model created. The different operations involved for CNN prediction are as follows:

Convolution operation

The convolution operation of CNN is used to extract the features from the training images and excludes the irrelevant noises. The convolution operation divides an image into tiny and smaller fragments so that the features of the images could be easily extracted and the irrelevant noises could be dropped down. This fragmented image is termed an image matrix (Im). The layer containing the N filter matrix (Fi) is slide over the image matrix throughout its width and height. Matrix multiplication of the image and filter matrix give the resultant matrix (Re) termed as convolution matrix. Mathematically,

$$Im \times Fi = Re \tag{1}$$

Activation operation

In neural networks, the activation operation of a node defines the output of that node for a given input or set of inputs. In this paper, a rectified linear unit (ReLU) is used as the activation operator. The ReLU is mathematically expressed as:

$$y = 0, if\ x < 0 \tag{2a}$$
$$y = x, if\ x \geq 0 \tag{2b}$$

ReLU, compared to sigmoid function or similar activation functions, allow faster and effective training of deep neural architectures on large and complex datasets [29].

Pooling operation

The pooling operation reduces the number of learning parameters and thereby the amount of computational work to be performed is reduced. This operation is useful to summarize the feature present in a region of the feature map generated by the convolution layer. Pooling operation in CNN is done mostly by average pooling and maximum pooling. In Average Pooling operation, the average value for patches of a feature map, and uses it to create a down sample. However the major disadvantage of using average pooling is that if there are more than one outlier it does not give the accurate result. This can be overcome with the use of maximum pooling where the maximum value for patches of a feature map is used [30].

Layer stacking

In layer stacking operation, the convolution operation, activation and pooling operation is repeated until the output obtained is a minimized matrix of the input image.

Fully connected layer

This is the last layer of a CNN model. This layer comprises neurons that are fully connected to the neurons from the previous layers. This is why this layer is called a fully connected (FC) layer. This layer is responsible for classifying and predicting the output or label of the input class.

Classification and Prediction

Classification is categorization and each neuron of the FC layer is mapped with a label. The FC layer predicts the label of the input class that has a maximum number of features similar to the testing images. In this study, SOFTMAX activation function is used to classify the label. The SOFTMAX activation function used for predicting a multinomial probability distribution [31]. The mathematical expression for SOFTMAX activation function is:

$$\sigma(z_i) = \frac{e^{(z_i)}}{\sum_{j=1}^{K} (e^{z_i})} \quad (3)$$

2.2. One way Analysis of Variance (One-way ANOVA)

ANOVA is the statistical analysis that helps to identify the factors that are making a significant impact on the performance of the model. The significance is measured by computing the p-value which indicates how well the dataset of the factors fit the model. Smaller p-value indicates better fit of the model [32]. One-way ANOVA compares the means of two or more groups for one dependent variable. The advantage of using one-way ANOVA is that the assumption of normal distribution is not required [33].

2.3. Machine learning (ML) models

The domain of ML is a subset of AI that is used for predicting by recognizing the pattern of the dataset used for creating the model [34]. ML models are broadly classified as supervised, unsupervised and reinforcement learning. In supervised learning techniques train the model based on the labeled input and output dataset. Unsupervised learning refers to the AI algorithms that identify patterns in data sets containing data points that are neither classified nor labeled. Reinforcement learning is based on regarding desired behavior and penalizing undesired behavior [35].

In the present study, different ML models are developed to correlate and predict the DP with CVD. Predicting CVD from different factors is a case of supervised learning. Therefore, supervised learning algorithms such as Decision tree (DT), k-Nearest Neighbor (kNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) are used to develop different ML models.

3. Case study

In this section of the paper, a brief description of the case study along with the different assumptions, dataset and data preprocessing is discussed.

3.1. Problem statement

The comprehensive intention of the present study is to focus on the relation between dental health and cardiac health. Cardiac health refers to the overall condition of the heart. A landmark study made in the year 1954, showed that poor oral health leads to systemic diseases such as heart diseases, diabetes, stroke etc . According to an estimate by the World Health Organization about 17.9 million people die each year of CVD which is approximately 32% of the total deaths [37]. The factors that increase the risk of CVD in individuals are raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity [38]. BMI is also a major factor in CVD [39] which can be used to replace the weight and height of patients. Therefore, BMI replaces the overweight and obesity factors according to the Eq. (5):

$$BMI = \frac{Weight}{Height^2} \tag{5}$$

where Height in Eq. (5) is taken in meters. On the basis of BMI, WHO has classified weight status into several categories which are shown in table 2.

Table 2: Different categories based on BMI

Sl. no.	**Weight status**	**BMI in $\frac{kg}{m^2}$**
1	Underweight	<18.5
2	Normal range	18.5 - 24.9
3	Overweight	25 - 29.9
4	Obese	>30
5	Obese I	30 - 34.9
6	Obese II	35 - 39.9
7	Obese III	≥40

The main aim of the present study is to develop ML models that are capable of predicting the CVD by taking into account the different factors. However all the factors identified as a cause for CVD are statistically scrutinized for DP. Those factors that showed significant relationship with DP aree

3.2. Dataset

In the present study, two datasets were used for developing the ML models. The first dataset is images of DP collected from different patients which are used to train the CNN model. The second dataset comprises 70000 sets of data showing the height, weight, systolic and diastolic blood pressure (BP), DP, cholesterol, glucose level, smoking habit, alcohol, activity and CVD. The unit of height and weight data are in meter (m) and kilogram (kg) respectively. As mentioned in section 3.1. that BMI is a major factor in causing CVD. Therefore CVD has replaced

the weight and height for analysis. The systolic and diastolic blood pressure is measured in mm of Hg column. The data for the remaining factors are categorical in nature. The DP data is divided into five categories as shown in table 1. The cholesterol and glucose data are divided into three categories viz. normal, above normal and severe. On the other hand, the data for smoking habit, alcohol, activity and CVD are divided into binary values i.e. yes or no.

3.3. Data preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. The steps involved for preprocessing are data cleaning, data transformation and data reduction. The dataset collected may have some irrelevant and missing data which can impact the result of the analysis. Hence, in this step the data are thoroughly checked and cleaned of irrelevant data. In data transformation the data are normalized so that all the data are scaled in a specific range. The data reduction is the process of data aggregating, data reduction and attribute selection [36].

On closely monitoring the second dataset, it is found that there are a lot of irregularities. Therefore, the dataset must be cleaned before using them for analyzing data. The criteria set to remove the irregularities are as follows:

I. Data for only those persons are considered whose age is in the range of 29 - 60 years.
II. If the height of the patients are more than 2 m (6.56 feet) or less than 1.25 m (4.1 feet) then those data are dropped from the dataset.
III. If the systolic BP is more than 250 mm of Hg or diastolic BP is more than 150 mm of Hg or systolic BP less than 80 mm of Hg or diastolic BP less than 50 mm of Hg then those data are dropped from the dataset because BP at this range is considered severely high and it indicates abnormality.
IV. If the BMI of a person is less than 10 then those data are dropped from the dataset as it indicates that the person is severely underweight and malnourished.
V. If the BMI of a person is more than 60 then those data are dropped from the dataset as it indicates that the person is severely obese.

Strictly abiding by the above mentioned criteria, out of the 70,000 datasets 1113 dataset are dropped and the remaining 68887 datasets are used for conducting the analysis. The data cleaning is followed by data transformation where the data except the categorical data are normalized according to the feature scaling method. Mathematically,

$$y = \frac{x - x_{(min)}}{x_{(max)} - x_{(min)}} \quad (4)$$

where y is normalized value of the data x, $x_{(max)}$ and $x_{(min)}$ are the maximum and minimum value of data for a factor.

4. Results and discussions

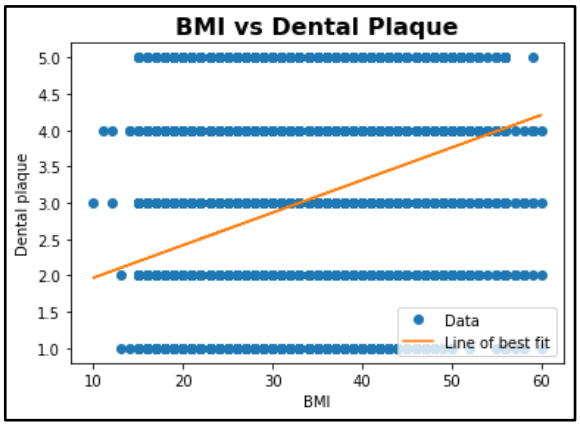
In this section of the paper, the results obtained from analyzing the datasets and a brief discussion is presented.

4.1. Results from one-way ANOVA

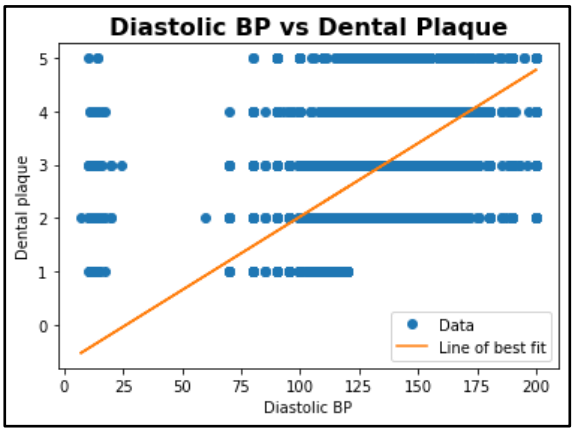
One-way ANOVA will be used to identify those factors that are significantly affecting the formation of DP. Scatterplot will be used to identify the relationship between the factors and DP. Only those factors that showed significant and positive relation with DP will be considered in the study.

4.1(a) Results from the regression analysis

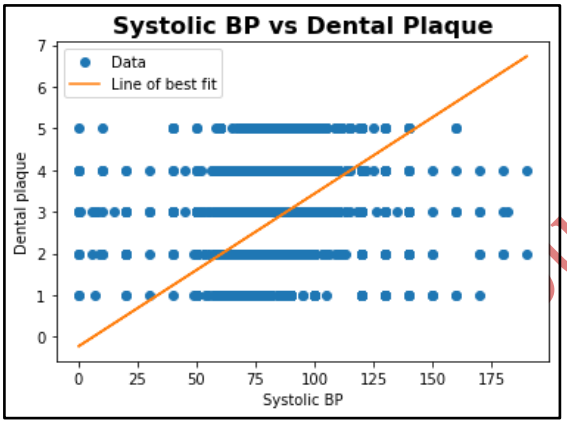
Regression analysis is conducted to determine the relationship between different factors responsible for causing CVD with DP. The regression plot for all the factors with DP are shown in figures 3.



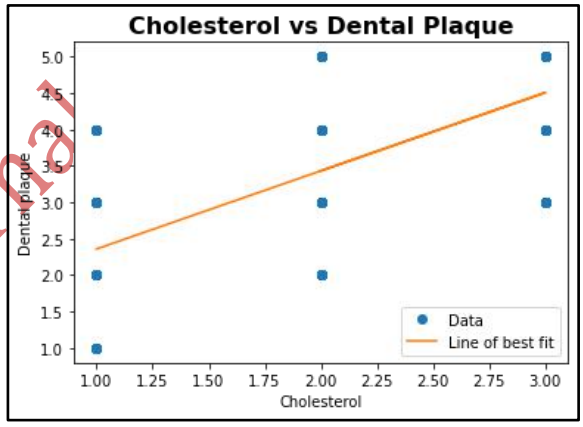
(i)



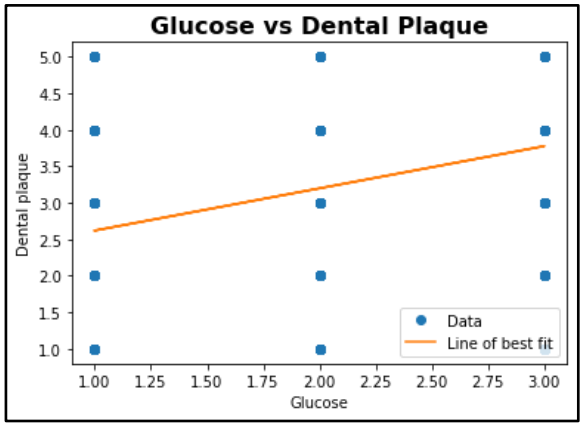
(ii)



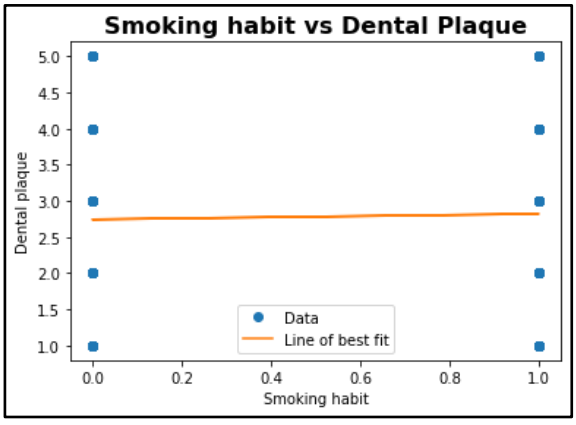
(iii)



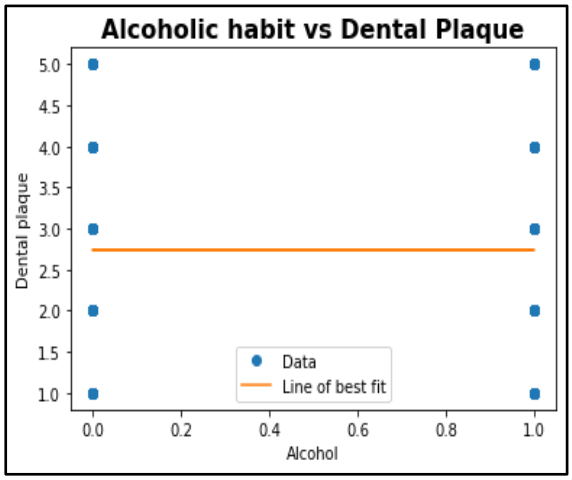
(iv)



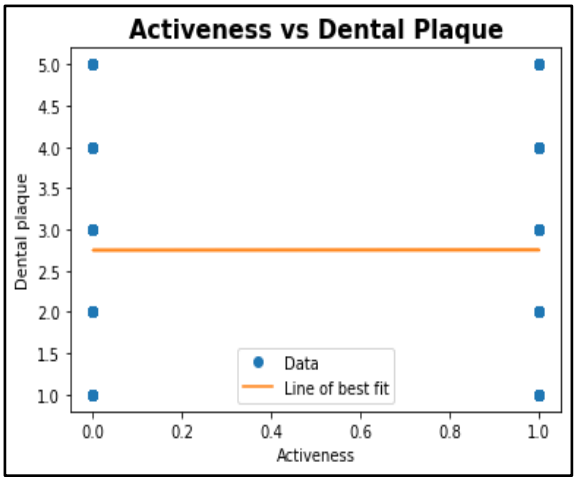
(v)



(vi)



(vii)



(viii)

Figure (3): Regression plot for (i) BMI, (ii) Diastolic BP, (iii) Systolic BP, (iv) Cholesterol, (v) Glucose, (vi) Smoking habit, (vii) Drinking habit and (viii) Activity

The coefficient of determination (R^2) and the equation of the regression line is tabulated in table 3.

Table 3: List of the R^2 values and the equation of regression line for the factors vs DP

Sl. no.	**Factors**	R^2	**Equation of regression line**
1	BMI	0.1395	$y = 0.0448 * x + 1.522$
2	Diastolic BP	0.1687	$y = 0.0276 * x - 0.7305$
3	Systolic BP	0.1934	$y = 0.0366 * x - 0.2273$
4	Cholesterol	0.383	$y = 1.0723 * x + 1.2874$
5	Glucose level	0.1389	$y = 0.5776 * x + 2.0422$
6	Smoking habit	0.0003	$y = 0.00772 * x + 2.7435$
7	Drinking habit	0.00	$y = 2.74$
8	Activity	0.0001	$y = 0.0001 * x + 2.74$

The correlation between features are shown in figure 4.

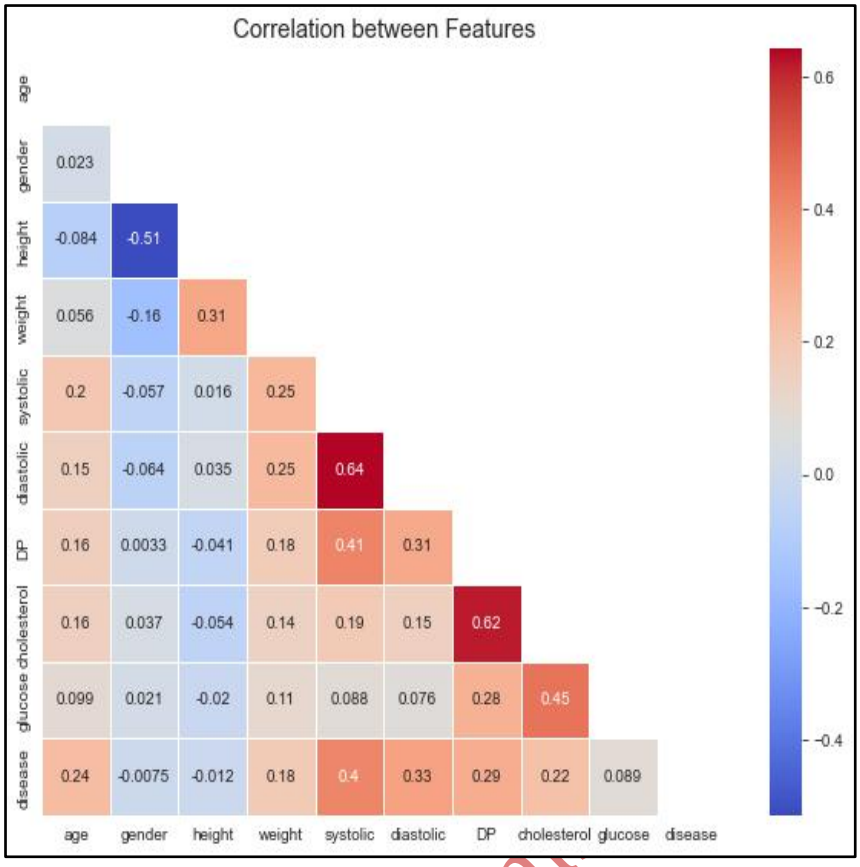


Figure 4: Heatmap showing correlation between features

Some of the points observed from the regression analysis are:

I. The factors such as BMI, diastolic and systolic BP, cholesterol and glucose level all show a positive relationship with DP.
II. The goodness of fit for cholesterol for the collected data is maximum in correlation with DP. This is due to the fact that the pathogenic microorganisms that exist in DP is the primary cause of elevated cholesterol [40].
III. All the factors that are somehow related to elevated cholesterol show a positive relationship with the formation of DP.
IV. Although smoking and drinking are some factors for predicting CVD, they are not related to the formation of DP. This is because smoking and drinking habits is one of the root causes of tartar formation [41].
V. Day to day activity is no way related to formation of DP.

4.1(b) Results from the one-way ANOVA analysis

From the regression analysis, it is observed that smoking and drinking habits and activeness have no or very little relation with formation of DP. Hence the three factors are dropped from further analysis. The remaining factors are tested for statistical significance using one-way ANOVA. It is conducted to determine the significant factors that not only affect the causing of CVD but also affects the formation of DP. For this purpose the significance level of the different factors viz. BMI, diastolic and systolic blood pressure, cholesterol, glucose level, smoking and drinking habit and activiness. The ANOVA result for the different factors is summarized in tables 4 - 8.

Table 4: ANOVA table for BMI w.r.t. DP

Sl. no.	Source	SS	DF	MS	F	p-value
1	BMI	4151.332	50	83.026633	62.732867	0.0
2	Within	91104.099	68836	1.323495	—	—

Table 5: ANOVA table for diastolic BP w.r.t. DP

Sl. no.	Source	SS	DF	MS	F	p-value
1	Diastolic BP	26630.135	112	237.77	238.28	0.0
2	Within	68625.296	68774	0.997838	—	—

Table 6: ANOVA table for systolic BP w.r.t. DP

Sl. no.	Source	SS	DF	MS	F	p-value
1	Systolic BP	12933.667	93	139.071690	116.21664	0.0
2	Within	82321.763	68793	1.196659	—	—

Table 7: ANOVA table for Cholesterol w.r.t. DP

Sl. no.	Source	SS	DF	MS	F	p-value
1	Cholesterol	36660.604	2	18330.302	21549.079	0.0
2	Within	58594.826	68884	0.850630	—	—

Table 8: ANOVA table for glucose level w.r.t. DP

Sl. no.	Source	SS	DF	MS	F	p-value
1	Glucose level	7515.944	2	3757.972	2950.372	0.0
2	Within	87739.486	68884	1.273728	—	—

Some of the points observed from the one-way ANOVA analysis are:

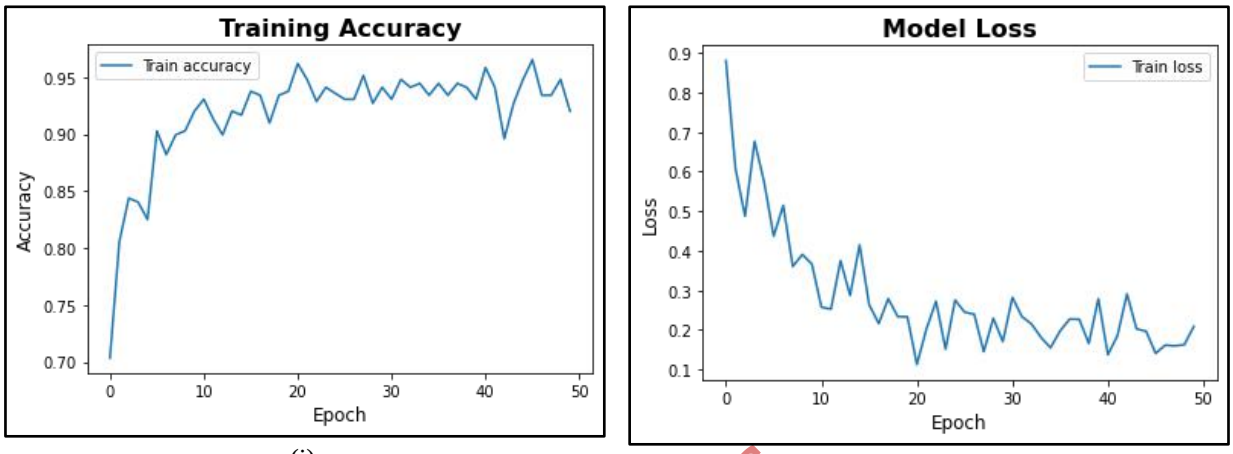
I. The factors BMI, diastolic and systolic BP, cholesterol and glucose level all show a significant relationship with DP. Hence the factors can be used for developing the machine learning models.

4.2. Results from the CNN model

In this section of the paper the performance and the validation of the CNN model is discussed in brief.

4.2(a). Performance of the CNN model

The CNN model is built using the Keras and Tensorflow library which is coded in python 3.8 and runs on a 64-bit windows 11 system with 8 GB RAM and i5, 1.6GHz processor. The parameters of the CNN model are optimized using Adam optimizer. The Adam optimizer inherits the positive attributes of the "Gradient Descent with Momentum" and "Root Mean Square Propagation" algorithms. The proposed CNN model is executed for 50 epochs with 9 steps per epoch and 1 validation step. The epoch that showed the least loss value is selected as the optimal model. The training accuracy and training loss graph obtained from the CNN model for detecting and categorizing DP is shown in figure 5.



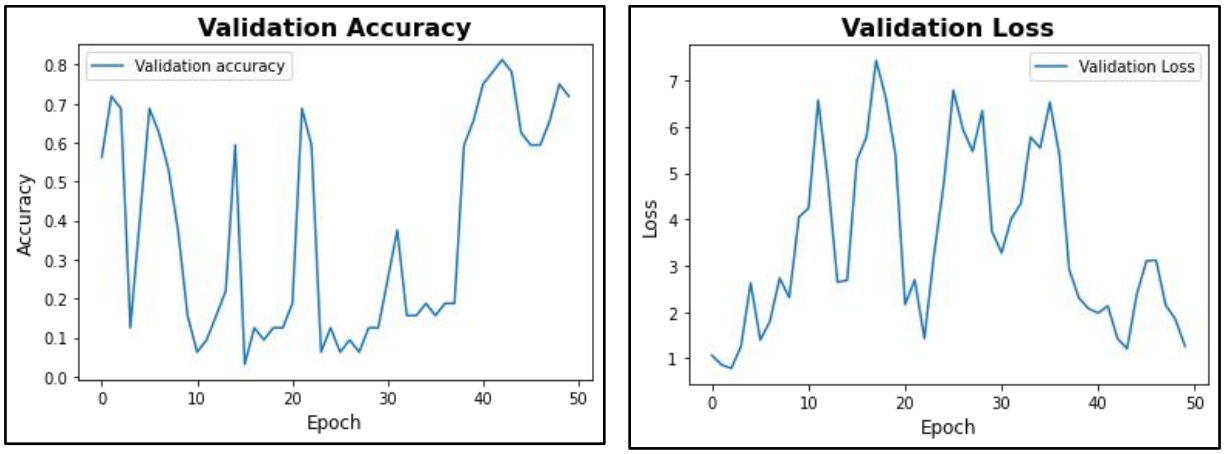
(i) (ii)

Figure (5): Training (i) Accuracy and (ii) loss curve of the CNN model

The training accuracy and loss computed for the 1st epoch of the CNN model is 0.7036 and 0.8806 respectively. However the training accuracy and loss values computed for the 50th epoch is 0.9201 and 0.2089 respectively.

4.2(b). Validation of the CNN model

The CNN model developed is validation with pictures of oral cavity with DP which were not used for developing the CNN model. The validation accuracy and validation loss graphs are shown in figure 6.



(i) (ii)

Figure (6): Validation (i) Accuracy and (ii) loss curve of the CNN model

The validation accuracy and loss computed for the 1st epoch is 0.5625 and 1.059 respectively which was enhanced to 0.7188 and 1.2601 respectively for the optimal epoch.

4.3. Machine learning models

For creating the ML models Pandas, NumPy, seaborn, matplotlib and scikit-learn are among the Python libraries imported. The dataset is split in a 75:25 ratio where 75% of the data were used for training the models and the remaining 25% data were used for testing the model. Feature Scaling is done to standardize the features by removing the mean and further scaling it to unit variance. After these procedures, the dataset is finally subjected to the machine learning algorithms namely XG Boost (XGB), Random Forest (RF), Decision Trees (DT), Logistic regression (LR), k-Nearest Neighbor (kNN) and Support Vector Machine (SVM) algorithms. The values of the training and testing accuracy, false negative, true positive and F1-score are tabulated in table 9.

Table 9: Summary table of the ML models without hyperparameter tuning

Sl. no.	Algorithm	Train accuracy	Test accuracy	F1 score	False negative	True positive
1	SVM	73.51	73.10	0.71	2846	5667
2	LR	72.62	72.60	0.71	2818	5695
3	XGB	76.06	73.24	0.71	2750	5763
4	RF	99.73	70.96	0.70	2660	5853
5	kNN	78.28	69.21	0.69	2724	5789
6	DT	99.73	63.07	0.63	3146	5367

In the next step hyperparameter tuning is carried out to decrease the overfitting and increase the accuracy of the models. The values of the training and testing accuracy, false negative, true positive and F1-score for the hyperparameter tuned ML models are tabulated in table 10.

Table 10: Summary table of the ML models with hyperparameter tuning

Sl. no.	Algorithm	Train accuracy	Test accuracy	F1 score	False negative	True positive
1	Tuned XGB	71.53	71.30	0.74	1553	6960
2	Tuned DT	72.70	72.97	0.73	2341	6172
3	Tuned RF	83.92	71.69	0.73	1987	6526
4	Tuned SVM	72.59	72.31	0.73	1913	6600
5	Tuned LR	70.84	70.76	0.73	1662	6851
6	Tuned kNN	99.73	71.71	0.71	2655	5858

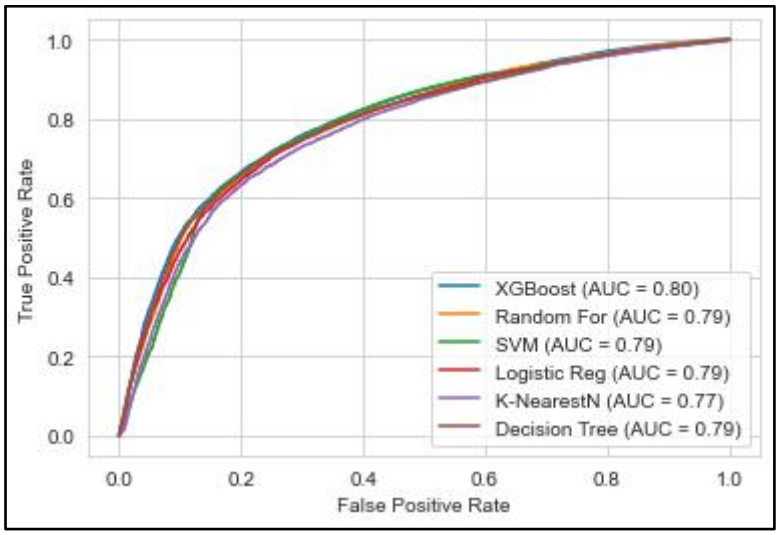


Figure (7): ROC curve for the ML models

After hyperparameter tuning the F1 score for the ML model created from XGB algorithm is increased and the overfitting is decreased. Then the receiver operating characteristic (ROC) curve which is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied is drawn. The ROC curve is shown in figure 7.

Feature importance was calculated for the ML model with the highest F1-score i.e the tuned XGB model. Feature importance is used to determine which factor is most dominant in predicting CVD. From figure 8, it is observed that the systolic BP is the most important factor in predicting CVD followed by the factors well above cholesterol, age, above cholesterol diastolic BP and then DP. Although DP is ranked the sixth important in predicting CVD, it is to be noted that systolic and diastolic BP and cholesterol shows significant and positive relation with DP.

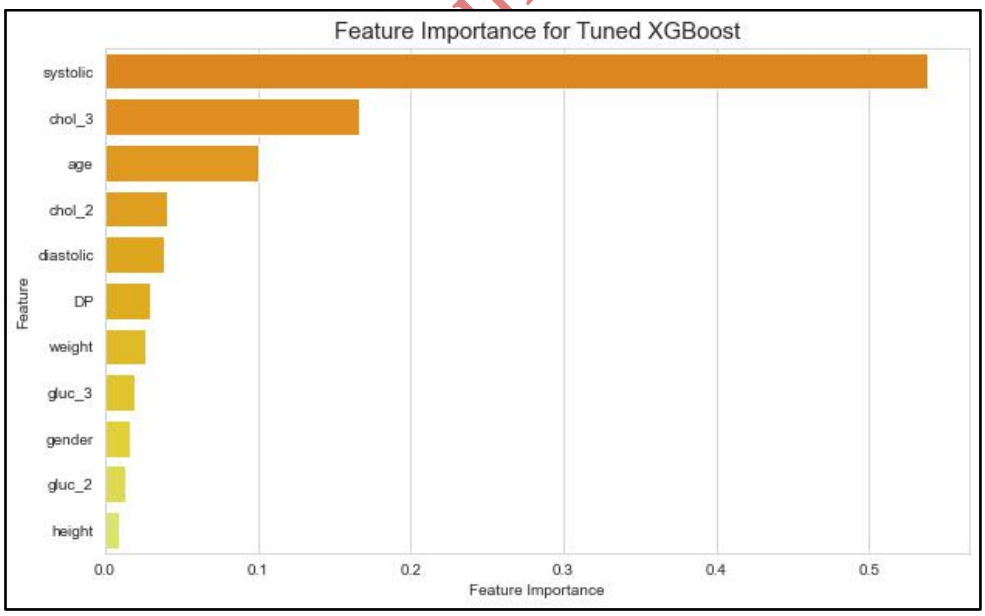


Figure (8): Feature importance for tuned XGB

5. Conclusion and future scope

The comprehensive intention of the present study is to develop an intelligent model that is capable of predicting the CVD based on the DP and other seismic factors. Although there is much research that involves the application of ML and AI in predicting CVD, those state-of-the art have not yet considered DP as one of the factors. The DP

is mostly formed due to consumption of foods with a high percentage of carbohydrate, sugary foods and drinks, fatty foods etc. Above that, this kind of food is one of the major parameters contributing to the increase of cholesterol, BP and weight which can eventually lead to CVD.

In this paper, for achieving the aim research is carried out in three phases. The first phase involves developing a deep learning model by CNN that can categorize DP into five classes based on the DP deposition. Keras and Tensorflow based CNN models are developed that have training and testing accuracy of 0.9201 and 0.7188 respectively. The second phase of the research involves identification of the factors from the existing literature that can cause CVD then correlating them to DP. In this process it was found that systolic and diastolic BP, cholesterol, BMI and glucose level shows positive and significant relation with CVD. Some factors such as smoking and drinking habits and activeness are essential factors for CVD, yet they are not or very little related to formation of DP. Hence the factors are dropped from developing the ML models. In the final phase, ML models are developed using the XGB, RF, DT, LR, kNN and SVM algorithms. The performance of all the algorithms are measured based on the training and testing accuracy, false negative and true positive values. Initially, the performance of SVM is the best whereas DT is the worst performed algorithm. Then the hyperparameter tuning of the ML models is done in order to increase the accuracy and decrease the overfitting. After hyperparameter tuning XGB is the best performing algorithm whereas kNN is the worst. The feature importance of the XGB model is computed that shows that systolic BP is the most important parameter followed by well above cholesterol, age, above cholesterol diastolic BP, DP, weight, well above glucose, glucose above normal level and height. Although DP is the sixth important factor in predicting CVD by XGB, yet factors like systolic and diastolic BP and cholesterol show significant and positive relation with DP.

It is practically not possible to involve all aspects of a research in a single paper which leads to some work to be performed in the future. In the future, exploratory data analysis can be done to get a better perspective of how DP is related with CVD. Moreover, artificial neural network model can be trained to predict CVD.

Acknowledgement:

We would like to express my heartfelt gratitude to all the tutors and mentors of On My Own Technology pvt. ltd. for extending their help in carrying out the particular project. It is because of their help that we are able to conduct the research. We shall remain ever grateful for their help and generosity. The authors would like to declare that no fundings in any form is received for carrying out this research work.

References

1. Kreth, J., Merritt, J., & Qi, F. (2009). Bacterial and host interactions of oral streptococci. *DNA and cell biology*, *28*(8), 397-403.
2. Shibly, O., Rifai, S., & Zambon, J. J. (1995). Supragingival dental plaque in the etiology of oral diseases. *Periodontology 2000*, *8*(1), 42-59.
3. Batty, G. D., Jung, K. J., Mok, Y., Lee, S. J., Back, J. H., Lee, S., & Jee, S. H. (2018). Oral health and later coronary heart disease: cohort study of one million people. European journal of preventive cardiology, 25(6), 598-605.
4. Kane, S. F. (2017). The effects of oral health on systemic health. Gen Dent, 65(6), 30-34.
5. Alpert, P. T. (2017). Oral health: the oral-systemic health connection. Home health care management & practice, 29(1), 56-59.
6. Marino, R., Albala, C., Sanchez, H., Cea, X., & Fuentes, A. (2015). Prevalence of diseases and conditions which impact on oral health and oral health self-care among older chilean. Journal of aging and health, 27(1), 3-16.
7. Valueva, M. V., Nagornov, N. N., Lyakhov, P. A., Valuev, G. V., & Chervyakov, N. I. (2020). Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Mathematics and Computers in Simulation, 177, 232-243.
8. Fukushima, K. (2007). Neocognitron. Scholarpedia, 2(1), 1717.

9. Datta, S., & Chaki, N. (2015, November). Detection of dental caries lesion at early stage based on image analysis technique. In 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS) (pp. 89-93). IEEE.
10. Imangaliyev, S., Veen, M. H., Volgenant, C., Keijser, B. J., Crielaard, W., & Levin, E. (2016, August). Deep learning for classification of dental plaque images. In International Workshop on Machine Learning, Optimization, and Big Data (pp. 407-410). Springer, Cham.
11. Prajapati, S. A., Nagaraj, R., & Mitra, S. (2017, August). Classification of dental diseases using CNN and transfer learning. In 2017 5th International Symposium on Computational and Business Intelligence (ISCBI) (pp. 70-74). IEEE.
12. Miki, Y., Muramatsu, C., Hayashi, T., Zhou, X., Hara, T., Katsumata, A., & Fujita, H. (2017). Classification of teeth in cone-beam CT using deep convolutional neural network. Computers in biology and medicine, 80, 24-29.
13. Lee, J. H., Kim, D. H., Jeong, S. N., & Choi, S. H. (2018). Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. Journal of dentistry, 77, 106-111.
14. Aberin, S. T. A., & de Goma, J. C. (2018, November). Detecting periodontal disease using convolutional neural networks. In 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM) (pp. 1-6). IEEE.
15. You, W., Hao, A., Li, S., Wang, Y., & Xia, B. (2020). Deep learning-based dental plaque detection on primary teeth: a comparison with clinical assessments. BMC Oral Health, 20(1), 1-7.
16. El Kholy, K., Genco, R. J., & Van Dyke, T. E. (2015). Oral infections and cardiovascular disease. Trends in Endocrinology & Metabolism, 26(6), 315-321.
17. National Institutes of Health. (2021). Oral Health in America: Advances and Challenges. US Department of Health and Human Services, National Institute of Dental and Craniofacial Research: Bethesda, MD, USA.
18. Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artificial intelligence in medicine, 16(1), 25-50.
19. Singh, G., Al'Aref, S. J., Van Assen, M., Kim, T. S., van Rosendael, A., Kolli, K. K., ... & Min, J. K. (2018). Machine learning in cardiac CT: basic concepts and contemporary data. Journal of Cardiovascular Computed Tomography, 12(3), 192-201.
20. Rana, A., Yauney, G., Wong, L. C., Gupta, O., Muftu, A., & Shah, P. (2017, November). Automated segmentation of gingival diseases from oral images. In 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT) (pp. 144-147). IEEE.
21. van Rosendael, A. R., Maliakal, G., Kolli, K. K., Beecy, A., Al'Aref, S. J., Dwivedi, A., ... & Min, J. K. (2018). Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry. Journal of cardiovascular computed tomography, 12(3), 204-209.
22. Dey, D., Gaur, S., Ovrehus, K. A., Slomka, P. J., Betancur, J., Goeller, M., ... & Norgaard, B. L. (2018). Integrated prediction of lesion-specific ischaemia from quantitative coronary CT angiography using machine learning: a multicentre study. European radiology, 28(6), 2655-2664.
23. Ren, R., Luo, H., Su, C., Yao, Y., & Liao, W. (2021). Machine learning in dental, oral and craniofacial imaging: a review of recent progress. PeerJ, 9, e11451.
24. Lee, E., Park, S., Um, S., Kim, S., Lee, J., Jang, J., ... & Jeong, T. (2021). Microbiome of Saliva and Plaque in Children According to Age and Dental Caries Experience. Diagnostics, 11(8), 1324.
25. Yauney, G., Rana, A., Wong, L. C., Javia, P., Muftu, A., & Shah, P. (2019, July). Automated process incorporating machine learning segmentation and correlation of oral diseases with systemic health. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3387-3393). IEEE.
26. Chen, J. I. Z., & Hengjinda, P. (2021). Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study. Journal of Artificial Intelligence, 3(01), 17-33.
27. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

28. Wlodarczak, P., Soar, J., & Ally, M. (2015, October). Multimedia data mining using deep learning. In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC) (pp. 190-196). IEEE.
29. Behnke, S. (2003). Hierarchical neural networks for image interpretation (Vol. 2766). Springer.
30. D. Song, Z., Liu, Y., Song, R., Chen, Z., Yang, J., Zhang, C., & Jiang, Q. (2018). A sparsity-based stochastic pooling mechanism for deep convolutional neural networks. Neural Networks, 105, 340-345.
31. E. Gao, B., & Pavel, L. (2017). On the properties of the softmax function with application in game theory and reinforcement learning. arXiv preprint arXiv:1704.00805.
32. Singh, R., Hussain, S. A. I., Dash, A., & Rai, R. N. (2020). Modelling and optimizing performance parameters in the wire-electro discharge machining of Al5083/B4C composite by multi-objective response surface methodology. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 42(6), 1-32.
33. Ross, A., & Willson, V. L. (2017). One-way anova. In Basic and advanced statistical tests (pp. 21-24). SensePublishers, Rotterdam.
34. Zhang, X. D. (2020). A matrix algebra approach to artificial intelligence.
35. Sasakawa, T., Hu, J., & Hirasawa, K. (2008). A brainlike learning system with supervised, unsupervised, and reinforcement learning. Electrical Engineering in Japan, 162(1), 32-39.
36. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 1-22.
37. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
38. Lu, Y., Hajifathalian, K., Ezzati, M., Woodward, M., Rimm, E. B., Danaei, G., & D'Este, C. (2014). Metabolic mediators of the eff ects of body-mass index, overweight, and obesity on coronary heart disease and stroke: a pooled analysis of 97 prospective cohorts with 1.8 million participants.
39. Alkhawam, H., Nguyen, J., Sayanlar, J., Sogomonian, R., Desai, R., Jolly, J., ... & Rubinstein, D. (2016). Coronary artery disease in patients with body mass index≥ 30 kg/m2: a retrospective chart analysis. Journal of Community Hospital Internal Medicine Perspectives, 6(3), 31483.
40. Katz, J., Flugelman, M. Y., Goldberg, A., & Heft, M. (2002). Association between periodontal pockets and elevated cholesterol and low density lipoprotein cholesterol levels. Journal of periodontology, 73(5), 494-500.
41. Murillo, G., Vargas, M. A., Castillo, J., Serrano, J. J., Ramirez, G. M., Viales, J. H., & Benitez, C. G. (2018). Prevalence and Severity of Plaque-Induced Gingivitis in Three Latin American Cities: Mexico City-Mexico, Great Metropolitan Area-Costa Rica and Bogota-Colombia. Odovtos-International Journal of Dental Sciences, 20(2), 91-102.