RISK-FACTOR BASED DIAGNOSIS FOR CHRONIC PERIODONTITIS USING MACHINE LEARNING MODELS

HTUN TEZA

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
(DATA SCIENCE FOR HEALTH CARE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2021**

Thesis
entitled

RISK-FACTOR BASED DIAGNOSIS FOR CHRONIC PERIODONTITIS USING MACHINE LEARNING MODELS

……
Mr. Htun Teza
Candidate

……
Anuchate Pattanateepapon, D.Eng.
(Electrical and Information Engineering Technology)
Major advisor

……
Prof. Ammarin Thakkinstian, Ph.D.
(Clinical Epidemiology & Community Medicine)
Co-advisor

……
Ratchainant Thammasudjarit, Ph.D.
(Computer Science)
Co-advisor

……
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

……
Asst. Prof. Oraluck Pattanaprateep, Ph.D.
(Pharmacy Administration)
Program Director
Master of Science Program in
Data Science for Health Care
Faculty of Medicine Ramathibodi
Hospital, Mahidol University

Thesis
entitled

RISK-FACTOR BASED DIAGNOSIS FOR CHRONIC PERIODONTITIS USING MACHINE LEARNING MODELS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Data Science for Health Care)
on
May 27, 2021.

……
Mr. Htun Teza
Candidate

……
Attawood Lertpimonchai, D.D.S, M.Sc., Ph.D. (Clinical Epidemiology)
Chair

……
Boonchai Kijsanayotin, M.D., Ph.D. (Health Informatics)
Member

……
Prof. Ammarin Thakkinstian, Ph.D. (Clinical Epidemiology & Community Medicine)
Member

……
Ratchainant Thammasudjarit, Ph.D. (Computer Science)
Member

……
Anuchate Pattanateepapon, D.Eng. (Electrical and Information Engineering Technology)
Member

……
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

……
Prof. Priyamitr Sritara,
M.D., FRCP., FACP., FRCP(T)
Dean
Faculty of Medicine,
Ramathibodi Hospital
Mahidol University

ACKNOWLEDGEMENTS

This thesis wouldn’t have been materialized without attentiveness and encouragement of my major advisor, Dr. Anuchate Pattanateepapon with his kind guidance, empowering support, and constructive feedback. Ajarn Eak has a significant contribution to the evolution of this project.

Gratitude must be expressed to my co-advisor, Professor Ammarin Thakkinstian, for her outstanding knowledge and assistance in this study. I am immensely grateful for her constant guidance, suggestions, and support.

I would also like to thank my external advisor Dr. Attawood Lertpimonchai, always with a warm smile, for his regular advice and guidance on the technical aspects of the underlying disease and the main characteristic features.

Additional thanks to Assistant Professor Oraluck Pattanaprateep, Dr. Ratchainant Thammasudjarit, both of whom had helped me migrate from clinical environment to research; and Sukanya Siriyotha, Pee Aom who was a willing and great help during the data exploration and cleaning phase.

Finally, I would like to express immeasurable appreciation and deepest gratitude to Ba Tin Win, Aunty War, Gee Daw Gyi and Gee Daw Nge for their endless love, support, and encouragement.

Htun Teza

RISK-FACTOR BASED DIAGNOSIS FOR CHRONIC PERIODONTITIS
USING MACHINE LEARNING MODELS

HTUN TEZA 6238135 RADS/M
M.Sc. (DATA SCIENCE FOR HEALTH CARE)

THESIS ADVISORY COMMITTEE: ANUCHATE PATTANATEEPAPON, D.Eng., AMMARIN THAKKINSTIAN, Ph.D., RATCHAINANT THAMMASUDJARIT, Ph.D.

ABSTRACT

Chronic periodontitis is one of the most common oral diseases in the world affecting 11.2% globally and 26% among Thai adults. Symptoms are negligible until it is too late and results in loss of tooth and quality of life. To diagnose chronic periodontitis, a chairside examination by a dentist or an oral hygienist is required. The process is time and resource-consuming, so a screening model to identify the risk of having chronic periodontitis in an examinee can be of assistance in reducing workload for the examiners.

Cross-sectional regression models are commonly applied using relevant demographic or risk behaviors as predictors. While logistic regression models are simple to apply or to interpret, their performance can be less optimal depending on feature selection and engineering. Machine learning models recently have been increasingly applied in medical and health-related fields due to their more complex yet powerful performances and their ability to handle high dimensional data as well as unstructured data.

In this study, screening models were applied such as mixed-effects logistic regression (MELR), recurrent neural networks (RNN), and mixed-effects support vector machine (MESVM). Using the Electric Generation Authority of Thailand (EGAT) cohort 2nd survey, the models were trained upon longitudinal data. Hyperparameters optimization is done for RNN and MESVM applying random-search followed by grid-search procedures. All models are evaluated with the same metrics; sensitivity, specificity, accuracy, positive likelihood ratio, positive prevalence value, negative prevalence value, Brier score, and F1 score to address class imbalance problems. It was observed that the MELR model (90.5% accuracy) performed better than the machine learning models (70.0% accuracy for RNN and 72.7% for MESVM).

IMPLICATIONS OF THE THESIS

Trained models could be applied in large-scale screening in a community such as public health missions as well as electronic health records after the models have been validated with external datasets to acceptable performances.

KEYWORDS: SEVERE CHRONIC PERIODONTITS / MIXED EFFECTS LOGISTIC REGRESSION / RECURRENT NEURAL NETWORKS / MIXED EFFECTS SUPPORT VECTOR MACHINE

92 pages

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ..... **III**

CHAPTER I BACKGROUND AND RATIONALE **1**

1.1 Background and Rationale 1

1.2 Research Question 4

1.3 Research Objectives 4

1.4 Expected Benefits 4

CHAPTER II LITERATURE REVIEW **6**

2.1 Epidemiology of Periodontitis 6

2.2 Risk Factors of Periodontitis 7

2.3 Predictive modelling of periodontitis 8

2.3.1 Statistical Modelling 9

2.3.2 Machine Learning 10

2.3.3 Deep Learning 16

2.4 Definition of Chronic Periodontitis 20

2.4.1 Classifications 20

2.4.2 Centre for Disease Control and Prevention – American Academy of Periodontology Definitions 21

2.5 Conceptual Framework 22

CHAPTER III METHODOLOGY **41**

3.1 Study Design and Setting 41

3.1.1 Inclusion Criteria 42

3.1.2 Exclusion Criteria 42

3.2 Data Collection 42

3.3 Features 43

3.3.1 Outcome of Interest 43

3.3.2 Features associated with Periodontitis 44

3.4 Sample size estimation 45

3.5 Data Analysis 46
3.5.1 Data Management 46
3.5.2 Data Preparation 49
3.6 Model Architecture 49
3.6.1 Feature Selection 50
3.6.2 Data Splitting 54
3.6.3 Model Development 55
3.6.4 Performance Evaluation 57
3.7 Limitations 58
CHAPTER IV RESULTS 61
4.1 Description of EGAT Study 61
4.2 Models 61
4.2.1 Mixed effects logistic regression 61
4.2.2 Recurrent Neural Network 62
4.2.3 Mixed effects – Support Vector Machine 64
CHAPTER V DISCUSSIONS 73
5.1 Minority positive class 74
5.2 Limitations of the current study 76
5.3 Application on mock data 78
5.4 Application in real life scenarios 79
5.5 Future Research and Study 80
5.6 Conclusion 81
REFERENCES 87
APPENDIX A ETHICAL CLEARNACE 91
BIOGRAPHY 92

LIST OF TABLES

Table 2.1 Performance of current predictive models 23
Table 2.2 Example of transformed data with 2 features and 2 Classes 24
Table 2.3 Centre for Disease Control and Prevention - American Academy of Periodontology (CDC-AAP) classification 25
Table 3.1 Calibration of periodontal examination (weight kappa ± 1mm) 59
Table 3.2 Labeling Criteria for the dataset 59
Table 3.3 Feature transformation 59
Table 4.1 Fixed Effects Coefficients and Odds Ratio Estimates for Significant Variables Retained in the Final Multivariate Mixed Effects Logistic Regression Model 67
Table 4.2 Performance of Mixed effects logistic regression 67
Table 4.3 Performance of overfit recurrent neural network 68
Table 4.4 Performance of final recurrent neural network 68
Table 4.5 Performance of overfit Mixed Effects – Support Vector Machine 68
Table 4.6 Performance of final Mixed Effects – Support Vector Machine 69
Table 4.7 Performance of all final models (Performance with 95% Confidence Interval) 69
Table 4.8 F1 score and Brier score of the models 69
Table 5.1 Subset of mock data samples 82
Table 5.2 Performance of different classification models on the mock data samples (0.35 as decision threshold) 83

LIST OF FIGURES

Figure 1.1 Periodontal Risk Assessment by Lang and Tonetti (2003) ……14
Figure 2.1 Oral Risk Factors applied in previous literatures and predictive models... 26
Figure 2.2 Demographic and Behavioral risk Factors applied in previous literatures and predictive models..... 26
Figure 2.3 Laboratorial features and biomarkers applied in previous literatures and predictive models..... 27
Figure 2.4 Number of papers in literature review, applying a particular model (Some papers apply multiple models, and each type is only counted once)..... 27
Figure 2.5 Distribution of models in literature review (Some papers apply multiple models, and all models are counted) 28
Figure 2.6 Flow chart for Mixed Effects Logistic Regression – Training Model Generation 28
Figure 2.7 Block Diagram for Mixed Effects Logistic Regression – Testing and Target Transformation..... 29
Figure 2.8 A few possible decision boundary (hyperplanes) for the dataset..... 29
Figure 2.9 Optimal decision boundary (hyperplane) with maximum margin 30
Figure 2.10 Soft margins allow misclassified data points. (The hyperplane is not optimal)..... 30
Figure 2.11 Kernel functions increase the dimension of the dataset, making it linearly separable. 31
Figure 2.12 Ordinary Least Squared Regression..... 31
Figure 2.13 Support Vector Regression 32
Figure 2.14 Soft Margin with Slacks..... 32
Figure 2.15 Expectation-Maximization Algorithm 33
Figure 2.16 Training Mixed Effects Machine Learning Regression..... 33
Figure 2.17 Mixed Effects Machine Learning Regression Framework 34
Figure 2.18 Training Mixed Effects Machine Learning Classification..... 34

LIST OF FIGURES (cont.)

Figure 2.19 Maximum of the absolute change in logit value 35
Figure 2.20 Flow chart for Mixed Effects Support Vector Machine – Training Model Generation 35
Figure 2.21 Block Diagram for Mixed Effects Support Vector Machine – Testing and Target Transformation 36
Figure 2.22 Perceptron of a neural network 36
Figure 2.23 Sigmoid curve or logistic curve 36
Figure 2.24 Architecture of a neural network (Left) and Training error in feed forward network (Right) 37
Figure 2.25 Architectures of a recurrent neural network 37
Figure 2.26 Illustration of a one-to-many recurrent neural network 38
Figure 2.27 Flow chart for Recurrent Neural Networks – Training Model Generation 38
Figure 2.28 Block diagram for Recurrent Neural Networks – Testing and Target Transformation 39
Figure 2.29 Distribution of labeling criteria in literature review (Some papers apply multiple criteria, and all criteria are counted) 39
Figure 2.30 Conceptual framework of chronic periodontitis screening models 40
Figure 3.1 Model Architecture 60
Figure 4.1 Model Development Diagram 70
Figure 4.2 Receiver operating curve of mixed effects logistic regression a) – on training data and right, b) on the testing data 70
Figure 4.3 Receiver operating curve of overfit recurrent neural network 71
Figure 4.4 Receiver operating curve of final recurrent neural network 71
Figure 4.5 Receiver operating curve of overfit Mixed Effects – Support Vector Machine 72
Figure 4.6 Receiver operating curve of final Mixed Effects – Support Vector Machine 72

LIST OF FIGURES (cont.)

Figure 5.1 Cost function for imbalanced class 84
Figure 5.2 Class weight-adjusted cost function..... 84
Figure 5.3 Sample weight-adjusted cost function 85
Figure 5.4 Screening system in action..... 85
Figure 5.5 Mobile application based on the mixed effects logistic regression model. 86

CHAPTER I
BACKGROUND AND RATIONALE

1.1 Background and Rationale

Periodontitis is one of the most common oral diseases and causes of tooth loss in adults.[1] It is the world's 6th most prevalent oral disease, affected around 743 million people worldwide. The prevalence was at 11.2% globally, and 15.0-20.0% of Asians.[2] According to the 8th Thailand national oral health survey (2017), the prevalence of periodontitis in Thai adults was 26%, and for the elderly was 36%. Periodontitis is a complex inflammatory disease that leads to the destruction of the supporting structures around the tooth, resulting in the loosening of the teeth and eventual tooth loss.[3] This leads to decreased occlusal ability, digestive ability, and effectively the patient's quality of life.

In addition to oral manifestations, previous studies found an association of chronic periodontitis with systemic diseases and conditions.[4] The association between atherosclerotic vascular diseases (ASVD) and periodontitis has been established.[5] Joint of European Federation of Periodontology and American Academy of Periodontology (EFP/AAP) Workshop on Periodontitis and Systemic Diseases reported that there is consistent and strong epidemiologic evidence that periodontitis increased risk for future cardiovascular diseases.[6] Chronic periodontitis and diabetes mellitus have bidirectional relationships; and it has also been reported that periodontitis and diabetes had significant direct and indirect effects mediated via each other on chronic kidney disease (CKD) incidence.[7] Relationships between periodontitis and other systemic disease, i.e., chronic obstructive pulmonary disease (COPD), rheumatoid arthritis (RA), Alzheimer's disease and erectile dysfunction, also have been reported. [8]

Severe chronic periodontitis is characterized by loss in alveolar bone height and radiographs are required to assess this sign. Symptoms of non-severe chronic periodontitis are quite negligible until it is too late, then it results in loosening and loss of the tooth. Diagnosis of less severe condition requires a dentist or dental hygienist to

manually measure the distance between the cementoenamel junction and the base of the periodontal pocket for all present teeth using periodontal probes. Such measure is gold-standard, but time and resource-consuming in multiple numbers of cases, for instance public health missions. Such a scenario can be more efficiently addressed by the presence of a screening system, reducing the number of examinees which the dentist must manually conduct comprehensive periodontal probing.

Risk scoring systems such as periodontal risk calculator (PRC)[9, 10] and periodontal risk assessment (PRA)[11] have been proposed by Page *et al.* and Lang *et al.*, respectively. PRC scores the patient with a risk score between 1-5 (1 being low risk and 5 being high) as well as a disease state score between 1-100. It uses 9 features to score, including the radiographic bone height. As seen in Figure 1.1., PRA categorizes the patient into three distinct classes, namely low risk individual, moderate risk individual and high-risk individual. While it uses 6 parameters to score, its parameters include clinical measurements such as presence of bleeding on probing and residual pockets. While the measures enable the process to be more objective, inclusion of clinical parameters restricts the applicability of the system without the presence of dental professionals.

By excluding oral examinations, other parameters such as demographics and risk behaviors are used to assess the risk. Risk prediction models are typically developed by statistical modelling and commonly applied are cross-sectional regression models. While logistic regression models are simple and efficient, they rely on a proper selection of the features, which means feature engineering is vital for the model. A common approach is to use a limited number of known risk factors and domain expert selected features. Supervised nature of the approach misses the opportunity to discover novel patterns, and limited model's performance leads to be suboptimal.

Nowadays, machine learning emerges as an alternative for risk prediction. Machine learning algorithms can have features needed for prediction learned from the available dataset.[12] Their abilities to handle high dimensions remove the necessity for a feature selection, as well as they can handle images or signal data as predictors in addition to the structured data. Machine learning models can learn with specified outcomes (supervised) or without specified outcomes (unsupervised) as well. While unsupervised models are applied to detect the patterns in the data, supervised machine

learning can be applied for both classification and regression tasks. A branch of machine learning called deep learning models are feature learning models, consisting of multiple layers of features, obtained by composing simple but nonlinear modules that each transforms the feature at one layer (beginning from input layer) into a feature at a higher, slightly more abstract layer, resulting in improved prediction from data.[13, 14]

Over traditional statistical modeling, machine learning models can also improve the performance by applying the hyperparameter optimization. Different data patterns require different sets of parameters, to minimize the loss of the classification model. For example, support vector machines can perform on non-linear relationships by applying soft margins, by allowing misclassifications or using kernels to make the classes linearly separable. Hyperparameters of deep learning models, such as activation functions, can be tuned to work with either linear or non-linear relationship between independent and dependent variables.

On the other hand, traditional statistical models have a descriptive model approach such as the relationship between the independent (age) and dependent (incidence of periodontitis) variables; hazard ratio of Cox regression and odd ratio of logistic regression. While this interpretability is preferable for clinical applications, machine learning models tend to have an algorithmic approach model, which performs better for prediction. High performance machine learning models such as deep learning and ensemble learning models are claimed to have “black box”, due to their lack of interpretability. Artificial neural networks have complex network with interconnected nodes or neurons, either passing information to another neuron or not due to being deactivated. Both algorithms, feed forward and backward propagations are going back and forth for all training samples to optimize the model for minimal loss. While these intricate connections and processes lead the model towards higher performance, the relationship between the input (age) and the output (incidence of periodontitis) of the model can no longer be interpreted. For a pure diagnostical purposes, such a model might be less acceptable in a clinical environment. But by applying it in such a way that it can help screen the patients so that they can be informed to emphasize their effort on oral hygiene and dental visits, it can be more of a practical purpose. Our challenges here are to see if the performance of machine learning models can be superior to a statistical

model for our study purpose and if increased performance will be desirable enough for the exchange with the interpretability of the model.

1.2 Research Question

Do machine learning models have better predictive performance than statistical models in screening of chronic periodontitis?

1.3 Research Objectives

The objectives of the study are -

1. Develop statistical and machine learning predictive models on longitudinal data for screening of severe chronic periodontitis.
2. Compare the performance of the predictive models between statistical model and machine learning models on longitudinal data for screening of severe chronic periodontitis.

1.4 Expected Benefits

By deploying a screening model in periodontal examinations, it should reduce the number of people requiring comprehensive periodontal probing, which would further reduce the time, resource requirements and workloads for the examiners. Longitudinal modelling would allow the models to learn the data patterns better than the cross-sectional regression model. After both internal and external validations, the resulting models with proficient performances could be deployed for screening purposes at surveys as well as applied on large longitudinal datasets such as electronic health records to monitor and recommend healthy oral and dental practices or visits at regular intervals.

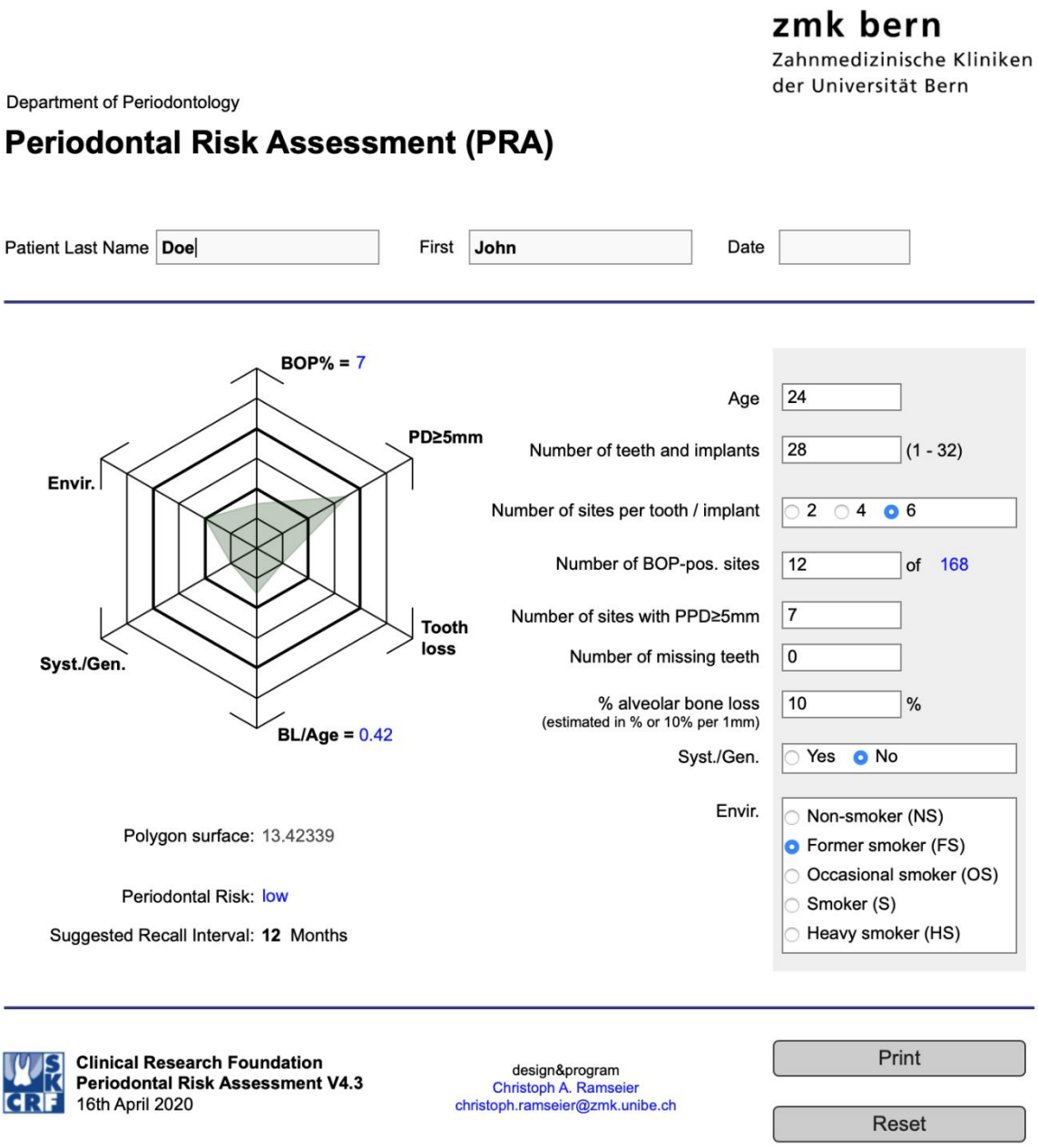

Figure 1.1 Periodontal Risk Assessment by Lang and Tonetti (2003)[15]

CHAPTER II
LITERATURE REVIEW

2.1 Epidemiology of Periodontitis

Periodontitis is the inflammation of periodontium, a disease involving the structure surrounding the tooth and it is considered one of the most common disease in humanity. From the 2009 and 2010 National Health and Nutrition Examination Survey (NHANES), over 47% of the U.S. adults who aged 30 years and older (64.7 million people), representing, had some form of periodontitis. And for adults 65 years and older, 70.1% had periodontal disease.[16]

While the bacterial plaque is considered to be the initiator of the condition, it always is present in the oral cavity as the dental plaque, in both healthy and compromised patients. It is formed from the acquired pellicle, which is a layer of saliva mainly consisting of glycoprotein, shortly after tooth brushing or oral hygiene methods. It helps with the adhesion of the bacteria to the tooth and the mass of bacteria proliferates in the dental plaque, forming bacterial or microbial plaque. With insufficient or improper oral hygiene practices, the plaque builds up to become the tartar or calculus, which further helps the adhesion of the bacterial plaque.

Fortunately, with the effective immune system, periodontal diseases will not develop as long as the balance between the microbial and host response is maintained. This balance can be broken either by the hyper-responsiveness or the high virulence of the bacteria, or by the decrease in host immune by systemic factors. Then the periodontium becomes inflamed and subsequent destruction of alveolar bone happens. But during the initial stages of the inflammation, the symptoms are less noticeable, so the process is encouraged by the patient's negligence of oral hygiene. The disease progresses and the patient may suffer from gingival bleeding with little provocation, gum swelling, dull pain, gingival abscess, and tooth mobility. This leads to tooth loss, decreasing the occlusal ability, digestive ability and, effectively, the patient's quality of life.[17]

2.2 Risk Factors of Periodontitis

Periodontitis is of complex etiological causes. While bacterial plaque is considered to be the initiator of the condition, by acting as the biofilm, there are other local factors such as mal-occlusion, dental restorations and oral prostheses, that encourage the formation of dental plaque. Also, other systematic factors, such as mal-nutrition and poorly controlled diabetes mellitus increase one's susceptibility to periodontal diseases. Oral habits such as smoking and betel quid chewing habits can increase one's risk while oral hygiene habits such as frequency of tooth brushing[18-23] and flossing removes the dental plaque, so reducing the risk of microbial plaque maturing. Smoking is a well-established risk factor, and it is also reported that severity of radiographical bone loss is enhanced by betel/pan chewing. The number of teeth and sometimes decayed, missing, and filled teeth index (DMFT) are also included as oral risk factors for chronic periodontitis. As shown in Figure 2.1, other factors such as tooth mobility, number of teeth with bleeding, and number of teeth which are mostly applied oral risk factors. Demographic risk factors such as smoking, age and sex appear in Figure 2.2.

Demographically, it is reported that the prevalence of periodontitis increases as one grows older.[19-30] Also, periodontitis has a higher prevalence in men (~57%) compared to women (~39%).[18-22, 24, 25] However, the paper also advices that different socioeconomic and behavioral factors between genders might have influenced, rather than the gender bias.[31] Income[18-20, 22] and education levels[18, 19, 22, 24, 25, 27-29, 32] are common features in predictive models. Also, family size[18, 21], body mass index[18, 20, 33], drinking habit[21, 24, 33], diabetes mellitus[19, 26, 34] and hypertension[18, 26, 35] are also suggested.

Several literatures study association between a limited number of potential biomarkers and chronic periodontitis, as shown in Figure 2.3. As mentioned, periodontitis is initiated by the microbial dental plaque and the host susceptibility for it. The presence of sub-gingival pathogens induces local inflammatory response and large number of leukocytes are exuded and migrated as the first line of defense. It is observed that the number of white blood cells (WBC) increases in patients with chronic periodontitis and the increased number of neutrophils and lymphocytes are statistically significant.[36] Immunoglobulin G, part of humoral immunity, is also observed to be

increased. Immunoglobulin G3 serum levels discriminate well between chronic periodontitis and healthy patients. While there is a local inflammation at the site of periodontitis, it is also studied that the patient with chronic periodontitis have low grade systemic inflammation. Interleukin-6 is produced at site of inflammation and considered to be dumped into systemic circulation, increasing interleukin-6 levels. Interleukin-6 also induces hepatic synthesis of C-reactive protein. It is observed that there is an association of chronic periodontitis with high Interleukin-6 levels and high C-reactive protein, measured by high sensitivity C-reactive protein test. Inflammation also adversely affects the lipoprotein levels, being observed the lower high-density lipoproteins levels (HDL) and higher low-density lipoproteins levels (LDL) in patients with chronic periodontitis.

2.3 Predictive modelling of periodontitis

As seen in Figure 2.4 and Figure 2.5, it is observed that logistic regression is mostly applied for predictive modelling. However, application of different criteria for labelling samples results in different performance of the same model. Also, different performance metrics applied by each study reduce comparability, as shown in Table 2.1.

Studies on prediction models try to compare between the performances of included different combinations of features (demographical features, risk behavior data, and oral features). For self-reportable models, questionnaires are used to collect the oral features instead of clinical examination. Eke et al. observes that including both demographic and oral features in the model performs better than only including demographic features. In addition to other features, Verhulst et al. also applies biomarker data of the saliva, resulting in higher and more balanced performance of the model among the reviewed models. However, while it performs better, identifying biomarkers from saliva such as protease and chitinase also consume resources. We need to balance our models between predictive power and required resource.

Nevertheless, the common goal of the majority of the studies is to diagnose periodontitis without clinical examination. By applying only self-reportable features such as demographics and risk behaviors, the resulted model can be applied with a rapid

non-invasive screening tool. With our study, we aim to improve model performance with machine learning models and hyperparameter optimizations.

2.3.1 Statistical Modelling

2.3.1.1 Logistic Regression

Data Transformation

Logistic Regression requires both the inputs and outputs of the model to be numerical. Therefore, for categorical data, feature transformation is required. For target variables, it is necessary to label the target variables' classes as [1] for positive class and [0] for negative class.

Methodology

Logistic regression is a statistical model, which applies the logistic function (sigmoid function) to determine the binary outcome of the sample in its basic form, although there are other complex adaptations of logistic regression for other purposes. In contrast to linear regression where dependent variable is linearly related to independent variable, the log-odds (logarithm of odds) of event is a linear combination of independent variables in logistic regression. It can be univariate (single predictor) or multivariate (multiple predictors). Depending on the number of outputs, it can be binomial (binary outcome), multinomial (more than two possible outcomes) or ordinal (dependent variables have ordinal nature). Logistic regression is usually the model of choice for the stepwise feature selection.

2.3.1.2 Mixed effects Logistic Regression

For training of a classification model, i.i.d assumption (independent, identically distributed) for the training dataset is made. Therefore, vanilla logistic regression cannot be applied for longitudinal datasets, where correlation between repeated measurements violates i.i.d. Mixed effect models are applied in such settings by considering as levels or hierarchy.

Mixed effect models, also called multilevel models, are statistical models considering both fixed and random effects. In biostatistical sense, fixed effects are population-average and random effects are subject-specific effects (also called latent variables, which are assumed to be unknown). Mixed effects models extend

the capability of the regression model, by recognizing that individuals in population are heterogenous. In mixed effects models, each subject is allowed to have their own subject-specific intercept and/or slope. Mixed effects logistic regression, like vanilla models, can also be applied for classification tasks. Flow chart and block diagrams for developing mixed effects logistic regression model as a classifier is shown in Figure 2.6. and 2.7.

$$y_{ij} = X_{ij}b + Z_{ij}u$$

- where

y = target variable (logit)

X = fixed effects feature

b = coefficient of feature X

Zu = random effects variable describing latent variables

i = cluster

j = observation of i^{th} cluster

2.3.2 Machine Learning

2.3.2.1 Support Vector Machine

Data Transformation

Support vector machine requires both the inputs and outputs of the model to be numerical. So, for categorical data, feature transformation is required. For target variables, it is necessary to label the target variables' classes as [1] for positive class and [-1] for negative class.

Methodology

Support vector machine is a type of supervised machine learning algorithm. Support vector machine works exclusively on binary classifications. While the two classes are separated with a decision boundary, such boundary can be drawn in thousands of ways as a few shown in Figure 2.8. The function of support vector machine is to search the best separating line, called the hyperplane, which leaves the maximum margin width from both classes. Support vector machine accomplishes this by considering only the support vectors, which are on the margin of the hyperplane, instead of considering all the data points, as shown in Figure 2.9. Support vectors are the data points that are closest to the other class in hard margins. For the output of the model,

the hyperplane is considered zero and the support vectors are considered [1] and [-1]. Theoretically, a data point can be on the zero plane, which makes it neither in the positive nor the negative class. Practically, only negative values are considered negative [-1] class and other values such as zero and the positive values are considered positive [1] class. Support vector machine performs optimally in linearly separable data.

However, real-life data are rarely linearly separable, due to outliers or noise data. As seen in Figure 2.10., soft margins are applied by considering other data points as support vectors allowing some data points to be on the other side of the hyperplane (misclassified) instead of using a hard margin which has low variance and high bias by overfitting to the training data. Alternatively, using kernel functions increases the dimension of the dataset. For example, in Figure 2.11., the two-dimensional dataset becomes three dimensions, which allows better separability by the linear decision plane.

And, since support vector machine separates using a linear plane, they are limited for binary classifications. However, several workarounds such as one-vs-all approach enables it to be applicable for multiclass classifications as well.

2.3.2.2 Support Vector Regression

Methodology

Support vector regression is an adaptation of support vector machine applying the concept of linear regressions. In ordinary least squared regression, the best fitted regression line is created from the data by minimizing the summation of squared error as shown in Figure 2.12.

$$y'_i = wx_i + b$$

- where

y'_i = regressed value for data point i

x_i = feature of data point i

w = weight or coefficient of feature x

b = bias of the regression line

$$error_i = y_i - y'_i$$

$$min \sum_{i=1}^{n} \|error_i\|^2$$

- where

y_i = actual value of data point i

$error_i$ = error of the regressor for data point i

In real life, the presence of noise data or outliers affects the regression line, and by extension, the error rate. In support vector regression[37], support vectors are determined to set the margin as in conventional support vector machine. Error is calculated only from the data points inside the margin thus ignoring the outliers. The width of the margin must be controlled, since a margin too wide will consider all data points with the model becoming influenced by noise and overfitted. On the other hand, small margin would not be able to learn from the data with the regression line becoming underfitted.

Therefore, for the support vector regression model, we would like to consider as much data points as we can while not becoming overfitted. As in Figure 2.13, the regression line (the hyperplane) and the margins are parallel, so the perpendicular distance between two parallel lines is widened as much as possible.

$$d = \frac{|y' - y|}{\sqrt{w^2 + 1}}$$

- where

d = perpendicular distance between the hyperplane and the margins

y = actual value of the support vector on the margin

y' = regressed value for the support vector

Since the perpendicular distance (d) is inversely proportional to weight (w), w is reduced instead of error as in linear regression. However, as d increases, more data points will be considered so risking overfitting. Therefore, the error for each data point is constrained under the amount of error we are willing to accept called epsilon (ϵ) and it is a hyperparameter.

$$min \sum \|w\|^2$$

$$\|error_i\| \leq \epsilon$$

Applying the concept of margins from conventional support vector machines, soft margins are applied in regression by considering some more data

points outside of the acceptable error (ϵ) as shown in Figure 2.16. By increasing the constraint, we let the model consider more data points called slacks. But since we do not want the model to consider too much data points, we penalize the model based on how much slacks we are giving ourselves.

$$min \sum \|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n}\|\xi_i\|$$

$$\|error_i\| \leq \epsilon + \|\xi_i\|$$

- where

ξ = the amount of slack allowed for the model

And C is also a hyperparameter how much we want to penalize for allowing slacks. Because we penalize only on the data points outside the epsilon zone, it is also known as epsilon insensitive loss. We cannot control how much slacks (may be too few or too many) but the amount of error we are willing to accept is set. This type of support vector regression is called Epsilon regression as shown in Figure 2.14. In another type of support vector regression called nu-regression[38], epsilon (ϵ) is not a hyperparameter but part of the penalty term. Here, ν (nu) is a hyperparameter which determines control the amount of slacks left outside the margin and the value lies between 0 and 1. Since increasing ϵ reduce ξ and the penalty on ϵ is reduced by ν value, ϵ is increased rather than ξ resulting in less slacks.

$$min \sum \|w\|^2 + C(\nu\epsilon + \frac{1}{n}\sum_{i=1}^{n}\|\xi_i\|)$$

2.3.2.3 Mixed Effects Machine Learning

As stated before, linear mixed models consider random effects different between each cluster.

$$y = Xb + Zu$$

Xb is the population average value and it accounts for within-cluster variation. Zu is the subject-specific value and it accounts for between-cluster variation. On the other hand, non-linear mixed models estimate the relationship between features and the target variable as non-linear, and machine learning models can be applied for such relationship.

$$y_{ij} = f(X_{ij}) + Z_{ij}u$$

- where

$f(.)$ = non-linear function

Classical machine learning classification and regression algorithms do not generate high quality models on correlated data so mixed effects machine learning models[39, 40] are developed as an extension of traditional machine learning methods. They are longitudinal/clustered supervised machine learning, as that of learning the two components of a non-linear mixed model separately through an iterative expectation maximization-like algorithm, in which the fixed-effect component is estimated using machine learning methods and the random-effect component is estimated using linear mixed model. By including random effects within the model, mixed effects machine learning is resistant to variabilities introduced by correlated data. Mixed effects machine learning can take advantage of dependencies between the observations to generate more robust and accurate models. It is to be noted that the applied machine learning model here should be a regression model.

Expectation-Maximization Algorithm

It is an iterative algorithm as shown in Figure 2.15.

Step 1. Given a set of incomplete data, consider a set of starting parameters.

Step 2. Expectation step (E — step): Using the observed available data of the dataset, estimate (guess) the values of the missing data.

Step 3. Maximization step (M — step): Complete data generated after the expectation (E) step is used in order to update the parameters.

Step 4. Repeat step 2 and step 3 until convergence.

Regression

Initial random effects are set as zero. Since we consider the target value to be the summation of fixed and random effects, fixed effects are calculated by subtracting random effects from the target and they are trained as the modified target value for the machine learning regressor. After training the machine learning model, the model is used to predict the value for each observation. The predicted values are subtracted from the target and the residuals are estimated to be the random effects used to train the linear mixed model. By the trained linear mixed model, new random effects

are re-estimated. The stopping criterion is set and until it is met, the fixed effects are calculated again by redacting random effects. Machine learning model is retrained, and the loop is continued as shown in Figure 2.16.

Stopping criteria are set in terms of maximum iterations and absolute change in likelihood of the mixed model. Recommended setting for maximum iterations value is not stated in the literature. Convergence in term of likelihood is set similar to statistical models as well where the iterations are proceeded until the change of the estimated parameter vector is negligible with respect to the accuracy of the estimates.[41] In STATA, tolerance for change is 1e-6 and maximum iteration is 300. In R(lme4), tolerance for change is 1e-6 and maximum iteration is 50.

When the model is applied, both the trained machine learning regressor and the mixed model are used to predict the values and they are summed. For samples not in the training data, the random effects are unknown therefore zero is used, which means they are predicted in terms of fixed effects only as shown in Figure 2.17.

Classification

For classification, the target value must be transformed into numerical or logit value manually since we are applying two regressor models. All initial processes are similar with regression framework, until the convergence criteria are met. This is considered as the inner loop or micro iteration. After the inner loop, fixed effects are predicted by the machine learning regressor and random effects are estimated by the linear mixed model. Both effects are summed to create the logit value for each observation. The logit value is transformed into probability and the probabilities are dichotomized using a decision threshold. The resultant classes are considered as the new target class.

The new targets are transformed into logit values again, and previously estimated random effects are removed from this to create new fixed effects. Machine learning regressor is trained again with updated fixed effects and the inner loop is restarted. This step is called the outer loop or macro iteration. The inner loop is repeated until the convergence criteria, and it leads to the outer loop again. The outer loop will have its own convergence criteria, and both will be repeated until both loops converge .

Convergence criteria for inner loop are the same as the regression framework. For the outer loop, there is no recommendation for maximum number of iterations as well so it must be set based on the computation resource and time resource availability. As shown in Figures 2.18 and 2.19, maximum of the absolute change in logit value is also monitored and the loop is continued as long as the value is more than the tolerance. During the application, the output of the model is calculated the same as before, but it is the logit value, so it is transformed into probability and dichotomized. Figures 2.20 and 2.21 visualize the flowchart and block diagram for developing a mixed effects machine learning framework with Support vector machine as the fixed effects estimator.

2.3.3 Deep Learning

Deep learning is a branch of machine learning methodologies whether it is supervised, semi-supervised or unsupervised. They are artificial neural networks at the basic level - including input layer, hidden layer, and the output layer - with single or multiple neurons at each layer depending on the purpose and architecture accordingly. While a basic artificial neural network including three basic layers are considered as multiple layer perceptron, a deep neural network includes more than three - i.e., more than one hidden layer. The larger the numbers of hidden layers in a neural network, the longer it takes the network to train and produce the output, but such architectures are considered to have better performances at solving more complex relationships between the independent and dependent variables. While statistical models and classical machine learning models' functionalities are limited to structured or tabulated data, members of the deep learning models can be applied with non-structured data such as image analysis and signal processing.

2.3.3.1 Artificial Neural Networks

Data Transformation

Artificial neural network requires both the inputs and outputs of the model to be numerical. So, for categorical data, feature transformation is necessary. For target variables, it is necessary to label the target variables' classes as [1] for positive class and [0] for negative class.

However, labeling of the target variable is also different based on the function of the model and the activation functions applied. For binary classification with sigmoid function, the samples are labeled 0 or 1. But SoftMax function requires separate target variables for each class, so the samples are labeled as [0,1] or [1,0]. In multi-class classifications where classes are mutually exclusive, it is necessary for the target to be one-hot-encoded such as [1,0,0], [0,1,0] and [0,0,1]. However, multi-label classifications where one sample can have multiple labels, targets are labeled as [1,1,1], [0,1,1], [1,0,1], [1,1,0], [0,0,1], [0,1,0], [1,0,0] and [0,0,0].

Methodology

Neural networks are loosely modeled after human brain, consisting of interconnected simple processing units, which learns from experience by modifying the connections. Neural networks are called deep learning as well, because of the presence of multiple hidden layers. While a neural network consists of multiple layers, the architecture can be categorized into three groups, input layer, hidden layers, and output layer.

The number of nodes (neurons) in the input layer are equal to all the features of the dataset or the features we selected for the prediction of the target variable. Neural networks require numerical values as input, so encoding for categorical variables is necessary. For ordinal variables, ordinal encoding is used, and one-hot-encoding is used for nominal variables, as shown in Table 2.2.

Hidden layer can be single or multiple, and these layers are where major computations of the neural network happens. As in Figure 2.22., a neuron in hidden layer uses the concept of perceptron, which is assigning weights to each input of the node. However, the weights of the inputs are not known at the beginning of the model, so random weights to the inputs and bias to the layers are assigned. The combination of weights, inputs and bias creates the linear relationship between the inputs and output of the node, an activation function is used to introduce non-linearity. For example, as shown in Figure 2.23., sigmoid function compresses the output value [$-\infty, \infty$] (x-axis) to [0, 1] (y-axis), and the output value is passed to the next layer, which can be either another hidden layer or the output layer.

The number of neurons in the output layer differs based on the function of the model and the activation functions applied. For regression, there is single

neuron, and no activation function is required. For classification, it can also be single neuron (single output) if the classification is binary, and the activation function is sigmoid. However, for multiclass or multilabel classifications and other activation functions such as SoftMax, multiple nodes (multiple output) in output layer are necessary. This process of passing from input layer to hidden layers to output layer is called “feed-forward” as seen in Figure 2.24. (Left).

However, since our initial weights are assigned at random, chances are the output value of the model is different to real value as it is in Figure 2.24. (Right). So, another process called “backward propagation” is used to correct this, by comparing the predicted value with the real value. The loss of this prediction is calculated, and the weights of the nodes are updated based on the nodes’ responsibility for the loss. The weights are increased or decreased to have the prediction closer to the ground value. This process of feed-forward, back-propagation is repeated for all samples in the dataset.

During the weight adjustment, some nodes get their weights changed into zero, which means the node will no longer contribute to the output. This is called “deactivated nodes”, and this allows the neural network to be applied without feature selection. Also, one of the biggest advantages of artificial neural networks is ability to model non-linear and complex relationship. However, neural networks are extremely complex and uninterpretable, so they are said to have a “Black box” as well.

2.3.3.2 Recurrent Neural Networks

Data Transformation

Recurrent neural networks, similar to artificial neural networks, requires both the inputs and outputs of the model to be numerical. For target variables, it is necessary to label the target variables’ classes as [1] for positive class and [0] for negative class, similar to artificial neural networks.

Also, unique for recurrent neural networks, the number of outputs can be as much as the number of time steps (observations) depending on the architecture. For multivariate models, the architecture can be many-to-one as well as many-to-many, as shown in Figure 2.25.

Methodology

Recurrent neural networks are considered part of the representative learning algorithms, specializing in temporal sequence. Recurrent networks remember the past and its decisions are influenced by what it learnt from the past. Therefore, the outputs of the model are not only influenced by the weights applied to the input like traditional neural networks, but also the hidden state vector, representing the context of prior input and/or output. The major application of recurrent neural networks is natural language processing and voice recognition, where the previous context is necessary.

As in Figure 2.26., the hidden state vector is initialized randomly and passed it into the activation function with the input. The activation is typically tanh function, which compress the output value [$-\infty$, ∞] (x-axis) to [-1, 1] (y-axis). The output of the function is passed to another activation function, sigmoid or SoftMax depending on the model, for the output of the observation. However, the same output of the tanh function also passes to the next tanh function together with the next observation of the input and it is repeated for all the observations. Therefore, the context of the previous observations is stored and passed along the time steps. Recurrent neural networks are unique in a way that the same weight is applied to all the inputs of the same parameter, but the different outcomes at different observations are resulted by the different hidden state vectors resulting from previous outcomes. Recurrent neural networks are trained with one sample at a time. Of the same sample, RNN cells train from one time-step to another. The output of the model is compared with the ground value, and the loss is calculated using loss function. The weights of the model are readjusted using backward propagation and gradient descent.

Normally, the loss value is decreased by using gradient descent. Backward propagation finds the derivatives of the networks by moving layer by layer from final layer. However, since activation functions such as sigmoid and tanh compress the output value, the gradient decreases exponentially as we propagate backwards towards the initial layers. Small gradient means the weights will not be updated as effectively by each training sessions. But the initial layers are important to recognize the core elements of the input data, and this ultimately leads to inaccuracy of the model. Such problem is susceptible by deeper neural networks (more layers), and in recurrent

neural networks solve this by applying more complex architecture, such as long short-term memory units. Flow chart and block diagram for developing recurrent neural network as the classification model is shown in Figures 2.27 and 2.28.

2.4 Definition of Chronic Periodontitis

2.4.1 Classifications

Periodontitis is characterized by loss in alveolar bone height, so it is inferred the loss in attachment of junctional epithelium, which is the clinical attachment level. The severity of periodontitis is considered by the increasing measurement of clinical attachment level. Clinical attachment level is measured from the cementoenamel junction to the junction epithelium (base of the periodontal pocket). American Academy of Periodontology (1994) classifies chronic periodontitis as slight (1-2 mm), moderate (3-4 mm), or severe ($\geq$ 5 mm).[42] Although the AAP 1999 definition was widely accepted in clinical circumstances, it was not uniformly adopted by periodontal research.

According to a systematic review of a common definition for periodontitis[7], while most of the studies relies on clinical examination, selected periodontal parameters are quite different. It was found that several parameters, such as clinical attachment level, periodontal pocket depth and bleeding on probing, are used separately or jointly to define periodontitis. Also, other measures, such as the cut-off points for the measurements and, the distribution of periodontally compromised teeth, are lacking in uniformity.

During the literature review, it is observed that Centre for Disease Control and Prevention - American Academy of Periodontology (CDC-AAP) classification for periodontitis is mostly applied, as seen in Figure 2.29. However, it is also observed that self-determined criteria to classify periodontitis are applied nearly as much, signifying the lack of uniformity in defining the condition. While World Health Organization's Community Periodontal Index for Treatment Needs (CPI-TN) is applied as well, it should be noticed that the index tries to identify the level of treatment needed for the patient explicitly instead of diagnosing the condition. Consensus report of 5th European Workshop in Periodontology proposes a new criterion for identifying periodontitis by

staging. The staging procedure takes three criteria in consideration: greatest decrease in clinical attachment level, radiographic bone loss and tooth loss due to periodontitis.

Due to lack of uniformity, the Centre for Disease Control and Prevention and American Academy of Periodontology proposed a new standard case definition for surveillance of periodontitis and the criteria are stated in Table 2.3[16, 43]. Several literatures have used Centre for Disease Control and Prevention - American Academy of Periodontology definitions.[18, 19, 22, 24, 25, 44]

2.4.2 Centre for Disease Control and Prevention – American Academy of Periodontology Definitions

Loss in alveolar bone height is the characteristic of chronic periodontitis and it is the measurement between cementoenamel junction to tip of the alveolar bone. Without radiographs, loss in attachment of junctional epithelium, called Clinical Attachment Level, is measured from the cementoenamel junction to the junctional epithelial attachment. Periodontal Pocket Depth is also an alternative measurement, which is the distance between coronal margin of the gingival sulcus to the base of the gingival sulcus or the periodontal pocket.

Periodontal pocket depth and clinical attachment levels are measured using periodontal probes such as the University of North Carolina-15 (UNC-15) probe. For every tooth present in the dentition excluding the third molar, six sites of the gingival sulcus are measured :

1. labial or buccal site,
2. labio-mesial or bucco-mesial site,
3. labio-distal or bucco-distal site,
4. palatal or lingual site,
5. palato-mesial or linguo-mesial site and
6. palate-distal or palate-distal site.

Applying these measurements, Centers for Disease Control and Prevention - American Academy of Periodontology criteria categorize periodontitis into four levels: none, mild, moderate, and severe periodontitis.

- Severe periodontitis : two or more interproximal sites with clinical attachment levels more than or equal to 6 mm that are not on the same

tooth AND one or more interproximal sites with periodontal pocket depth more than or equal to 5 mm.

- Moderate periodontitis : two or more interproximal sites with clinical attachment levels more than or equal to 4 mm, OR two or more interproximal sites with periodontal pocket depth more than or equal to 5 mm, not on the same tooth.
- Mild periodontitis : two or more interproximal sites with clinical attachment levels more than or equal to 3 mm, AND two or more interproximal sites with periodontal pocket depth more than or equal to 4 mm, not on same tooth or one site with periodontal pocket depth more than or equal to 5 mm.
- None / Healthy periodontium.

2.5 Conceptual Framework

According to the research question and objectives of our study, a conceptual framework for developing algorithms screening severe chronic periodontitis is constructed in Figure 2.30.

Table 2.1 Performance of current predictive models

Paper	**Best performing model**	**Performance Metrics**								
		Sens.	Spec.	Acc.	Prec.	AUC	PPV	NPV	Corr. Coef.	MSE
Leite et al. [18]	LR	67.57	67.50	--	--	0.670	--	--	--	--
Cyrino et al. [19]	LR	54.4	94.3	--	--	0.833	--	--	--	--
Thakur et al. [34]	ANN	--	--	--	--	--	--	--	0.82072	0.079978
Shankarapillai et al. [26]	ANN	--	--	--	--	--	--	--	0.97809	0.13281
Zhan et al. [25]	LR	80.0	72.7	--	--	0.830	74.6	78.5	--	--
Özden et al. [27]	SVM	--	--	--	0.98	--	--	--	--	--
Özden et al. [27]	ANN	--	--	--	--	--	--	--	0.4061	--
Özden et al. [27]	DT	--	--	--	0.98	--	--	--	--	--
Lai et al. [20]	LR	63.5	68.6	65.8	--	0.712	61.6	70.3	--	--
Javali et al. [21]	LR	--	--	61	--	0.7509	--	--	--	--
Eke et al. [22]	LR	93.5	29.2	--	--	0.79	--	--	--	--
Wu et al. [24]	LR	--	--	--	--	0.93	--	--	--	--
Verhulst et al. [29]	LR	80	88	--	--	0.91	93	69	--	--

Abbreviations –
Acc. = Accuracy; AUC = Area under receiver operating characteristic (ROC) curve; ANN = Artificial neural networks; Corr. Coef. = Correlation coefficient; DT = Decision tree; LR = Logistic regression; MSE = Mean squared error; NPV = Negative predictive value; PPV = Positive predictive value; Prec. = Precision; Sens. = Sensitivity; Sens. + Spec. = Sensitivity + Specificity; Spec. = Specificity; SVM = Support vector machine

Table 2.2 Example of transformed data with 2 features and 2 Classes

Age	**Education Level (Ordinal)**	**Education**	**Occupation**	**Occupation_ doctor**	**Occupation_ engineer**	**Occupation_ programmer**	**Label**	**Class**
	Ordinal Categorical variable	Ordinal Encoding	Nominal Categorical variable		One Hot Encoding		Binary Categorical Variable	Binary Encoding
35	Primary School	0	Programmer	0	0	1	Periodontitis	1
25	Bachelor's degree	3	Doctor	1	0	0	Healthy	0
21	Middle School	1	Doctor	1	0	0	Healthy	0
30	Middle School	1	Programmer	0	0	1	Periodontitis	1
29	High School	2	Doctor	1	0	0	Periodontitis	1
22	High School	2	Engineer	0	1	0	Healthy	0
22	High School	2	Doctor	1	0	0	Healthy	0
29	Bachelor's degree	3	Engineer	0	1	0	Healthy	0
33	High School	2	Engineer	0	1	0	Periodontitis	1
29	High School	2	Programmer	0	0	1	Periodontitis	1

Table 2.3 Centre for Disease Control and Prevention - American Academy of Periodontology (CDC-AAP) classification

Case	***Definition***
No periodontitis	No evidence of mild, moderate, or severe periodontitis
Mild periodontitis	≥2 interproximal sites with clinical attachment level ≥3 mm, and ≥2 interproximal sites with periodontal pocket depth ≥4 mm (not on same tooth) or one site with periodontal pocket depth ≥5 mm
Moderate periodontitis	≥2 interproximal sites with clinical attachment level ≥4 mm (not on same tooth), or ≥2 interproximal sites with periodontal pocket depth ≥5 mm (not on same tooth)
Severe periodontitis	≥2 interproximal sites with clinical attachment level ≥6 mm (not on same tooth) and ≥1 interproximal site with periodontal pocket depth ≥5 mm

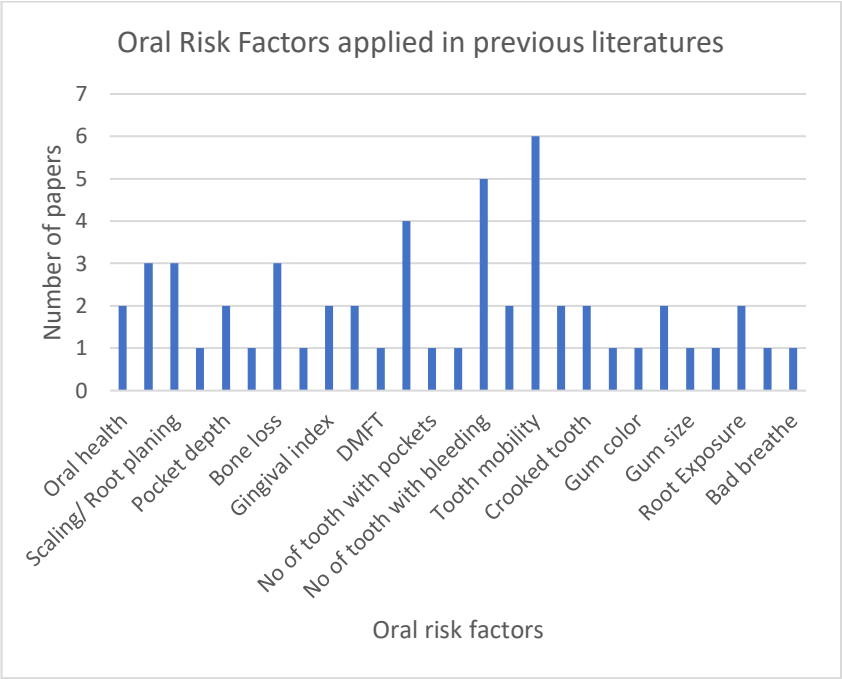

Figure 2.1 Oral Risk Factors applied in previous literatures and predictive models

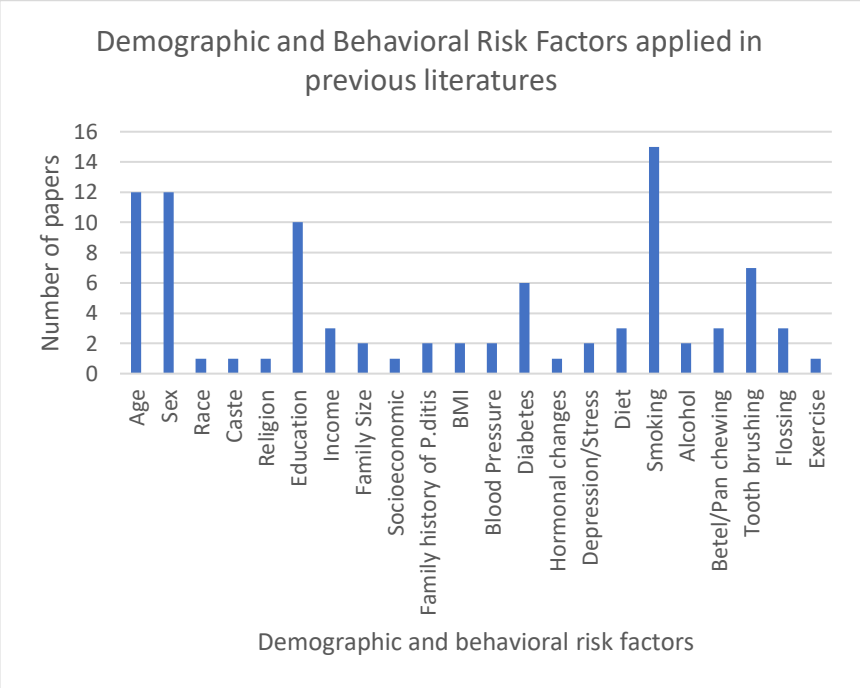

Figure 2.2 Demographic and Behavioral risk Factors applied in previous literatures and predictive models

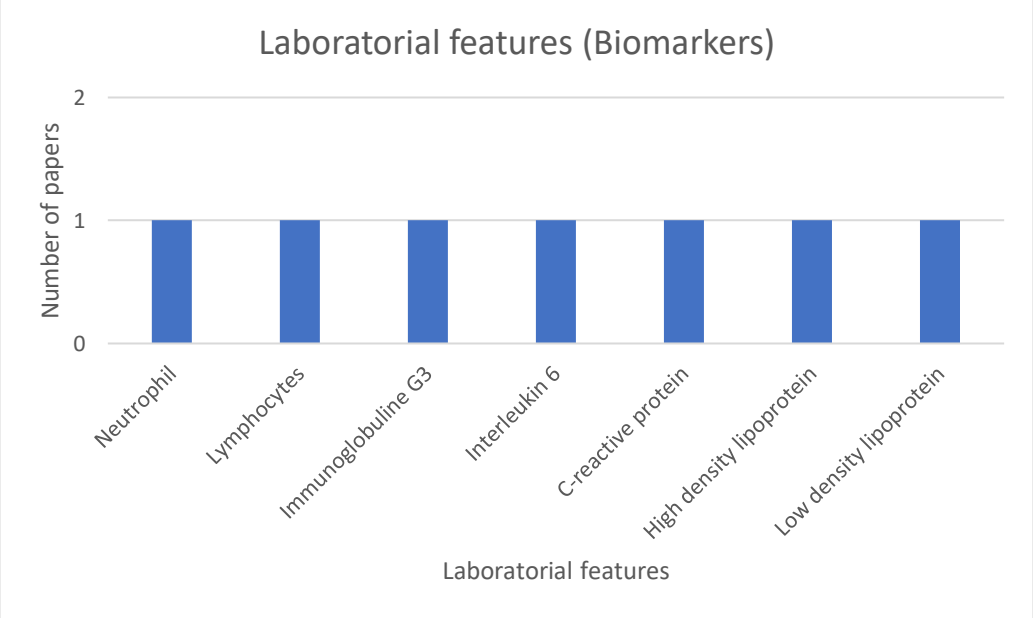

Figure 2.3 Laboratorial features and biomarkers applied in previous literatures and predictive models

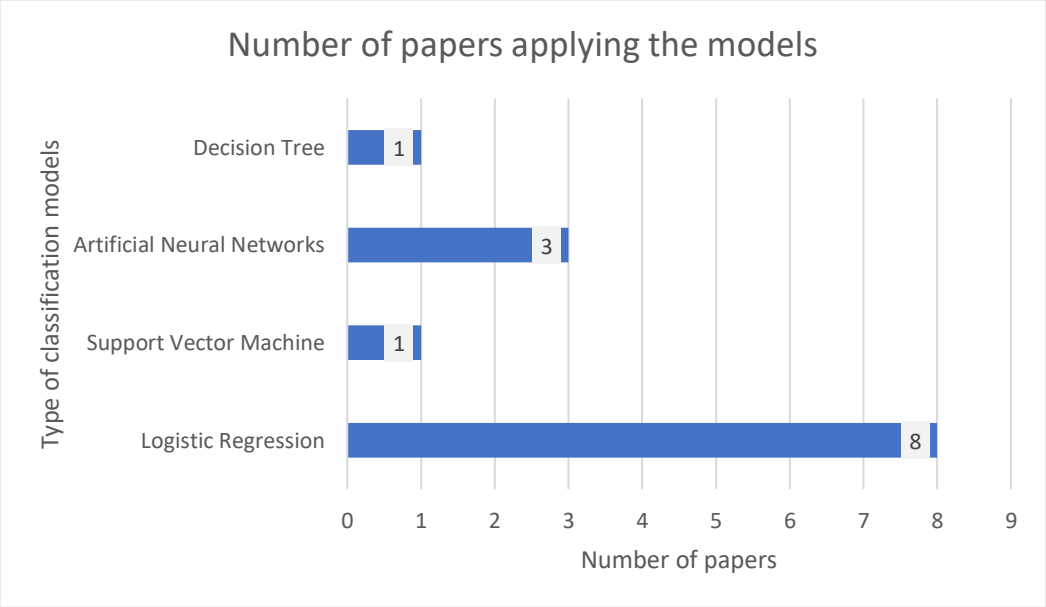

Figure 2.4 Number of papers in literature review, applying a particular model (Some papers apply multiple models, and each type is only counted once)

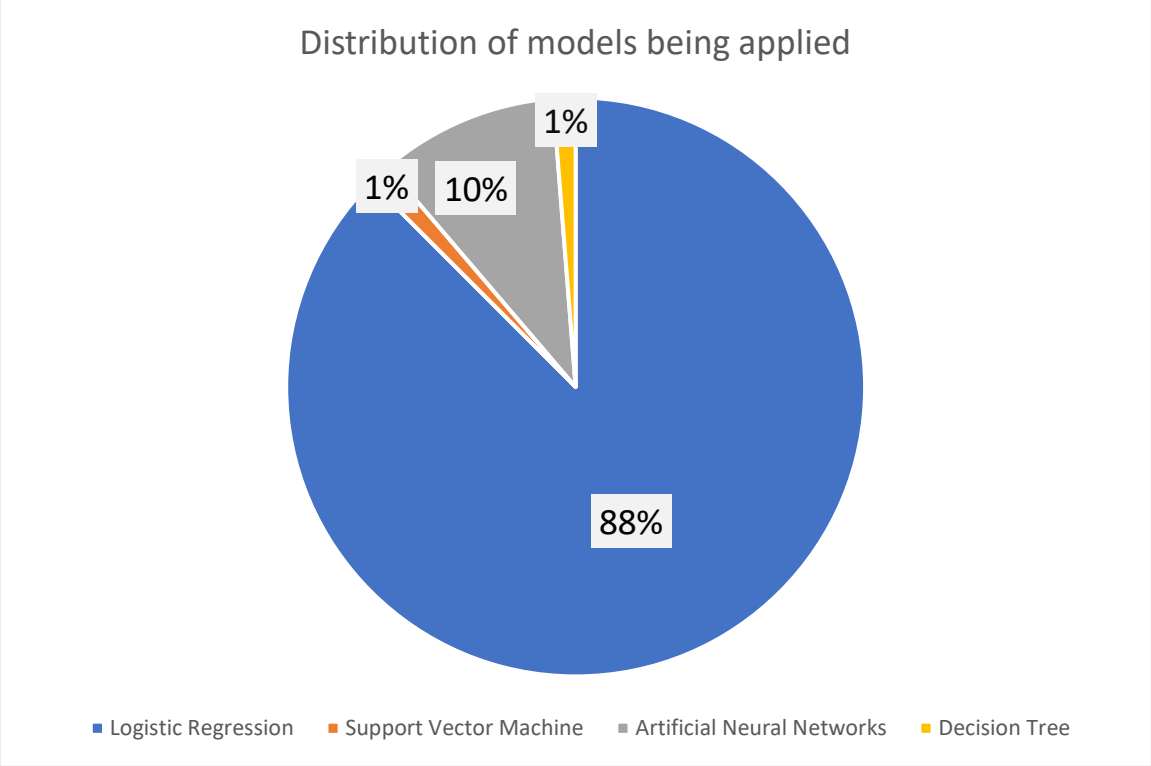

Figure 2.5 Distribution of models in literature review (Some papers apply multiple models, and all models are counted)

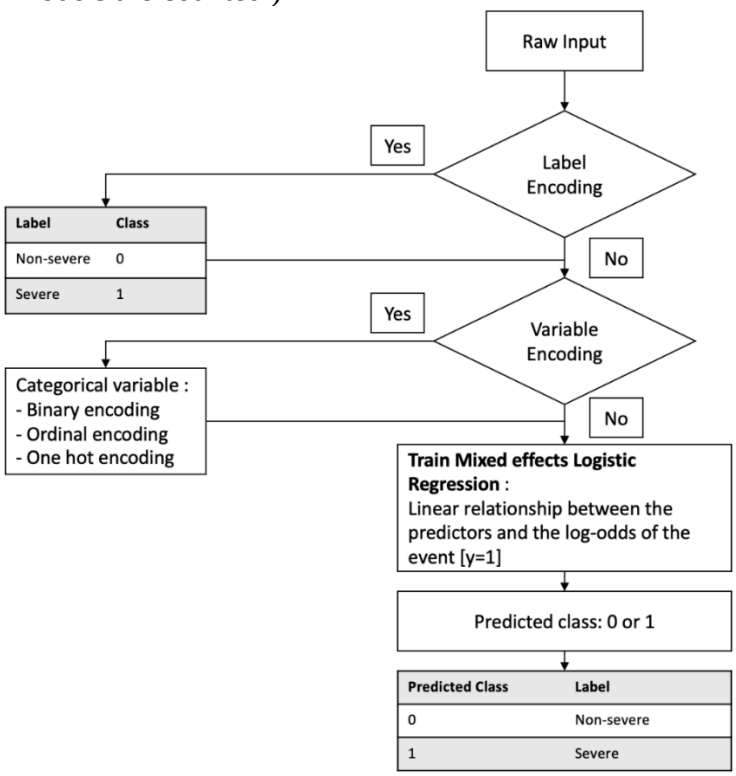

Figure 2.6 Flow chart for Mixed Effects Logistic Regression – Training Model Generation

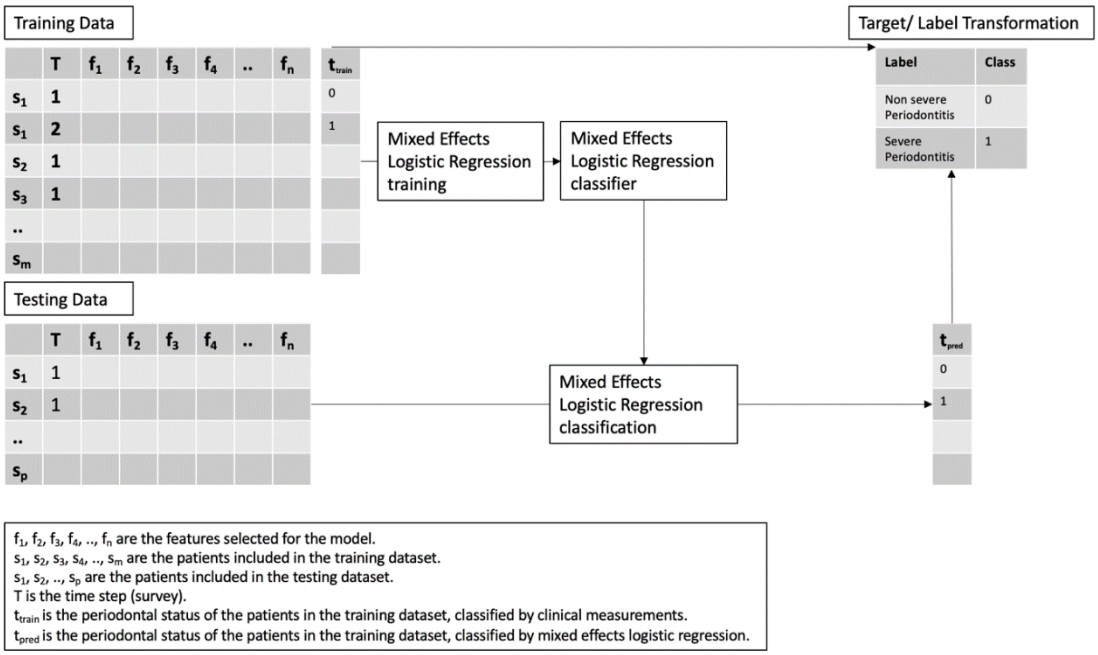

Figure 2.7 Block Diagram for Mixed Effects Logistic Regression – Testing and Target Transformation

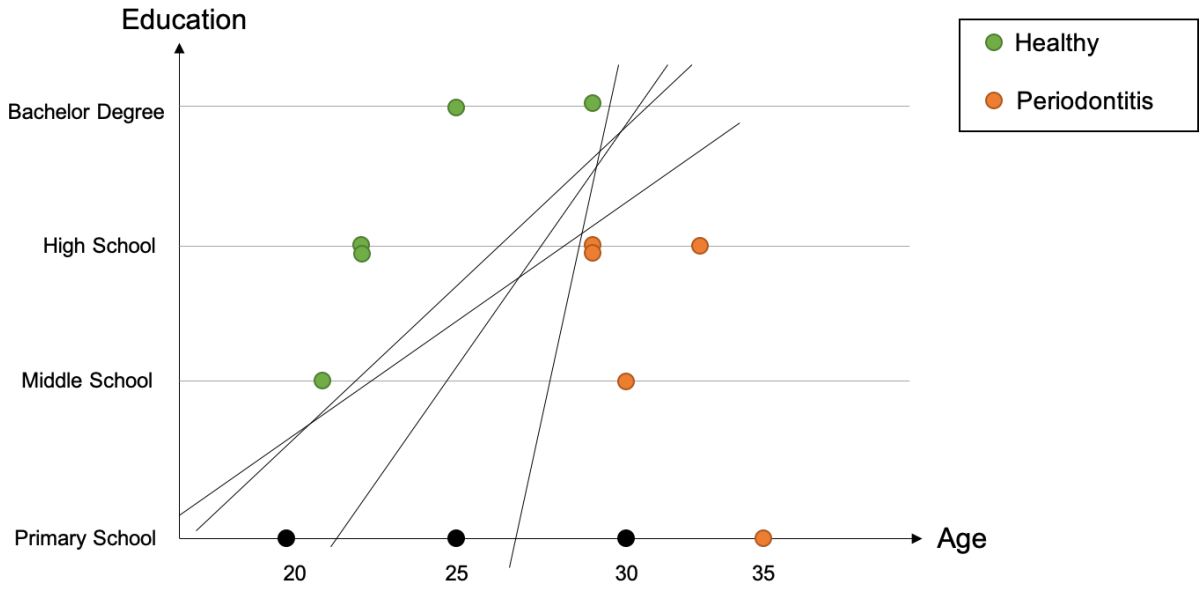

Figure 2.8 A few possible decision boundary (hyperplanes) for the dataset

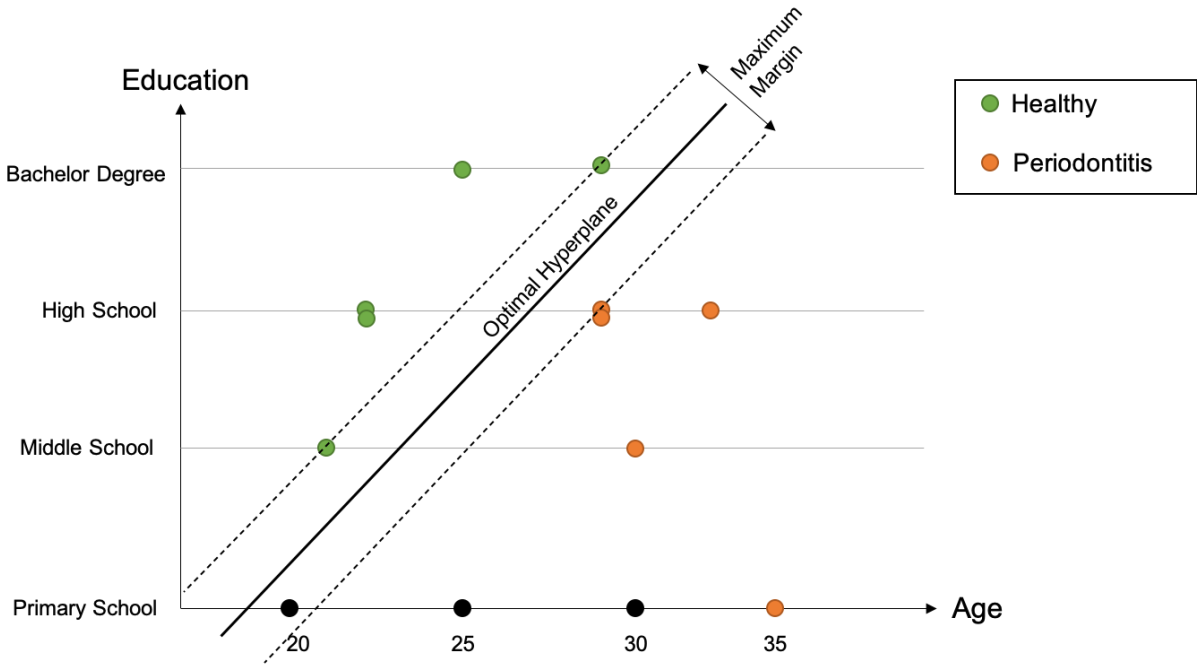

Figure 2.9 Optimal decision boundary (hyperplane) with maximum margin

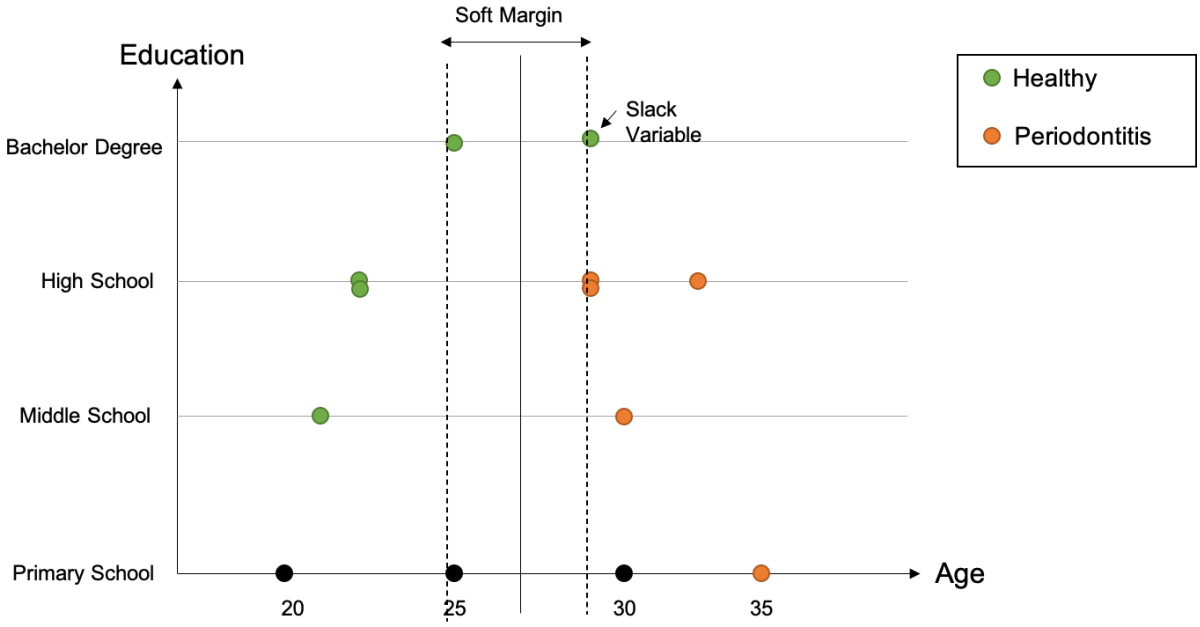

Figure 2.10 Soft margins allow misclassified data points. (The hyperplane is not optimal)

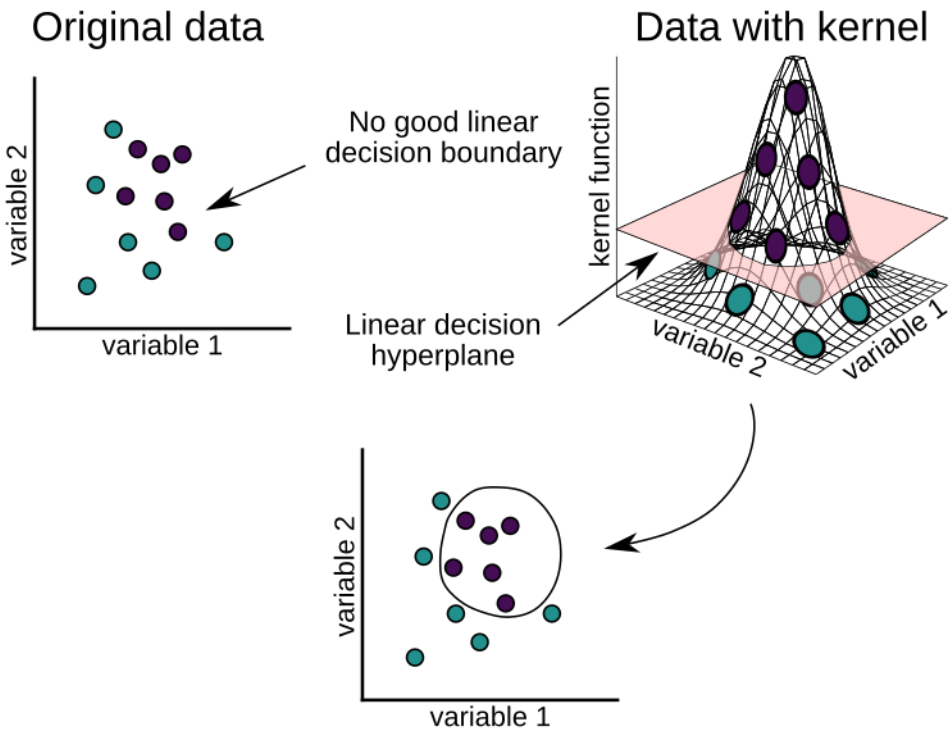

Figure 2.11 Kernel functions increase the dimension of the dataset, making it linearly separable.

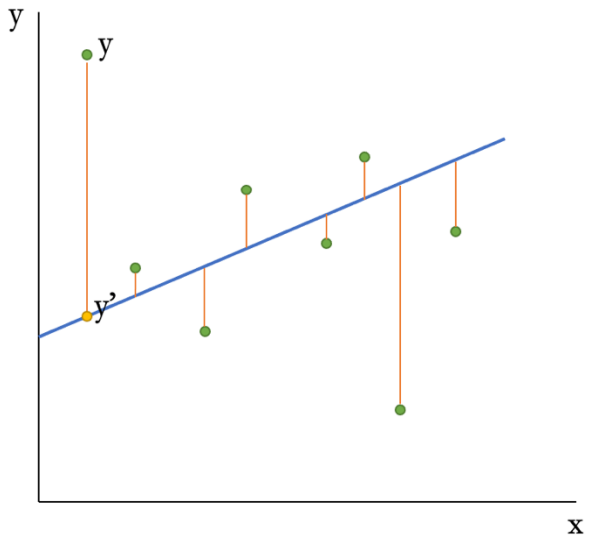

Figure 2.12 Ordinary Least Squared Regression

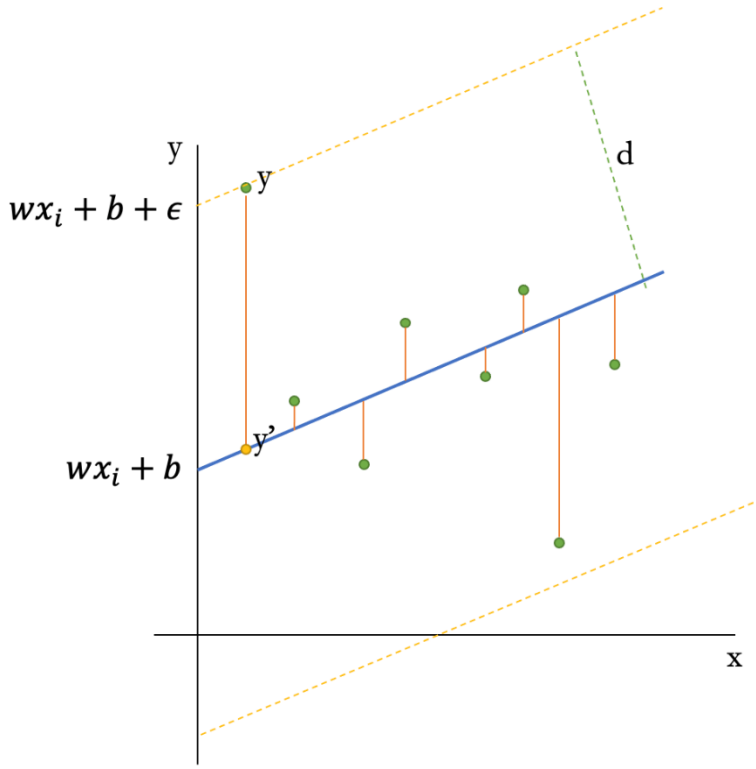

Figure 2.13 Support Vector Regression

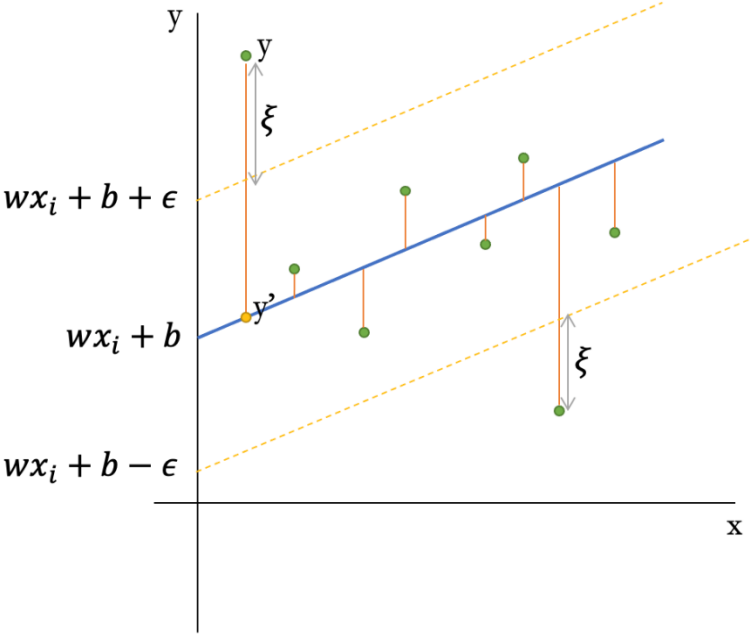

Figure 2.14 Soft Margin with Slacks

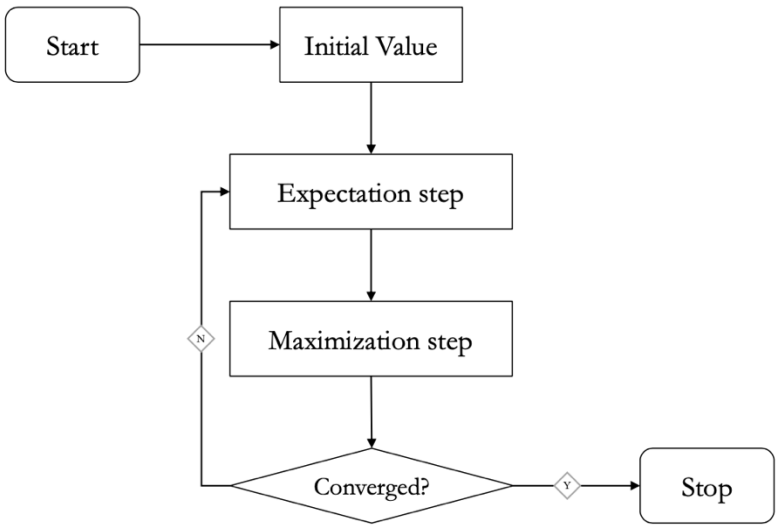

Figure 2.15 Expectation-Maximization Algorithm

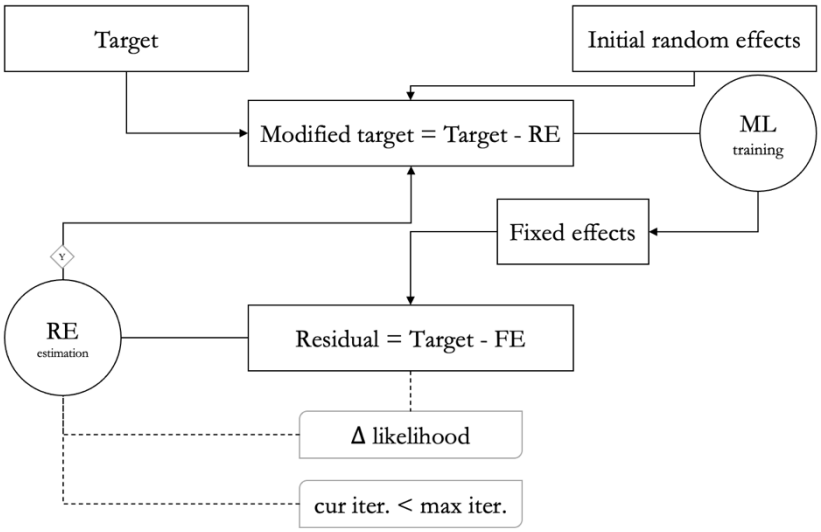

Figure 2.16 Training Mixed Effects Machine Learning Regression

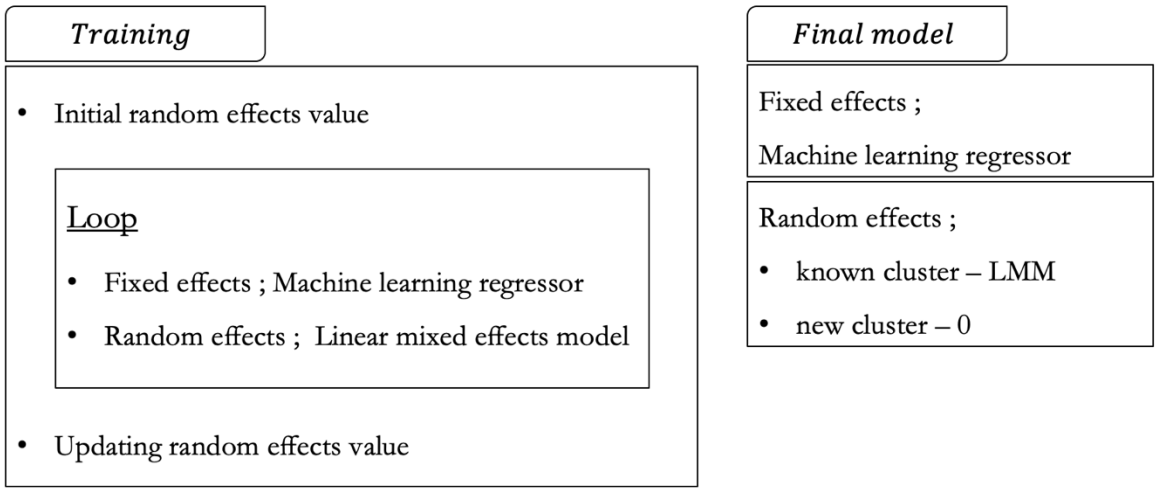

Figure 2.17 Mixed Effects Machine Learning Regression Framework

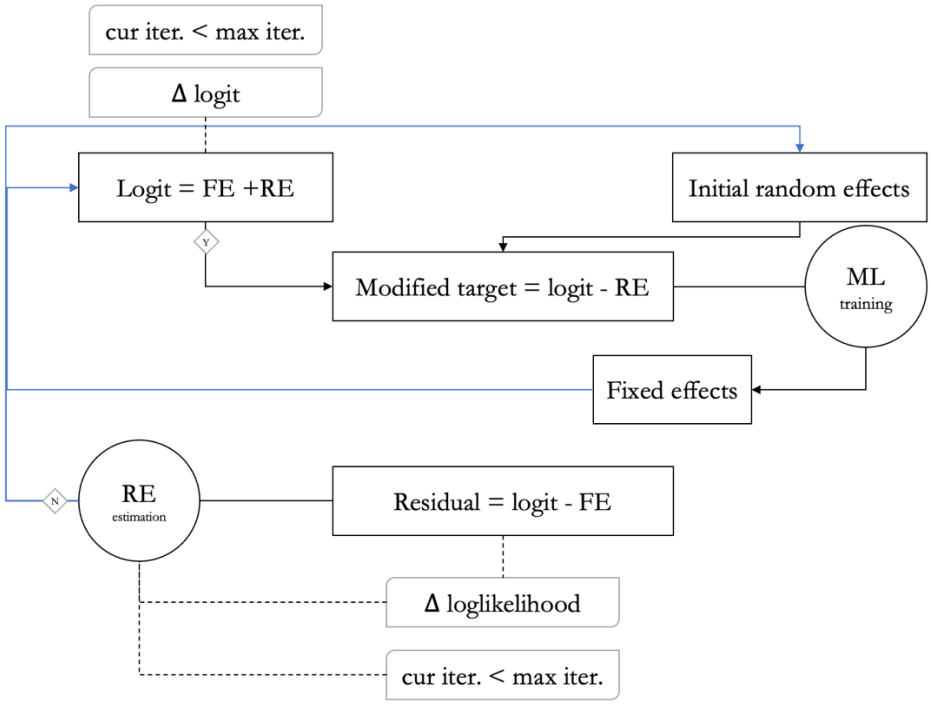

Figure 2.18 Training Mixed Effects Machine Learning Classification

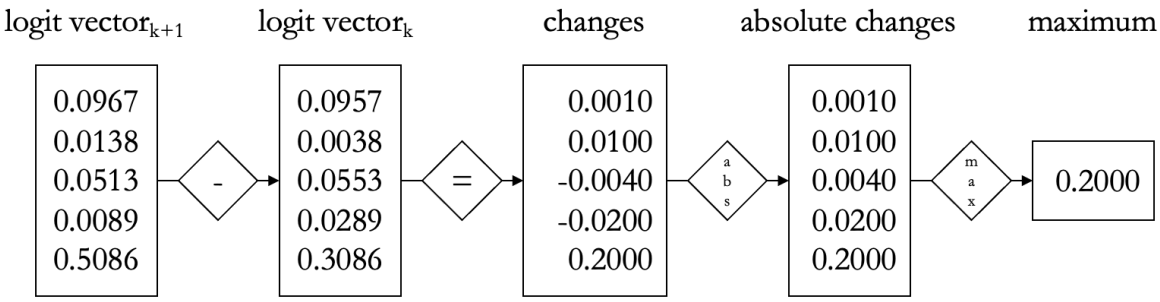

Figure 2.19 Maximum of the absolute change in logit value

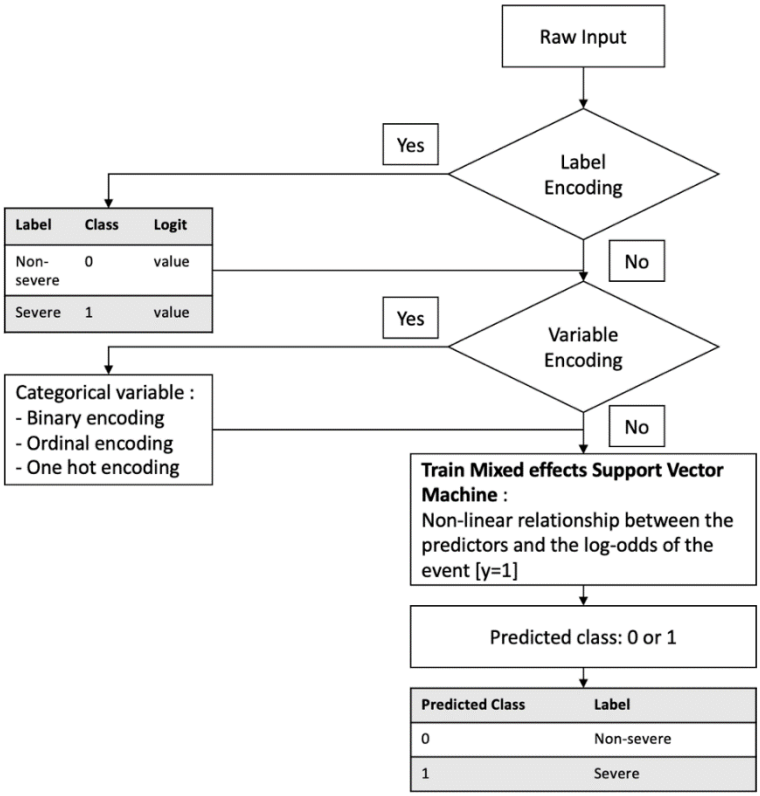

Figure 2.20 Flow chart for Mixed Effects Support Vector Machine – Training Model Generation

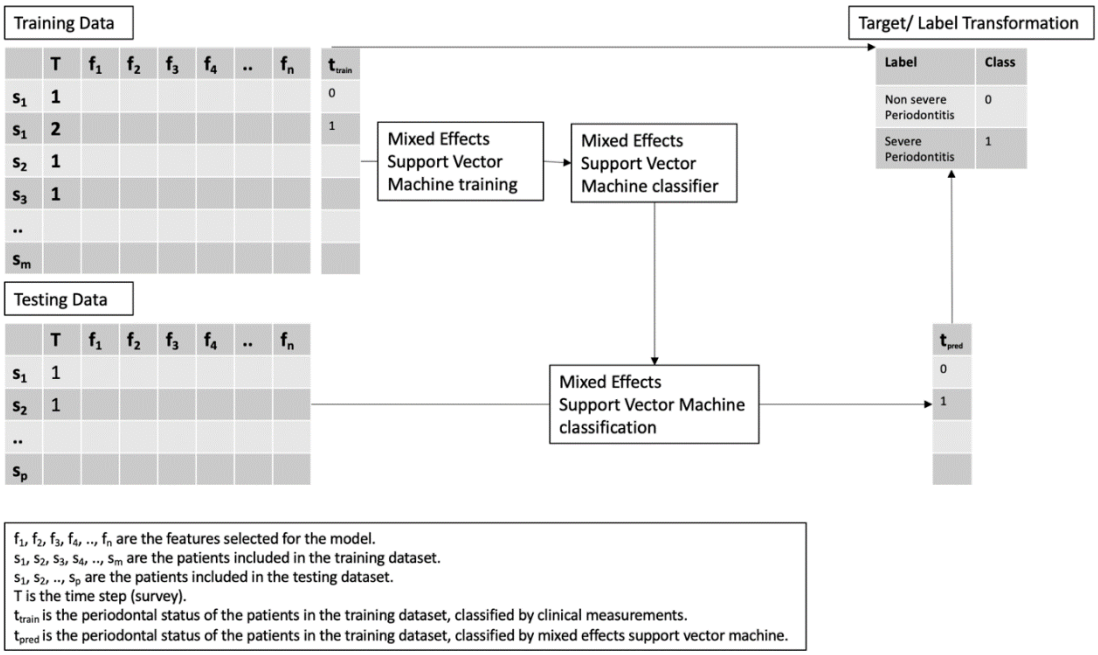

Figure 2.21 Block Diagram for Mixed Effects Support Vector Machine – Testing and Target Transformation

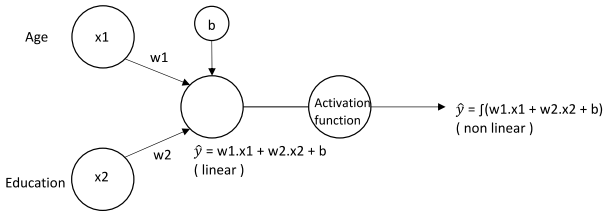

Figure 2.22 Perceptron of a neural network

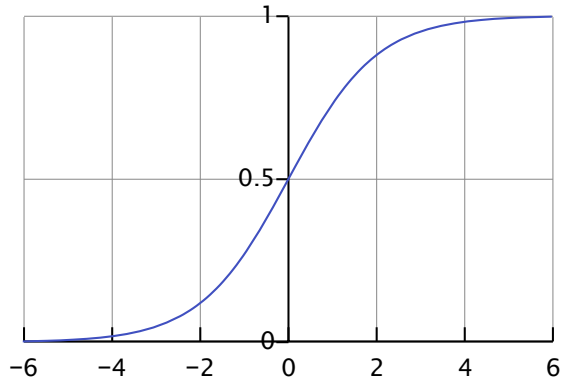

Figure 2.23 Sigmoid curve or logistic curve

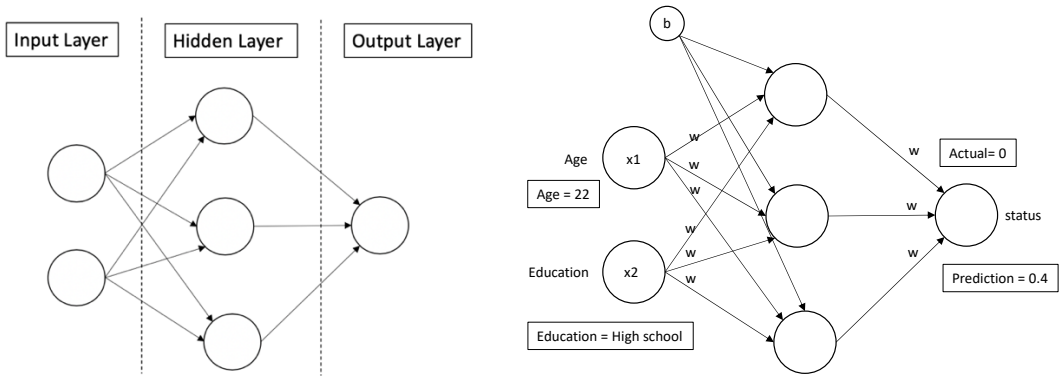

Figure 2.24 Architecture of a neural network (Left) and Training error in feed forward network (Right)

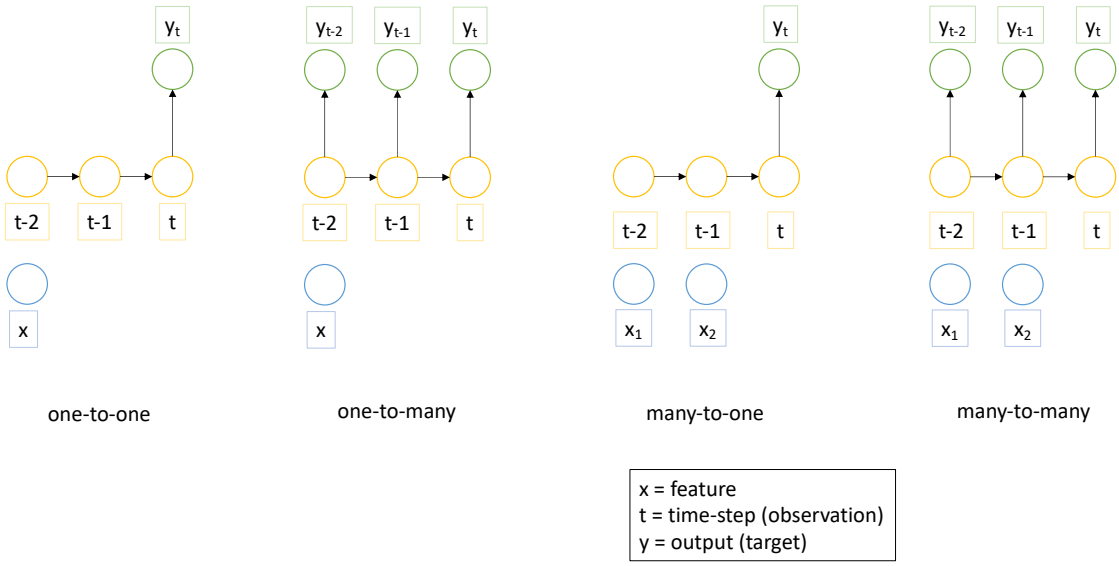

Figure 2.25 Architectures of a recurrent neural network

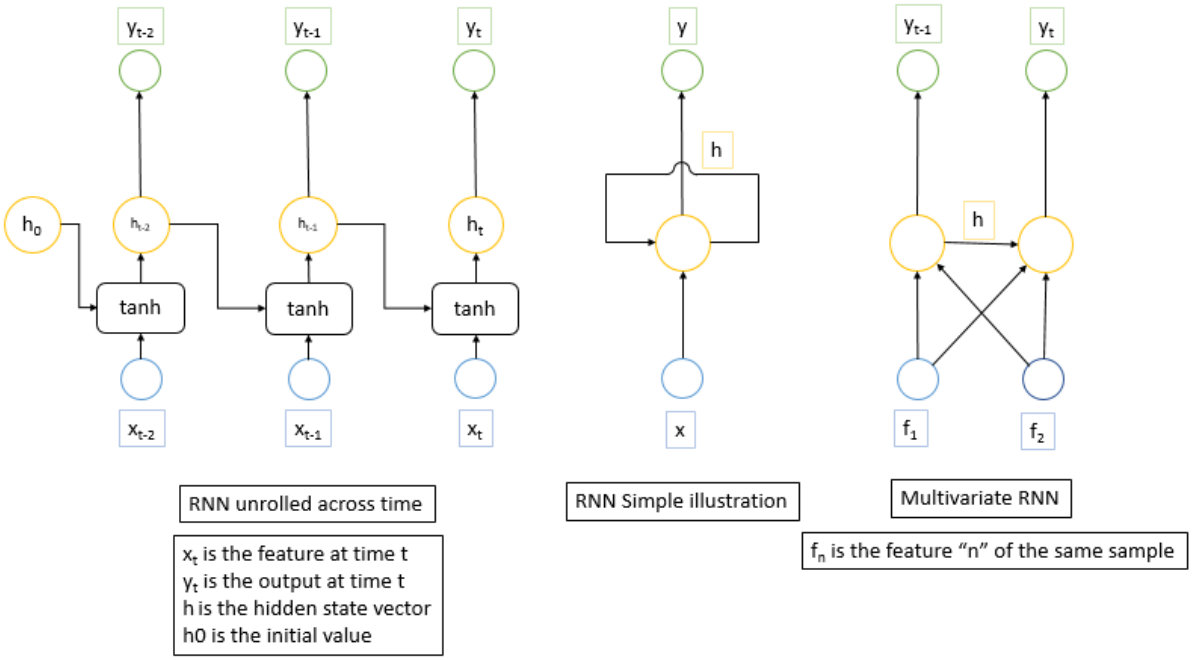

Figure 2.26 Illustration of a one-to-many recurrent neural network

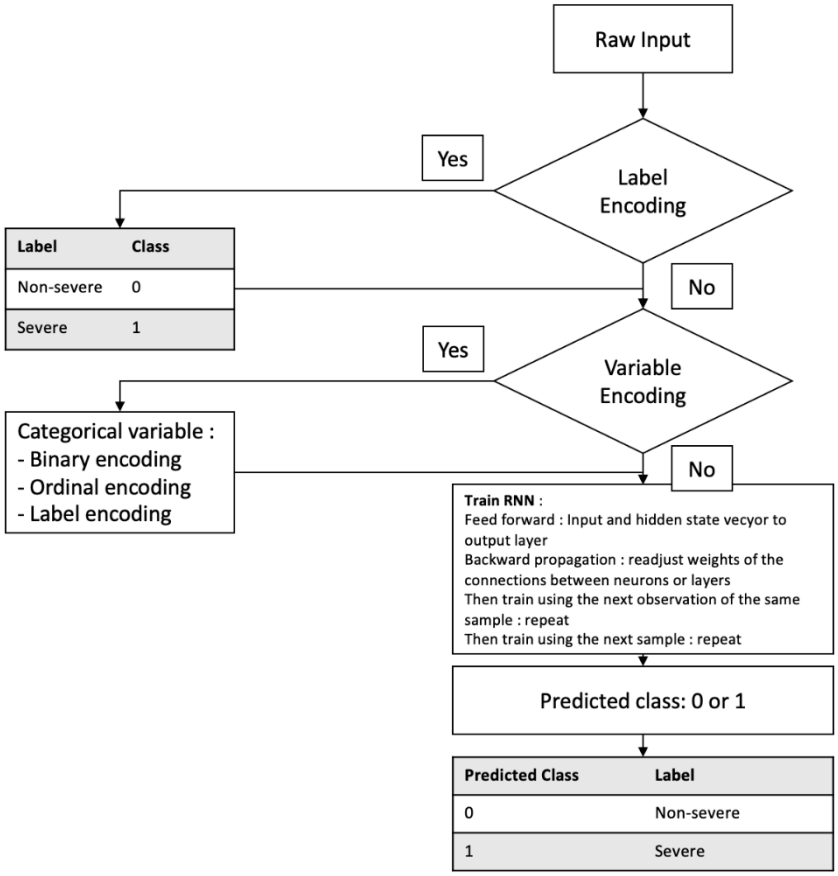

Figure 2.27 Flow chart for Recurrent Neural Networks – Training Model Generation

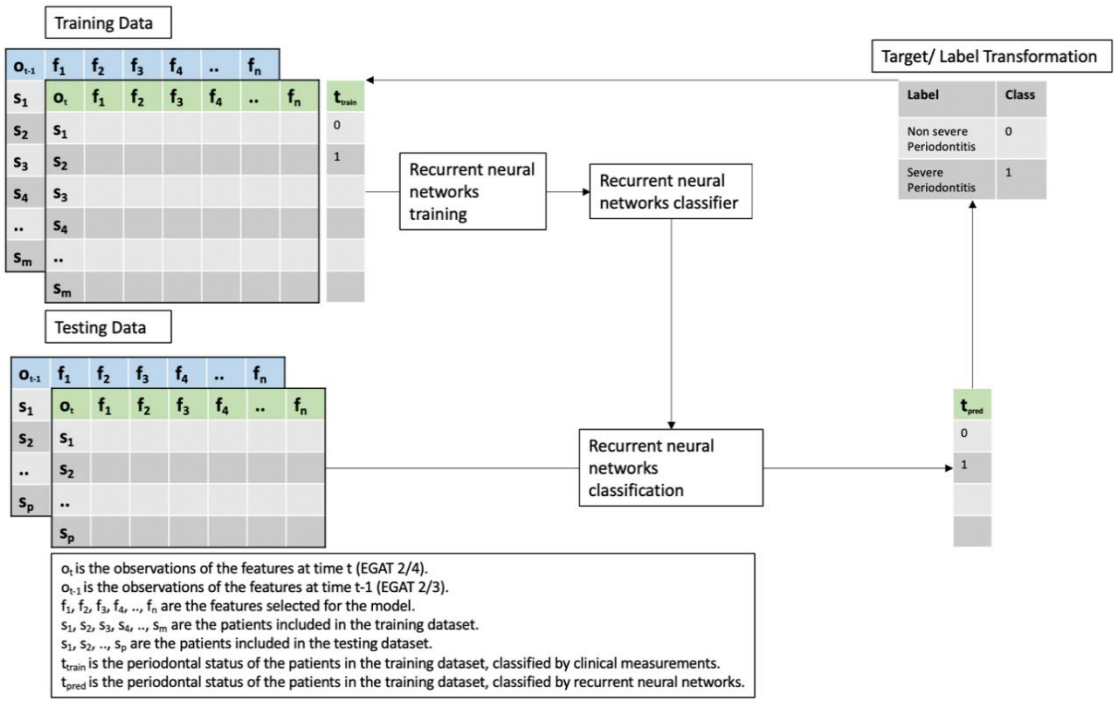

Figure 2.28 Block diagram for Recurrent Neural Networks – Testing and Target Transformation

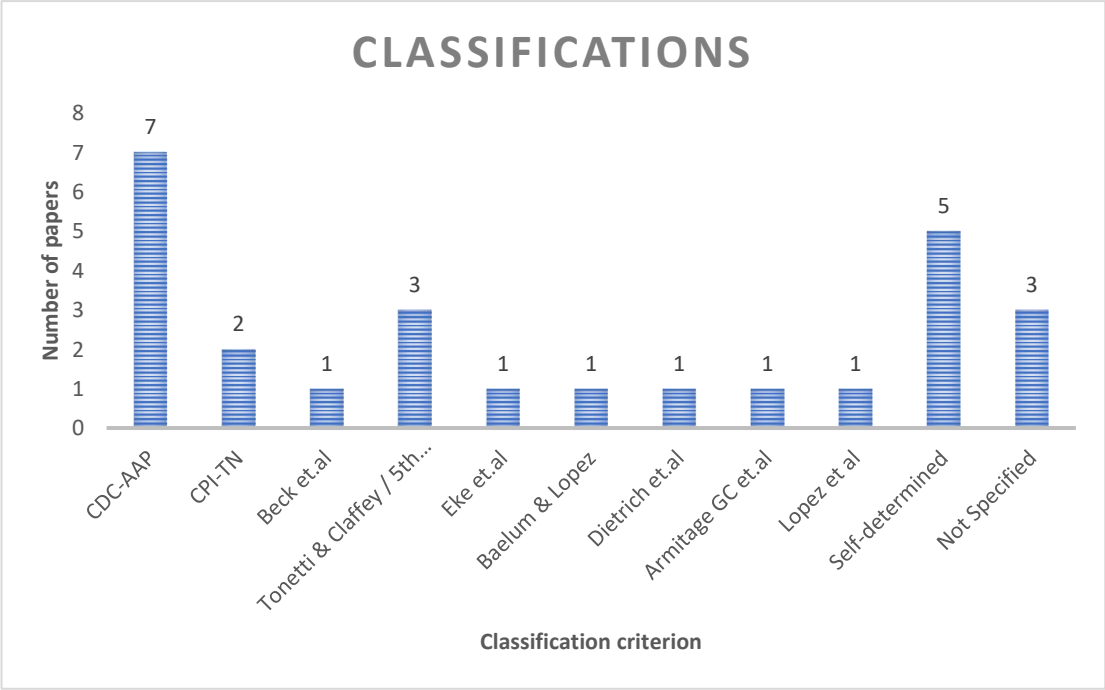

Figure 2.29 Distribution of labeling criteria in literature review (Some papers apply multiple criteria, and all criteria are counted)

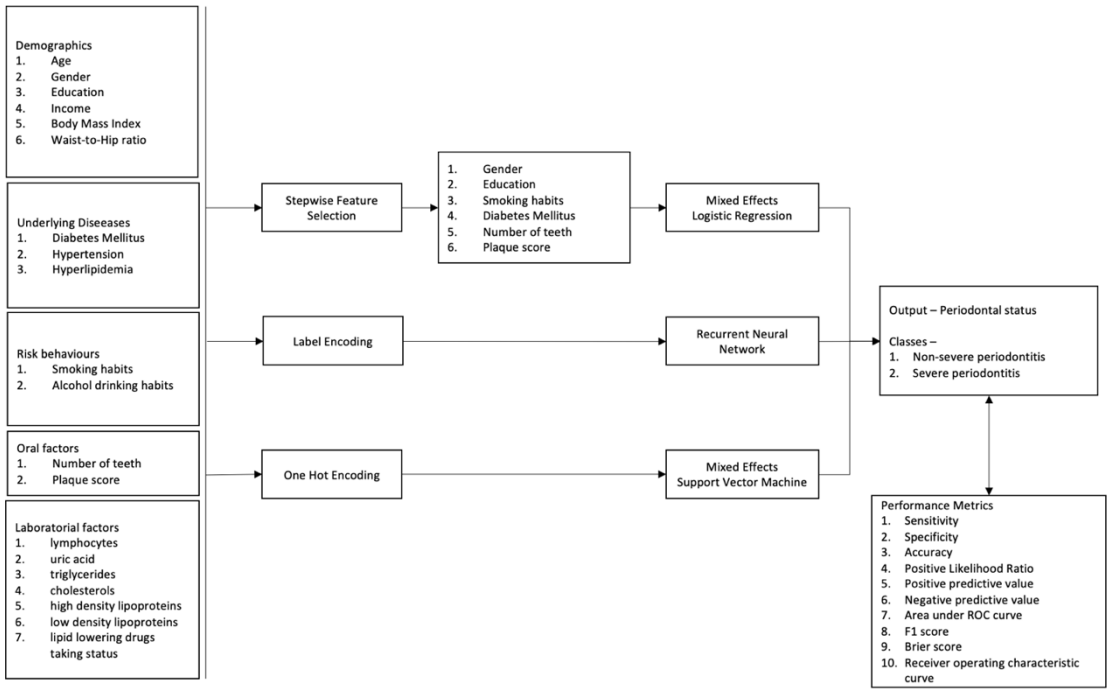

Figure 2.30 Conceptual framework of chronic periodontitis screening models

CHAPTER III
METHODOLOGY

3.1 Study Design and Setting

This study is a sub-cohort of prospective cohort study, namely Electric Generating Authority of Thailand (EGAT) cohort, by retrieving 5-years follow up period. Details about EGAT cohort are referenced[45], but in short, EGAT project contains three parallel cohorts, also known as EGAT1, EGAT2 and EGAT3. Each cohort begin in 1985, 1998 and 2009 respectively. Each follow up is examined every 5 years, except for 12 years gap between 1st survey (1985) and 2nd survey (1997) of EGAT1. In the 3rd survey (2002) of EGAT1, periodontists collaborated with the cohort by including half-mouth examination in the study. In 2003, 2nd survey of EGAT2, also known as EGAT 2/2, started including full-mouth examination. EGAT 2/3 (2008) and 2/4 (2013) included more questionnaires about oral health and habits.

This study was conducted applying EGAT2 cohort. The EGAT 2/3 and 2/4 are used as the training and testing datasets. EGAT 2/3 and EGAT 2/4 are defined as the patient characteristics 5 years before and now respectively. All models predict the periodontal status of the samples into two classes (severe chronic periodontitis and none or non-severe chronic periodontitis). Rationale and setting details of the research objective are as follows.

Rationale

We aimed to screen the periodontal examinees by predicting the probability of having severe chronic periodontitis without the need for comprehensive periodontal probing. From literature reviews and expert opinions, the features that are correlated with periodontitis were selected, such as demographics, underlying diseases, risk behaviors, oral and laboratorial features. Selected features were applied for the models as predictors. Periodontal status was the target variable for all models.

Setting

For longitudinal modelling, we applied both EGAT 2/3 and EGAT 2/4 dataset. The models applied were mixed effects logistic regression, recurrent neural networks, and mixed effects support vector machines The model performances were measured by six performance metrics: sensitivity, specificity, area under receiver operating curve, positive likelihood ratio, positive predictive value, negative predictive value, F1 score, and Brier score. The models were compared against each other.

3.1.1 Inclusion Criteria

All available subjects in EGAT 2/3 were included unless they met the exclusion criteria. For EGAT 2/4, only subjects followed up from EGAT 2/3 were included unless they met the exclusion criteria.

3.1.2 Exclusion Criteria

Some subjects were not present in ALL periodontal examinations due to (1) refusal to participate, (2) systemic conditions which required antibiotic prophylaxis before dental procedure including congenital heart disease or valvular heart disease, previous history of bacterial endocarditis or rheumatic fever, total joint replacement, and end-stage renal disease, and (3) fully edentulous subjects. Such subjects were excluded for all models.

3.2 Data Collection

In each survey, general demographic data (age, gender, educational level, income, marital status), behavioral data (smoking status, alcohol consumption, exercise/physical activity), family history of illness, underlying diseases (diabetes mellitus, hypertension) were collected by self-administered questionnaires. Physical examinations, i.e., blood pressure (BP), heart rate, blood glucose level, weight, height, and waist & hip circumference, were performed by clinicians and trained personnel from Ramathibodi Hospital. Laboratory tests under fasting state were carried out included glucose, low-density lipoprotein (LDL), high-density lipoprotein (HDL), immunoglobulin G3, interleukin 6, and a complete blood count (CBC).

3.3 Features

3.3.1 Outcome of Interest

The outcome of interest was the periodontal status of the subject at the period of examination. The subjects were labelled as “severe” or “non-severe” periodontitis and severe periodontitis, according to the periodontitis definition of the Center for Disease Control and Prevention – American Academy of Periodontology (CDC-AAP), which defined “severe periodontitis” as harboring two or more interproximal sites with clinical attachment level ≥ 6 mm that are not on the same tooth and one or more interproximal sites with periodontal pocket depth ≥ 5 mm.

3.3.1.1 Periodontal Examinations

Periodontal examinations including periodontal pocket depth and gingival recession were carried out on all fully erupted teeth, except third molars and retained roots. Periodontal pocket depth is the measurement from coronal margin of gingival margin to the tip of a periodontal probe, and gingival recession is the measurement from coronal margin of gingival margin to the cementoenamel junction. The parameters were measured applying a periodontal probe - University of North Carolina 15 (PCP-UNC15) on six sites, i.e., mesial, mesio-buccal, mesio-lingual, disto-buccal, disto-lingual, and lingual site of the gingival sulcus per tooth. These measurements were made in millimeters and were rounded to the nearest whole millimeter.

Calibration and standardization for periodontal measurements were implemented among six to eight examiners before the survey. The weighted kappa coefficients (±1 mm) were used to determine the agreement of inter-examiner and intra-examiner (Table 3.1). Between each pair of examiners, the kappa ranged from 0.72 to 1.00 for periodontal pocket depth and 0.67 to 1.00 for clinical attachment level/ gingival recession. The weighted kappa coefficients (±1 mm) within each examiner ranged from 0.85 to 1.00 for the periodontal pocket depth and from 0.80 to 1.00 for the clinical attachment level.

3.3.1.2 Periodontal classification

Due to the absence of homogenous classification for chronic periodontitis, we labeled the samples of our dataset based on the CDC-AAP classification. To classify a sample as chronic periodontitis, clinical attachment level is required, so it is calculated. The subtraction of gingival recession from pocket depth results in the measurement from the cementoenamel junction to the tip of the periodontal probe, hence clinical attachment level is resulted. Whereas the CDC-AAP criteria has four classes of periodontitis (non, mild, moderate, and severe), we categorized our samples into two, severe periodontitis and non-severe periodontitis (non, mild, and moderate) as shown in Table 3.2.

3.3.2 Features associated with Periodontitis

3.3.2.1 Self-administered data

Demographics

Demographic data such as age, gender, education, and income were reported by the individual themselves using case-report forms.

Risk behaviors

Subjects were categorized into (1) non-smoker (2) ex-smokers and (3) current-smokers, based on multiple questionnaires such as past/current smoking habits, quantity and duration of smoking, age at start or quit smoking.

Alcohol drinking habits were also classified as similar, based on history of alcohol consumption, along with frequency, duration, and type of alcohol.

Oral factors

Oral and dental examinations were carried out by experienced periodontists from the Department of Periodontology, Faculty of Dentistry, Chulalongkorn University in mobile dental units. Number of teeth and oral hygiene index (plaque score) were measured as part of oral parameters.

3.3.2.2 Physical Examinations

Body measurements

Height was measured in centimeters and weight was measured in kilograms, while being dressed in normal clothing with shoes taken off. Waist and hip circumferences were measured in centimeters with measuring tapes by trained personnel. Body mass index (BMI) was calculated from the recorded weight in

kilograms divided by squared height in meters. Waist-to-hip ratio was calculated from the recorded waist circumference in centimeters divided by hip circumference in centimeters.

Underlying conditions

Underlying conditions were identified from physical and laboratorial examinations, along with prescribed medications. Diabetes mellitus was diagnosed if an individual had fasting blood sugar (FBS) ≥ 126 mg/dl or had been taking anti-diabetic drugs. Hypertension was diagnosed if the participant had systolic blood pressure (SBP) ≥ 140 mmHg or diastolic blood pressure (DBP) ≥ 90 mmHg or had been taking prescribed anti-hypertensive drugs. Dyslipidemia was identified if the subject has high-density lipoprotein (HDL) < 40 mg/dl in male or HDL < 40 mg/dl in female OR low-density lipoprotein (LDL) ≥ 160 mg/dl OR triglyceride ≥150 mg/dl OR used any lipid-lowering medications.

3.3.2.3. Laboratorial Examinations

Blood samples were collected after 12-hour overnight fasting. Blood glucose was measured by plasma samples in mg/dl (Peridochrome, Boehringer Mannheim, Mannheim, Germany). High-density lipoproteins and low-density lipoproteins were measured in mg/dl using enzymatic-calorimetric assays (Boehringer Mannheim, Mannheim, Germany). immunoglobulin G3 in mg/dl, interleukin 6 in mg/dl and a complete blood count (CBC) was measured in count per micro liter.

3.4 Sample size estimation

There is no explicit guideline for sample size estimation for machine learning model. According to this literature review[46], the researchers recommend number of sample size for developing a clinical prediction model should be :

$$n = \frac{Z^2 p(1-p)}{d^2}$$

- where

n = number of sample size

d = absolute margin of error

p = anticipated outcome proportion

We aimed for margin of error(d) $\leq$ 0.05 and applied the Z value of 1.96. The prevalence of severe periodontitis in Thai adult population is 26%. So, we anticipated the outcome proportion in our study population(p) to be 0.3. At least 322.69 ~ 323 subjects including 97 subjects with severe periodontitis was required for our models.

Available sample size was explored. EGAT 2/3 (2008) has 2,271 subjects and 2,016 subjects are followed up in EGAT 2/4 (2013). We consider our study to have enough sample size to train and test our models

3.5 Data Analysis

In this section, we report the process of data pipeline – data management presenting how the data was collected as part of electricity generating authority of Thailand (EGAT) and transformed into the interested study factors. Data preparation reports how the data was manipulated to be applied as predictors for the classification model.

3.5.1 Data Management

3.5.1.1 Data Acquirement

Demographic and medical records

Demographic and medical data were retrieved from the EGAT databases. These were merged with the Excel worksheets of the civil registrations for the additional data.

Periodontal databases

Periodontal databases were constructed, all periodontal parameters for EGAT 2/3 and 2/4 were computerized as follows:

Building the periodontal databases

Databases were constructed using the Epidata version 3.1, separately by EGAT 2/3 and 2/4, because some variables were differently measured for each survey. Data entry systems were designed with “tooth by tooth” system. Users had to entry all parameters of one tooth including periodontal pocket depth, and gingival recession, before moving on to the next tooth. If a particular tooth is missing, the system

would not allow users to entry any data for that tooth. In addition, databases were encoded with specified value or range for each variable to prevent error during data entry.

Data entry (Periodontal parameters)

Data from case record forms (CRF) were manually checked by a data manager before entering the data. Legibility of handwriting, minor missing data and consistency of all parameters were revised. If handwriting is not clear, the query will be done directly to the recorder. Then data were independently entered twice by two persons. These two data sets were then validated, any inconsistence was checked and corrected. Finally, all records were manually checked and edited based on the original CRF, again.

3.5.1.2 Data Cleaning

Selected features and data were retrieved from the main databases. The variables were renamed systematically across both datasets in order to combine them all together. Then data cleaning was performed by the data cleaning team consisting of Prof. Ammarin Thakkinstian, Dr. Anuchate Pattanateepapon, Dr. Attawood Lertpimonchai, and Dr. Htun Teza. Regular meeting at least twice a month was organized to solve any incorrectness or unclear data. Data were summarized and cross-checked using pandas library and python 3.8. Any inconsistency or outliners were verified and checked with the CRFs to check data validity. All variables, except gender and height, were assigned as the time-varying variables for necessary models.

Gender

Gender is considered to be consistent across all datasets. Inconsistent data value between observations is validated by original case-report form.

Date of examination

The time length between the date of examination and the date of birth is calculated for the age of the subject at time of examination. The date has to be during the survey period and the values that are not or missing are recoded as the middle time of the survey period.

Date of birth

Similar with gender, date of birth is also assumed to be consistent across all dataset. However, when discrepancies are observed, civil

registration databases are also used as the source. Between the three datasets, the majority value for date of birth is selected.

Education

The level of education cannot be decreased. Illogical declinations are detected and decided by the team.

Risk behaviors

Smoking and alcohol drinking habits were classified within each period with multiple questions in questionnaire, then the datasets were merged. The values were checked to be logical, such as "current smoker" cannot become a "never smoker" in the next observation. If inconsistency is present, the decision will be made by the team.

Body Measurements

Height, weight, waist and hip were summarized and checked for outliers (i.e., exceeds mean ± 4SD). If outliers presented, the original CRF is checked. The change of the value overtime would also be checked after merging the datasets. Substantial change of weight, waist and hip would be list, and then, its possibility would be validated by comparing with other relevant variables.

Blood pressure

To determine the data validity of blood pressure, guidelines such as : presence of data for both systolic and diastolic blood pressure, within proper range of the value, and SBP value being higher than DBP were used.

Laboratory results

All laboratory results, which were reported in the continuous data, were checked for outliers (i.e., exceeds mean ± 4SD). If outliers exist, the likelihood of the value will be discussed and decided by the team. Illogical values were recoded to be missing values.

3.5.1.3 Carried forward/backward methods

To replace missing data for some variables, the forward/backward carry over methods were used. For example, carried backward method means that never smokers in EGAT 2/4 were imputed in EGAT 2/3 as "never smoker" as well.

3.5.2 Data Preparation

3.5.2.1 Feature Transformation

Logistic regression, recurrent neural networks and mixed effects support vector machine require the input of the models to be numerical values. In Table 3.3, categorical variables were encoded based on the type of categorical variable. Binary variables were encoded as 0 and 1. Ordinal variables such as education level were encoded using ordinal encoding, and for nominal categories, label encoding was used for recurrent neural networks and one-hot-encoding for mixed effects support vector machine.

3.5.2.2 Target Labelling

CDC-AAP criteria uses both the measurement of clinical attachment level and periodontal pocket depth to classify as periodontitis as stated in Table 3.2. Subjects that are eligible for "Severe" criterion of the classification were labelled as "Severe", and the rest are labelled as "Non-severe". During the encoding, "Non-severe" subjects were encoded as 0 and "Severe" subjects were encoded as 1.

3.6 Model Architecture

The model architectures for the statistical model (mixed effects logistic regression) and machine learning models (recurrent neural networks and mixed effects support vector machine) are constructed as follows.

Mixed effects logistic regression

Developing environment is Stata/SE 16.0 and more details in section 3.6.3.1. Cutoff point is 0.35 and more details in section 3.6.3.2. Feature selection is done by stepwise forward selection and more details in section 3.6.1. The model performance is evaluated by sensitivity, specificity, accuracy, area under receiver operator characteristics curve, positive likelihood ratio, positive prevalence value and negative prevalence value, F1 score and Brier score, more details in section 3.6.4.

Recurrent Neural Network

Developing environment is Python and more details in section 3.6.3.1. Cutoff point is 0.35 and more details in section 3.6.3.2. No feature selection is done and more details in section 3.6.1. Hyperparameter tuning is done with random search

followed by grid search and more details in section 3.6.3.3. The model performance is evaluated by sensitivity, specificity, accuracy, area under receiver operator characteristics curve, positive likelihood ratio, positive prevalence value and negative prevalence value, F1 score and Brier score, more details in section 3.6.4.

Mixed effects support vector machine

Developing environment is R and more details in section 3.6.3.1. Cutoff point is 0.35 and more details in section 3.6.3.2. No feature selection is done and more details in section 3.6.1. Hyperparameter tuning is done with random search followed by grid search and more details in section 3.6.3.3. The model performance is evaluated by sensitivity, specificity, accuracy, area under receiver operator characteristics curve, positive likelihood ratio, positive prevalence value and negative prevalence value, F1 score and Brier score, more details in section 3.6.4.

3.6.1 Feature Selection

From all the features measured in EGAT2 cohorts and datasets, 21 features associated with chronic periodontitis were nominated as observed from literature reviews and as recommended by experts' opinion in periodontology.

1. Demographics – age, gender, education level, income, body mass index, and waist to hip ratio.

2. Underlying diseases – diabetes mellitus, hypertension, dyslipidemia, and chronic kidney disease.

3. Risk behaviors – smoking and alcohol drinking habits.

4. Oral features – number of present/remaining teeth and plaque score.

5. Laboratorial features – lymphocytes, uric acid, triglycerides, cholesterols, high density lipoproteins, low density lipoproteins, and lipid lowering drugs taking status.

Feature selection, also known as variable selection, is a procedure of nominating a subset of relevant independent variables to apply as predictors in model construction. While several deep learning procedures are representative learning, where the irrelevant features are weighted less or none at all, the process reduces the dimension of the training dataset, subsequently computational resource and the training time requirements. It also reduces the risk of the model overfitting on the training dataset,

allowing the models to have relatively low bias and high variance. Feature selection methods can be grouped into three categories:

1. Filter methods
2. Wrapper methods and
3. Embedded methods.

Filter methods select the variables regardless of the model, by testing for difference in variance or correlation/association between independent (age) and dependent (chronic periodontitis) variables. The selected variables are used as the predictors for the classification or regression model. These methods are considered to be robust against overfitting and have less computational time. However, since they consider one-to-one relationships, such methods tend to select redundant variables (weight and body mass index) by not accounting for interaction between variables. Chi-square tests and analysis of variance (ANOVA) tests are considered as filter methods.

Unlike filter methods, wrapper methods evaluate subsets of variables, allowing to detect the possible interactions. It has the greedy approach, evaluating all possible combination of variables. Applying to a specific model, candidate variables are added one by one, or applied as a whole and removed one by one. On a chosen model fit criterion such as Akaike information criterion (AIC), the variables are chosen if their presence as predictor improves the fit of the model. However, the computational cost is high on datasets with many features. Also, this procedure requires a model to be tested on the fit for the dataset, therefore it is considered to have high chance of overfitting. Wrapper methods include stepwise regression methods such as forward selection and backward elimination.

Embedded methods are proposed to combine the advantages of two prior methods. These methods are included as part of a model training procedure. They calculate the importance of a feature in making prediction. Tree-based methods report the contributions of each feature while regularization methods such as LASSO and Ridge decreases the coefficients of the less relevant variables to reduce its contribution towards final prediction. Like filter methods, these methods are considered to be robust against overfitting, while they also consider the interaction between the features like wrapper methods.

3.6.1.1 Statistical Model

Appropriate feature selection is required for statistical modelling which is Mixed effects logistic regression. Part of filter methods of feature selection, stepwise selection can be applied in different ways, such as forward selection or backward elimination. In forward selection, the initial model is built with one variable, adding one by one. Using a model fit criterion, the variable is selected if its inclusion gives the most statistically significant improvement of the fit. After selection of second variable, all the remaining variables are tested again as the candidate for the third variable. This procedure is repeated until including more variables do not improve the model.

In backward elimination, the initial model is built with all available variables, removing one after another. Here, the variable is eliminated if the absence of it gives the most statistically insignificant deterioration of the fit. This procedure repeated until removing more variables results in statistically significant deterioration. Combination of both prior methods, called bidirectional elimination, tests for both including and excluding the variable at each step.

Other than testing for fit of the model, p-value is the common statistical entry and exit criteria of the variables. Multivariate regression models are applied as the model and the threshold is set for including or excluding the variable. Unlike conventional statistically significant value of 0.05, 0.1 is the typical value used and the variables with less p-value are included in the model for current step. Similarly, variables with p-value more than 0.1 are excluded.

For our study, univariate mixed effects models for each were developed from the nominated variables, and they were ranked in increasing order based on their statistical significance which is p-value of Wald chi-squared test less than 0.1. It resulted in 15 significant variables out of initial 21. Afterwards, multivariate models were built by including one variable by one beginning from the most significance (least p-value). If the significant variable is no longer significant in multivariate regression, it will not be included in the subsequent regression with next significant variable. It resulted in six final variables being included with the final multivariate mixed effects logistic regression model. They were gender, education level, smoking habit, diabetes mellitus, number of present/remaining teeth, and plaque score.

3.6.1.2 Machine Learning Models

Machine learning models generally can handle higher data dimensions than their statistical counterparts, so feature reduction can be skipped for these models. However, we had made several attempts to observe if machine learning models would gain any advantage by removing several features from the initial 21 nominees.

Stepwise feature selection

By applying filter method of feature selection, 15 features nominated by univariate mixed effects logistic regression and 6 features included in final model were considered as additional subsets of features to train the machine learning models

Recurrent feature elimination

Recurrent feature elimination is a wrapper-type feature selection algorithm. A machine learning model is designated and applied at the core of the selection method. It is considered being wrapped and used to select the predictors. When compared with the filter method of feature selection where the features are scored and selected based on them, recurrent feature elimination is a wrapper method that uses backward elimination filter-based method internally.

For our study, support vector machine with linear kernel[47] was applied as the wrapped machine learning core. While trying to figure out the decision threshold for the dataset, weights or w for each feature was calculated. Such weights were ranked and the one with the smallest value is removed. This process is iterated until there is only one feature left in the model. The performance of the support vector machine model was compared for all versions with varying number of predictors and the one with the highest prediction performance was considered as the optimum feature subset for the machine learning model.

Random forest feature importance

Random forest models are part of ensemble machine learning models, applying bagging algorithm with multiple decision trees. Decision trees are supervised machine learning algorithms, built using recursive partitioning, more commonly known as “Divide and Conquer” approach. The decision tree splits the dataset into smaller subsets, and those subsets are split again into even smaller subsets, until each leaf are homogenous (single class) or stopping criteria is specified. Random

forest uses bootstrapping to create small subsets of the dataset to build the individual trees. Firstly, to validate the trees, out-of-box samples are allotted for each tree. From the remaining dataset, the features are picked without replacement and the samples are picked with replacement.

Decision trees in random forest are built in the same process as singular decision tree, except they are trained with smaller dataset and varying combination of features, so they are uncorrelated to each other. In each split of every decision tree, all the features included in the model are evaluated on their ability to split the mother node in pure daughter nodes. This ability is measured in impurity indexes such as Gini impurity index for classification or Variance reduction for regression tasks. Considering the decision trees in the random forest are built with varying feature sets, the impurity indexes can be averaged for each feature across the forest. This is embedded method of feature selection and the features with the least average impurity index values are considered to have better power of splitting the samples into separate classes. The nominated features were ranked and separate machine learning models were trained with increasing number of features – from the most important feature to least. It resulted in 21 models - first model with only one but most important feature, second model including the two most important features to the last model with all 21 features included.

Findings

For all machine learning models, the hyperparameters were set constant so that we could compare the performance based on the difference of predictors only. We observed that including a smaller number of features as predictors results in comparably similar or inferior performance. Due to the inherent nature of machine learning models to handle higher dimensions, recurrent neural networks' application of dropout layers and our plans to apply further hyperparameter optimizations for the machine learning models, we decided to skip the feature elimination step, i.e., include all features, for all machine learning models.

3.6.2 Data Splitting

The total samples are split in 80% for model training and 20% for model performance testing as per Pareto principle. Since we are working with panel data, extra

caution is taken to avoid situations where different observations of the same individual appear in both datasets.

3.6.3 Model Development

Seed of 1996 was set in all developing environments for reproducibility. For Python, multiple seeds were required to be set for different libraries with built-in random value generators such as NumPy and TensorFlow in addition to the overall environment seed, but the same value of 1996 was still given for all. All models were developed using 64bit 2.3 GHz Dual-Core Intel Core i5 processor.

3.6.3.1 Developing Environments

Mixed effects logistic regression

The model was developed in STATA/SE (Special Edition) version 16.0 for 64-bit Intel processors. Built-in library of "melogit" — Multilevel mixed-effects logistic regression was used to fit models for binary and binomial responses which is appropriate for our objective. Mixed effects model with random intercept was applied with random effects for each subject.

Recurrent Neural Network

The model was developed in Python 3.8.2 using Spyder integrated development environment 4.2.5 version. During the model development process, several libraries were applied along the data pipeline. For dataframe management and manipulation, NumPy version 1.19.2 and pandas version 1.2.3 are applied. Scalers and sample weights were created using scikit-learn version 0.23.2 and recurrent neural networks were developed using Keras version 2.4.3 and TensorFlow version 2.31. For data visualizations, Matplotlib version 3.3.1 was used.

Mixed effects – Support Vector Machine

The model was developed in R version 4.02 using R Studio 1.3.1056 version. Support vector machine was applied as machine learning regressor in mixed effects machine learning model. Several packages were applied for the data management and model development process. readstata version 0.9.2 was used for importing STATA datasets. For the mixed effects – support vector machine, e1071 version 1.7-4 was used to model SVM and lme4 version 1.1.-26 to estimate the random

effects. pROC version 1.16.2 and epiR version 2.0.19 were used to evaluate the model performance.

3.6.3.2 Screening cut-off points

All models applied were logit-based models, thus they outputted the log-odds of the event. The value was further transformed into probability with the values ranging from 0 (zero probability of having severe chronic periodontitis) to 1(100% probability of having severe chronic periodontitis). Since the goal of our study is binary classification, the probability was dichotomized using a decision threshold. The default value is 0.5, but for our study, we applied the value of 0.35 to reflect the prevalence of the event (severe chronic periodontitis) in our dataset which is 34.6%. Although, it should be noted that this decision threshold influences the number of false positives or false negatives, affecting the ability of the classifier to overestimate or underestimate the condition. For instance, by applying a lower decision threshold, subjects with a lower probability are determined as positive, which would increase the false positives but reduce the number of people incorrectly identified as negatives who will miss the chance of an early diagnosis.

3.6.3.3 Hyperparameter optimization

For hyperparameter tuning process, the multiple sets were trained and evaluated with bootstrapped samples using random search function, which was followed by grid search function of scikit learn library.

Recurrent neural networks

The main hyperparameters tuned are -

1. Number of hidden layers
2. Number of nodes or neurons in each hidden layer and
3. the learning rate of the optimizer.

To compare the performance of the models with different combinations of aforementioned hyperparameters, basic specifications for other parameters were set constant to find the best performing model on the data. For all feature sets, 20% of training data was used for validation. Simple RNN layer cells and Tanh activation functions were used for all nodes in the hidden layers. Dropout rate of

0.2 was put between every hidden and output layers so 20% of all connections between nodes were randomly deactivated, thus it was not a fully connected model. One node in the output layer; sigmoid activation function was used for binary classification. Binary Cross Entropy was used for loss function and accuracy is the monitor metric. Batch size of 64 was applied for Mini-batch Gradient Descent optimization. 1000 epochs with early stopping were used for time and computation resource constraints. The outputs of the model were dichotomized using 0.35 according to the prevalence of severe chronic periodontitis in the dataset.

Number of hidden layers, number of nodes in each hidden layer and learning rate of the optimizers were tuned for the optimal performance metrics. Models were trained with various combinations of only one hidden layer to six hidden layers, nodes in each layer ranging from 21 to 80 and the learning rate from 1 to 0.001.

Mixed effects support vector machine

For mixed effects machine learning, support vector machine was applied as the machine learning regressor to estimate the fixed effects portion of the framework. Overall hyperparameters of the framework were set constant to be 10 maximum macro iterations with 0.01 tolerance and 50 maximum micro iterations with 0.001 tolerance. Initial random effect of zero was set. Instead, the hyperparameters of support vector regressor were tuned.

Since we applied nu regression architecture, the main parameters optimized are –

1. Kernels
2. C value
3. gamma value when applicable and
4. nu value.

Models were trained with various combinations of nu-value 0.1 to 0.6; linear, gaussian, polynomial kernels with C value 0.1 to 0.0001. Gamma value of 0.1 to 0.9 were also applied when radial and polynomial kernels were applied.

3.6.4 Performance Evaluation

3.6.4.1 Performance metrics and measurements

The framework of the model is shown in Figure 3.1. The performances of all models were evaluated using accuracy, sensitivity, specificity, positive likelihood ratio, positive predictive value, negative predictive value, C-statistics (area under receiver operating characteristic curve), and receiver operating characteristic curve. 95% confidence interval is calculated for these metrics. In addition, F1 score or balanced F score, a measure of model accuracy better suited for unbalanced datasets such as ours, and Brier score were calculated as well.

3.6.4.2 Evaluating statistical and machine learning models

The classification models were evaluated based on prognosis accuracy (i.e., sensitivity, specificity and accuracy) calculated upon the constructed confusion matrices and we considered the models with the metrics above 70% to be of acceptable performance. 95% confidence intervals of the metrics were calculated in the R environment using epiR library. The library calculates the range using –

$$confidence\ interval = \bar{X} \pm Z\frac{e}{\sqrt{n}}$$

- where

$\bar{X}$ = the mean value

e = the standard deviation

Z = z-value for selected confidence interval (1.96 for 95% and 2.576 for 99%)

n = the number of observations.

Discrimination and calibration abilitiies of the classifiers were evaluated using area under receiver operating curve (AUC) and Brier score. For both metrics, the values range between 0 and 1; while higher is better for AUC, the opposite is true for Brier's score. In addition, accuracy and Brier score metrics were monitored for the performance of the same model on different datasets – training and testing – and evaluated the overfit problem common to machine learning models.

3.7 Limitations

From literature reviews, it was observed that including oral features in the models predict better than not including them. The Electricity Generation Authority of Thailand (EGAT) dataset does not include much oral features, such as tooth mobility, bleeding on stimulation and more. While we tried to compensate the issue by deploying

more complex and higher performing models, the good models should perform better with such features.

Table 3.1 Calibration of periodontal examination (weight kappa ± 1mm)

	Periodontal pocket depth		**Clinical attachment level/ Gingival Recession**	
	Inter-examiner	Intra-examiner	Inter-examiner	Intra-examiner
EGAT 2/3	0.77 – 0.89	0.87 – 0.91	0.67 - 0.94	0.90 - 0.96
EGAT 2/4	0.74 - 1.00	0.87 - 1.00	0.78 - 1.00	0.87 - 1.00

Table 3.2 Labeling Criteria for the dataset

Label	**Case**	**Definition**
Non-severe periodontitis	No periodontitis	No evidence of mild, moderate, or severe periodontitis
	Mild periodontitis	≥2 interproximal sites with clinical attachment level ≥3 mm, and ≥2 interproximal sites with periodontal pocket depth ≥4 mm (not on same tooth) or one site with periodontal pocket depth ≥5 mm
	Moderate periodontitis	≥2 interproximal sites with clinical attachment level ≥4 mm (not on same tooth), or ≥2 interproximal sites with periodontal pocket depth ≥5 mm (not on same tooth)
Severe periodontitis	Severe periodontitis	≥2 interproximal sites with clinical attachment level ≥6 mm (not on same tooth) and ≥1 interproximal site with periodontal pocket depth ≥5 mm

Table 3.3 Feature transformation

Feature	**Original Form**		**Encoding**	**Model Required Form**
	Type	**Possible value**		**Encoded value**
Demographics				
Age	continuous	≥43	Similar to original form	
Gender	categorical	Male, Female	Binary Encoding	0, 1
Education	categorical	Less than secondary school, vocational or diploma, higher bachelor's degree, missing value	Ordinal Encoding	0, 1, 2
Income	categorical	< 20,000, 20,000 – 49,999, >50,000	Ordinal Encoding	0, 1, 2
Body Mass Index	continuous	~	Similar to original form	
Waist-to-hip ratio	continuous	~	Similar to original form	
Underlying diseases				
Diabetes Mellitus	categorical	negative, positive	Binary Encoding	0, 1
Hypertension	categorical	negative, positive	Binary Encoding	0, 1
Hyperlipidemia	categorical	negative, positive	Binary Encoding	0, 1
Risk behaviors				

Smoking habit	categorical	non-smoker, ex-smoking, current smoker	Ordinal Encoding	0, 1, 2
Alcohol drinking habit	categorical	Never drinker, ex-drinker, current drinker	Ordinal Encoding	0, 1, 2
Oral features				
Number of teeth	continuous	≥1 & ≤28	Similar to original form	
Plaque score	continuous	0~100	Similar to original form	
Laboratorial factors				
Lymphocytes	continuous	~	Similar to original form	
Uric acid	continuous	~	Similar to original form	
Triglycerides	continuous	~	Similar to original form	
Cholesterols	continuous	~	Similar to original form	
High density lipoproteins	continuous	~	Similar to original form	
Low density lipoproteins	continuous	~	Similar to original form	
Lipid lowering drugs taking status	categorical	negative, positive	Binary Encoding	0,1

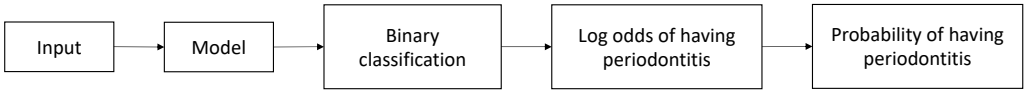

Figure 3.1 Model Architecture

CHAPTER IV
RESULTS

4.1 Description of EGAT Study

As shown in Figure 4.1, 2271 subjects were included in the EGAT 2/3 survey, and 2016 subjects were examined in EGAT 2/4 survey. Splitting using Pareto's principle, 80% was applied as training data and it included 1817 subjects. For testing dataset, there were 454 subjects included. 69.99% male and 30.01% female were included in the training while 73.01% male and 26.99 female observations were included in the testing dataset. For both training and testing datasets, around 33% held a high school diploma and around 38% of the observations held bachelor's degree. 29.01% of training and 30.96% of the testing populations were ex-smokers, and non-smoking samples were 53.61% and 53.19%, respectively. 13.13% of training and 13.76% of testing observations had diabetes mellitus underlying. Average number of teeth was 23.29 and 23.45 for training and testing data with the average plaque score of 71.02 and 70.42, respectively. Training dataset had 1,094 observations (34.64%) with severe chronic periodontitis present while testing data had 267 observations (34.41%), thus both datasets were considered as similar distributions.

4.2 Models

4.2.1 Mixed effects logistic regression

4.2.1.1 Data Manipulation

Dataframe was managed in the long format where repeated measures of the same individual were recorded in separate row. Within 1817 distinct training subjects, 1817 subjects were observed in 2008 and 195 subjects in 2013.

4.2.1.2 Feature selection result

Appropriate feature selection is required for statistical modelling, and stepwise forward selection was done.

For the multivariate regression, final model included six variables – gender, education level, diabetes mellitus, smoking habits, number of present/remaining teeth and plaque score. Fixed effects coefficients of the included variables are stated in Table 4.1. The output of the model was dichotomized using the prevalence of severe periodontitis in our dataset which is 35%.

4.2.1.3 Performance

Application of the final model as the risk prediction model had great performances. Without considering the known random effects of the training samples, the model identifies 91.3% of positive cases and 90% of negative cases correctly with 90.5% overall accuracy. It had 82.9 and 95.2 positive and negative predictive value respectively. The positive likelihood ratio is 9.18 and the area under receiver operating curve was 0.98. It is good discrimination ability, allowing the model to have high sensitivity and specificity simultaneously. Figure 4.2. presents the receiver operating characteristic curves of the model on training and testing data.

When the same model was applied on the testing dataset, the model performed similarly with discriminative power of 0.98. Discriminative power was evaluated using area under receiver operating characteristic curve and values over 0.9 is considered outstanding. It was 91.5% accurate with 89.5 sensitivity and 92.5 specificity. Positive likelihood ratio of 11.9, positive predictive value of 86.2 and negative predictive value of 94.4 were observed. F1 score of the model on the testing data was 0.8782. The performance of the model is shown in Table 4.2.

4.2.2 Recurrent Neural Network

4.2.2.1 Data Manipulation

Data frame was managed in the cube format which is similar to the long format except the repeated measurements of each individual are stacked in the third dimension. Since all training individuals are required to have equal timesteps for recurrent neural networks, subjects with only one measurement were dropped and only 1345 distinct subjects were left.

For neural networks, numerical inputs were required so discrete or continuous variables were included as they are after applying MinMaxScaler to bound the values between 0 and 1. Label encoding was applied for categorical variables with more than binary class.

4.2.2.2 Feature Selection Result

All 21 features were applied as the input of the model and dropout layers were applied between each hidden layer instead of manual feature selections.

4.2.2.3 Hyperparameter Optimization

Out of 2690 training data, only 858 records had chronic severe periodontitis, so class imbalance problem was anticipated. Therefore, class weights of 0.734 and 1.568 for negative and positive classes were calculated using scikit-learn package. However, Keras considers the concept of class to be ambiguous in 3 and more dimensional data so the sample weights were applied using class weight values as a workaround.

For the hyperparameter tuning, basic specifications were set to find the best performing model on the data. For all feature sets, 20% of training data was used for validation. Simple RNN layer cells and Tanh activation functions were used for all nodes in the hidden layers. Dropout rate of 0.2 was allocated between every hidden and output layers so 20% of all connections between nodes were randomly deactivated, so it was not a fully connected model. One node in the output layer; sigmoid activation function was used for binary classification. Binary Cross Entropy was used for loss function and accuracy was the monitor metric. Batch size of 64 was applied for Mini-batch Gradient Descent optimization. 1000 epochs with early stopping were used for time and computation resource constraints. The outputs of the model were dichotomized using 0.35 according to the prevalence of severe chronic periodontitis in the dataset. Number of hidden layers, number of nodes in each hidden layer and learning rate of the optimizers were tuned for the optimal performance metrics. The parameters were tuned for two different purposes and more details are presented in the following subsection.

4.2.2.4 Performance

Recurrent neural network with three hidden layers and 70 Simple RNN nodes in each layer was applied and learning rate of 1 was used to optimize model loss. The resulting model was 92.3% accurate overall with 87.4% sensitivity and 94.7% specificity. Along with 88.4% positive predictive value and 94.1% negative predictive value, the model had 16.3 positive likelihood ratio. AUC measures the probability that a model can correctly discriminate between randomly selected individuals with or without the event and 0.95 means the model was very proficient.

However, as shown in Table 4.3., when the same model was applied on the testing data, the performance diminished overall with 65.2 accuracy, 42.7 sensitivity and 75.7 specificity. The discrimination became 0.65 which is very poor. Comparing the positive likelihood ratio of 1.7, positive predictive value of 45.3 and negative predictive value of 73.8 to respective performances on training dataset, the model can be considered overfit. Brier score is the measure of average difference between the observed and predicted probability by the model and the scores were 0.0625 on training but 0.2770 on testing. Figure 4.3 presents the receiver operating characteristic curves of the overfit model on training and testing data.

Instead, a new set of hyperparameters was searched with the condition that we allow ±5% discrepancy in accuracy performance between two datasets. The final model had four hidden layers with 62, 72, 72 and 62 RNN nodes in feed forward order and learning rate of 0.01 for optimizer. As seen in Table 4.4., it is evident that the performance of the model is inferior compared to the preceding models. Area under receiver operating curve of 0.75 was considered only moderate but the model was no longer overfit to the training data. Receiver operating characteristic curves of the final model on training and testing data are compared in Figure 4.4.

4.2.3 Mixed effects – Support Vector Machine

4.2.3.1 Data Manipulation

Data frame was managed in the long format same as mixed effects logistic regression models. For support vector machines, numerical inputs were required so one hot encoding is applied for categorical variables. No additional data scaling was done other than default parameter in the e1071 library.

4.2.3.2 Feature Selection Result

No additional feature reduction was done after initial 21 features selected by literature reviews and expert opinion. On the contrary, one hot encoding the categorical variables with more than binary class resulted in additional input features totaling 26 variables.

4.2.3.3 Hyperparameter Optimization

Hyperparameters of the overall model were set to be 10 maximum macro iterations with 0.01 tolerance and 50 maximum micro iterations with 0.001 tolerance. Initial random effect of zero was set. Instead, the hyperparameters of support vector regressor were tuned. Kernels, C value and gamma value when applicable were also tuned. Nu regression was applied while optimizing multiple nu values. The parameters were tuned for two different purposes and more details are presented in the following subsection.

4.2.3.4 Performance

Support vector regressor with nu value of 0.4 was applied. Radial kernel with 0.2 gamma value and C value of 0.1 was set. Resulting model performed very good with overall accuracy of 98.4%. The metrics were 99.7% sensitivity (true positive rate) and 97.7% specificity (true negative rate). The model were 43.3 times more likely to correctly identified the true positive subjects as positive than incorrectly consider the negative patients as such. Discriminative power of 0.99 can be considered very proficient.

However, when validated by the testing dataset as in Table 4.5, the model performance was reduced greatly to AUC value of 0.62 when is considered poor. Figure 4.5. presents the receiver operating characteristic curves of the overfit model on training and testing data. The overall accuracy was 62% with only 38.1% of positive predictions and 80.1% negative predictions were correctly predicted. Brier scores for training and testing data were 0.0403 and 0.2668 respectively so the model is considered overfit to the training dataset so new hyperparameter sets were searched.

Nu-regression with nu value of 0.5 and radial kernel was applied. C-value of 0.1 and gamma value of 0.3 was set and the resulting model was considered as the optimized model with balanced performances. Area under receiver operating curve of 0.76 is only moderate but when compared to performances on the testing data, it was observed that the model was no longer overfit to the training data.

Receiver operating characteristic curves of the final model on training and testing data are compared in Figure 4.6. The performances of the final mixed effects – support vector machine is shown in Table 4.6.

Table 4.7.presents comparison of all final classification models (Mixed effects logistic regression, Recurrent neural networks, and Mixed effects support vector machine). F1 score was measured for all models because we consider the positive class to be importance in our unbalanced data set. The score is the harmonic mean between precision (the ratio of correctly predicted positive to all predicted positive samples) and recall (the ratio of correctly predicted positive to all positive samples); in other words, the metrics measures how many positive instances it classifies correctly (precise) and how much the classifier does not miss positive instances (robust). Brier score is a measure of average difference in predicted probability. The metric ranges from 0 to 1, with lower value being preferred. It also is a measure of accuracy albeit it is not sensitive to decision threshold. Like accuracy metric, this value can be compared between performance of the model upon different datasets observing the model's ability to generalize on unseen or non-training data. F1 score and Brier score are reported in Table 4.8.

Table 4.1 Fixed Effects Coefficients and Odds Ratio Estimates for Significant Variables Retained in the Final Multivariate Mixed Effects Logistic Regression Model

Variables	Covariates	Coefficient (SE)	Odd ratios (95% CI)	P-value
Gender	Male	0.97 (0.23)	2.63 (1.68 to 4.10)	< 0.001
	Female	ref	ref	
Education	< High school	2.04 (0.38)	7.68 (3.62 to 16.30)	< 0.001
	Vocational School	1.35 (0.35)	3.86 (1.93 to 7.72)	< 0.001
	Bachelor's degree	0.29 (0.35)	1.34 (0.68 to 2.64)	< 0.001
	> Bachelor's degree	ref	ref	0.393
Smoking	Non-smoker	ref	ref	
	Ex-smoker	0.73 (0.21)	2.09 (1.38 to 3.17)	0.001
	Current smoker	1.68 (0.25)	5.38 (3.28 to 8.83)	< 0.001
Diabetes Mellitus	Positive	0.50 (0.22)	1.66 (1.07 to 2.57)	0.024
	Negative	ref	ref	
Number of teeth	-	-0.06 (0.02)	0.94 (0.91 to 0.97)	< 0.001
Plaque score	-	0.03 (0.004)	1.03 (1.02 to 1.03)	< 0.001

Abbreviation: CI: Confidence Interval; SE: Standard Error; ref: Reference covariate group.

Table 4.2 Performance of Mixed effects logistic regression (decision threshold – 0.35)

	On Training data (95% CI)	On Testing data (95% CI)
%Sensitivity	91.3 (89.5 – 93.0)	89.5 (85.1 – 92.9)
%Specificity	90.0 (88.7 – 91.3)	92.5 (89.9 – 94.6)
%Accuracy	90.5 (89.4 – 91.5)	91.5 (89.3 – 93.3)
AUC	0.98 (0.98 – 0.98)	0.98 (0.98 – 0.99)
Positive Likelihood Ratio	9.18 (8.05 – 10.46)	11.93 (8.77 – 16.25)
%Positive Predictive Value	82.9 (80.7 – 85.0)	86.2 (81.6 – 90.1)
%Negative Predictive Value	95.2 (94.1 – 96.1)	94.4 (92.0 – 96.2)

Table 4.3 Performance of overfit recurrent neural network (decision threshold – 0.35)

	On Training data (95% CI)	**On Testing data** (95% CI)
%Sensitivity	87.4 (85.0 – 89.6)	42.7 (36.0 – 49.7)
%Specificity	94.7 (93.5 – 95.6)	75.7 (71.5 – 79.6)
%Accuracy	92.3 (91.3 – 93.3)	65.2 (61.4 – 68.8)
AUC	0.95 (0.94 – 0.96)	0.65 (0.61– 0.70)
Positive Likelihood Ratio	16.3 (13.5 – 19.8)	1.7 (1.4 – 2.2)
%Positive Predictive Value	88.4 (86.1 – 90.5)	45.3 (38.3 – 52.4)
%Negative Predictive Value	94.1 (93.0 – 95.2)	73.8 (69.5 – 77.7)

Table 4.4 Performance of final recurrent neural network (decision threshold – 0.35)

	On Training data (95% CI)	**On Testing data (95% CI)**
%Sensitivity	63.1 (59.7 – 66.3)	58.2 (51.3 – 64.9)
%Specificity	73.3 (71.2 – 75.3)	73.5 (69.2 – 77.5)
%Accuracy	70.0 (68.3 – 71.8)	68.6 (64.9 – 72.1)
AUC	0.75 (0.73 – 0.77)	0.73 (0.68 – 0.77)
Positive Likelihood Ratio	2.36 (2.16 – 2.59)	2.20 (1.82 – 2.66)
%Positive Predictive Value	52.5 (49.4 – 55.6)	50.8 (44.4 – 57.3)
%Negative Predictive Value	80.9 (78.9 – 82.8)	78.9 (74.7 – 82.7)

Table 4.5 Performance of overfit Mixed Effects – Support Vector Machine (decision threshold – 0.35)

	On Training data (95% CI)	**On Testing data (95% CI)**
%Sensitivity	99.7 (99.1 – 99.9)	69.6 (62.8 – 75.8)
%Specificity	97.7 (96.9 – 98.3)	52.0 (47.4 – 56.5)
%Accuracy	98. 4 (97.8 – 98.9)	57.2 (53.4 – 61.0)
AUC	0.99 (0.99 – 1.0)	0.62 (0.58 – 0.66)
Positive Likelihood Ratio	43.3 (32.2 – 58.2)	1.45 (1.27 – 1.65)
%Positive Predictive Value	95.6 (94.1 – 96.8)	38.1 (33.1 – 43.2)
%Negative Predictive Value	99.8 (99.5 – 100)	80.1 (75.3 – 84.4)

Table 4.6 Performance of final Mixed Effects – Support Vector Machine (decision threshold – 0.35)

	On Training data (95% CI)	**On Testing data (95% CI)**
%Sensitivity	52.8 (49.5 – 56.0)	46.1 (39.1 – 53.2)
%Specificity	82.7 (80.9 – 84.4)	78.2 (74.2 – 81.8)
%Accuracy	72.7 (71.0 – 74.4)	68.6 (65.0 – 72.1)
AUC	0.76 (0.75 – 0.77)	0.70 (0.68 – 0.73)
Positive Likelihood Ratio	3.05 (2.72 – 3.43)	2.11 (1.69 – 2.64)
%Positive Predictive Value	60.5 (57.0 – 63.8)	47.2 (40.1 – 54.4)
%Negative Predictive Value	77.8 (75.9 – 79.6)	77.4 (73.4 – 81.0)

Table 4.7 Performance of all final models (Performance with 95% Confidence Interval) (decision threshold – 0.35)

Metrics\Models	**Mixed effects Logistic Regression**		**Recurrent Neural Networks**		**Mixed effects Support Vector Machine**	
	Train	Test	Train	Test	Train	Test
%Sensitivity	91.3 (89.5 – 93.0)	89.5 (85.1 – 92.9)	63.1 (59.7 – 66.3)	58.2 (51.3 – 64.9)	52.8 (49.5 – 56.0)	46.1 (39.1 – 53.2)
%Specificity	90.0 (88.7 – 91.3)	92.5 (89.9 – 94.6)	73.3 (71.2 – 75.3)	73.5 (69.2 – 77.5)	82.7 (80.9 – 84.4)	78.2 (74.2 – 81.8)
%Accuracy	90.5 (89.4 – 91.5)	91.5 (89.3 – 93.3)	70.0 (68.3 – 71.8)	68.6 (64.9 – 72.1)	72.7 (71.0 – 74.4)	68.6 (65.0 – 72.1)
AUC	0.98 (0.98 – 0.98)	0.98 (0.98 – 0.99)	0.75 (0.73 – 0.77)	0.73 (0.68 – 0.77)	0.76 (0.75 – 0.77)	0.70 (0.68 – 0.73)
Positive Likelihood Ratio	9.18 (8.05 – 10.46)	11.93 (8.77 – 16.25)	2.36 (2.16 – 2.59)	2.20 (1.82 – 2.66)	3.05 (2.72 – 3.43)	2.11 (1.69 – 2.64)
%Positive Predictive Value	82.9 (80.7 – 85.0)	86.2 (81.6 – 90.1)	52.5 (49.4 – 55.6)	50.8 (44.4 – 57.3)	60.5 (57.0 – 63.8)	47.2 (40.1 – 54.4)
%Negative Predictive Value	95.2 (94.1 – 96.1)	94.4 (92.0 – 96.2)	80.9 (78.9 – 82.8)	78.9 (74.7 – 82.7)	77.8 (75.9 – 79.6)	77.4 (73.4 – 81.0)

Table 4.8 F1 score and Brier score of the models

Metrics\Models	**Mixed effects Logistic Regression**		**Recurrent Neural Networks**				**Mixed effects Support Vector Machine**			
			Overfit model		Final model		Overfit model		Final model	
Metrics\Data	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
F1 score	0.8694	0.8782	0.8792	0.4396	0.5731	0.5427	0.9759	0.4922	0.5636	0.4665
Brier score	0.0610	0.0578	0.0625	0.2770	0.1809	0.1866	0.0403	0.2668	0.1978	0.2000

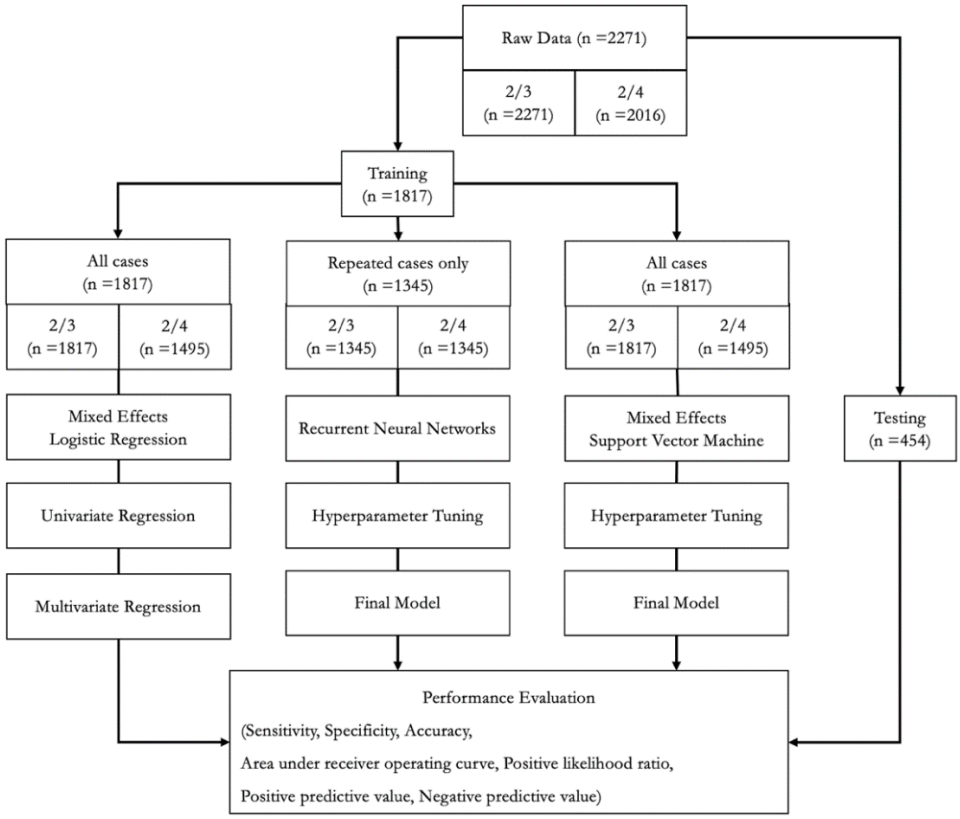

Figure 4.1 Model Development Diagram

a) b)

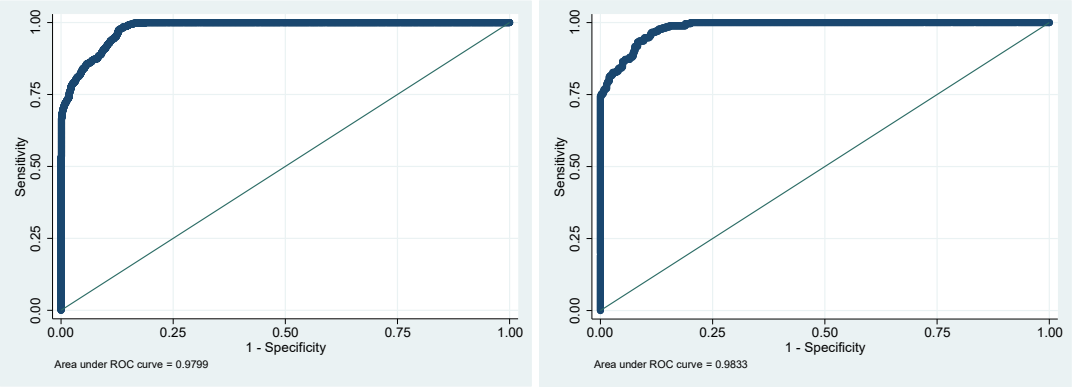

Figure 4.2 Receiver operating curve of mixed effects logistic regression a) – on training data and right, b) on the testing data

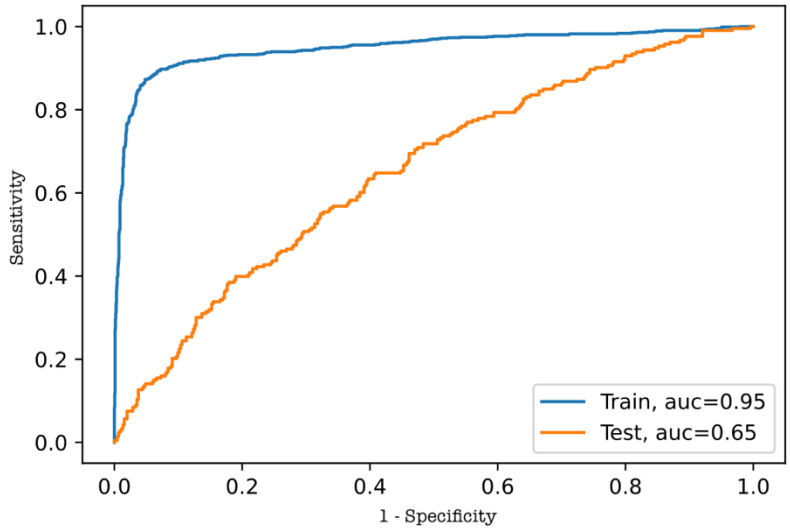

Figure 4.3 Receiver operating curve of overfit recurrent neural network

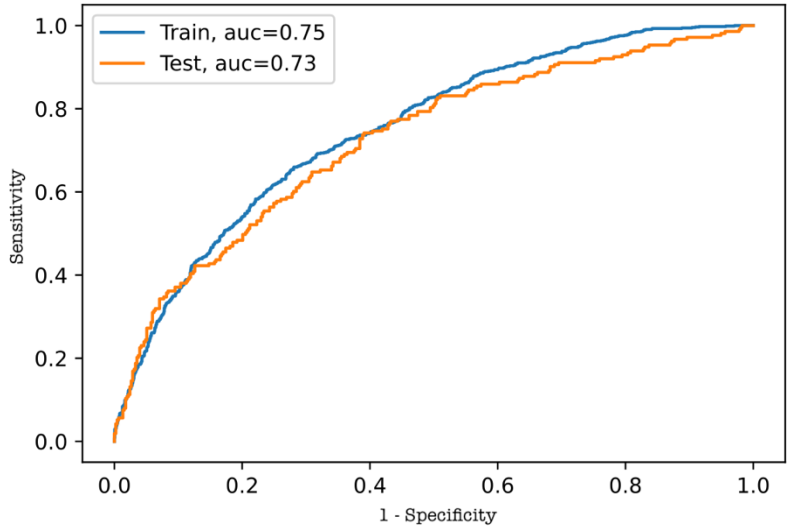

Figure 4.4 Receiver operating curve of final recurrent neural network

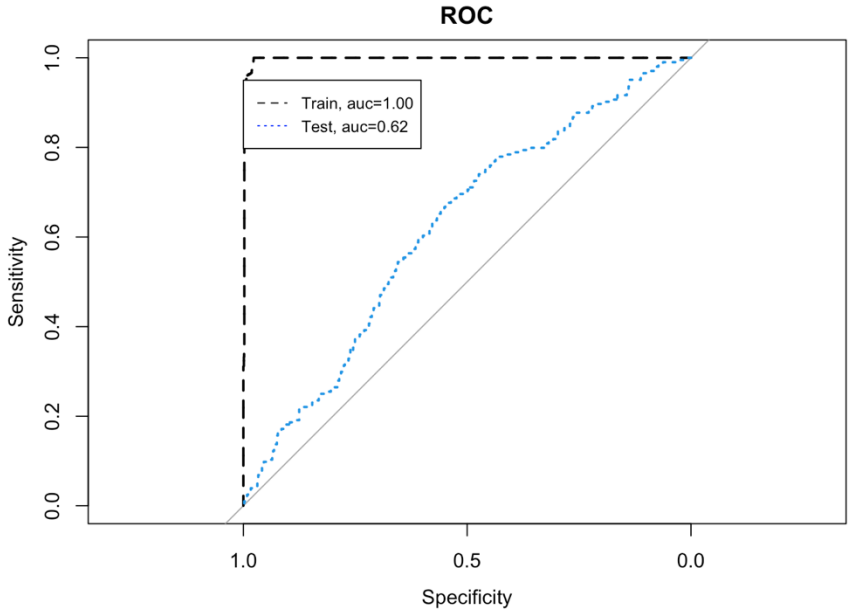

Figure 4.5 Receiver operating curve of overfit Mixed Effects – Support Vector Machine

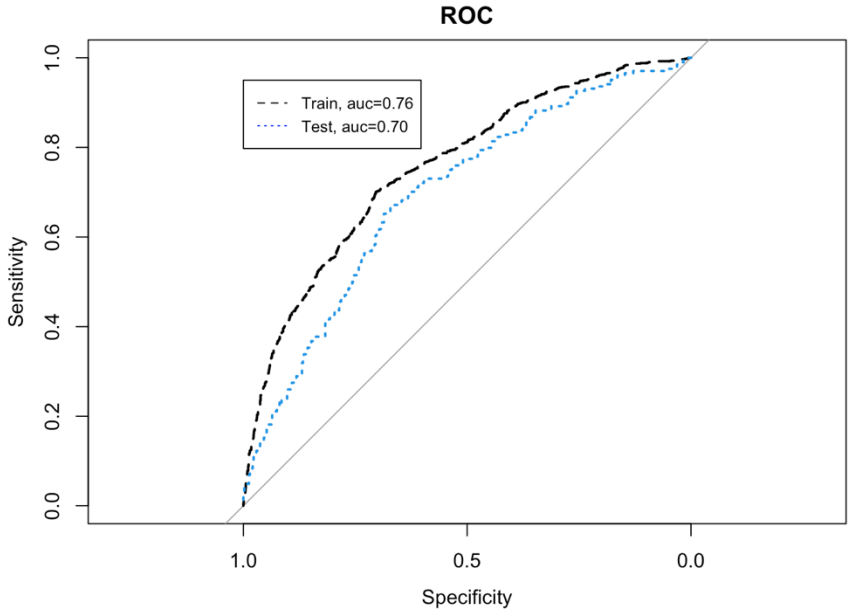

Figure 4.6 Receiver operating curve of final Mixed Effects – Support Vector Machine

CHAPTER V
DISCUSSIONS

Previous studies had applied cross-sectional regression models to predict the probability of having chronic periodontitis. With our study, data from longitudinal studies including samples with repeated measurements were used as training data for both the statistical model and machine learning models. It was observed that mixed effects logistic regression model had highest performance when compared to the machine learning models as well as cross-sectional logistic regression models.

From the literature review, Verhulst et al.[44] was considered to have the best performance with 80% and 88% for sensitivity and specificity respectively, and area under receiver operating curve value of 0.91. While the model applied salivary biomarker data such as Chitinase and Protease activity in addition to demographics and oral features, our mixed effects logistic regression model had comparably better performance with 89.5% and 92.5% for sensitivity and specificity, and discriminative power of 0.98. Considering our model considered only six features such as gender, education level, smoking habit, diabetes mellitus, dental plaque score and number of present teeth in dentition, this resulting superior performance owes to the longitudinal data.

Mixed effects logistic regression models are considered to be superior to simple logistic regression models because they consider random effects or subject specific effects by considering multiple observations of the same subject in addition to fixed effects estimated by conventional models. Machine learning models failed to live up to the expectations, observing less stellar performances because they were unable to take advantage of their complex algorithms and advantages. For example, while the machine learning models can handle data with higher dimensions and correlated features, there were only 21 predictor features even when all relevant features nominated by the expert committee was included. Machine learning models can handle unstructured data such as images and signals but our data at hand were panel data

including sociodemographic features. All models learned solely from tabulated structured data therefore although the machine learning models included more features compared to the statistical mixed logistic regression model, they had inferior or poor performances even after hyperparameter optimization procedure. We considered several possible factors affecting the performance of the models.

5.1 Minority positive class

In many real-life problems, imbalanced datasets happen due to multiple reasons such as selection of survey population done correctly or not. The class imbalance problem can be better understood as three separate problems, which are –

1. assuming that a performance metric is appropriate when it is not,
2. assuming that the test distribution matches the training distribution when it is not,
3. assuming that there is enough minority class when it is not.

Provost F. (2000)[48] states that two fundamental assumptions are made in traditional classifiers. The first is that the goal of the classifiers is maximum accuracy (minimum error rate); the second is that the class distribution of the training and test datasets is the same. Under these two assumptions, predicting everything as the majority class for an imbalanced dataset is often the right thing to do.

Within 776 observations of our 454 testing subjects, 509 observations were negative. Considering if a classification model predicts all observations to be negative, 267 observations will be incorrectly identified as negative (false negative), and 509 observations will be correctly identified as negative thus true negative. While there will be zero observations for both correctly identified positives (true positive) and incorrectly identified positives (false positive), we will have 65.6% accuracy rate regardless of the usability of such model.

Sensitivity is the ratio of true positive to all positive observations where specificity is the same but for negative cases. Positive predictive value is the ratio of true positive to all observations predicted as positive and the same goes for negative predictive value with negative cases. In the stated scenario, the sensitivity of this model will be zero in contrast to 100% specificity, meaning the model is unusable. Similarly,

negative predictive value will be 65.6% and positive predictive value is zero as well since the model cannot detect positive (non-disease) subjects.

Therefore, in machine learning algorithms, hyperparameter tuning is done during the training process to optimize the performance of the model. Depending on the chosen set of hyperparameters, the model can become overfit to the training dataset like several reported models in chapter 4. The model pays a lot of attention to random noises in the training data, so they fail to generalize on the data it has not seen before, and they are considered as high variance. As a result, they perform very well on the training dataset but high error rate on the testing dataset. On the contrary, the model can become biased by paying very little attention to the training data, resulting in oversimplified models. They lead to high error rate on both training and testing data. Further, hyperparameter optimization process is done to balance between bias-variance tradeoff by comparing the model performance on both training and testing data.

Even with appropriate optimization, the training data distribution should reflect the true distribution or prevalence of the condition, so that the model can learn to generalize and perform similarly on new subjects as well. That also applies with the data splitting where the testing data distribution should reflect the training data. According to the 8th Thailand national health survey (2017), 26% of Thai adults and 36% of Thai elderly people had severe chronic periodontitis. Our surveys included subjects who are the current employees of Electricity Generating Authority of Thailand with the mean age of 54.4 (43.7 – 75.3) and our training dataset reflected to 1,094 (34.6%) from 3,158 observations having severe chronic periodontitis which can be considered consistent. The testing data included 267 positive observations (34.4%) out of 776, which also matched appropriately.

Ling (2010)[49] states that the imbalanced class problem becomes meaningful only if one or both two assumptions above are false; that is, if the cost of different types of error (false positive and false negative in the binary classification) is not the same, or if the class distribution in the test data is different from that of the training data. The first problem was effectively dealt with cost-sensitive models. In recurrent neural networks, the amount of error in each subject is evaluated with a loss function such as binary cross entropy as shown in Figure 5.1, and the overall error of the model is considered the cost of the model. During the training process, for each set of weights

and biases for the hidden layers, the cost value is calculated, and these sets are adjusted to decrease the cost value as much as possible. Thus, by adjusting the loss value for misclassification, we can guide the model into more balanced performance instead of preferring the majority class. In Figure 5.2, class weights were applied to make it more expensive to misclassify a minority class into majority class than a majority into minority, which would further encourage prediction of everything as the majority class. Since we had similar class distribution for all our datasets, we could disregard the second problem as well. Then the literature suggests inadequate number of samples in the minority class for the classifier to learn adequately, which means we had a problem of insufficient or small training class which is different from imbalanced class problem. It could only be addressed by collecting more minority class subjects.

5.2 Limitations of the current study

To adjust for the second assumption made above, class weights were planned to be applied for recurrent neural networks. Keras library is a python library with TensorFlow backend, a major utility for training neural networks and deep learning and our source of choice for the recurrent neural networks. This library considers the concept of class to be ambiguous in data with 3 or more dimensions, which is the input data dimension for recurrent neural networks. Thus, sample weights based on class weights were applied instead as shown in Figure 5.3. For mixed effects support vector machine, e1071 library being applied for support vector machine in the model does not have an option to adjust for class weights. However, observing that mixed effects logistic regression does not require adjusting class weights and recurrent neural networks having similar problems even with class weights applied, we considered the poor performances were the problem of insufficient positive class rather than imbalanced class problem.

In mixed effects logistic regression, observations of different subjects are used by logistic regression to estimate the fixed effects or population average effects of the selected predictor variables (gender, education, number of teeth) on the target variable and multiple records of the same subject are used to adjust for subjects specific or random effects. While the mixed models accept only one observation as well as

repeated measurements, recurrent neural networks require all the training subjects to have exactly same number of timesteps. For other applications of recurrent neural networks such as natural language processing, padding and masking techniques are applied to adjust, but it is not done in our study. Therefore, we had to remove subjects with only one observation from the training and testing data, resulting in decreased number of subjects in comparison with other models. We considered this to be one of the major factors affecting the performance of our neural network.

Main advantage of recurrent neural networks is the ability to consider previous timesteps in terms of hidden vectors together with current features. However, since we had only two timesteps, the first timestep was basically a multilayer perceptron (simple artificial neural networks) mapping from features to the target variable at the first timestep. The second timestep would include the context from the first timestep, yet it was observed that the performance of the recurrent neural networks was inferior compared to mixed effects logistic regression model. Typically, the problem with similar models is that the model forgetting over long sequences but here we believe small number of timesteps as well as small training class resulted in poor performance of the model.

For our machine learning models, we did not do further dimensional reduction over expert opinions and decision with the advisor team. Mixed effects logistic regression, the statistical model required feature reduction, since including too much could result in overfitting. However, we need to balance the appropriate number of features since not including all features correlated with the target will result in inferior performance of the model. While we do not have a set limit on numbers of included parameters within the model, several rules of thumb such as one predictor parameter for ten events (one in ten rule), one in twenty rule and one in fifty rules have been suggested.[50] Here we applied stepwise forward selection with statistical significance of 0.1 for univariate and 0.05 for multivariate regression. Of course, this approach is not without its drawbacks, since stepwise method is considered unstable[51] in a sense that addition or removal of a covariate can result in varying p-value of the parameter, including scenarios where they become insignificant in multivariate regressions. However, we may consider our mixed effects logistic regression to have appropriate performance without overfitting or inferior predictive ability.

For sigmoid-based classification models, the output of the models are probabilities of having the positive class. Therefore, we must select a threshold on which we would dichotomize the value. The default value would be 0.5, but currently the decision threshold was 0.35 to reflect the prevalence of the condition in our data (34.6%). However, we may adjust the threshold to overestimate or underestimate since the cost of having more false predictions is different based on the problem. By lowering the decision threshold, the model will overestimate by considering subjects with lower probability to be positive, which means that it will result in less false negatives and more false positives. We are willing to accept more false positive subjects since we do not want to miss the opportunity of early diagnosis by getting a false negative in the screening step. The follow up examination is what we are trying to circumvent, however the screening system will reduce the overall workload necessary regardless as shown in Figure 5.4. We need to balance between demerits of following up and demerits of not following up.

5.3 Application on mock data

To exercise in applying our classifiers for the screening purposes, the models were applied with selected samples. Four mock samples who were present at both surveys were selected and a subset of their features which were applied by mixed effects logistic regression model are shown in Table 5.1. Four subjects had different disease progression over different observations,

1. continuing healthy periodontium,
2. persisting severe chronic periodontitis,
3. developing over time and
4. recovering over time.

The selected mock population had 25% female and 3 subjects were 75% non-smokers. All subjects had at least a bachelor's degree, and none had diabetes mellitus. Average number of present teeth in the first survey was 23.5 and in the second, it was 22.25 teeth with two subjects losing dentition over time. The female subject had decreased oral hygiene over time from 22.7% to 31.8% of tooth surfaces with dental plaque adhesion in second survey but still had a better oral hygiene compared to the male subjects with

average of 63.12 plaque score. Models with selected final sets of hyperparameters were performed on the mock samples to evaluate their performance.

The characteristics of the subjects underwent necessary feature selection or feature encodings before being given to the models. Since all models were sigmoid or logit based, they output the log odds having severe chronic periodontitis which was transformed into probability, followed by dichotomization with a decision threshold. Predicted periodontal status and the probabilities outputted by the model are reported in Table 5.2.

Brier score measures the average discrepancy in outputted probabilities of the model in a form similar to mean squared error in regression problems albeit with probabilities. Mixed effects logistic regression model had Brier score of 0.1664, recurrent neural networks had Brier score of 0.1690 and mixed effects support vector machine performed the worst with Brier score of 0.261. It should be noted that our selection for cutoff point influenced the performance of the models. For subject C at EGAT 2/4 (predicted probability of 0.35) and subject D at EGAT 2/3 (predicted probability of 0.36), mixed effects logistic regression would have incorrectly identified as negative if we do not reduce the value to 0.35. Machine learning models especially mixed-effects support vector machine tends to underestimate, i.e., predict lower probabilities overall. As stated before, the decision threshold should be manipulated as necessary to be suitable for our goals.

5.4 Application in real life scenarios

Logistic regression models have been traditionally applied as scoring systems. Since logistic regressions are linear relationship of predictor features to the log-odds, the intercept of the model with the coefficients of each features multiplied with the features of a subject can output the logit of the subject, which in turn can be converted to the probability of having the condition. To assess the risk score for developing severe periodontitis,

Risk score = -3.93 + (0.97 x male)

+ (2.04 x education < High school)

+ (1.35 x education Vocational School)

+ (0.29 x education Bachelor's degree)
+ (0.73 x Ex-smoker) + (1.68 x Current smoker)
+ (0.50 x diabetes mellitus)
+ (-0.06 x number of teeth) + (0.03 x plaque score)

– where the covariate should be replaced with 1 if applicable and 0 if else. From the risk score, the subject's risk of developing the condition can be calculated as $\frac{e^{Risk\ score}}{1+e^{Risk\ score}}$.

With machine learning models, the concepts of coefficients are ambiguous to calculate manually. Instead, the models are outputted as a file format such as flask, pickle, or hierarchy data format (.hdf5/ .h5py). The model can be imported in web services such as Amazon Web Service (AWS) or Heroku to deploy. Advantage of this approach is that the web application can be built to be visually appealing and easily applicable by the intended users. The complex applications are done in the background and additional processes such as data scraping and preprocessing from electronic medical records can be automated as well.

With necessary internet connectivity, the model can be updated in the backend with new data that can also be collected with a web application. With new data, the effects of each predictor known as coefficients or weights can be readjusted or updated with new evidence. Similar systems can be built for logistic regression models as well, but unlike machine learning models, all the previous training data must be stored and trained together with the new data so that the coefficient can be updated. Of course, the validity of the user-inputted data would be a concern to be included as the training data as well as there will be privacy concern for valid data. Easy access of risk assessment programs can lead to overuse or apprehension of those who might not be the target population.

5.5 Future Research and Study

For both training and testing of our models, we applied data from the same source. For proper model evaluation, external validation using data from other centers or surveys is required. Data from other Thai populations as well as different countries or ethnicities should be used to evaluate the capability of the model to generalize. If

necessary, the models should be updated applying new data, especially the recurrent neural networks which we considered to be suffering from insufficient training class and insufficient timesteps to make full use of its unique capability.

With appropriate or acceptable performance, the models should be able to deploy, so that they can help screening the situations where large numbers of people are to be periodontally examined such as public health missions. Application programming interfaces (APIs) can be used to scan the health information systems to screen the patients ahead of time. Web or desktop applications can be deployed at the stations where history taking interviews are done. As shown in Figure 5.5, mobile applications should help the staffs to apply while on the go, or even let the examinees apply by themselves. Additional functionality such as scoring oral hygiene, recommending oral hygiene regiments and dental treatment visits could be bundled together in such applications to encourage or positively reinforce for better oral health knowledge and practices.

5.6 Conclusion

While applying several classification models as the screening models for chronic periodontitis, machine learning models such as recurrent neural networks and mixed effects support vector machine failed to perform better than the statistical models. For better applications of the machine learning models, we will have to address the limitations as stated before. In addition, including the different type of data such as orthopantomograms would take advantage of the capabilities of the machine learning models as well as increase their classification performances. Currently, the mixed effects logistic regression model resulted in superior performance with 11.93 positive likelihood ratio. Since the model was trained and tested with the data from the same survey, external validation using other population is necessary. With acceptable performances, such screening model will be able to save time, material, and human resources necessary to manually measure 168 individual sites per every examinee.

Table 5.1 Subset of mock data samples

ID	Survey	Sex	Education	Smoking	Diabetes Mellitus	Number of teeth	Plaque score	Diagnosis
A	2/3	Female	> Bachelor's degree	Non-smoker	Negative	22	22.7	Negative
A	2/4	Female	> Bachelor's degree	Non-smoker	Negative	22	31.8	Negative
B	2/3	Male	Bachelor's degree	Non-smoker	Negative	21	85.7	Positive
B	2/4	Male	Bachelor's degree	Non-smoker	Negative	21	95.2	Positive
C	2/3	Male	Bachelor's degree	Non-smoker	Negative	26	100	Negative
C	2/4	Male	Bachelor's degree	Non-smoker	Negative	25	74	Positive
D	2/3	Male	Bachelor's degree	Ex-smoker	Negative	25	44	Positive
D	2/4	Male	Bachelor's degree	Ex-smoker	Negative	21	50	Negative

Table 5.2 Predicted periodontal status and probabilities outputs of three classification models on the mock data samples (0.35 as decision threshold)

ID	Survey	True Diagnosis	MELR	RNN	MESVM
A	2/3	Negative	N (0.01)	N (0.06)	N (0.24)
A	2/4	Negative	N (0.01)	N (0.03)	N (0.23)
B	2/3	Positive	P (0.85)	P (0.75)	P (0.36)
B	2/4	Positive	P (0.88)	P (0.53)	P (0.36)
C	2/3	Negative	P (0.50)	P (0.45)	N (0.31)
C	2/4	Positive	P (0.35)	N (0.23)	N (0.31)
D	2/3	Positive	P (0.36)	P (0.57)	N (0.30)
D	2/4	Negative	P (0.46)	N (0.29)	N (0.31)

Abbreviations-

MELR = Mixed Effects Logistic Regression

MESVM = Mixed Effects Support Vector Machine

N = None or Non-severe chronic periodontitis

P = Severe chronic periodontitis

X (0.0) = Predicted diagnosis (probability of having severe chronic periodontitis)

RNN = Recurrent Neural Networks

Subject	Class	Binary Cross-entropy $loss = -y(log\hat{y}) - (1-y)(\log(1-\hat{y}))$
A	0	$class\ zero\ loss = -\log(1-\hat{y})$
B	0	$class\ zero\ loss = -\log(1-\hat{y})$
C	0	$class\ zero\ loss = -\log(1-\hat{y})$
D	0	$class\ zero\ loss = -\log(1-\hat{y})$
E	1	$class\ one\ loss = -log\hat{y}$

$$Cost = \text{Subject A loss} + \text{Subject B loss} + \text{Subject C loss} + \text{Subject D loss} + \text{Subject E loss}$$
$$= 4\ (class\ zero\ loss) + class\ one\ loss$$

Figure 5.1 Cost function for imbalanced class

Subject	Class	Loss	$Class\ weight = \{0:1, 1:4\}$
A	0	$class\ zero\ loss$	$\times 1$
B	0	$class\ zero\ loss$	$\times 1$
C	0	$class\ zero\ loss$	$\times 1$
D	0	$class\ zero\ loss$	$\times 1$
E	1	$class\ one\ loss$	$\times 4$

$$weighted\ cost = 4(class\ zero\ loss) + 4(class\ one\ loss)$$

Figure 5.2 Class weight-adjusted cost function

Subject	Class	Loss	$Sample\ weight = \{A: 1, B: 1, C: 1, D: 1, E: 4\}$
A	0	*class zero loss*	$\times 1$
B	0	*class zero loss*	$\times 1$
C	0	*class zero loss*	$\times 1$
D	0	*class zero loss*	$\times 1$
E	1	*class one loss*	$\times 4$

$$weighted\ cost = 4(class\ zero\ loss) + 4(class\ one\ loss)$$

Figure 5.3 Sample weight-adjusted cost function

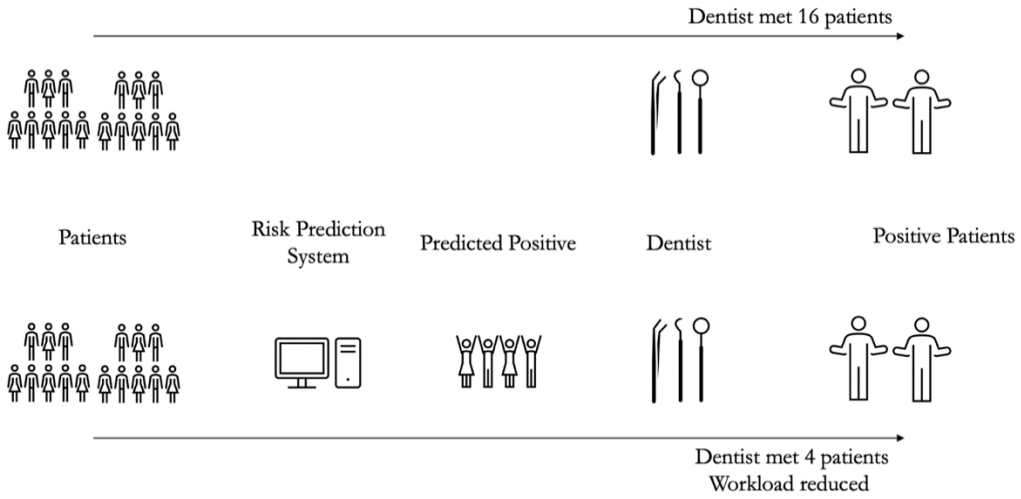

Figure 5.4 Screening system in action

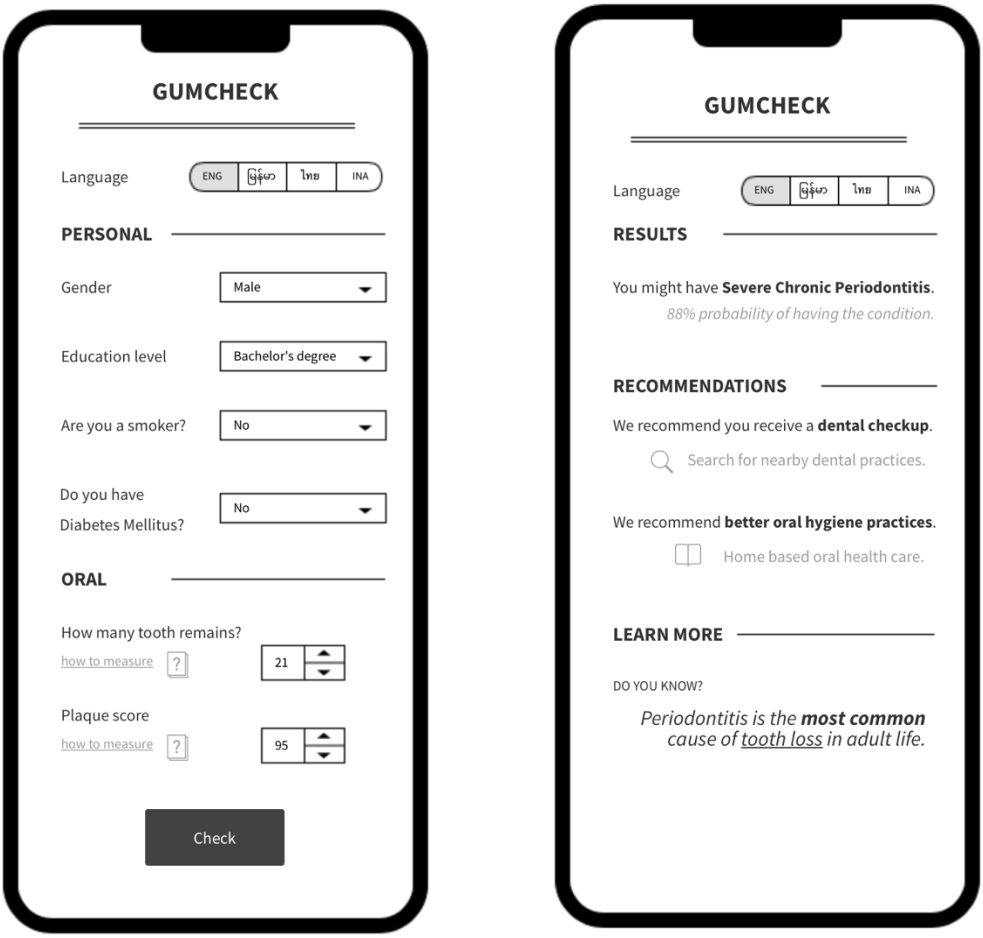

Figure 5.5 Mobile application based on the mixed effects logistic regression model

REFERENCES

1. Phipps KR, Stevens VJ. Relative contribution of caries and periodontal disease in adult tooth loss for an HMO dental population. J Public Health Dent. 1995;55(4):250-2.
2. Corbet EF, Leung WK. Epidemiology of periodontitis in the Asia and Oceania regions. Periodontology 2000. 2011;56(1):25-64.
3. Tonetti MS, Jepsen S, Jin L, Otomo-Corgel J. Impact of the global burden of periodontal diseases on health, nutrition and wellbeing of mankind: A call for global action. Journal of Clinical Periodontology. 2017;44(5):456-62.
4. Linden GJ, Lyons A, Scannapieco FA. Periodontal systemic associations: review of the evidence. Journal of Clinical Periodontology. 2013;40(s14):S8-S19.
5. Mattila KJ, Nieminen MS, Valtonen VV, Rasi VP, Kesäniemi YA, Syrjälä SL, et al. Association between dental health and acute myocardial infarction. Bmj. 1989;298(6676):779-81.
6. Tonetti MS, Van Dyke TE. Periodontitis and atherosclerotic cardiovascular disease: consensus report of the Joint EFP/AAP Workshop on Periodontitis and Systemic Diseases. J Clin Periodontol. 2013;40 Suppl 14:S24-9.
7. Lertpimonchai A, Rattanasiri S, Tamsailom S, Champaiboon C, Ingsathit A, Kitiyakara C, et al. Periodontitis as the risk factor of chronic kidney disease: Mediation analysis. J Clin Periodontol. 2019;46(6):631-9.
8. Monsarrat P, Blaizot A, Kémoun P, Ravaud P, Nabet C, Sixou M, et al. Clinical research activity in periodontal medicine: a systematic mapping of trial registers. J Clin Periodontol. 2016;43(5):390-400.
9. Page RC, Krall EA, Martin J, Mancl L, Garcia RI. Validity and accuracy of a risk calculator in predicting periodontal disease. J Am Dent Assoc. 2002;133(5):569-76.
10. Persson GR, Mancl LA, Martin J, Page RC. Assessing periodontal disease risk: a comparison of clinicians' assessment versus a computerized tool. J Am Dent Assoc. 2003;134(5):575-82.
11. Lang NP, Tonetti MS. Periodontal risk assessment (PRA) for patients in supportive periodontal therapy (SPT). Oral Health Prev Dent. 2003;1(1):7-16.
12. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798-828.
13. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
14. Kwon O, Na W, Kim YH. Machine Learning: a New Opportunity for Risk Prediction. Korean Circ J. 2020;50(1):85-7.
15. Periodontal Risk Assessment perio-tools.com [Available from: https://www.perio-tools.com/pra/en/

16. Eke PI, Dye BA, Wei L, Slade GD, Thornton-Evans GO, Borgnakke WS, et al. Update on Prevalence of Periodontitis in Adults in the United States: NHANES 2009 to 2012. J Periodontol. 2015;86(5):611-22.
17. Al-Harthi LS, Cullinan MP, Leichter JW, Thomson WM. The impact of periodontitis on oral health-related quality of life: a review of the evidence from observational studies. Aust Dent J. 2013;58(3):274-7; quiz 384.
18. Leite FRM, Peres KG, Do LG, Demarco FF, Peres MAA. Prediction of Periodontitis Occurrence: Influence of Classification and Sociodemographic and General Health Information. J Periodontol. 2017;88(8):731-43.
19. Cyrino RM, Miranda Cota LO, Pereira Lages EJ, Bastos Lages EM, Costa FO. Evaluation of self-reported measures for prediction of periodontitis in a sample of Brazilians. J Periodontol. 2011;82(12):1693-704.
20. Lai H, Su CW, Yen AM, Chiu SY, Fann JC, Wu WY, et al. A prediction model for periodontal disease: modelling and validation from a National Survey of 4061 Taiwanese adults. J Clin Periodontol. 2015;42(5):413-21.
21. Javali S, Sunkad M, Wantamutte A. Prediction of risk factors of periodontal disease by logistic regression: a study done in Karnataka, India. International Journal Of Community Medicine And Public Health. 2018;5:5301.
22. Eke PI, Dye BA, Wei L, Slade GD, Thornton-Evans GO, Beck JD, et al. Self-reported measures for surveillance of periodontitis. J Dent Res. 2013;92(11):1041-7.
23. Ababneh KT, Abu Hwaij ZM, Khader YS. Prevalence and risk indicators of gingivitis and periodontitis in a multi-centre study in North Jordan: a cross sectional study. BMC Oral Health. 2012;12:1.
24. Wu X, Weng H, Lin X. Self-reported questionnaire for surveillance of periodontitis in Chinese patients from a prosthodontic clinic: a validation study. J Clin Periodontol. 2013;40(6):616-23.
25. Zhan Y, Holtfreter B, Meisel P, Hoffmann T, Micheelis W, Dietrich T, et al. Prediction of periodontal disease: modelling and validation in different general German populations. J Clin Periodontol. 2014;41(3):224-31.
26. Shankarapillai R, Mathur L, Ananthakrishnan Nair M, Rai N, Mathur A. Periodontitis risk assessment using two artificial neural networks-A pilot study. International Journal of Dental Clinics. 2010;2:36-40.
27. Ozden FO, Özgönenel O, Özden B, Aydogdu A. Diagnosis of periodontal diseases using different classification algorithms: a preliminary study. Niger J Clin Pract. 2015;18(3):416-21.
28. Torrungruang K, Tamsailom S, Rojanasomsith K, Sutdhibhisal S, Nisapakultorn K, Vanichjakvong O, et al. Risk Indicators of Periodontal Disease in Older Thai Adults. Journal of periodontology. 2005;76:558-65.
29. Verhulst M, Teeuw W, Bizzarro S, Muris J, Naichuan S, Nicu E, et al. A rapid, non-invasive tool for periodontitis screening in a medical care setting. BMC Oral Health. 2019;19.
30. Tadjoedin F, Fitri AH, Kuswandani S, Sulijaya B, Soeroso Y. The correlation between age and periodontal diseases. Journal of International Dental and Medical Research. 2017;10:327-32.
31. Ioannidou E. The Sex and Gender Intersection in Chronic Periodontitis. Front Public Health. 2017;5:189.

32. Javed F, Tenenbaum HC, Nogueira-Filho G, Qayyum F, Correa FO, Al-Hezaimi K, et al. Severity of periodontal disease in individuals chewing betel quid with and without tobacco. Am J Med Sci. 2013;346(4):273-8.
33. Suvan J, Petrie A, Moles DR, Nibali L, Patel K, Darbar U, et al. Body mass index as a predictive factor of periodontal therapy outcomes. J Dent Res. 2014;93(1):49-54.
34. Thakur A, Guleria P, Bansal N, editors. Symptom & risk factor based diagnosis of Gum diseases using neural network. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence); 2016 14-15 Jan. 2016.
35. Macedo Paizan ML, Vilela-Martin JF. Is there an association between periodontitis and hypertension? Curr Cardiol Rev. 2014;10(4):355-61.
36. Nibali L, D'Aiuto F, Griffiths G, Patel K, Suvan J, Tonetti MS. Severe periodontitis is associated with systemic inflammation and a dysmetabolic status: a case-control study. J Clin Periodontol. 2007;34(11):931-7.
37. Drucker H, Burges C, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Adv Neural Inform Process Syst. 1997;28:779-84.
38. Schölkopf B, Bartlett P, Smola A, Williamson R. Shrinking the tube: a new support vector regression algorithm. Proceedings of the 11th International Conference on Neural Information Processing Systems; Denver, CO: MIT Press; 1998. p. 330–6.
39. Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation. 2014;84(6):1313-28.
40. Hajjem A, Larocque D, Bellavance F. Generalized mixed effects regression trees. Statistics & Probability Letters. 2017;126:114-8.
41. Moddemeijer R, editor On the convergence of the iterative solution of the likelihood equations2006.
42. Armitage GC. Development of a classification system for periodontal diseases and conditions. Ann Periodontol. 1999;4(1):1-6.
43. Page RC, Eke PI. Case definitions for use in population-based surveillance of periodontitis. J Periodontol. 2007;78(7 Suppl):1387-99.
44. Verhulst MJL, Teeuw WJ, Bizzarro S, Muris J, Su N, Nicu EA, et al. A rapid, non-invasive tool for periodontitis screening in a medical care setting. BMC Oral Health. 2019;19(1):87.
45. Vathesatogkit P, Woodward M, Tanomsup S, Ratanachaiwong W, Vanavanan S, Yamwong S, et al. Cohort profile: the electricity generating authority of Thailand study. Int J Epidemiol. 2012;41(2):359-65.
46. Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. Bmj. 2020;368:m441.
47. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics. 2018;19(1):432.
48. Provost F, editor Machine Learning from Imbalanced Data Sets 1012008.
49. Ling CX, Sheng VS. Class Imbalance Problem. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer US; 2010. p. 171-.

50. Sneyd JR. Interactive Textbook on Clinical Symptom Research: Methods and Opportunities. BJA: British Journal of Anaesthesia. 2003;90(4):532-.
51. Steyerberg EW, Eijkemans MJC, Harrell Jr FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine. 2000;19(8):1059-79.

APPENDIX A
ETHICAL CLEARNACE

Human Research Ethics Committee, Faculty of Medicine Ramathibodi Hospital, Mahidol University
270 Rama 6 Rd. Phayatai Ratchathewi Bangkok 10400 Tel.(660)2012175, 2011544, 2010388
Website: https://med.mahidol.ac.th/research/ethics
E-mail: raec.mahidol@gmail.com

COA. MURA2020/1560

Title of Project (English)	Clinical Prediction of Chronic Periodontitis Using Machine Learning
Type of Review	Expedited
Principal Investigator	Htun Teza, BDS.
Official Address	Department of Clinical Epidemiology and Biostatistics Faculty of Medicine Ramathibodi Hospital Mahidol University
Co-investigator (s)	1. Anuchate Pattanateepapon, MSc. 2. Ammarin Thakkinstian, Ph.D. 3. Prin Vathesatogkit, M.D. 4. Attawood Lertpimonchai, Ph.D.
Approval includes	1. Submission Form Protocol Version 1 Date 21/08/2020 2. Certificate in Ethics Training

Institutional Review Boards in Mahidol University are in full compliance with International Guidelines for Human Research Protection such as Declaration of Helsinki, The Belmont Report, CIOMS Guidelines and the International Conference on Harmonization in Good Clinical Practice (ICH-GCP)

Date of Approval	**September 29, 2020**
Date of Expiration	**September 28, 2021**

Signature of Chair.....

(Asst. Prof. Chusak Okascharoen, M.D., Ph.D.)

This certificate is subject to the following conditions:
1) Approval is granted only for the project with details described in submitted proposal
2) Submission of modification to the approved project is needed before implementation
3) A yearly progress report is required for renewing of approval
4) Written notification is required when the project is complete or terminated

BIOGRAPHY

NAME	HTUN TEZA
DATE OF BIRTH	5 September 1996
PLACE OF BIRTH	Nay Pyi Taw, Myanmar
INSTITUTIONS ATTENDED	UNIVERSITY OF DENTAL MEDICINE YANGON, 2012-2018 Bachelor of Dental Surgery MAHIDOL UNIVERSITY, 2019-2021 Master of Science (Data Science for Health Care)
SCHOLARSHIP RECEIVED	Faculty of Medicine Ramathibodi Hospital, 2019 Faculty of Graduate Studies Mahidol University, 2019-2021
HOME ADDRESS	Bo Tar Yar Road, Pyinmana, Nay Pyi Taw, Myanmar
PUBLICATION / PRESENTATION	CLINICAL PREDICTION OF CHRONIC PERIODONTITIS 7th Regional Conference on Graduate Research, Sripatum University January 16, 2021