May 2017

Development and Evaluation of an Interdisciplinary Periodontal Risk Prediction Tool Using a Machine Learning Approach

Neel Anil Shimpi
University of Wisconsin-Milwaukee

DEVELOPMENT AND EVALUATION OF AN INTERDISCIPLINARY PERIODONTAL RISK PREDICTION TOOL USING A MACHINE LEARNING APPROACH

by

Neel Shimpi

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in

Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

May 2017

ABSTRACT
DEVELOPMENT AND EVALUATION OF AN INTERDISCIPLINARY PERIODONTAL RISK PREDICTION TOOL USING A MACHINE LEARNING APPROACH

by

Neel Shimpi

The University of Wisconsin-Milwaukee, 2017
Under the Supervision of Professor Susan McRoy, PhD

Periodontitis (PD) is a major public health concern which profoundly affects oral health and concomitantly, general health of the population worldwide. Evidence-based research continues to support association between PD and systemic diseases such as diabetes and hypertension, among others. Notably PD also represents a modifiable risk factor that may reduce the onset and progression of some systemic diseases, including diabetes. Due to lack of oral screening in medical settings, this population does not get flagged with the risk of developing PD.

This study sought to develop a PD risk assessment model applicable at clinical point-of-care (POC) by comparing performance of five supervised machine learning (ML) algorithms: Naïve Bayes, Logistic Regression, Support Vector Machine, Artificial Neural Network and Decision Tree, for modeling risk by retrospectively interrogating clinical data collected across seven different models of care (MOC) within the interdisciplinary settings. Risk assessment modeling was accomplished using Waikato Environment for Knowledge Analysis (WEKA) open-sourced tool, which supported comparative assessment of the relative performance of the five ML algorithms when applied to risk prediction.

To align with current conventions for clinical classification of disease severity, predicting PD risk was treated as a ‘classification problem’, where patients were sorted into two categories

based on disease severity and 'low risk PD' was defined as no or mild gum disease ('controls') or 'high risk PD' defined as moderate to severe disease ('cases'). To assess the predictive performance of models, the study compared performance of ML algorithms applying analysis of recall, specificity, area under the curve, precision, F-measure and Matthew's correlation coefficient (MCC) and receiver operating characteristic (ROC) curve. A tenfold-cross validation was performed. External validation of the resultant models was achieved by creating validation data subsets applying random selection of approximately 10% of each class of data proportionately.

Findings from this study have prognostic implications for assessing PD risk. Models evolved in the present study have translational value in that they can be incorporated into the Electronic Health Record (EHR) to support POC screening. Additionally, the study has defined relative performance of PD risk prediction models across various MOC environments. Moreover, these findings have established the power ML application can serve to create a decision support tool for dental providers in assessing PD status, severity and inform treatment decisions. Further, such risk scores could also inform medical providers regarding the need for patient referrals and management of comorbid conditions impacted by presence of oral disease such as PD. Finally, this study illustrates the benefit of the integrated medical and dental care delivery environment for detecting risk of periodontitis at a stage when implementation of proven interventions could delay and even prevent disease progression.

Keywords: Periodontitis, Risk Assessment, Interprofessional Relations, Machine learning, Electronic Health Records, Decision Support Systems

To my

Grandfather Shri. Kashinath M. Shimpi

and

Grandmother Late Smt. Shailaja K. Shimpi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAPD	=	American Academy of Periodontology
ACO	=	Accountable Care Organization
AHA	=	American Heart Association
ANN	=	Artificial Neural Network
ARFF	=	Unordered Instant Attribute Matrix
ATP	=	Adult Treatment Panel
AUC	=	Area Under ROC curve
B	=	Buccal surface of a tooth
B-ANN	=	Bagging employing Artificial Neural Network
B-DT	=	Bagging employing Decision Tree
B-LR	=	Bagging employing Logistic Regression
B-NB	=	Bagging employing Naïve Bayes
B-SVM	=	Bagging employing Support Vector Machine
BMI	=	Body Mass Index
BP	=	Blood Pressure
CDC	=	Centers for Disease Control and Prevention
CDST	=	Clinical Decision Support Tool
CEJ	=	Cemento-Enamel Junction
CFS	=	Correlation Based Feature Selection
CGL	=	Computational Geometric Learning
CI	=	Confidence Interval
DB	=	Distobuccal surface of a tooth
DD	=	Dental Data
DL	=	Distolingual surface of a tooth
DP	=	Dental Practice
DRS	=	DentoRisk
DT	=	Decision Tree
DW	=	Data warehouse
EHR	=	Electronic Health Record
FN	=	False Negative
FP	=	False Positive
FS	=	Feature Selection
FQHC	=	Federally Qualified Health Centers
HDL	=	High Density Lipid
HIDEP	=	Health Improvement in Dental Practice Model
HTAN	=	Hyperbolic Tangent Function

IE	=	Interdisciplinary Environment
iEHR	=	Integrated Medical-Dental Records
IG	=	Information Gain
IRB	=	Institutional Review Board
L	=	Lingual surface of a tooth
LD	=	Laboratory Data
LDL	=	Low Density Lipid
LD-Bgl	=	Blood glucose levels
LD-Lp	=	Lipid Panels
LF	=	Laboratory Findings
LMA	=	Levenberg Marquadt Algorithm
LR	=	Logistic Regression
MB	=	Mesiobuccal surface of a tooth
MCC	=	Matthew's Coefficient Correlation
MD	=	Medical data
Ml	=	Mesiolingual surface of a tooth
ML	=	Machine Learning
MOC	=	Model of Care
MP	=	Medical Practice
NB	=	Naïve Bayes
NHANES	=	National Health and Nutritional Examination Survey
NIDCR	=	National Institute of Dental and Craniofacial Research
NIH	=	National Institute of Health
OHIS	=	Oral Health Information Suite
OHS	=	Oral Health Status
PAT	=	Periodontal Assessment Tool
PCP	=	Primary Care Provider
PD	=	Periodontitis
POC	=	Point-of-care
PPD	=	Periodontal Probing Depth
PR	=	Patient reported data
PR-D	=	Patient reported- demographics
PR-S	=	Patient reported-social history
PR-M	=	Patient reported-Medicaid/Medicare status
PR-O	=	Patient reported- oral hygiene practices
PR-Dm	=	Patient reported- Diabetes history
PRA	=	Periodontal Risk Assessment
PRC	=	Periodontal Risk Calculator
RABIT	=	Risk Assessment Based Individualized Treatment
RBF	=	Radial Basic Function

RF	=	Random Forests
ROC	=	Receiver Operating Characteristic
RP	=	Recall-Precision
SCG	=	Scaled Gradient Conjugate
S.D.	=	Systemic Diseases
SD	=	Standard Deviation
SVM	=	Support Vector Machine
T2DM	=	Type II Diabetes Mellitus
TC	=	Total Cholesterol
TG	=	Triglyceride
TN	=	True Negative
TP	=	True Positive
U.S.	=	United States
UniFe	=	University of Ferrara model
WEKA	=	Waikato Environment for Knowledge Analysis
WHO	=	World Health Organization

ACKNOWLEGEMENTS

First and foremost I would like to express my special appreciation and thanks to my advisor Dr. Susan McRoy, who has been a tremendous mentor for me. I appreciate all her contributions of time, guidance and support to make my PhD experience productive and stimulating. I would like to gratefully acknowledge the guidance and encouragement of my doctoral committee members: Dr. Huimin Zhao, Dr. Min Wu and Dr. Amit Acharya. To Dr. Huimin Zhao, thank you for creating a deep interest in data mining, for being a great mentor and for providing guidance and constant feedback. To Dr. Min Wu, thank you for providing me with valuable feedback and insightful discussions about the research. To Dr. Amit Acharya, thank you for being a great inspiration and mentor and for constant support and encouragement. The joy and enthusiasm of all my committee members for their research was contagious and a great motivation for me.

I would like to extend my gratitude to Marshfield Clinic Research Institute and University of Wisconsin-Milwaukee. My sincere thanks to Dr. Ingrid Glurich, Rajesh Koralkar, Harshad Hegde**,** Dixie Schroeder and all my co-workers for providing constant encouragement during the entire journey of my PhD. I would like to thank Shane Haensgen from doctoral services for helping me with the formatting and Debra Abanathy for assisting with the administrative process.

I am extremely thankful to my mom Ashwini Shimpi, dad Anil Shimpi, my uncle Dr. Rajendra Shimpi, sister Dr. Ashlesha Shimpi and all my family members for their continued love and support. I would like to thank all my friends, teachers, professors and all those who have influenced my life.

CHAPTER 1

INTRODUCTION, STATEMENT OF PROBLEM AND SIGIFICANCE OF THE RESEARCH

1.1 INTRODUCTION

Periodontitis (PD), like many other chronic diseases, has subtle symptomology that becomes apparent after much damage has been done to the underlying bone [1]. Due to its chronic nature, the disease progresses continuously without causing any severe discomfort in the oral cavity [1]. If left untreated or diagnosed at an advanced stage; this chronic inflammatory process may lead to severe PD, causing irreversible damage to the periodontium (including gums, supporting bone, periodontal ligament and cementum) and eventually tooth loss [2]. Notably PD also represents a modifiable risk factor that may reduce the onset and progression of some systemic diseases (S.D.), including diabetes [2]. A better appreciation of the systemic effects and well-known periodontal risk factors along with behavioral factors has shifted the focus positing that collectively, the sum of risk contributed by a combination of individual factors provides better predictive power than with any single risk factor.

PD is a major public health concern which profoundly affects oral health and concomitantly, general health of the population worldwide [3]. In examining the prevalence of oral diseases in the United States (U.S.), the 2015 report from the Centers for Disease Control and Prevention (CDC) proclaimed that about 44.7% of the U.S. population more than 30 years of age and 66 % of the population more than 65 years of age has some form of periodontitis [4]. Similarly, incidence of PD in patients with existing systemic disease such as Type 2 diabetes has

been shown to exceed prevalence in the general population [5]. Severe periodontitis prevalence is estimated to impact 5-20% of most adult populations worldwide [6]. The cost of treating PD ranges from $500 to $10,000 depending on the severity of disease [7]. Notably, a study reported that a periodontal intervention in individuals who were recently diagnosed with Type 2 diabetes reduced the total healthcare cost by $ 1,799 over two years [8].

Historically, dental practice has been confined to delivery of oral and maxillofacial care. However, with increasing scientific evidence supporting oral and systemic disease associations, a new era has begun that casts dental and medical providers as proactive participants in establishing care for patients with chronic diseases including PD and diabetes [9]. This paradigm shift in delivering holistic, patient-centered care or whole-person care has necessitated development of interprofessional collaboration among dental and medical providers within an interdisciplinary environment (IE). One such development is the integrated medical-dental electronic health record (iEHR) that facilitates improved care coordination and information sharing amongst the providers [10][11] This information generated in the IE also presents with opportunities to explore the data. Secondary use of the data stored in the electronic health record (EHR) has emerged as a powerful approach to stratify patients for risk of diseases or comorbidities [12].

Using a risk based approach the healthcare providers can assess the patient's current disease and risk of developing future disease [13]. In support of this concept, the Veterans Affairs Dental Longitudinal Study examined clinical records and radiographs of 523 subjects to evaluate the validity of risk prediction using a computer-based tool and concluded that risk scores correlated strongly with periodontal status [14]. The study also posited that the use of a risk assessment tool over time may be beneficial in terms of achieving uniformity, accuracy,

informed clinical decision making, improved oral health outcomes and reduction in need for complex therapy.

A critical step in periodontal health management is development of a logical and properly-sequenced protocol consistent with the existing comorbidities [1]. However, simultaneously characterizing relative potential for PD severity that are more congruent with systemic diseases will play a crucial role in the long term management of PD as well as S.D., especially since treatment protocols vary with PD severity, type, and existing systemic diseases such as diabetes. Due to the substantial potential impact on quality of life and overall health, systematic assessment of risk for PD should form a standard component of periodontal assessment, but currently remains a gap in clinical care. This study proposes to examine plausibility of constructing predictive models in various models of care settings (MOC) for projecting chronic disease risk by extracting relevant information from the routinely collected clinical encounter data within the EHR.

1.2 STATEMENT OF PROBLEM

Interdisciplinary efforts for assessing PD risk even in an integrated medical and dental care delivery environment remains a gap. The persistent challenge in health care is failure to detect the risk of periodontitis at a stage when implementation of proven interventions could delay and even prevent PD and systemic disease progression. Prior knowledge of the medical and dental factors that predict the complexity of PD risk will allow clinicians to better prevent the periodontium from destruction thereby provide better management of the local as well as systemic inflammation.

Extracting relevant information from the medical and dental records of the patients to determine the PD risk in an IE would be time-consuming and computationally intensive for health care providers due to the large data generated at point-of-care (POC). A reason attributing to this could be because of the lack of oral health education among medical students. The All Schools Summary Report of 2012 that aggregated data from graduating students from 126 U.S. medical schools reported that only about 3% and 16% of medical students were 'excellently trained' and 'well trained', respectively, to address oral/dental health topics in their health-related school [15]. Due to this situational information overload, the healthcare providers may overlook risk factors, misinterpret the synergistic effects of the etiological factors and may not assess the risk of PD. The inconsistencies and shortcomings of these practices support the need for constructing and deploying a predictive model at POC, which is time-efficient, easy to use, accurate and will facilitate the provider in identifying the PD risk in an IE.

The aim of this study is to develop a robust, valid and practical means of assessing PD risk that can be applied in healthcare settings that use EHR. To construct such a predictive model,

identification of risk factors and methodology for identifying and extracting relevant information from the clinically/demographically captured data is essential. Although there is a benefit of using this clinical knowledge to assess the risk of periodontitis due to its established evidence between the risk factors and the disease progression, it is not known whether the resulting model will adequately represent the complexity of periodontitis processes that underlie the insidious pathogenesis. Not all risk factors for periodontitis may have been identified, and predictive features may remain to be identified, tested and modeled. Sorting and testing large numbers of features with potential association and building a model is possible by using the big data available in the large healthcare system's enterprise data warehouse. This study proposes use of a machine learning approach that will examine evidence-based candidate risk factors previously identified in existing models and vet new novel risk factors identified through machine learning approaches in order to achieve a more enhanced and comprehensive model that will optimize PD prediction.

1.3 SIGNIFICANCE OF RESEARCH

Practice patterns for oral and medical healthcare delivery, their respective individual reimbursement systems and the current state of dental and medical academic practice reflects sustained siloing of medical and dental healthcare delivery models PD is currently managed using a reparative model, wherein, the dental providers focus on treatment of obvious PD conditions that requires immediate intervention with less focus on prevention of future disease [8]. This reparative approach disregards individual variation in susceptibility (due to oral hygiene habits, tobacco use among others) and risk for disease (caused due to presence of co-morbid conditions such as diabetes) resulting in delivery of optimal treatment only to patients who require immediate treatment [16]. Establishing interdisciplinary care for improving healthcare practice and expanding access to preventive oral health care through primary care providers (PCPs) has been proposed by the National Academy of Medicine and others [17]. Notably, patients visit their PCPs with higher periodicity compared to frequency of visits to dental providers [18]. Consequently, there is a need for a paradigm shift to a transformational approach encompassing assessment of patient's risk for developing future disease in dental as well as medical settings. Disparities in access to dental care for low socio-economic populations persist, which also supports the need for broader consideration of PD risk. Due to lack of oral screening in medical settings and access barriers, risk of developing PD is often overlooked in this population. Availability of clinical decision support tools (CDST) within the EHR with capacity to use available patient data to access periodontal risk will support evaluation of these patients and other individuals who are at a high risk of developing PD due to underlying comorbid conditions. This new paradigm will not only allow employing preventive methods but also make sure that the individuals receive the most appropriate treatment, thus transitioning the

reparative model to a wellness model. In addition to this, it is presumed that the predictive model will influence multiple levels of care and education including personal level, organizational level and at community level [19].

Chapter 2

OBJECTIVES AND RESEARCH QUESTIONS

2.1 OBJECTIVES

The main objective of this study is to develop a robust, valid and practical means of assessing PD risk that can be applied in healthcare settings that use EHRs.

For an effective realization of the main objective of this study, the following sub-objectives are established:

1. To identify the various existing risk factors that have previously been proposed in evidence-based literature.

2. To develop a comprehensive list of relevant data variables captured in medical and dental settings in the EHR that may be vetted in a more comprehensive model.

3. To develop machine learning (ML) algorithms with a data-driven approach such that after being trained and tested on existing knowledge, it will predict the future behavior of the existing periodontal situation with an acceptable accuracy.

4. To evaluate and compare the resultant predictive models for their ability to predict the risk of developing PD.

2.2 RESEARCH QUESTIONS

2.2.1 RESEARCH QUESTION 1

What are the significant clinically captured medical and dental factors that contribute to risk of developing periodontitis?

The study will explore medical factors such as duration of diabetes, oral hygiene techniques and laboratory values along with other risk factors which have not been previously vetted in a single multivariate model. Feature selection/engineering methods will be used to identify and remove irrelevant and redundant attributes from data that do not contribute to the accuracy of the predictive models.

2.2.2 RESEARCH QUESTION 2

What is the relationship between the type of model (classification approach) and values of different measures of performance, including recall, precision, etc.?

The study will utilize retrospective data and compare the relative accuracy among different models created through machine learning approaches. The models will be validated and performance metrics such as sensitivity, specificity, precision, recall among others and statistical validations will be conducted and will assess the tradeoffs among the best candidate models.

2.2.3 RESEARCH QUESTION 3

What is the relationship between the different methods of combining the predicted classifications of different models and the values of different measures of performance?

The study will create various prototypical models through assessment of relative risk contributed by clinical and demographic or other relevant variables available in the EHR and model variables. This study will compare the majority voting and a more optimistic strategy, such as “at least two”?

2.2.4 RESEARCH QUESTION 4

What is the relationship between the subset of data variables used and the values of different measures of performance?

This study will compare the subsets such as: only medical, only dental, both medical and dental, medical with extra patient reported dental, medical without any patient reported data.

Chapter 3

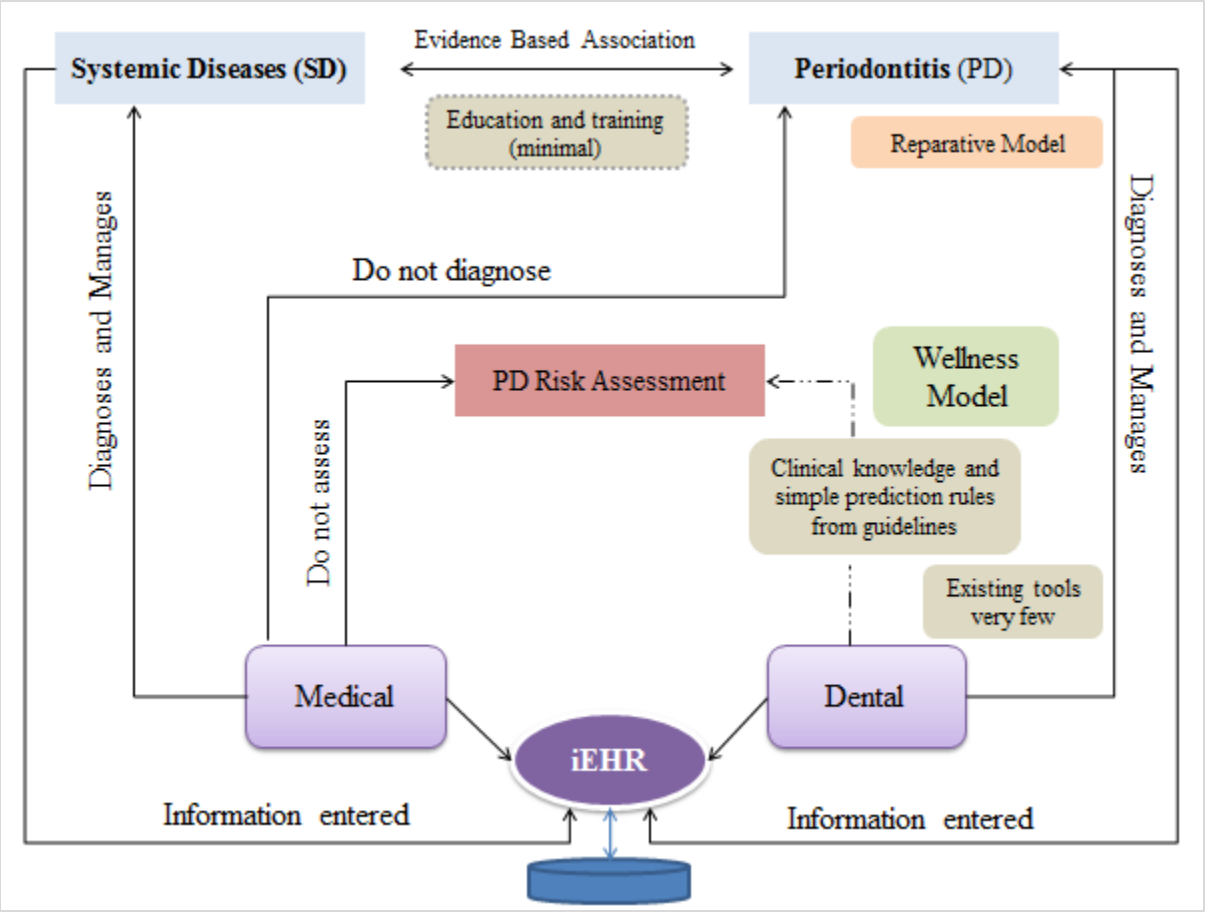
CURRENT STATE OF ART

3.1 CURRENT WORKFLOW OF AN INTERDISCIPLINARY ENVIRONMENT AND PROPOSED WORKFLOW

Although significance of using periodontal risk assessment tools in clinical practice has been placed, studies have also shown that like many other dental diseases, periodontal disease are still managed using a reparative model, wherein, the dental providers focus on clinically obvious conditions that requires immediate intervention and giving less attention on preventing future disease [8][20]. Practice patterns in medical and dental settings often exhibit organizational silos. **Figure 1** shows a generalized model created based on the literature to portray the interdisciplinary framework and workflow for determining periodontal risk in an IE. The right side of the figure shows the dental section of the organization and the left shows the medical section. Dental providers and medical providers are critical members of a collaborative team approach. It is assumed that health care providers from various specialties are working synergistically in an interdisciplinary environment, across dental-medical domain; however, the respective professions still tend to work independently, without little interaction [21]. Currently, the dental provider and medical provider communicate with each other through shared messaging, exchanged notes, EHR, phone, fax and email among others. The medical provider diagnoses S.D., while the dental providers diagnose PD. PD risk assessment is not performed by the medical providers while very few tools are available for dental providers to assess the PD risk. Evidence based studies show PD-S.D. relationships that would require sharing knowledge

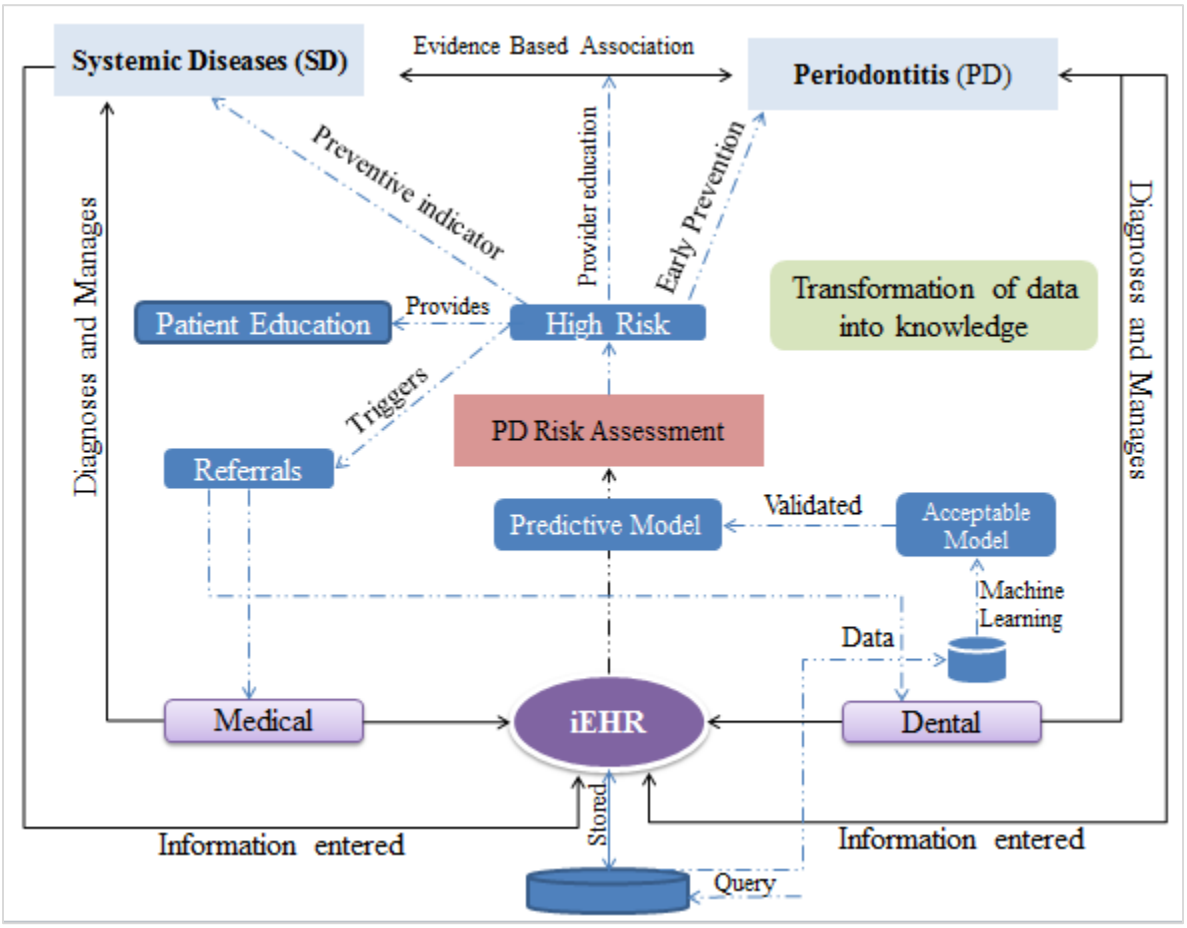
among the medical and dental providers, however, this sharing may be difficult as studies also show lack of interdisciplinary education and training [21].

Figure 1: An Interdisciplinary framework and current workflow of determining Periodontitis (PD) risk in an Interdisciplinary Environment (IE).



The dotted line shows a possible use of few periodontal tools that may be available for the dental providers which aligns to a wellness model. Although, a medical provider and dental provider have access to the iEHR, there is no systematic attempt to communicate with each other about the existence of systemic disease and PD risk, unless the medical provider checks the periodontal status by accessing the dental records of the patient and vice versa.

Figure 2 represents the proposed interdisciplinary framework for PD risk predictive model in an IE



The framework in **Figure 2** shows the predictive model as a key component. Similar to figure 1, the right side shows the dental setting and left side shows the medical setting. The dotted lines show the proposed framework. A database query will provide information which will be then utilized to build model by using machine learning algorithms. The accepted resultant model after validation will be considered as the potential predictive model at POC. A high risk result through the predictive model will support 5 components : a). Trigger an interdisciplinary reference, b) provide patient education; c) act as a preventive indicator for modification of management of S.D., if required; d) will act as a preventive indicator in terms of

dental treatment and management and e) enhance education and training amongst provides. The framework will support effective communication amongst the medical and dental providers.

The next section reviews the recent literature describing the risk and preventive factors for assessing periodontitis; various PD risk assessment tools, machine learning methods use for developing assessment tools; IE and iEHR, and current workflow and documentation of PD in typical EHR.

3.2 REVIEW OF LITERATURE

3.2.1. PATHOGENESIS OF PERIODONTITIS

The term ‘periodontitis’ originates from Greek words: “peri” meaning “around”, odous (GEN odontos) meaning “tooth” and the suffix itis meaning inflammation [22]. The underlying cause of periodontitis has historically been attributed to dental plaque accumulations, subsequent colonization and infection by periodontal pathogens and concomitant inflammatory processes. Historical case reports by Hippocrates in 467-300 B.C and Albucasis in 936-1013 A.D provided early illustration of the association between calculus and PD [23].

The American Academy of Periodontology (AAPD) submitted a report in 2014 on the various risk factors for periodontal diseases, based on prior studies [24]. This report gave an overview of the local and systemic risk factors that may contribute in developing PD. The local factors comprised of variables such as plaque and calculus, tooth occlusion etc., while systemic factors consisted of diabetes and osteoporosis among others [25]. The study identified tooth brushing and flossing as the critical factors in preventing PD. Correspondingly, removal of plaque and calculus by professional cleaning was also considered important to improve periodontal health [26] . Incorporating the patient self-report on impact of environmental factors including tobacco use, frequency of tooth brushing, presence of co-morbid conditions such as diabetes will not only create alertness amongst the care team which then provides opportunities for patient education and deciding on the informs best treatment modality for the patient thereby, increasing the quality of care.

Specific areas in the mouth where oral hygiene is impaired such as in areas of dental calculus deposition, poor margins of the tooth crowns among others helps in retention of dental

plaque in these areas. The accumulation of plaque in supra-gingival and sub-gingival area of tooth promotes growth of microbial organisms which are complex and mixed. A variety of microorganisms contribute to the pathogenesis of PD, however a large proportion of the microbial flora still remains uncharacterized [27]. Growing literature reveals that there is a positive association of gram negative and anaerobic bacteria of matured subgingival plaque and PD [28][29][30]. Some of the putative pathogens that commonly cohabits these subgingival sites may include *Porphyromonas gingivalis, Tannerella.forsythensis, Aggregatibacter.actinomycetemcomitans* and spirochete *Tannerella.denticola* [31][32]. The increase in microbial growth and food debris in the subgingival areas deepens the crevices between the gums and the tooth root resulting in formation of soft tissue pockets and periodontal tissue breakdown leading to formation of periodontal pockets [33].

Proactive bacterial removal from the teeth by good oral hygiene practices have shown to play a very important role in preventing PD. Studies have demonstrated that cleaning of teeth every 48 hours can maintain the gingival health [34]. Studies have also shown that suspension of oral hygiene behavior such as tooth brushing and flossing can predispose the gums to local gingivitis within 4 to 11 days and to generalized gingivitis within 2-3 weeks [35][36].

Inflammation consequential to microbial infection can destruct the periodontal ligament and the surrounding supporting alveolar bone, while simultaneously triggering the inflammatory processes in the body. In the future, information about the patient's microbiome may also be shown to be a factor [37][38][39][40][41][42].

3.2.2 ORAL-SYSTEMIC ASSOCIATIONS

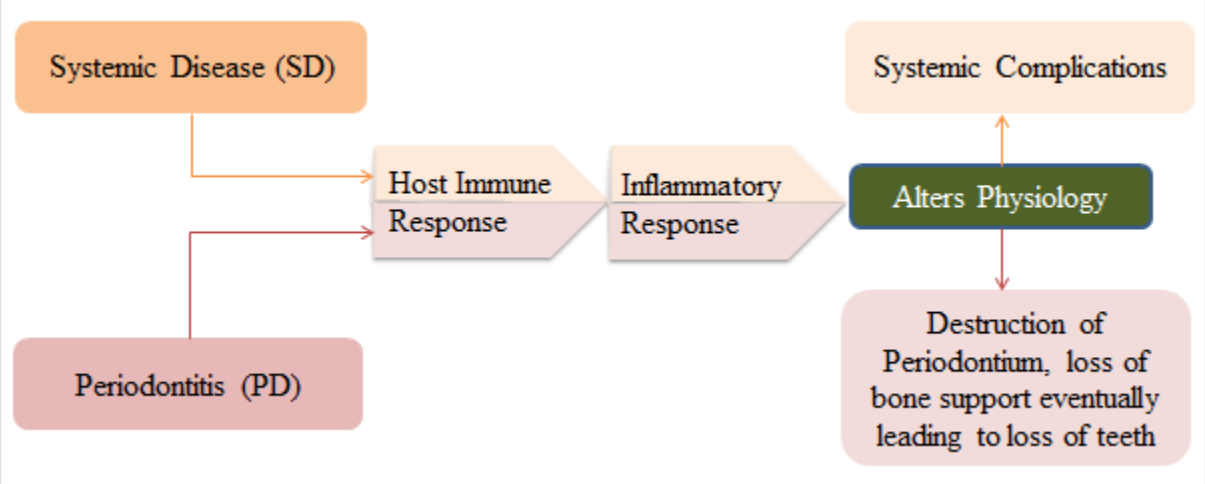
PD is a chronic disease, which is advancing globally across all geographic regions and contributing to the global burden of chronic diseases [43]. Along with the local (including plaque retention, calculus, tooth anatomy, tooth occlusion, fillings and restorations among others) and systemic factors (including diabetes mellitus, chronic smoking among others) contributing to the initiation and progression of periodontal disease, studies have pointed out that osteoporosis and psychological factors including stress are also associated with periodontal disease [43].

Notably, over the past few years, new findings have enhanced the understanding of pathogenesis and etiology of PD [23]. Longitudinal and cross-sectional studies support a biological link between diabetes and PD. Further, studies have demonstrated that poor glycemic control can contribute to poor periodontal health [44]. The World Health Organization (WHO) has estimated that by year 2030, the number of individuals with diabetes would reach at least 366 million [45]. The estimation indicates that 366 million patients will also be at increased risk for developing or exacerbating existing PD by 2030.

Recent studies have demonstrated that a variety of inflammatory markers (such as Interleukin 1, C-reactive protein) associated with systemic diseases (such as diabetes, cardiovascular disease, among others,) are correlated to increasing severity of PD [20-21]. Similarly it has been reported that the incidence of PD in patients with existing systemic disease exceeds estimated prevalence for the general population [46].

The bulk of evidence points to de-regulation of host-immune response leading to chronic inflammation and thus progressing the disease. **Figure 3** shows the pathway of pathogenesis in PD and S.D.

Figure 3: Pathway of pathogenesis in PD and Systemic Diseases (S.D.)



A MEDLINE search conducted on published articles for relationships between diabetes and PD since 2000 and effects of periodontal infection on glycemic control and diabetes complications since 1960 supported diabetes having an adverse effect on periodontal health and infection [13]. The summary of evidence in [13] shows that the diabetes-related variables considered in different studies included glycemic control, duration of diabetes, severity of diabetes based on presence of complications and fasting blood glucose levels. The PD status measures that were considered were gingival bleeding, pocket depth, loss of periodontal attachment, radiographic bone loss, juvenile periodontal score, modified gingival index, Russell's periodontal index, periodontal disease rate (proportion of teeth affected by periodontal disease).

In 1998, Garcia et al [14] explored the relationship between PD and common SD such as diabetes, respiratory diseases, cardiovascular disease and osteoporosis. It was recognized that community-acquired pneumonia and lung abscesses can be the result of anaerobic bacterial infection and dental plaque might be a source of these bacteria. The study also identified that cigarette smoking, genetic predisposition and other environmental factors such as second-hand

smoke to be risk factors for development of respiratory diseases. The study concluded that there is a weak connection between respiratory and cardiovascular disease and PD. The study also showed that there is evidence of connection between periodontal disease and osteoporosis, however large scale studies need to be conducted to better understand the relationship between the two diseases.

Hypertensive management in dental office is very important for improved monitoring and dental treatment. Prevailing evidence suggests that diastolic hypertension is usually seen in patients before the age of 50 years; while patients more than 50 years of age predominantly suffer from systolic hypertension [47]. Evidence suggests association between high diastolic pressure and deep periodontal pockets. A previous study conducted among approximately 1200 patients revealed that patients with diastolic blood pressure >90mm Hg had deep periodontal pockets and exhibited some form of periodontitis [48].

Obesity is characterized by the abnormal or excessive deposition of fat in the adipose tissue [49]. Consequences of obesity often lead to negative effects on health including periodontal health. Although the role of obesity in periodontal inflammation is yet not defined, studies have considered obesity as one of the multifactorial effect increasing PD risk [24].

3.2.3 ACCESS TO DENTAL/ORAL CARE

A significant barrier to accessing preventive oral care is access to dental insurance [50]. State and federal leaders have been unable to provide lower-income individuals with affordable dental coverage [51][52]. Individuals who qualify for public insurance in the U.S. experience access barriers in identifying dentists willing to provide care. More than half of the dental specialists in the U.S. refuse to treat patients enrolled in Medicaid due to insufficient reimbursement rates [53]. North Dakota, Wyoming, South Dakota and Wisconsin were among the states experiencing the largest increases in dental service inflation from 2003 to 2013 [54]. The objectives enumerated in the 2010 Surgeon General report strongly support increased dental screening and better access to dental care [55]. The Federally Qualified Health Centers (FQHC), which were added to the Medicare benefit in October of 1991 advances the access to care by providing services through community health centers, migrant health centers, homeless health centers and public housing to the medically underserve populations/areas, migrant agricultural workers, homeless individuals and families among others [56]. The FQHC are also required to provide dental screenings, according to the federal statue, to determine the need and delivery of dental care [50].

The introduction of health care reform in the U.S. supports increased innovation in the care delivery models, including better integration of providers through an Accountable Care Organization (ACO) [57]. However, it has been reported that dental care often excluded from ACO's largely because of a lack of integrated health information technology [57]. Other reasons that ACOs are not including dental care are mentioned in the 2016 report of the Health Policy Institute. One factor mentioned is the lack of healthcare provider's understanding the link between oral and systemic health outcomes. The report also gives an example of the benefits of

health integration where the researchers found that there were improved health outcomes with little or no increased cost when an integrated treatment was given to patients that included the primary care, mental health, specialty health and behavioral health.

3.2.4 PERIODONTAL POCKET DEPTH (PPD)

Periodontal pocket depth (PPD) is considered to be an essential part of comprehensive periodontal examination [58]. Probing at six surfaces of each tooth also known as the six point probing is a standard of care in dental practice. Each tooth is divided into 6 surfaces for measurement purpose (Mesiolingual (ML), Distolingual (DL), Mesiobuccal (MB), Distobuccal (DB), Lingual (L) and Buccal (B)) as shown in **Appendix B**. Measurements are recorded for all the teeth surfaces in a periodontal chart in an EHR most often during a comprehensive dental visit. The probing depth measurement provides the dental providers with the information of absence or presence of denuded root surfaces and periodontal pockets. For example, a probing depth of more than 4mm from cemento-enamel junction (CEJ) is associated with a deeper pocket.

3.2.5. CURRENT STATE OF ART-RISK ASSESSMENT TOOLS FOR PERIODONTAL CONDITIONS

The WHO defines risk "as any attribute, characteristic or exposure of an individual that increases the likelihood of developing a disease or injury" [59]. Consistent with these definitions and importance of risk in periodontal care, the AAPD has stated that "the clinical use of risk assessment will become a component of all comprehensive dental and periodontal evaluations as well as part of all periodic dental and periodontal examination"[60]. In 2008, the AAPD defined risk assessment as "the process by which qualitative or quantitative assessments are made of the likelihood for adverse events to occur as a result of exposure to specified health hazards or by absence of beneficial influences" [60].

Although the most recognized sign of PD such as gingival enlargement and inflammation, bleeding on probing and loss of attachment of the alveolar bone indicate the ongoing PD activity, these signs do not predict future disease activity and are simply cumulative measures of the past disease activity [61]. In spite of this limitation, many clinicians use the extent and severity of PD assesses the risk of developing PD. That is, patients are assumed to be at low risk when there is very little or no destruction of periodontium while patients who are showing the signs and symptoms of PD destruction are assumed to be at a high risk of developing periodontitis.

The AAPD defines risk assessment as "the process by which qualitative or quantitative assessments are made of the likelihood for adverse events to occur as a result of exposure to specified health hazards or by the absence of beneficial influences" [24]. Due to the multifactorial etiology of PD, AAPD guidelines recommend assessing the PD risk at patient level and at clinical level.

Although more than 50 studies have been published in the past 20 years, little has been published about risk assessment tools related to periodontal disease. A search conducted in MEDLINE between 1996 to 2016 for published articles on humans, English-language, and "(Periodontal Diseases"[Mesh]) AND "Risk Assessment"[Mesh] AND tool*) OR (("Risk Assessment/methods"[MeSH Terms] AND periodontal diseases) AND ("Periodontal Diseases"[Mesh]) AND predict*)) OR (("Probability"[Mesh] AND) AND (("Periodontal Diseases"[Mesh]) AND "Algorithms"[Mesh]) yielded 41 citations of which 12 described current literature on computer based PD risk assessment. **Table 1** Summarizes the research papers published around various periodontal risk assessment tools along with the data variables used, leveraging previously-summarized risk assessment models for PD.

Table 1: Summary of the research papers published around various periodontal risk assessment tools along with the data variables used, leveraging previously summarized risk assessment models for PD

Year Published	Ref	Author	Name of the risk model, if any	Number of variables used	Sample size and population	Approach
2002	[62]	Page et al	PRC (Periodontal Risk Calculator)/ PreVisor	9	523	Mathematically driven algorithm
2003	[63]	Lang et al	PRA (Periodontal Risk Assessment)	6	Parameters based on evidence-based	Functional diagram based on retrospective model
2005	[64]	Page et al	PAT (Periodontal assessment tool) of OHIS (Oral Health Information Suite)	23	523	Mathematically driven algorithm
2007	[65]	Sanderberg et al	HIDEP (Health Improvement in Dental Practice Model	17	750	Computerized tool with predefined risk groups. Used to determine Dental caries risk as well as periodontal risk
2007	[66]	Chandra	Modified PRA, retrospective model	8	26	Functional diagram based on retrospective model
2008	[67]	Eickholz et al	Modified PRA	10	100	Poisson's regression
2008	[68]	Jansson et al	used PRA	6	20	Hexagon diagram proposed by
2009	[69]	Trombelli et al	UniFe (University of Ferrara)	5	107	Predefined parameter scores
2010	[70]	Leininger et al	Modified PRA	6	30	Functional diagram based on retrospective model
2010	[71]	Lindskog et al	DRS(DentoRisk)	6	183	Linear regression, multivariate linear regression
2010	[72]	Shankarapillai et al	None	16	230	Artificial Neural Network (Comparison of LMA and SCG)
2011	[73]	Costa et al	used PRA	6	164	Functional diagram based on retrospective model
2013	[74]	Teich	RABIT (Risk Assessment-Based Individualized Treatment)	N/A	N/A	Conceptual- practice management for recall of periodontal treatment
2013	[75]	Lu et al	Modified MPRA	7	158	Functional diagram based on retrospective model
2014	[76]	Busby et al	OHS (Oral Health Status) incorporates PreVisor	23	25	Online tool based on PreVisor OHIS suite

3.2.6 RISK ASSESSMENT USING MACHINE LEARNING APPROACHES

The prospects of secondary use of clinical data has increased considerably with the increase in adoption of EHRs [77][78][79][80]. The secondary use of data provides with a huge amount of data that presents incredibly new opportunities to make contributions to the healthcare field such as clinical decision support tools, reminder-alert systems among others [81]. Perhaps one of the remarkable breakthroughs in healthcare applications is the use of machine learning (ML) [82]. ML performs its tasks in two different types of spaces that includes spaces 'X' and $'\vartheta'$ consisting of instances and ML models, respectively. Depending on the training set, $\{x^{(\gamma)}\}_{\gamma=1}^{n}$ C X , and outcome variable Y, supervised machine learning fits a pre-defined function to given training set $\{x^{(\gamma)}, y^{(\gamma)}\}_{\gamma=1}^{n}$ C $Xx\ Y$ and tries to find the function y = f(x, θ) where $\theta \in \vartheta$.

There is a mounting evidence exhibiting use of machine learning algorithms in medical domain, however there are very few studies that have used machine learning approaches in dental domain [83][84]. Presently, two of the closest periodontal disease studies to the current study are perhaps a pilot study that assessed periodontitis risk by comparing the Levenberg Marquadt algorithm (LMA) and the Scaled Conjugate Gradient (SCG) algorithm [72], secondly is the diagnosis of periodontal diseases using different classification algorithms [38].

The first study is described in the Artificial Neural Network (ANN) section. The second study compared the performance of three algorithms Decision Tree (DT), Support Vector Machine (SVM) and ANN) by utilizing 100 training and 50 test sets of 150 patients [38]. A total number of 11 dental variables were used including two dental indices to measure the health of the periodontium. These variables, however, did not incorporate any systemic diseases. The outcome measures used the AAPD classification of periodontal diseases [85]. The results of the study show that DT and SVM outperformed ANN markedly [38]. The study highlighted using

DT due to its ability of support for easy interpretation and conversion of complex processes into simple decision making.

A systematic search was undertaken to characterize the utilization of various algorithms used for determining the risk of patients for developing prediabetes [86]. The results of the review conducted by seeking relevant literature in PubMed since 1946, EMBASE since 1974 and Grey Literature showed that 18 tools met the criteria [86]. Of these 18 tools to detect risk for prediabetes, 11 tools used logistic regression, 6 used decision tree and one used SVM. The tools were validated as follows: 7 by external dataset, 14 by bootstrapping and cross validation on the internal dataset and one using a partially independent dataset. It was also noted that none of the tools used multiple imputation which are believed to provide better results and discriminating capacity compared with simpler tools [87].

Some of the most common algorithms used in medical field for predictive modelling are:

3.2.6.1. BAYES THEOREM (E.G. NAÏVE BAYES)

This is a simple and intuitive method for conditional probability and useful for very large data sets [88]. In this study, the hypothesis for a patient having high or low risk of PD (H) given evidence E, which is combination of data variables (E1, E2, E3….En), the posterior probability is calculated by

$$\Pr[H|E] = \frac{\Pr[E1|H] \times \Pr[E2|H] \times \ldots\ldots\ldots \Pr[E190|H] \times \Pr[H]}{\Pr[E]} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where,

Pr [H|E] is the posterior probability of class (high risk, low risk) given evidence (n data variables)

Pr [H] is the prior probability of class

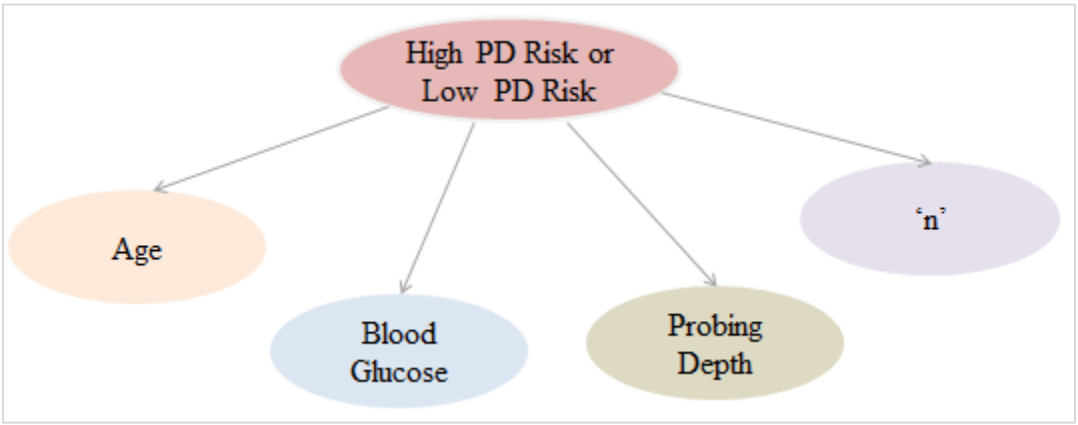
Pr [E|H] is the likelihood which is the probability of evidence given class

Pr [E] is the prior probability of evidence

Naïve Bayes methods apply Bayes theorem with a "naïve" assumption of independence between every pair of data variable [88][89].

The structure of NB classifier as a directed acyclic graph is shown in **Figure 4** which consists of one parent node (class: High PD risk or a Low PD risk) and several child nodes that represent the data variable nodes (such as age, blood glucose level, probing depths and 'n' (other study data variables).

Figure 4: Structure of Naïve Bayes



Due to the large and growing amount of data within the EHR, a lot of hidden information and patterns from the data need to be mined. A usual approach in predicting a disease risk is calculating probabilities using Bayes theorem [90]. Some recent research on data mining utilized medical records to determine the risk of developing cardiovascular disease. One such study utilized naïve Bayes to determine the cardiovascular disease risk by using 22 data variables [91]. The study results show that the sensitivity was 84.3% and specificity was 86.19% while the overall accuracy was 85.9% for determining the risk. The systematic review conducted on application of NB in predicting diseases (including brain disease, breast cancer, prostate cancer,

glaucoma severity, toothache disease, systemic sclerosis among others) reported that 23 studies including a total of 53,725 patients showed NB had a best performance in predicting diseases [92]. Moreover, the systematic review also reported that 80% of the studies utilizing the NB algorithm reported an accuracy of more than 75% and an Area under the curve (AUC) higher than 80% (for 6 out of the 11 articles). Different studies that compared naïve Bayes with other algorithms have shown that the accuracy for naïve Bayes was better than any of the other algorithms [86-88]. Notably, this algorithm is successful in practice even though feature independence is generally considered as an inaccurate assumption [93]. Studies have shown that use of Bayes' theorem has simplified the use of diagnostic information and facilitates graphical, intuitive and information updating [94].

A study [95] compared Bayesian regularization with LMA on a social data to comprehend their predictability. The study results showed that Bayesian methods outperformed the LMA and obtained highest correlation coefficient between the real and predicted data sets. The studies also emphasized on using Bayesian technique in predicting situations as they build on robustness of model and optimize the network architecture [96]. Studies have shown that Bayesian approach can solve the over fitting problem easily and is one of the algorithm with highest predictive power [97][90][98].

3.2.6.2.DECISION TREES (DT)

Another category of models that seem appealing to solve the PD risk problem are the decision trees which are known as "divide and conquer" algorithms [88]. Based on the literature, notably, this algorithm has the main advantage of interpretability and being easy to understand

due to its capacity to display a range of possible outcomes along with consequent decisions [88]. The most well-known algorithm for building the decision trees is the J48.5.

The decision tree classifies the instances and the data variables selected is placed at root node to construct a branch. This process is repeated recursively using instances that reach the branch. For deciding which attribute to split on, information gain is considered to be effective and is defined as a measure of entropy. Entropy I of dataset D is calculated as follows [99]:

$$Entropy(H, L) \equiv \sum_{i=1}^{2} -p_i \log_2 p_i \quad \text{.....(2)}$$

where p_i represents the proportion of dataset in E that belong to class i. $'c'$ is the number of classes which is 2 in this study, 'H' is the number of cases and 'L' is the number of controls. The information gain of attribute A of E, $Gain(S, A)$, is defined as:

$$Gain(E, A) \equiv Entropy(E) - \sum_{v \in Values(A)} \frac{|E_v|}{|E|} Entropy(E_v) \quad \text{.....(3)}$$

where $Values(A)$ represents all the values that attribute A can take.

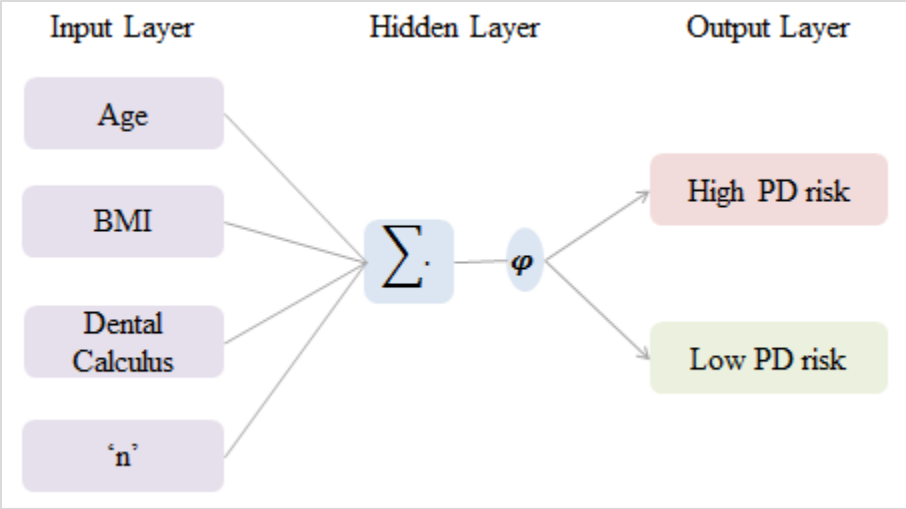
Studies have shown that utilizing DT yields clinically applicable decision rules that can be easily applied in healthcare sector to better manage disease risks [100]. Moreover, these can also be adapted to the real world applications. A recent study employed DT to a publicly available, Canadian Inpatient dataset and demonstrated the benefits of using DT to manage hospital readmission risks [101]. The study arbitrarily limited the tree to a depth of six levels and the model performance was compared using area under the receiver operating characteristic curve and lift curves. Although the area under the receiver operating curve (ROC) curve performance was low (0.612), the DT algorithm showed a substantial interpretability and adaptability [97].

3.2.6.3.NEURAL NETWORKS

Based on the literature, significantly, neural networks are considered to be nonlinear and flexible modeling techniques. Consequently, these also define the relationships between the attributes and can be used to improve accuracy [88]. Neural networks are used in inputs with high dimensional and discrete data and output with a vector valued or discrete data. Artificial neural networks (ANN) have been used in medical domains [102].

A study used [72] two algorithms, LMA and Scaled Conjugate Gradient (SCG) to classify patients with major PD and minor PD. The study used data of 230 patients and sixteen variables including history of diabetes and hypertension as medical variables and scored the PD risk on a scale of 1 to 5. The results of the study showed that LMA (a variant of backpropagation with an Artificial Neural Network) outperformed SCG. An additional factor of a type of chew tobacco showed an increase in sensitivity and accuracy of the model. In contrast, SCG was found to be more effective and faster than backpropagation of ANN and computational geometric learning (CGL) in [103]. The study was performed on patients between 15 to 60 years. Some of the factors such number of missing teeth, bleeding on probing and furcation involvement previously used in some of the periodontal risk assessment tools were not utilized in this study [62][63][64][65][66][67]. **Figure 5** shows the theoretical model of neural network for PD risk assessment tool

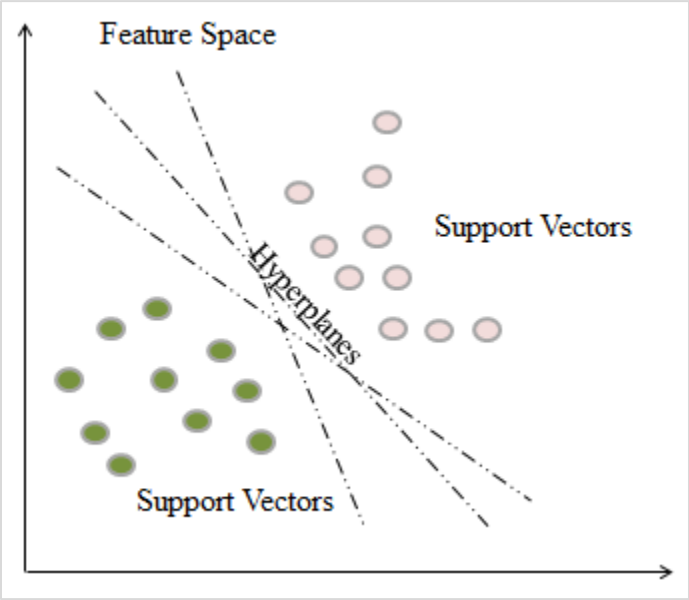
Figure 5: Theoretical model of neural network in PD risk assessment



3.6.2.4.SUPPORT VECTOR MACHINES (SVM)

These algorithms are a composite of linear modeling and instance-based learning [88]. According to [88], SVM algorithms build a linear discriminant function that includes extra nonlinear terms in the function, which improves performance for some problems. Notably, using "kernel trick" nonlinear relationships can be represented [104]. This algorithm works on finding an efficient way of separating hyper planes in n-dimensional space [105]. The algorithms are shown to be effective in many areas of medical domain such as for image recognition, follicle ovarian follicle detection among others [106][107] . Each data variable is plotted as a point that takes the value of a particular coordinate in n-dimensional space. The hyperplane is then defined that segregates the class into two, for example low PD risk and high PD risk. **Figure 6** shows an example of linear SVM separating the data variables into their respective classes.

Figure 6: Example of Linear Support Vector Machine



Models that are built on EHR data have been shown to be more amenable to the existing workflows. However, the data generated in the EHR can lead to bias in predictive modelling and poor performance due to missing data and imbalances of classes of interest [108]. A multilevel framework was proposed in a study to simultaneously classify large datasets and reduce the effects of missing data by using SVM [108].

3.6.2.5.LOGISTIC REGRESSION (LR)

The LR uses a sigmoid function which represents S-shaped curve with a value between 0 and1 and is given by

$$1/ (1 + e^{-value}) \quad \text{.....(4)}$$

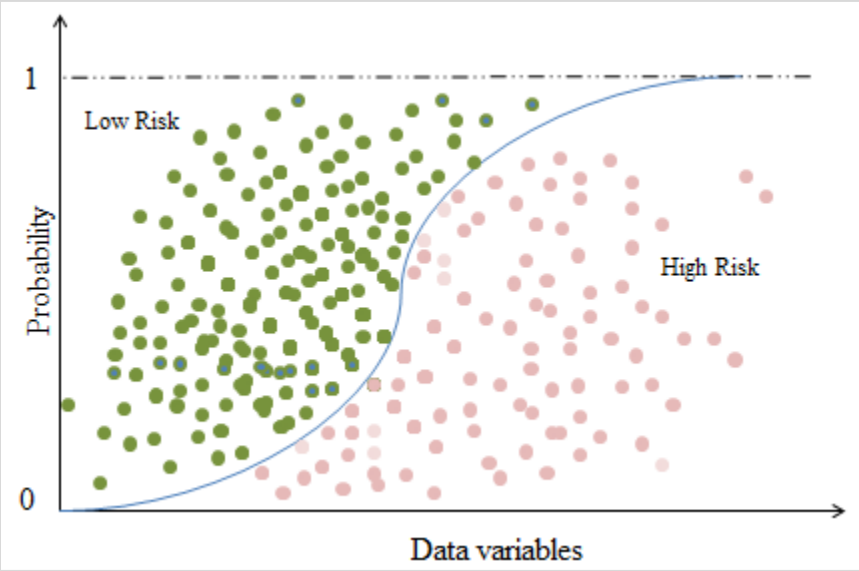
where e the base of natural logarithms. In logistic regression the input values for the variables in various MOCs are combined linearly either by weights or my coefficient values to predict the

low or high risk of PD (binary value). In case of PD risk the logistic regression equation will be given as

$$\text{PD risk (P)} = e^{(\beta 0 + \beta 1 * V)} / (1 + e^{(\beta 0 + \beta 1 * V)}) \quad \text{.....(5)}$$

where P is the predicted output, $\beta 0$ is the bias and $\beta 1$ is the coefficient for the single input V. **Figure 7** shows an example of the decision boundary of logistic regression. The decision boundary runs from 0 to 1.

Figure 7: Example of the decision boundary of Logistic regression

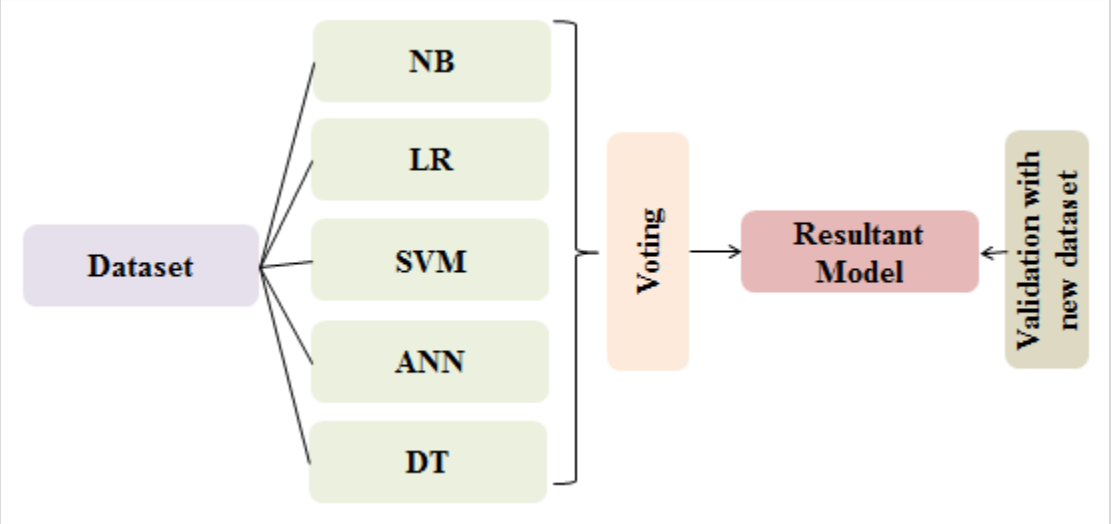


The summary of evidence of a case-control study conducted on 1433 patients built a predictive logistic regression model for assessing risk for chemotherapy-related hospitalizations displayed a sensitivity of 49% and a specificity of 85% with a 95% Confidence Interval (CI) between 41%–57% and 81%–89% respectively [109]. The study findings also showed a time-window bias that will show a high risk of chemotherapy-related hospitalizations for patients who are on longer duration of chemotherapy. In order to correct for this bias, expanding the maximum likelihood equation can be highly effective [110].

3.6.2.6.ENSEMBLE METHODS

Ensembles help in improving prediction accuracy by averaging multiple models [88]. This gives as advantage of decreasing the sampling variance of the final model.

Figure 8: Theoretical model for PD risk using Ensemble



A better performing ensemble method is known to have a significantly lower error compared to individual models because multiple models generated by random sampling of the supplied dataset are merged in bagging. A study utilized ensembles of ANN to diagnose lung cancer cells and results showed a reduction in the magnitude of error from 6.4% to 17.5% [111]

3.6.2.7.IMBALANCED DATA

Data imbalance commonly permeates biomedical informatics studies. Evidence shows that application of ML is inclined towards prediction of majority class in terms of imbalanced datasets [112][113]. Few attempts have been made by studies to show the class imbalance problems. One such study showed that balanced training datasets resulted in highest balanced accuracy, Matthew Correlation Coefficient (MCC) and area under ROC curves when

undersampling of the major class were done [114]. Performance metrics such as MCC are used in many studies utilizing predictive modelling [115][116]. A recent reported that MCC was the single best performance measure and encompasses all metrics of the confusion matrix [117].

In another study, the authors used weighted SVM for datasets with imbalanced classes and compared SVM-based algorithms. The study results showed that SVM-based algorithms produced fast, more accurate and robust classification results. Specifically, the study highlighted multilevel weighted SVM having a better performance that the regular SVM. Similarly, the use of repeated random sub-sampling was shown to be effective for imbalanced data [118]. Class imbalance occurs when one class is significantly more than the other class [118]. For instance, if PD high risk class is 1% and PD low risk class is 99%, presenting these to the classifiers can have undesirable results, as the classifier can get good accuracy by simply selecting the most frequent class. It was also noted in the study that the unbalanced-class for PD ranges between 0.01% and 29%. In such cases, the sensitivity will be 0% while the specificity will be over 90%. The authors of study, used a large dataset of 7,995, 048 records and performed two experiments including comparison between Random Forests and other classifiers with sub-sampling and the other one without sub-sampling [118]. The results of the study show that repeated sub-sampling with a RF ensemble learning method outperformed other classifiers such as SVM, bagging and boosting in terms of ROC and better resolved the class imbalance. The results of the study performed by [118]showed similar results with better performance of ensemble learning as compared to logistic regression (LR) and SVM [119]. Moreover, ensemble-based methods are more amenable as they combine multiple base classifiers to reduce the bias or variance or both [120]. Another study proposed the use of cost-sensitive algorithms to solve the problem of class imbalance [121]. The errors of over-prediction and under-prediction can incur different costs. In

such cases, uses of cost-sensitive algorithms have proven to decrease these errors [104]. The authors in a study proposed a method for minimizing the average incorrect prediction cost when there is asymmetric cost structure [104].

A risk assessment process involves estimating the probability of an adverse event due to various risk factors. Based on the literature review, achieving better results for predicting disease risk for informed treatment decision and prognosis may require comparing various algorithms together to select the best performing model.

Chapter 4

RESEARCH METHODS AND DESIGN

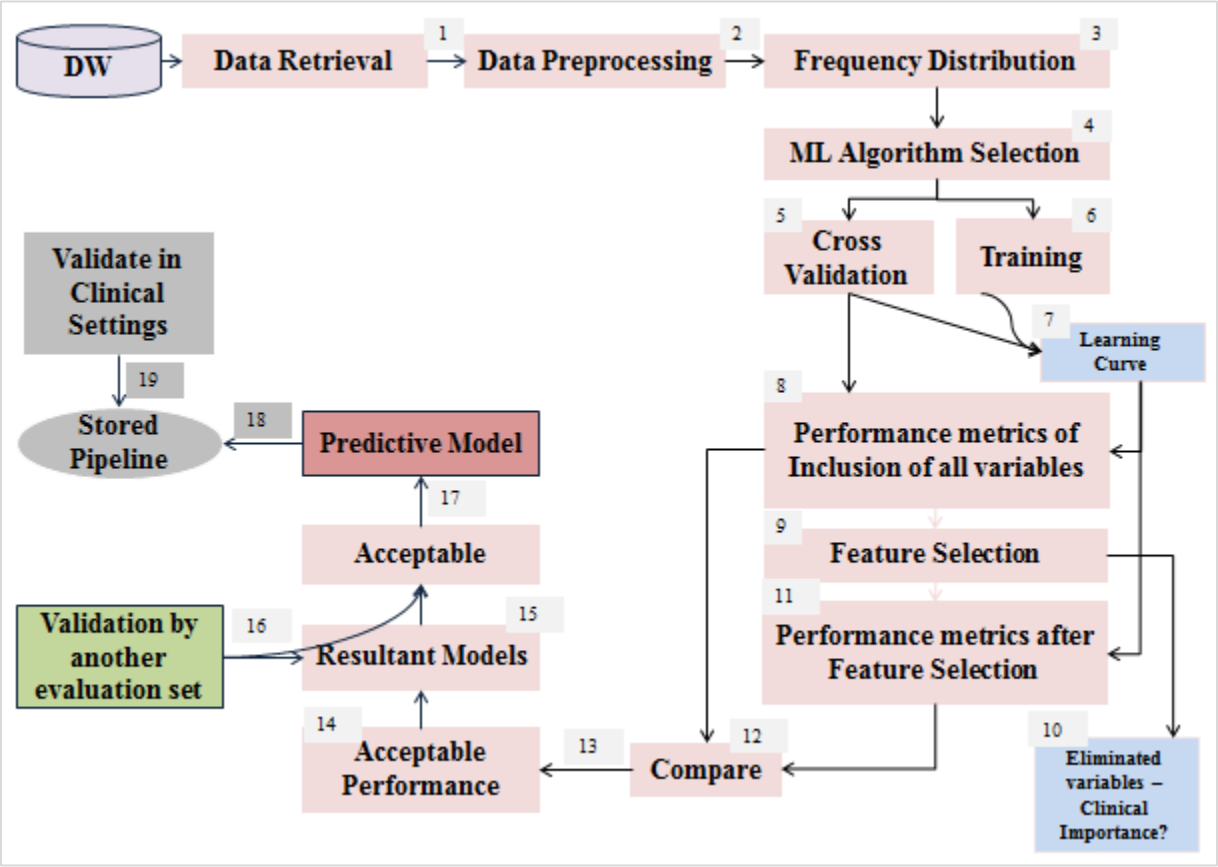
4.1 INSTITUTIONAL REVIEW BOARD

The Marshfield Clinic Research Foundation Institutional Review Board (IRB) reviewed and approved the study using expedited review. IRB oversight was deferred by University of Wisconsin-Milwaukee to Marshfield Clinic Research Foundation.

4.2. CONCEPT DESCRIPTION

To align with current conventions for clinical classification of disease severity by the epidemiological definition by NIDCR, predicting PD risk was treated as a 'classification problem', where patients were sorted into two categories based on disease severity and 'low risk PD' was defined as no or mild gum disease ('controls') or 'high risk PD' defined as moderate to severe disease ('cases') [122]. **Figure 9** shows a generalized pipeline of steps for developing predictive model for PD risk assessment.

Figure 9 Generalized Pipeline for predictive modeling



4.3. DATA RETRIEVAL

The objective of this study was to design a predictive tool to support identification of at-risk PD patients in an interdisciplinary environment. To achieve the objective of the study, retrospective structured data were mined from the MCHS's enterprise data warehouse (MCDW) inclusive of the 6-year temporal window coinciding with the introduction of the dental component of the iEHR in 2010 through 2016 of patients with ages between 18-89 years. MCHS is one of largest private practice groups in the U.S. providing multispecialty care through a network of over 50 regional medical clinics and 10 dental clinics across a broad largely rural service area spanning central, northern and western Wisconsin [10]. The first effort was to review the literature and find out the established medical-dental risk factors. Similarly, a

comprehensive list of all the data that was captured in a routine screening or office visit in the iEHR was developed. The study ran a set of keywords from the initial search strategy on the MEDLINE database with four criteria that included English language, articles on humans, last 20 years and availability of an abstract. Articles from the MEDLINE database were screened and the relevant were accepted for full review.

The operational definitions of all the variables collected at the beginning of the study are specified in appendix 'A'. Achievement of the CDST was contingent on modelling: Patients with medical and dental records in the iEHR.

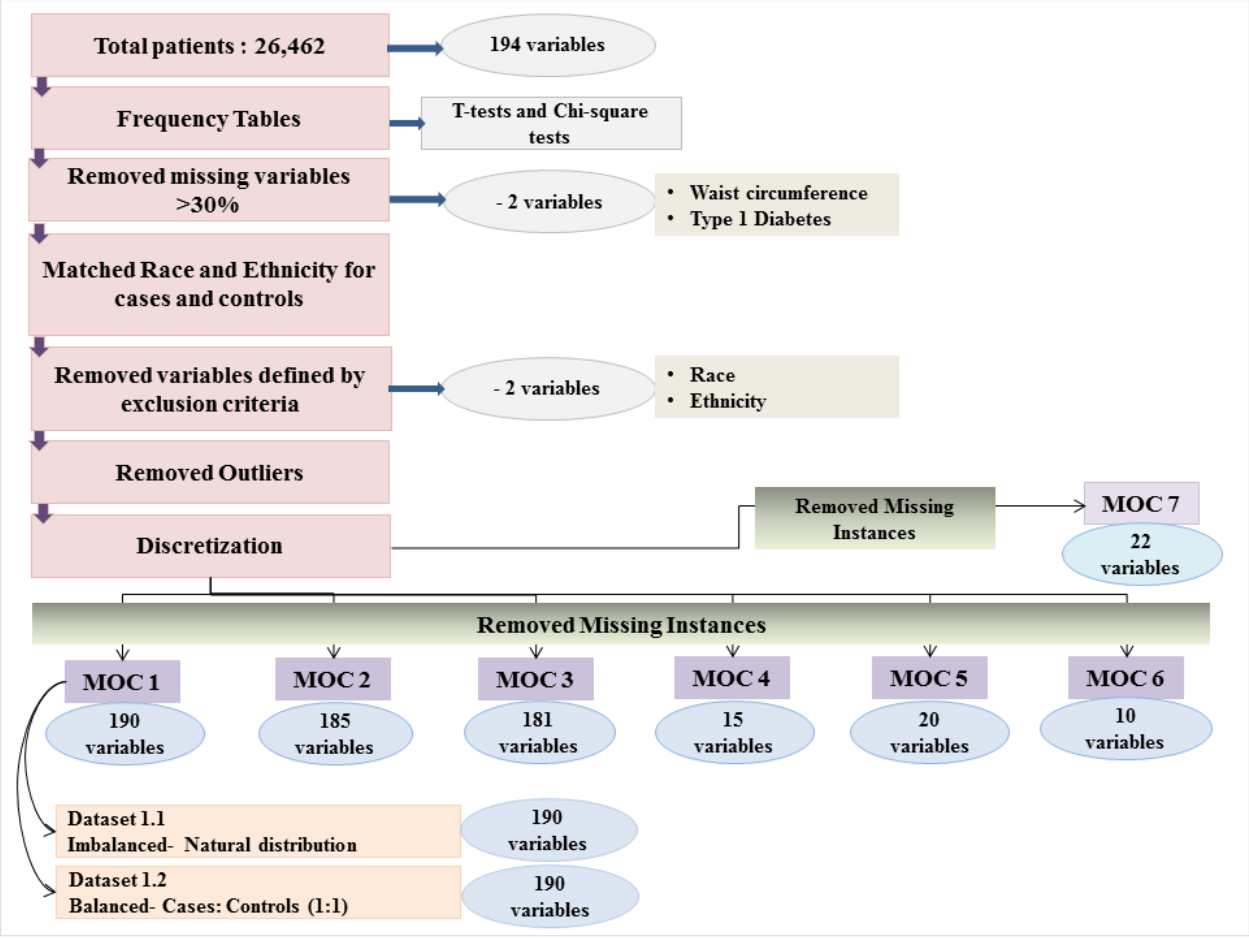
4.4. DATA PREPARATION

The very stable population residing within the MCHS service area comprises of approximately 160,000 patients, > 95% (approximately 155,200) White/Caucasian race and non-Hispanic/Latino ethnicity. Hence the racial and ethnic inclusion of the patient in this study was limited to White/Caucasian, non-Hispanic/Latino individuals since other races and ethnicities were underrepresented for 6 datasets. Deletion methods were used for handling missing data wherein any attribute with more than 30% of missing data variables were excluded while instances (each patient) that had missing values on the remaining attributes were removed from the dataset [123]. The datasets that were created did not have any missing value.

Frequency distribution analysis was first performed to summarize the data in form of histograms and frequency of the distributions. Chi-square tests were used for categorical (nominal) variable and T-tests were used for continuous (numerical) variables. Data were analyzed and outliers were removed according to the standard lab procedures using IBM SPSS statistics software (version 24) [124][93][125]. Outliers were determined by using Tukey's

method of leveraging the Interquartile range [126]. Cohen defines statistical power as the probability that a test will "yield statistically significant results" i.e. the probability that the null hypothesis will be rejected when the alternative hypothesis is true [127]. Based on literature review, this study assumes that having an accuracy of 85± 5 will be sufficient to say that the model is acceptable. The minimum sample size calculations were performed by NCSS statistical software [128]. A paired two-tailed test was performed on AUC of the algorithms tested in the various MOCs to assess the significance.

Figure 10: Data preparation process



The datasets were then divided into 7 datasets with different variables captured in the iEHR representing the seven identified MOCs. Instances having missing values were removed. These

data sets then underwent stratified sampling where, the instances were first stratified by class. The data sets were again divided into training/testing and evaluation sets. To generalize the machine learning algorithm and determine that the algorithm is performing well, the resultant models were also tested on a 10% new data created by randomly selecting (from each class proportionately) from the total data from each of the MOCs. For this study, the new data referred to as the "external evaluation data set". Table 2 shows the characteristics of the datasets:

4.5. DATA PREPROCESSING

Given the retrospective data, the task of classifying patients as low or high PD risk, a flat file with class as "high" or "low" was used. The data sets were represented in an unordered instances-attribute matrix (ARFF file). The data for each patient was defined as "one instance" whereas each variable was defined as "one attribute".

An ARFF file example is shown below

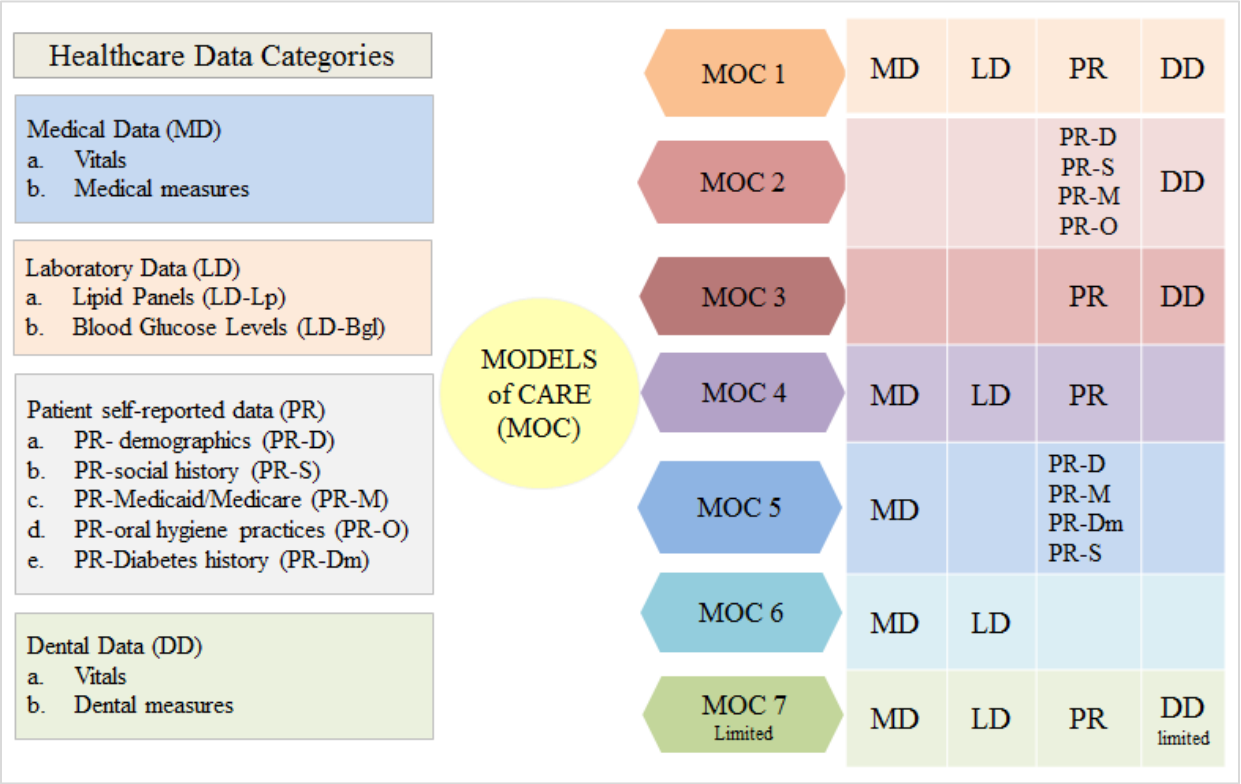
```
% ARFF file for PD risk data with some numeric and nominal features
%
@relation PD risk

@attribute age (numeric)
@attribute diabetes {type1, type2, prediabetes} (nominal)
@attribute 'n' numeric
@attribute PD risk {high, low}
@data
%
% 9499 instances *
45, type2, 3, ……, n, high
```

70, prediabetes, 9, …….., n, low

4.6. IDENTIFICATION OF MODELS OF CARE AND HEALTHCARE DATA CATEGORIES

The fundamental goal of any healthcare organization is the improve patients health by increasing the efficiency in healthcare systems and establishing models of care according to the delivery of services [20]. The MOC framework incorporates various practice care that are shared across the healthcare systems ensuring that the patient gets the proper care with the available resources. To improve clinical care and outcomes and promote the effectiveness of evidence-based clinical practice amongst the interdisciplinary team, six models of care (MOC 1 to MOC 6) were identified. The rationale for selecting different care models was based on the availability of data from various sources such as laboratory findings, patient self-reported data, dental practice and medical practice. A separate MOC 7 was a smaller attempt to look into the future possibility of care, if medical providers initiate oral health screening including dental calculus measurement and counting the number of present teeth in the oral cavity of the patient.

Figure 11: Various models of care and healthcare data categories



For this study, the healthcare data categories and MOCs were defined as following:

4.6.1 HEALTHCARE DATA CATEGORIES

4.6.1.1 MEDICAL DATA (MD)

Medical data included data information relevant to medical practice (including vitals: Height, Weight, systolic and diastolic blood pressure and Body Mass Index (BMI) and medical measures including diagnosis of diabetes Type 1, diabetes Type 2 and prediabetes)

4.6.1.2 DENTAL DATA (DD)

Dental data included data information relevant to dental practice (including vitals: systolic and diastolic blood pressure, dental measures including oral hygiene status diagnosed by dental provider and periodontal pocket depth.

4.6.1.3 LABORATORY DATA (LD)

Laboratory data included data information that was documented in laboratory settings (such as lipid profiles including High Density Lipids, Low Density Lipids, triglycerides and total cholesterol, random blood glucose levels).

4.6.1.4 PATIENT SELF-REPORTED DATA (PR)

Patient self-reported included data information that was reported by the patients in clinical settings (such as tobacco use, height, weight, duration of diabetes, number of teeth present, frequency of tooth brushing and flossing).

4.6.2 MODELS OF CARE

4.6.2.1 MOC 1: INTERDISCIPLINARY MODEL

The 'interdisciplinary' MOC examined the utility of predictive model in an interdisciplinary setting for a) patients who are medical as well as dental patients of the same healthcare organization and b) patients who have detailed documents of previous medical and dental records from same or different organization recorded in the current iEHR. Newhouse et al defined interdisciplinarity as " the coordinated and coherent linkages between disciplines resulting in reciprocal interactions that overlap disciplinary boundaries generating new common methods, knowledge, or perspectives [21]."Interdisciplinary model" included data relevant to MD, LD, PR and DD.

4.6.2.2 MOC 2: DENTAL ONLY

The ‘dental only’ MOC examined the utility of predictive model specifically for dental practice and to improve preventive care in terms of PD risk assessment for a) patients who are solely dental patients and who do not have medical visits and b) patients who visit dental center, however do not have any medical record. “Dental only model” included data relevant to PR-D, PR-S, PR-M, PR-O and DD.

4.6.2.3 MOC 3: DENTAL WITH PATIENT REPORTED MEDICAL

The ‘dental with patient reported medical’ MOC examined the utility of predictive model to improve preventive care in terms of PD risk assessment for patients who visit dental center but have a primary care provider from a different organization. “Dental with patient reported medical model” included data relevant to PR and DD.

4.6.2.4 MOC 4: MEDICAL WITH PATIENT REPORTED DENTAL

This ‘ medical with patient reported dental’ MOC examined the utility of predictive model to improve preventive care in terms of PD risk assessment for a) patients who visit medical center but do not have dental provider and b) patients who visit medical center and have visit dental providers from other organization. “Medical with patient reported dental model” included data relevant to MD, LR and PR.

4.6.2.5 MOC 5: MEDICAL ONLY

This ‘medical only’ MOC examined the utility of predictive model to improve preventive care in terms of PD risk assessment for patients who visit medical center but do not have a dental provider. “Medical only model” included data relevant to PR-D, PR-M, PR-Dm, PR-S and MD.

4.6.2.6 MOC 6: MEDICAL WITHOUT PATIENT REPORTED DATA

This ‘medical without patient reported data’ MOC examined the utility of predictive model to improve preventive care in terms of PD risk assessment for patients who do not report on any information. “Medical without patient reported data” included data relevant to MD and LD.

4.6.2.7 MOC 7: MEDICAL MODEL WITH LIMITED DENTAL PARAMETER

This ‘medical model with limited dental parameter’ MOC examined the utility of predictive model to improve preventive care in terms of PD risk assessment for patients who visit medical centers where medical providers start collecting information on limited parameters of DP such as dental calculus and number of teeth present. “Medical model with limited dental parameter” included data relevant to MD, LD, DD (including dental calculus and number of teeth present) and PR. This model was limited to 4000 patients.

4.7 EXPERIMENTS

4.7.1. A COMPARISON OF THE EFFECT OF IMBALANCED AND BALANCED DATASETS ON PERFORMANCE METRICS

To establish a common ground for other MOCs regarding the impact of imbalanced class, on model performance, the first test was to evaluate the performance based on application to the natural distribution (imbalanced class distribution) of the classes within the population versus the performances on a balanced class. Dataset were further divided into 2 subsets, one with a case: control ratio of 1:1.25 and other with a case: control ratio of 1:1 by under sampling. The five algorithms i.e. NB, LR, SVM, ANN and DT were employed on both the datasets. In this study, a balanced dataset was created by random undersampling of the training data such that there was a similar proportion of cases and controls.

4.7.2. PERFORMANCE METRICS

The performances of classifiers were evaluated and compared using the following metrics: accuracy, precision, specificity, and sensitivity (recall) as well as by plotting ROC and Mathews correlation coefficient (MCC). The algorithms were evaluated through creation of a confusion matrix and measurement of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. For each classifier the confusion matrix was formed and then metrics were calculated as follows:

To assess the prediction model performance of different algorithms, the study used stratified 10 fold cross validation and compared ML algorithms using the following measures:

4.7.2.1.AREA UNDER THE CURVE ROC (AUC)

In regards to the PD risk prediction, the study used the definition of the Area under the ROC (receiver operative curve) (AUC) as defined by Hand and Till [129] for binary classification and is given by the following equation

$$\text{AUC}= \frac{\frac{[S0 - n0\,(n0+1)]}{2}}{n_0\, n_1} \quad \ldots\ldots (6)$$

where S_0 is the sum of ranks of class and is given by $=\sum r_i$, where r_i is the rank of the ith class, n_0 and n_1 are the numbers of 'PD high risk' and 'PD low risk', respectively, and where r_i is the rank of the ith 'PD high risk' in the ranked list.

4.7.2.2. SENSITIVITY/RECALL

Sensitivity (also called recall) is the ratio of the number of correctly classified 'PD high risk' from a given 'PD high risk' and the total number of 'PD high risk' and 'PD low risk'

$$\text{Recall/ Sensitivity (Se)} = \frac{TP}{TP+FN} \quad \ldots\ldots (7)$$

where TP=true positive, FN=False negative

4.7.2.3. PRECISION

Precision is the ratio of the number of correctly classified 'PD high risk' from a given 'PD high risk' and the total number of 'PD high risk' and misclassified 'PD low risk' as 'PD high risk'

$$\text{Precision}= \frac{\text{TP}}{\text{TP+FP}} \quad \ldots\ldots (8)$$

where TP= true positive and FP= false positive

4.7.2.4. SPECIFICITY

Specificity is the ratio of the number of correctly classified 'PD low risk' from a given 'PD low risk' and the total number of 'PD low risk' and misclassified 'PD high risk' as 'PD low risk'

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad \text{.....(9)}$$

where TN=true negative, FP= false positive

4.7.2.5. ACCURACY

Accuracy is the ratio of the number of correctly classified 'PD high risk' and 'PD low risk' from the 'PD high risk' and the total number of correctly classified and misclassified as 'PD high risk' and 'PD low risk'

$$\text{Accuracy} = \frac{TN+TP}{\text{TP+TN+FP+FN}} \quad \text{.....(10)}$$

where TN=true negative, TP=true positive, FN=false positive and FN=false negative

4.7.2.6. F-MEASURE

F-measure is the weighted average of precision and recall and is given by

$$\text{F-measure} = 2 * \left[\frac{(Precision * Recall)}{(Precision+Recall)}\right] \quad \text{.....(11)}$$

4.7.2.7. MATTHEW'S CORRELATION COEFFICIENT (MCC)

Matthew's Correlation Coefficient (MCC) considers the accuracy and error rates of high PD risk and low PD risk and is calculated by the following equation:

$$\text{MCC} = \frac{TP*TN - FP * FN}{\sqrt{(TP+FP)(TP+FP)(TN+FP)(TN+FN)}} \quad \text{.....(12)}$$

where TN=true negative, TP=true positive, FN=false positive and FN=false negative

4.7.2.8. RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

Receiver Operating Characteristic (ROC) curve illustrates the performance of the outcome variable in a graphical format. The false positive rate (1-specificity) on X-axis was plotted against true positive rate (sensitivity) on Y-axis. For evaluating the performance of diagnostic tests for each of the algorithm on the MOC, ROC curves for each algorithm were created for all the models.

4.7.3. FEATURE SELECTION

To optimize the classifier and identify the representative subset of attributes, multivariate filter i.e. correlation- based feature selection (CFS) method and univariate filters such as information gain with ranker method was employed. For MOC 1, information gain with a ranker method was utilized while CFS with best search method was utilized for other MOC models. CFS has been utilized as a benchmark method for examining inherent predictive ability of each feature along with the degree of redundancy between them [130][131]. It is assumed that a feature subset is highly representative when the features are highly correlated with the predictive class, yet uncorrelated with each other [131].

4.7.4. FEATURE SELECTION AND OPTIMIZED REPRESENTATION OF TEETH SURFACES FOR PERIODONTAL PROBING DEPTH

Feature selection and optimized representation of teeth surfaces for periodontal probing depth. In this experiment feature selection based on information gain in conjunction with ranker method was applied to imbalanced datasets. The purpose of feature selection was two folded. One was to optimize the size by generating a representative set and second was to search the most significant tooth surfaces for PPD. Features were eliminated in two steps. A feature-selection was performed on all the variables in the MOCs that eliminated redundant variables and resulted in a representative set based on their relative contribution to risk for PD emergence.

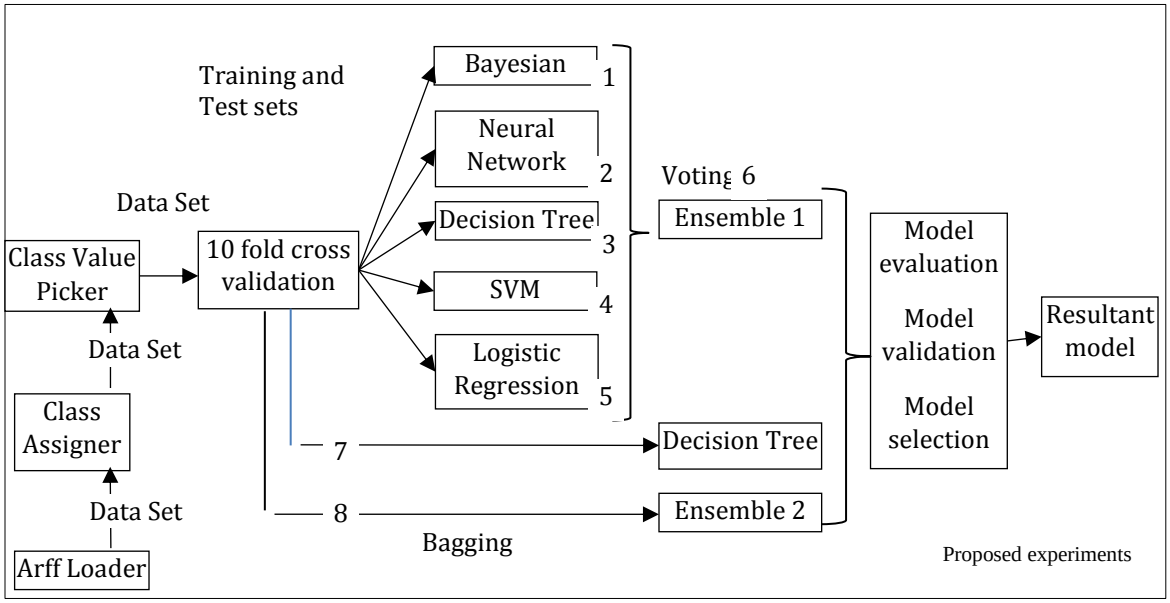
4.7.5. LEARNING CURVES

To understand the performance of a model in predicting the outcome with the corresponding increase in number of instances used to train it, learning curves were generated and the training performances and cross validation performances of learning curves were evaluated for all the models. The learning curve was also used to find the smallest sample size that can be used to train an algorithm, yielding an accuracy similar to the one achieved by using the entire dataset [132].

4.7.6. INTER-COMPARISON OF PERFORMANCE OF VOTING, BAGGING AND DECISION TREE.

Bagging and voting was used for building ensembles that uses different sets of training data using a single learning method. Figure 12 shows the experimental framework for comparing various ensembles of classifiers.

Figure 12: Experimental framework for comparing various ensembles of classifiers



4.7.7. VALIDATION OF THE RESULTANT MODEL BY AN EXTERNAL EVALUATION SET

This work validated the tool by creating a dataset of comparable population by creating a new subpopulation drawn from the population seen at the healthcare organization. To validate the predictive performance of the resultant model, a new subset of 10% of the total data set (drawn from the same data warehouse) was utilized to evaluate the resultant model. This subset was called as "external evaluation set"**.** To generalize the machine learning algorithm and determine that the algorithm is performing well, the resultant models were also tested on a 10% new data created by randomly selecting (from each class proportionately) from the total data from each of the MOCs. **Table 2** show the experiments performed for various MOCs.

Table 2: Experiments performed in various models of care							
Experiment	**MOC 1**	**MOC 2**	**MOC 3**	**MOC 4**	**MOC 5**	**MOC 6**	**MOC 7**
a. Balanced-Imbalanced	✓						
b. Performance metrics on dataset with all variables	✓	✓	✓	✓	✓	✓	✓
c. Feature selection	✓	✓	✓	✓	✓		
d. Feature selection for periodontal probing depth	✓	✓	✓				
e. Learning Curves	✓	✓					
f. Comparison with ensemble	✓						
g. Evaluation of external dataset on resultant model	✓	✓	✓	✓	✓	✓	✓

4.7.8. DISCRETIZATION

Discretization is the process of transforming continuous valued attributes to discrete ones. Discretization was done for some of the variables, including high HDL, LDL, TC and TG according to ATP III classification [133]. HDL was categorized as 'high' ($\geq$ 60), 'normal' (40-59), and 'low' (<40). LDL were grouped as <100 as optimal, 100-129 as near optimal, 130-159 as borderline high, 160-189 as high, $\geq$ 190 as very high. TC was defined as <200 as desirable, 200-239 as borderline high and $\geq$ 240 as high. TG was categorized as 'Ideal', 'Borderline High', 'High' and 'Very High'. BMI was categorized into 'Underweight' with a principal cut-off point as less than18.50, 'normal' with a range between 18.50 and 24.99, 'overweight' as 25.00-29.99 and 'obese' as BMI of more than or equal to 30.00. Similarly, systolic blood pressure and diastolic blood pressure were grouped according to the American Heart Association (AHA)

[134]. The systolic blood pressure was categorized as ‘normal’ having systolic blood pressure less than 120 mm of Hg, ‘prehypertension’ between 120-139 mm of Hg, ‘Hypertension Stage 1’ between 140-159, ‘Hypertension Stage 2’ with 160 mm of Hg or higher; while diastolic was categorized ‘normal’ as less than 80, ‘prehypertension’ between 80-89, ‘Hypertension Stage 1’ between 90-99 and ‘Hypertension Stage 2’ 100 mm of Hg or higher.

4.7.9. MACHINE LEARNING

4.7.9.1. POTENTIAL SUPERVISED LEARNING ALGORITHMS

To test generalizability of ML in classifying patients for PD risk, this study focused on a set of widely used supervised methods, including five approaches to algorithmic derivation: Naïve Bayes (NB), Logistic Regression (LR), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (DT). Ensemble methods such as Voting and Bagging were used for the interdisciplinary dataset. The study utilized ML algorithms available in the Waikato Environment for Knowledge Analysis (WEKA) open-sourced tool version 3.8.1 [135].

4.7.9.2. DATA PARTITIONING

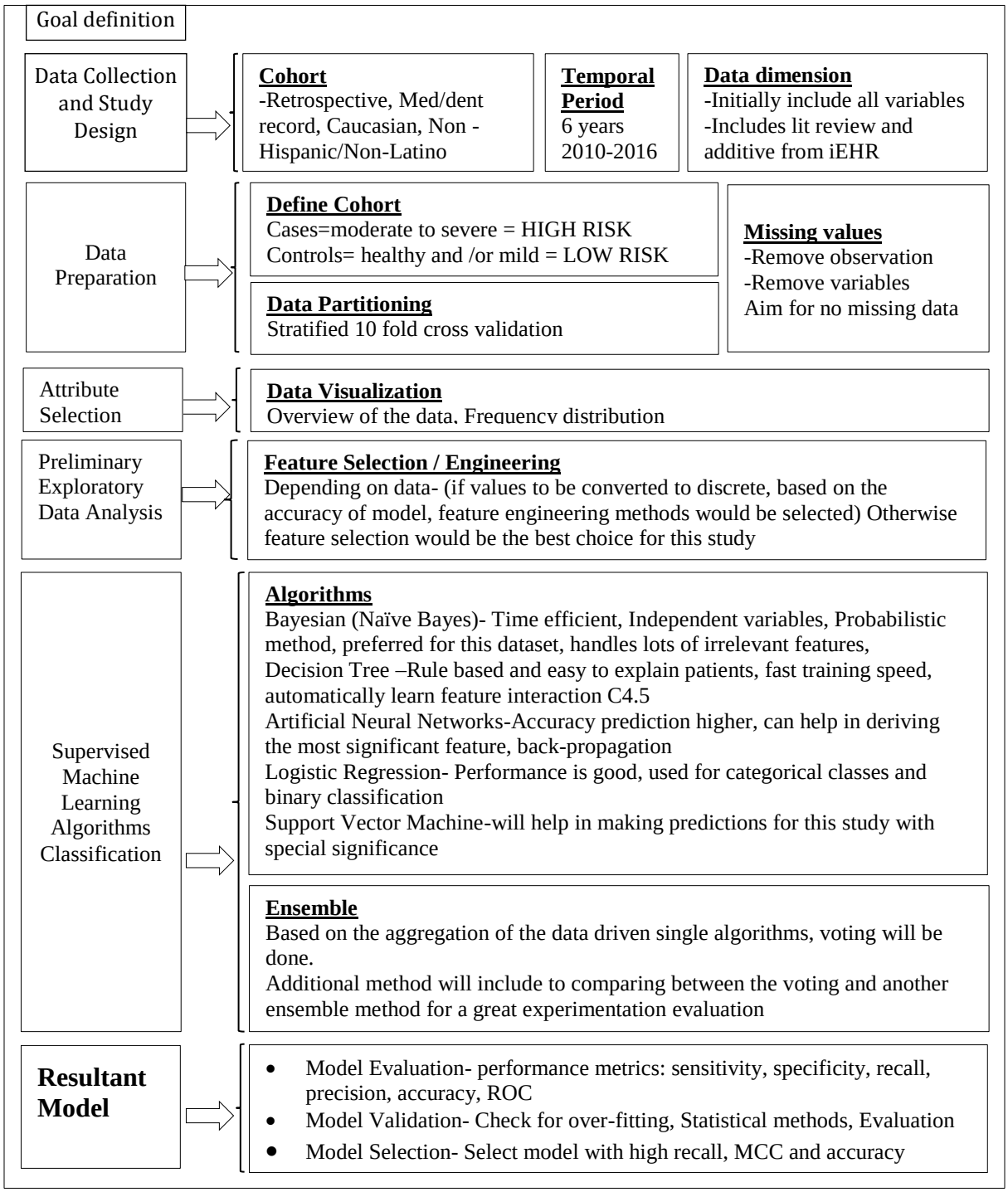
A stratified 10-fold cross validation was conducted, thus partitioning the data set into 10 equal size data subsets. In this, the data (d) was randomly partitioned into 10 non-overlapping data subsets (d1, d2, d3….d10). At each iteration i (from 1 to 10) a single data subset was retained as a validation data set (di) for testing the model and the remaining 9 data subsets (d\di) were used as training dataset to train the classifier [88]. The cross-validation process was repeated 10 times for all the algorithms in the seven different MOC models, giving an opportunity for each data subset to act as a validation dataset once. The results of the 10 fold cross validation were averaged to produce a single estimation [88].

4.7.9.3. VALIDATION AND MODEL SELECTION

Model selection is aimed at finding the right level of model complexity that balances bias and variance in order to achieve high predictive accuracy. Over-fitting is one of the major concerns in predictive analytics because it reduces the model's ability to predict new data accurately [88]. Assessing over-fitting was achieved by comparing the performance of the overall accuracy on

the training and cross validation sets. The best predictive model for the tool was also determined by selecting the model with the highest AUC and MCC.

Figure 13 shows the steps in detail for building the PD model



Chapter 5

RESULTS

5.1 RESULTS OF THE LITERATURE REVIEW

A comprehensive review of literature geared towards the existing risk factors for developing PD was conducted. In this review, about 40 articles describing associations of PD and systemic disease along with factors that were used in previous PD risk assessment tools were identified and grouped into categories that were similar to the ones that were captured in the MCHS iEHR for ease of retrieval from the MCDW. For example, these included demographics, dental variables, and co-morbid conditions such as diabetes among others. The factors previously established as candidates for association with PD as shown in Figure 4 were used as variables as inclusion criteria and additional factors as shown in Figure 5, such as patient's status for Medicare and Medicaid, dental calculus, diabetes category, oral hygiene status, number of teeth present, tooth brushing and flossing frequency, periodontal pocket depth, body mass index (BMI), blood pressure (systolic and diastolic), and lipid profiles [such as high density lipids (HDL), low density lipids (LDL), total cholesterol (TC), triglyceride (TG)], were specified for collection and analysis in the dataset.

Figure 14: PD Risk factors from literature review

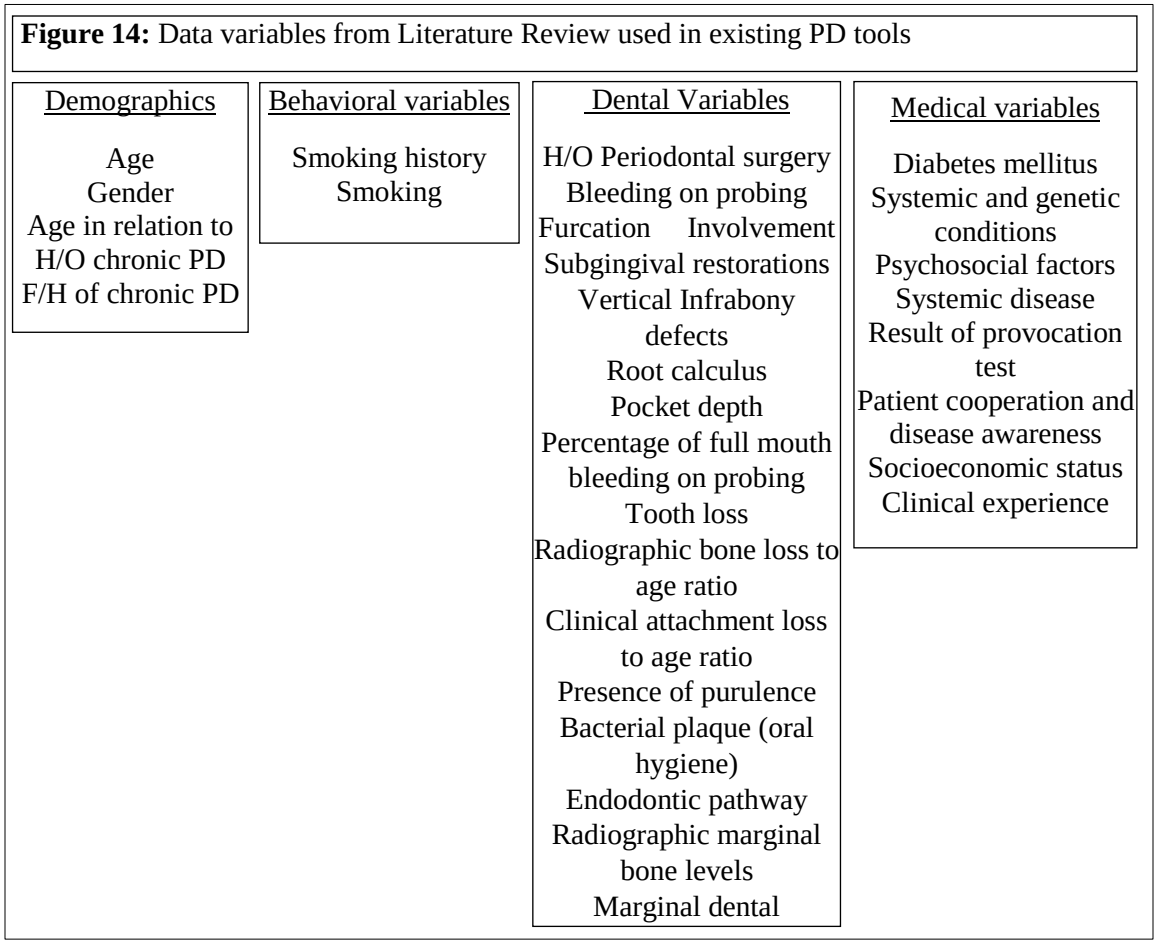
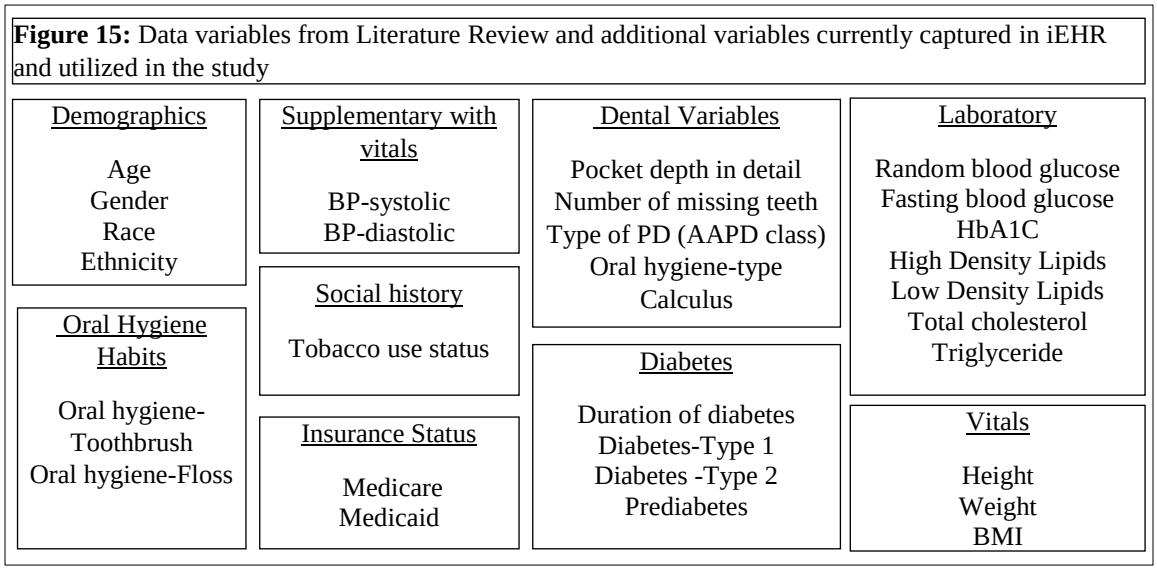


Figure 15: Data variables captured in iEHR

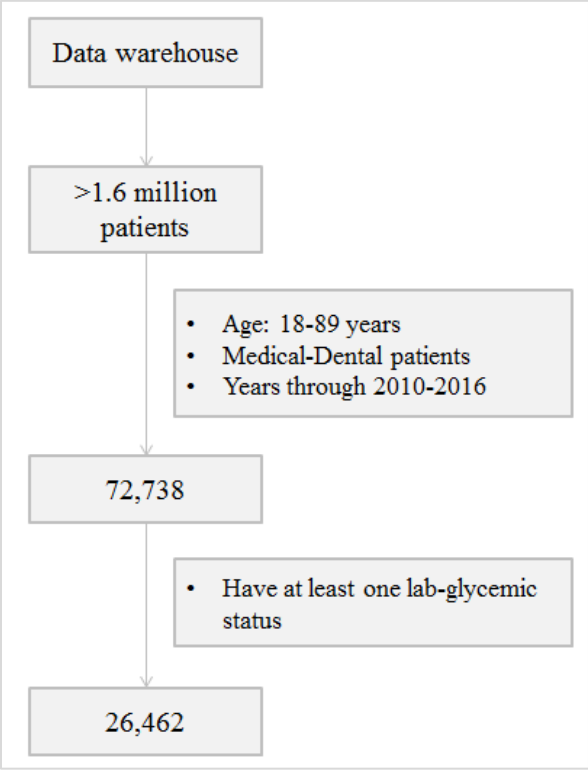


5.2. RESULTS OF DATA MINING ACTIVITY

From a cohort of more than 1.6 million medical-dental patients, 72,738 patients were medical as well as dental patients. Evidence base strongly supports bidirectional exacerbation between Type 2 diabetes (T2DM) and PD and hence an inclusion criterion of having at least one glycemic values was applied to this cohort of 72,738 patients. This yielded a total of 26, 462 patients.

For the purpose of this study all the PPD of all the present teeth of the patients retrieved were treated as separate variables. All the third molars were excluded from this dataset. In absence of teeth, the probing depth was considered as zero. Overall, for 28 teeth there were 168 variables.

Figure 16: Data retrieval process- inclusions and exclusions



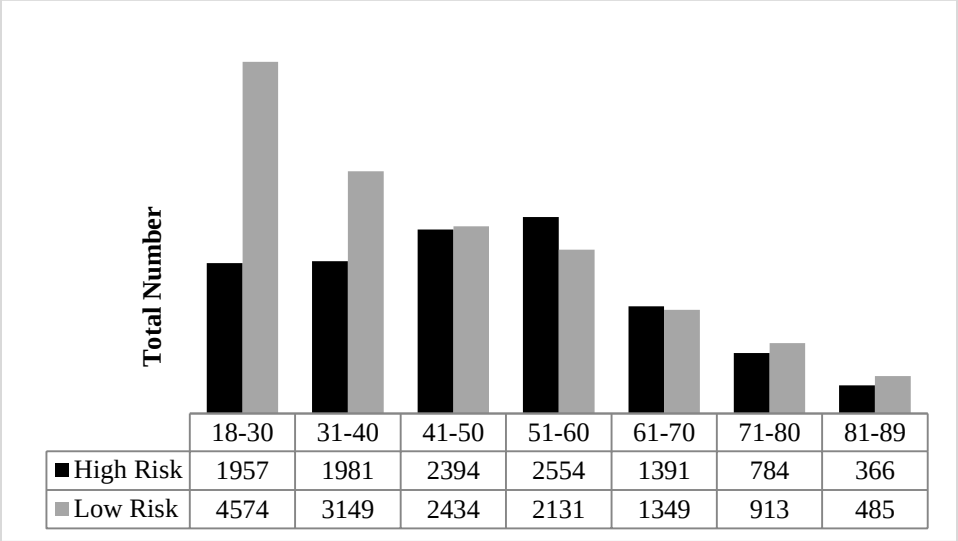
5.3 DEMOGRAPHICS

5.3.1. AGE DISTRIBUTION

(Total patients: 26,462)

The overall mean age of the patients was 45.33 years (SD 17.17 years)

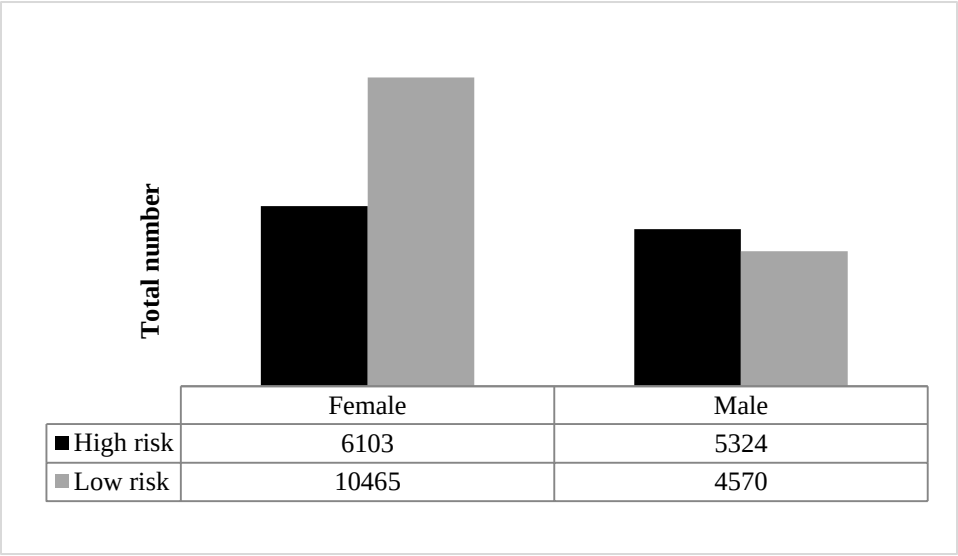
Figure 17: Age distribution of the overall study cohort



	18-30	31-40	41-50	51-60	61-70	71-80	81-89
High Risk	1957	1981	2394	2554	1391	784	366
Low Risk	4574	3149	2434	2131	1349	913	485

5.3.2. GENDER DISTRIBUTION

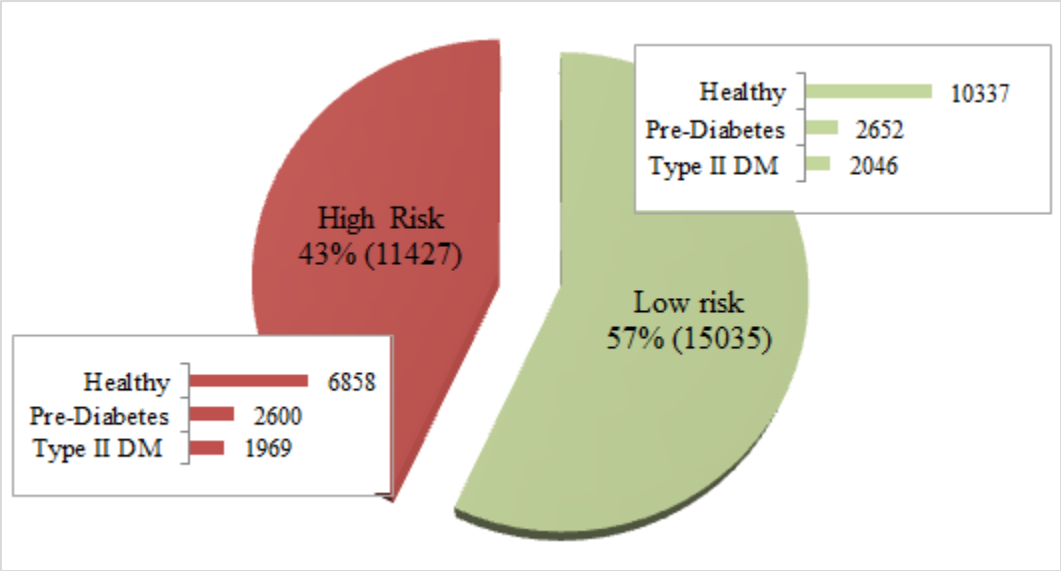
Figure 18: Gender distribution of the overall study cohort



	Female	Male
High risk	6103	5324
Low risk	10465	4570

5.3.3. PREVALENCE OF PD

Figure 19: Prevalence of PD and Type II Diabetes (T2DM) in study cohort



Data characteristics

Data characteristics of various models of care including total number of instances, total instances used for training and testing, total number of variables and total number of instances used in evaluation sets are shown in **Table 3**.

Table 3: Characteristics of the datasets in various models of care

Datasets	Total instances (No of patients)	Total instances used for Training/Testing	Total variables/ Attributes	Evaluation sets (10% of total instances)
MOC1 (D1)	11,048	9,944	190	1,104
MOC2 (D2)	16,768	15,092	181	1,362
MOC3 (D3)	13,525	12,173	185	1,352
MOC4 (D4)	15,705	14,135	15	1,570
MOC5 (D5)	19,972	14,135	20	1,997
MOC6 (D6)	22,085	22,085	10	None
MOC7 (D7)	4,000	4,000	22	None

The variables that were included in various datasets are shown in **Table 4**

Table 4: Characteristics of variables included in the datasets prepared for the MOCs

Variables	Char*	Model of Care						
		1	2	3	4	5	6	7
Age	Numeric	✓	✓	✓	✓	✓		✓
Gender	Nominal	✓	✓	✓	✓	✓		✓
Medicaid	Numeric	✓	✓	✓	✓	✓		✓
Medicare	Numeric	✓	✓	✓	✓	✓		✓
Tobacco	Nominal	✓	✓	✓	✓	✓		✓
BP	Nominal	✓	✓	✓	✓	✓	✓	✓
BMI	Nominal	✓	✓		✓	✓	✓	✓
Height	Numeric	✓		✓		✓		✓
Weight	Numeric	✓		✓		✓		✓
Random Blood Glucose	Numeric	✓			✓	✓	✓	✓
HDL	Nominal	✓			✓	✓	✓	✓
LDL	Nominal	✓			✓	✓	✓	✓
Total cholesterol	Nominal	✓			✓	✓	✓	✓
Triglyceride	Nominal	✓			✓	✓	✓	✓
Tooth brushing	Numeric	✓	✓	✓		✓		✓
Flossing	Numeric	✓	✓	✓		✓		✓
Oral hygiene status	Nominal	✓	✓	✓				✓
Dental calculus	Numeric	✓	✓	✓				✓
PPD	Numeric	✓	✓	✓				
Missing teeth	Numeric	✓	✓	✓				✓
Present teeth	Numeric					✓		
T2DM diagnosis	Nominal	✓		✓	✓		✓	✓
T2DM patient reported	Nominal			✓		✓		
Duration of T2DM	Numeric	✓		✓	✓	✓	✓	✓
Prediabetes	Nominal	✓			✓	✓	✓	✓

5.4. MODEL OF CARE 1: INTERDISCIPLINARY MODEL

5.4.1 PATIENT CHARACTERISTICS

The overall mean age of patients was 47.36 ± 16.60, with 65% of patients being female. Of these 7,315 were Medicaid patients and 3,578 were Medicare patients. Mean brushing frequency was 1.6 ± 0.60. **Table 5** shows the frequency distribution of the variables used in MOC 1. The distribution of the patients for presence or absence of diabetes was: 1887 'Type 2 Diabetes, 2834 'Pre-diabetes' and 6327 'No diabetes'. A majority of patients (1369/1887; 72.5%) diagnosed with diabetes had a duration of diabetes <1 year. Approximately 130 (3%) of patients had documentation of poor oral hygiene while the rest were categorized as good, fair and excellent. Of the 4,000 patients, 1128 (28%) were current smokers, 1464 were former smokers (36.7%) and 1411 (35%) never smoked tobacco. The mean random blood glucose level for this cohort was 112 ± 46.36. More than half of the patients showed a normal blood pressure levels. Approximately, 5,813 patients were obese with a BMI of more than 30, whereas about 3,060 patients were overweight with a BMI between 25 and 29.99. There was significant difference ($p<0.0001$) between the control (low risk) and cases (high risk) for most of the variables utilized in the study and is described in Table 4.

Table 5: Frequency distribution of the variables used in the Interdisciplinary Model of Care (MOC1)

Variables	Categories	High Risk N (%) = 4766	Low Risk N (%) = 6282	P-Value
Age*		50.03±15.62	45.30 ± 16.98	<0.0001
Gender^	Female	2695 (24.39%)	4500 (40.73%)	<0.0001
	Male	2071 (18.75%)	1782 (16.13%)	
Medicaid^	Yes	3312 (29.98%)	4802 (43.46%)	<0.0001
	No	1454 (13.16%)	1480 (13.40%)	
Medicare^	Yes	1855 (16.79%)	2113 (19.13%)	<0.0001
	No	2911 (26.35%)	4169 (37.74%)	
Height*		168.73 ± 10.23	166.69 ± 9.39	<0.0001
Weight*		90.35 ± 24.03	88.17 + 24.07	<0.0001

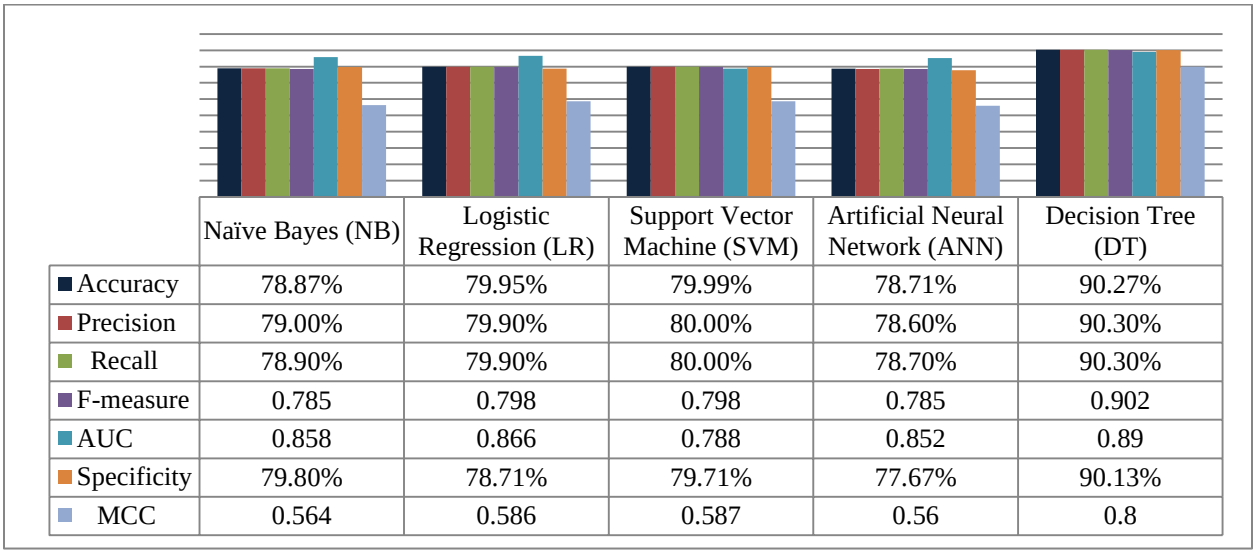
Tooth brushing^	0	175 (1.58%)	173 (1.57%)	<0.001
	1	1879 (17.01%)	2173 (19.67%)	
	2	2533 (22.93%)	3765 (34.08%)	
	3	179 (1.62%)	171 (1.55%)	
Frequency of tooth Flossing^	0	2348 (21.25%)	2859 (25.88%)	<0.001
	1	1794 (16.24%)	2605 (23.58%)	
	2	443 (4.01%)	517 (4.68%)	
	3	181 (1.64%)	301 (2.72%)	
Tobacco use status^	Current	1598 (14.47%)	1725 (15.62%)	<0.001
	Former	1635 (14.80%)	2199 (19.91%)	
	Never	1532 (13.87%)	2358 (21.35%)	
Duration of diabetes*		17.93 + 52.50	15.42 + 48.86	<0.0096
High Density Lipids (HDL)^	Healthy	1021 (9.24%)	3155 (28.56%)	<0.001
	Low	2381 (21.55%)	1529 (13.83%)	
	High	1364 (12.35%)	1598 (14.47%)	
Low Density Lipids (LDL)^	Optimal	2175 (19.69%)	3054 (27.64%)	<0.0001
	Near Optimal	1519 (13.74%)	2009 (18.18%)	
	Borderline	763 (6.90%)	887 (8.03%)	
	High	237 (2.15%)	264 (2.34%)	
Triglycerides (TG)^	Healthy	3240 (29.32%)	4512 (40.83%)	<0.0001
	Borderline	765 (6.92%)	884 (8%)	
	High	757 (6.85%)	882 (7.98%)	
	Very High	4 (0.036%)	4 (0.036%)	
Total Cholesterol (TC)^	Desirable	3370 (30.50%)	4609 (41.71%)	0.4449
	Borderline	1059 (9.59%)	1289 (11.67%)	
	High	337 (3.50%)	384 (3.47%)	
Random Blood Glucose*		114.78 + 47.80	110.56 $\pm$ 45.39	<0.0001
Oral hygiene^	Poor	986 (8.92%)	657 (5.95%)	<0.001
	Fair	55 (0.50%)	74 (0.67%)	
	Good	1221 (11.05%)	2778 (25.14%)	
	Excellent	986 (8.92%)	657 (5.95%)	
Dental calculus^	0	47 (0.43%)	115 (1.04%)	<0.001
	1	2050 (18.56%)	4316 (39.07%)	
	2	1850 (16.75%)	1548 (14.01%)	
	3	819 (7.41%)	303 (2.74%)	
No of missing teeth*		3.26 $\pm$ 4.70	3.65 $\pm$ 6.01	<0.0002
BP-diastolic^	<80	3018 (27.32%)	4262 (38.57%)	<0.0001
	80-89	1350 (12.21%)	1633 (14.78%)	
	90-99	326 (2.96%)	55 (0.49%)	
	>100	72 (0.06%)	332 (0.30%)	
BP-systolic^	<120	1732 (15.68%)	2724 (24.65%)	<0.0001
	120-139	2275 (20.60%)	2860 (25.89%)	
	140-159	638 (5.78%)	584 (5.28%)	
	>160	121 (1.09%)	114 (1.03%)	
BMI^	Underweight	45 (0.04%)	68 (0.61%)	<0.0001
	Normal	830 (7.51%)	1232 (11.15%)	

	Overweight	1379 (12.49%)	1681 (15.21%)	
	Obese	2512 (22.73%)	3301 (29.88%)	
Diabetes categories^	Type II Diabetes	903 (8.17%)	984 (8.91%)	<0.001
	Prediabetes	1347 (12.19%)	1487 (13.46%)	
	No Diabetes	2516 (22.77%)	3811 (34.49%)	
TOTAL		43% (4,766)	57% (6,282)	
*Indicate numerical value, ^ indicates categorical value				

5.4.2 RESULTS OF COMPARISON AND EFFECT OF IMBALANCED AND BALANCED DATASETS ON PERFORMANCE METRICS.

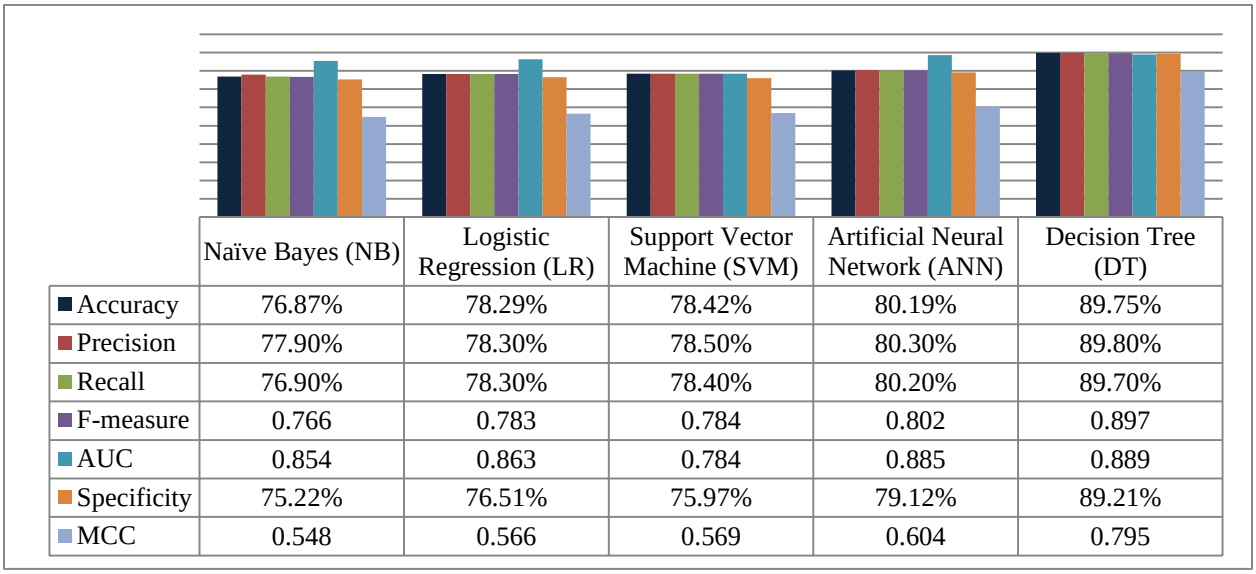
The results of performance measures in terms of accuracy, precision, recall, F-measures, specificity and MCC showed that DT demonstrated higher analytic accuracy in classifying the patients with high and low PD risk as compared to NB, LR, SVM and ANN. The recall and precision for DT imbalanced dataset were 90.30% (95% CI=89.57% to 91.10%) and 90.13% (89.17% to 91.04%), respectively. The learning rate for ANN was at 0.3 with a momentum of 0.2 and training time of 500. The time taken to build the ANN model was 4,468 seconds. The AUC for DT was 0.89 followed by LR, NB, ANN and SVM. A paired t-test showed that there was no significant difference between NB, LR, SVM and ANN. Figure 20 shows the results of the performance metrics of MOC 1.

Figure 20: Performance analysis of results of Interdisciplinary Model of Care (MOC 1) with 190 variables and an imbalanced dataset



	Naïve Bayes (NB)	Logistic Regression (LR)	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Decision Tree (DT)
Accuracy	78.87%	79.95%	79.99%	78.71%	90.27%
Precision	79.00%	79.90%	80.00%	78.60%	90.30%
Recall	78.90%	79.90%	80.00%	78.70%	90.30%
F-measure	0.785	0.798	0.798	0.785	0.902
AUC	0.858	0.866	0.788	0.852	0.89
Specificity	79.80%	78.71%	79.71%	77.67%	90.13%
MCC	0.564	0.586	0.587	0.56	0.8

Figure 21: Performance analysis of results of MOC 1 with 190 variables and a balanced dataset

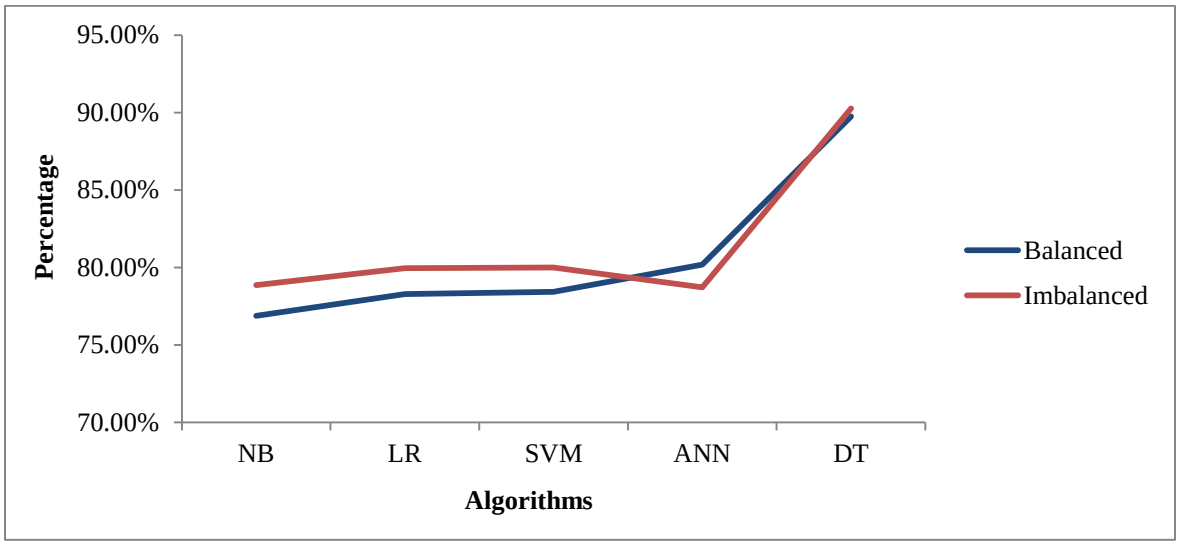


	Naïve Bayes (NB)	Logistic Regression (LR)	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Decision Tree (DT)
Accuracy	76.87%	78.29%	78.42%	80.19%	89.75%
Precision	77.90%	78.30%	78.50%	80.30%	89.80%
Recall	76.90%	78.30%	78.40%	80.20%	89.70%
F-measure	0.766	0.783	0.784	0.802	0.897
AUC	0.854	0.863	0.784	0.885	0.889
Specificity	75.22%	76.51%	75.97%	79.12%	89.21%
MCC	0.548	0.566	0.569	0.604	0.795

Overall, the accuracy, recall, precision, F-measure, AUC, specificity were slightly low for balanced dataset as compared to the imbalanced dataset. DT model indicated the highest accuracy. In terms of specificity and sensitivity, DT outperformed other models. The AUC for ANN and DT was almost similar. There was a significant difference for the AUC when SVM

was compared to NB, LR, SVM, ANN and DT ($p<0.001$). **Figure 22** shows the results of the comparison made between imbalanced and balanced set showed that the total accuracy of imbalanced dataset was higher than balanced dataset. The accuracy for imbalanced dataset was higher as compared to the balanced dataset, except for ANN model, where the total accuracy of balanced dataset was slightly higher than the imbalanced dataset.

Figure 22: The results of comparison between imbalanced and balanced dataset showed that the total accuracy of imbalanced dataset was higher than balanced dataset.

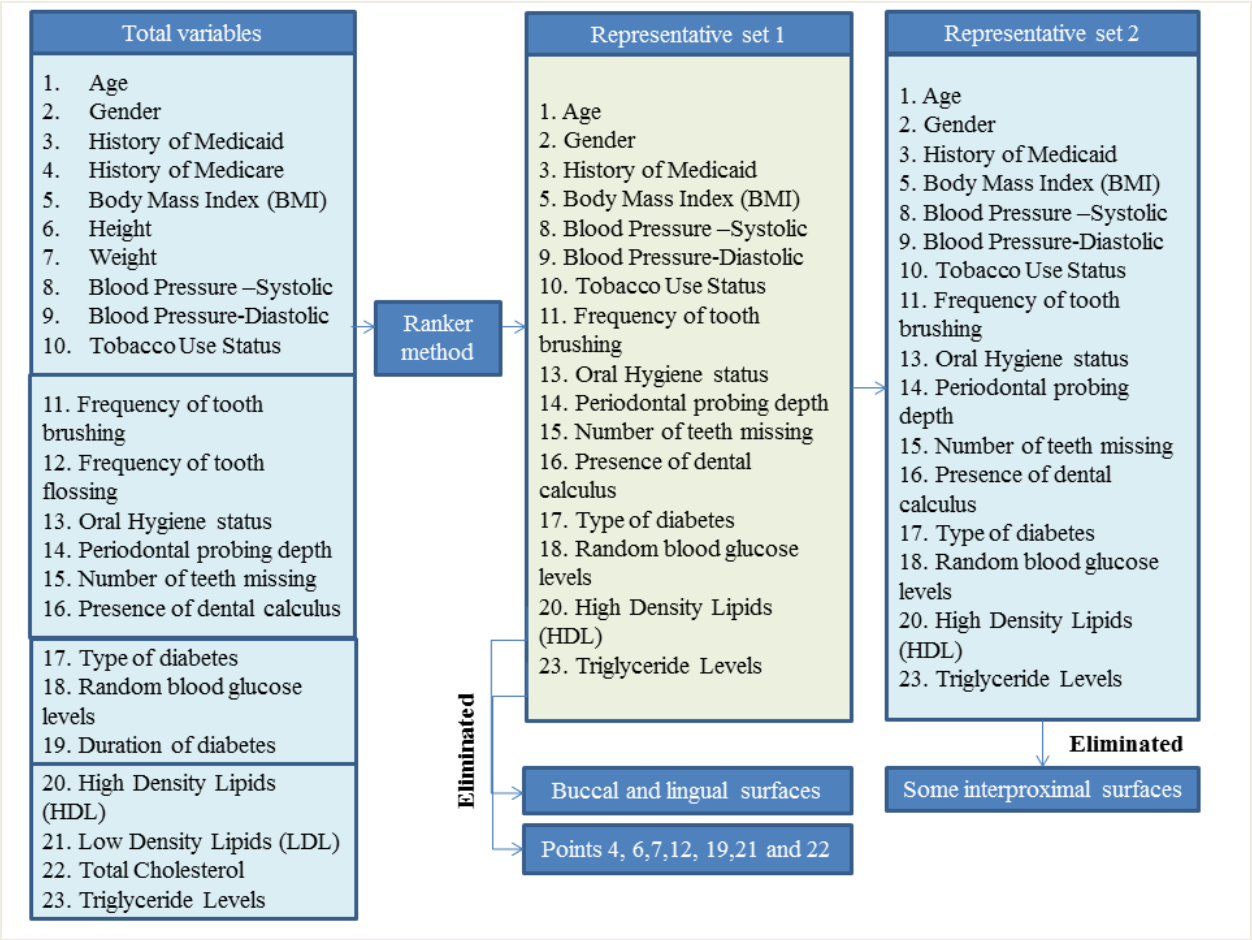


5.4.3 RESULTS OF FEATURE SELECTION

The rationale supporting model generation by feature selection was to establish the relative contribution of different tooth surfaces and the clinical attributes that define the best model for predicting a higher accuracy informed by the least number of features. This method eliminated the variables in two steps as shown in Figure 20

As shown in **Figure 23**, the first step of feature selection eliminated Medicare status, height, weight, frequency of flossing, LDL, total cholesterol and duration of diabetes. In the second step none of the clinical variables were eliminated, however 44 teeth surfaces were eliminated.

Figure 23 Elimination process by feature selection in Model of Care 1 (MOC 1)



The results of performance metrics after features selection for imbalanced dataset are shown in **Table 6**

Overall, an increase in recall was seen for all algorithms after feature selection. Correspondingly, the specificity increased noticeably in ANN, SVM and DT. There was a decrease in value of MCC for NB, SVM and LR, however showed an increase in ANN. MCC

and F-measure remained constant for DT before and after feature selection. DT outperformed other algorithms in terms of sensitivity 90.20% (95% CI 89.34-90.87) and specificity 90.25% (95% CI 89.29-91.15). There was an increase in the F-measure of ANN and NB from 0.785 to 0.841 and 0.785 to 0.800, respectively.

Table 6: Results of classifiers of MOC 1 imbalanced dataset after feature selection

ML	Accuracy % (95%CI)	Precision % (95%CI)	Recall % (95%CI)	Specificity % (95%CI)	F-measure x(95%CI)	AUC x(95% CI)	MCC x(95%CI)
NB	80.07% (79.09-81.05)	80.80% (80.79-80.81)	80.42% (79.41-81.41)	79.49% (78.17-80.70)	0.800 (0.79-0.81)	0.856 (0.857-0.858)	0.609 (0.57-0.63)
LR	78.68% (77.83-79.53)	78.60% (78.59-78.61)	80.78% (79.98-80.43)	78.35% (77.24-79.43)	0.785 (0.77-0.79)	0.855 (0.856-0.854)	0.560 (0.53-0.59)
SVM	79.00% (78.05-79.94)	79.00% (78.99-79.01)	81.27% (80.48-82.04)	80.90% (79.82-81.94)	0.787 (0.77-0.79)	0.808 (0.809-0.807)	0.566 (0.53-0.59)
ANN	84.13% (83.36-84.9)	84.10% (84.09-84.11)	84.10% (83.33-84.87)	84.35% (83.58-85.12)	0.841 (0.83-0.85)	0.895 (0.897-0.893)	0.674 (0.65-0.69)
DT	90.19% (89.24-90.89)	90.20% (90.19-90.21)	90.20% (89.34-90.87)	90.25% (89.29-91.15)	0.902 (0.91-0.89)	0.901 (0.902-0.899)	0.799 (0.77-0.81)
DT-Pruning	91.19% (90.52-91.87)	91.20% (91.18-91.22)	91.20% (91.10-92.15)	91.18% (91.08-92.12)	0.912 (0.90-0.92)	0.914 (0.915-0.913)	0.819 (0.79-0.83)

The AUC markedly increased from 0.852 to 0.895 for ANN and similarly, from 0.788 to 0.808 for SVM. Precision for SVM decreased marginally, whereas there was increase in precision for all other algorithms. Correspondingly, the total accuracy increased slightly for NB and ANN. The ranking performed by two tailed t-test ($\alpha = 0.05$) in terms of total accuracy in WEKA experimental ranked the algorithms in a descending order of their performance as DT>ANN> NB>SVM> LR. A cross-validated t-tests (alpha <0.05) on AUC, precision, recall, F-measure and MCC showed that DT and ANN outperformed NB, LR and SVM.

The results of balanced dataset after feature selection are shown in Figure 21.

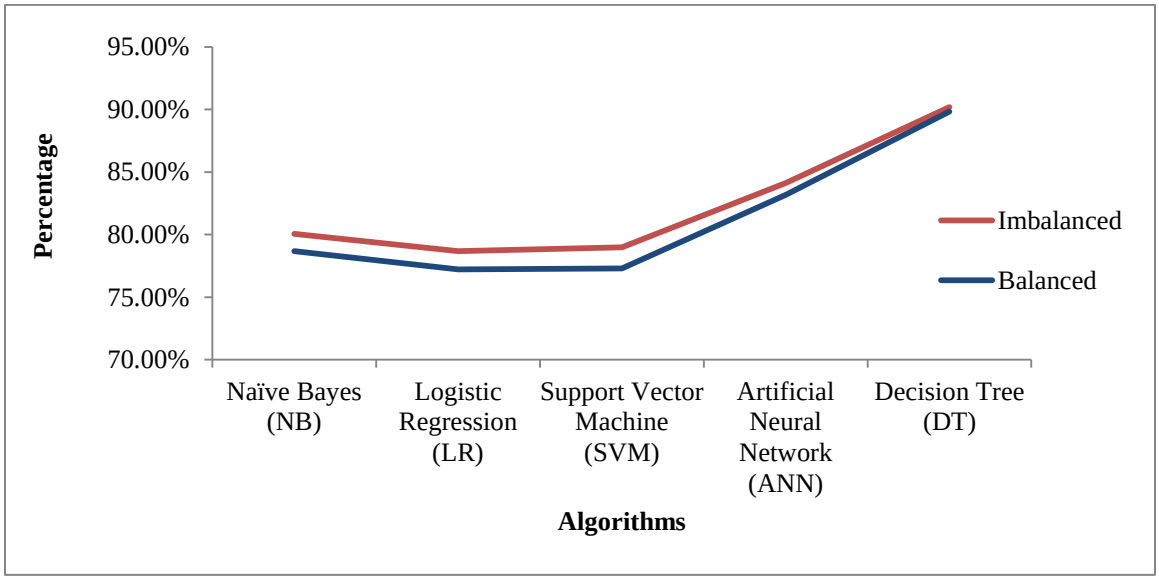
Overall, the precision of balanced dataset was lower than imbalanced dataset for all algorithms. The AUC of ANN was noticeably higher in balanced dataset as compared to the imbalanced dataset. In terms of sensitivity and specificity for SVM, the imbalanced dataset outperformed the balanced dataset. The MCC values for DT in the balanced and imbalanced dataset were almost equal. Correspondingly, the AUC for DT for both the dataset was almost similar.

Table 7: Results of classifiers of MOC 1 balanced dataset after feature selection

ML	**Accuracy % (95%CI)**	**Precision % (95%CI)**	**Recall % (95%CI)**	**Specificity % (95%CI)**	**F-measure x(95%CI)**	**AUC x(95% CI)**	**MCC x(95%CI)**
NB	78.69% (78.19-80.15)	79.20% (78.59-80.81)	78.70% (78.41-80.41)	77.51% (77.17-78.70)	0.786 (0.785-0.787)	0.863 (0.861-0.65)	0.579 (0.56-0.58)
LR	77.23% (76.67-78.42)	77.30% (77.19-78.12)	77.20% (76.98-78.41)	77.32% (76.24-78.43)	0.772 (0.771-0.773)	0.855 (0.856-0.854)	0.545 (0.53-0.55)
SVM	77.31% (76.03-78.67)	77.40% (76.89-78.04)	77.30% (76.48-78.04)	78.12% (77.80-79.92)	0.773 (0.771-0.774)	0.773 (0.771-0.774)	0.547 (0.53-0.56)
ANN	83.18% (82.36-83.95)	83.20% (83.09-83.11)	83.20% (82.32-83.85)	83.15% (82.57-84.11)	0.832 (0.830-0.850)	0.902 (0.901-0.903)	0.664 (0.63-0.67)
DT	89.83% (88.24-91.89)	89.80% (89.19-90.21)	89.80% (88.34-90.87)	88.16% (88.26-89.12)	0.898 (0.897-0.899)	0.899 (0.888-0.900)	0.797 (0.78-0.80)
DT-Pruning	89.92% (89.78-91.16)	90.20% (90.18-90.22)	90.20% (90.18-90.22)	89.99% (89.05-90.24)	0.902 (0.901-0.903)	0.901 (0.900-0.902)	0.794 (0.78-0.80)

Overall, the total accuracy for balanced dataset was lower as compared to the imbalanced dataset (as shown in Figure 22)

Figure 24: Comparison of total accuracy for balanced and imbalanced dataset in MOC 1 after feature selection showed that the imbalanced dataset performed better than balanced dataset.



A ranking for teeth surfaces based on their information gain was first plotted in a perio chart in descending order. Various surfaces of the teeth were color coded based on the information gain. **Figure 25** shows the periodontal chart and tooth surfaces based on the descending order of information gain.

Figure 25: Periodontal chart and tooth surfaces based on the descending order of information gain

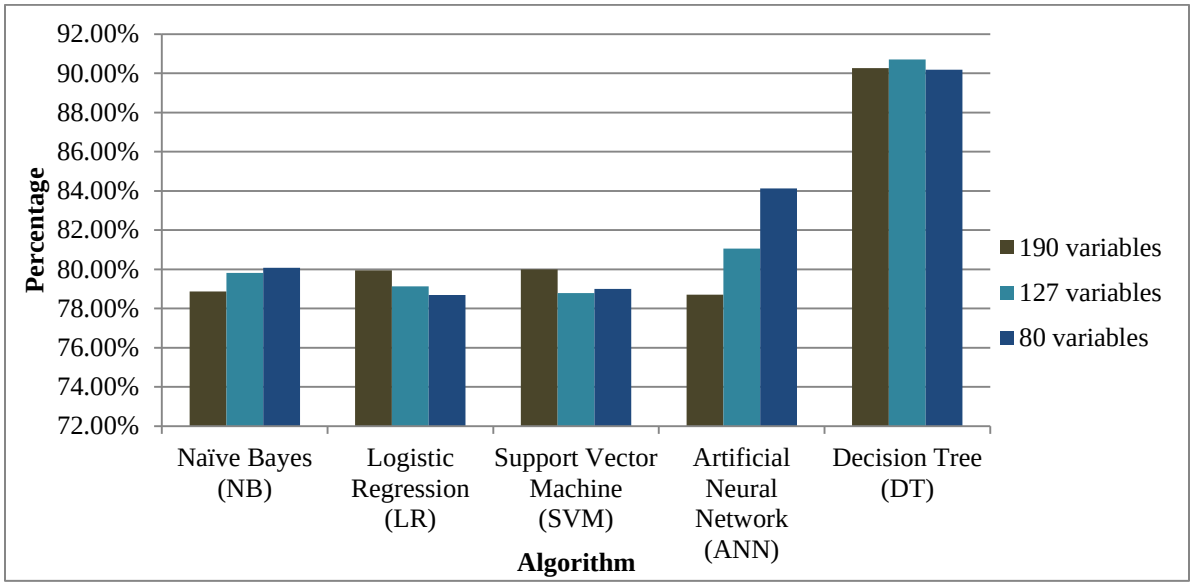
	ML	ML	ML	ML	ML	ML	ML		ML	ML	DL	ML	ML	DB	ML	
	MB	DB	DL	MB	MB	MB	DL		DL	MB	DB	DB	DB	DL	MB	
	L	DL	DB	DL	DB	DL	DB		DB	DB	ML	DL	DL	ML	DL	
	DB	MB	MB	DB	DL	DB	MB		MB	DL	MB	MB	MB	MB	DB	
	DL	L	L	L	L	L	L		L	L	L	L	L	L	L	
	B	B	B	B	B	B	B		B	B	B	B	B	B	B	
1	2	3	4	5	6	7	8		9	10	11	12	13	14	15	16
32	31	30	29	28	27	26	25		24	23	22	21	20	19	18	17
	B	B	B	B	B	L	L		L	B	B	B	B	B	B	
	L	L	L	L	L	B	B		B	L	L	L	L	L	L	
	MB	MB	MB	DB	DL	ML	ML		ML	ML	ML	MB	MB	MB	MB	
	DB	DB	DL	DL	DB	MB	DL		MB	MB	DL	DL	DB	DB	DB	
	DL	DL	DB	MB	ML	DL	DB		DB	DL	DB	DB	DL	DL	DL	
	ML	ML	ML	ML	MB	DB	MB		DL	DB	MB	ML	ML	ML	ML	

ML= Mesiolingual, MB= Mesiobuccal, DL=Distolingual, DB= Distobuccal, L=Lingual and B=Buccal. The arrow points to the descending order of the gain ratio of the teeth surfaces for upper and lower jaw. The numbers indicate the teeth number.

The first elimination in feature selection resulted in exclusion of the orange color (56 attributes) i.e. all the lingual and buccal surfaces of the teeth. The second elimination resulted in removal of blue color (44 attributes) including mesiolingual (ML), distolingual (DL) and distobuccal (DB) surfaces of all mandibular anterior teeth and mesiobuccal (MB) surface of lower central incisors. The interproximal surfaces including MB, DB and DL of maxillary incisors, first premolars and second premolars were excluded in second elimination of the feature selection. Similarly the MB of left maxillary canine and DL of right maxillary canine were removed.

The first elimination of features resulted in 127 features. The second elimination of features included the blue color tooth surfaces in Figure 19 resulted in the 80 attributes. The total accuracy for imbalanced datasets for all the algorithms with first and second elimination was compared with the total accuracy of the original dataset and is shown in Figure 20.

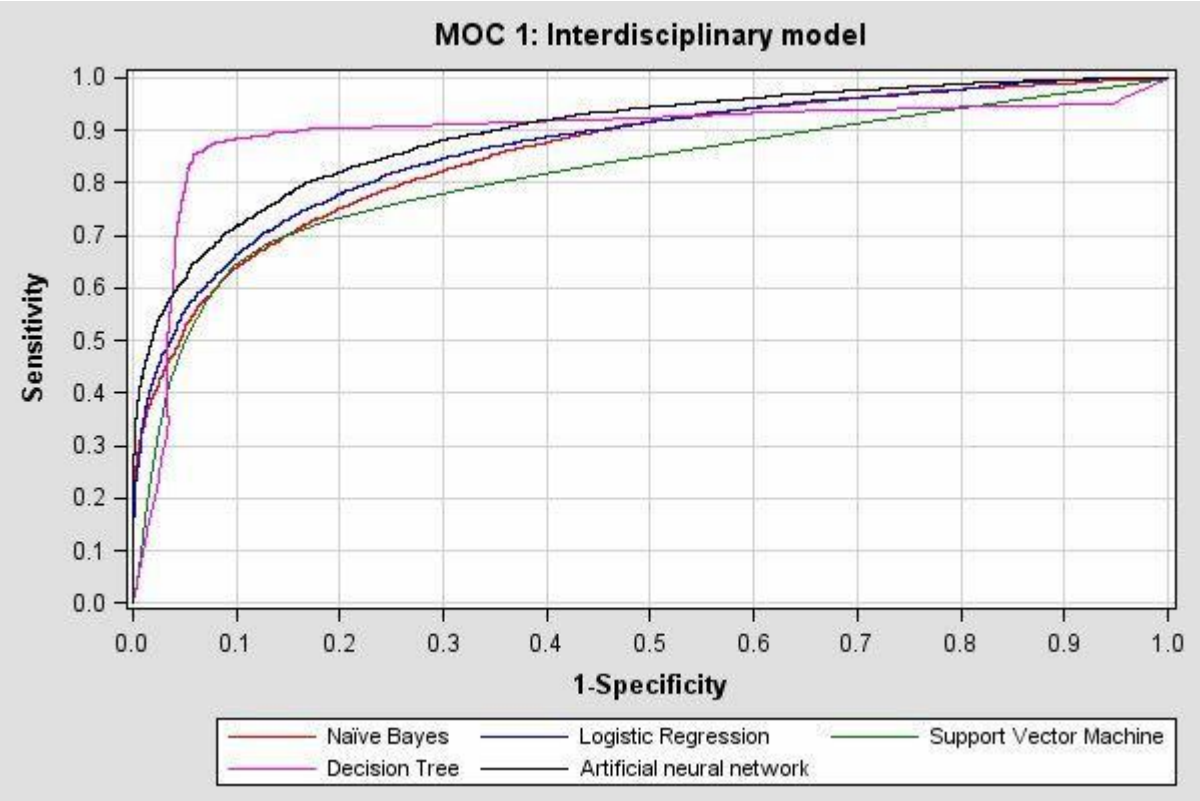
Figure 26: Comparison of Imbalanced dataset for number of variables before and after feature selection in MOC 1



The results of first elimination of the feature selection show slight increase in the accuracy in terms of total accuracy for DT, ANN and NB. However, there was a decrease in accuracy for LR and SVM. The accuracy levels were consistent for DT with a slightly higher accuracy with first elimination of the features. A comparison was also made for the total accuracy for imbalanced and balanced set after feature selection. The results show that total accuracy for imbalanced dataset were higher than balanced dataset.

To assess and compare the performance of the classifiers over their entre operating range, a plot of sensitivity versus 1-Specifity was plotted for all the algorithms for imbalanced dataset. Figure 24 shows the distribution of ROC curves for NB, LR, ANN, SVM and DT.

Figure 27: Receiver Operating Characteristic curves for algorithms in MOC 1



The ROC for DT is higher than ANN, NB, LR and SVM. The ROC of DT begins at (0, 0) and eventually bends towards the right at (0.02, 0.2) and then runs vertical to (0.05, 0.85) indicating more true positives than false positives and correspondingly signaling a greater noise. LR, ANN and NB show slightly symmetric curve as compared to DT. The ROC curves of algorithms in descending order of their performance are DT>ANN>NB> SVM.

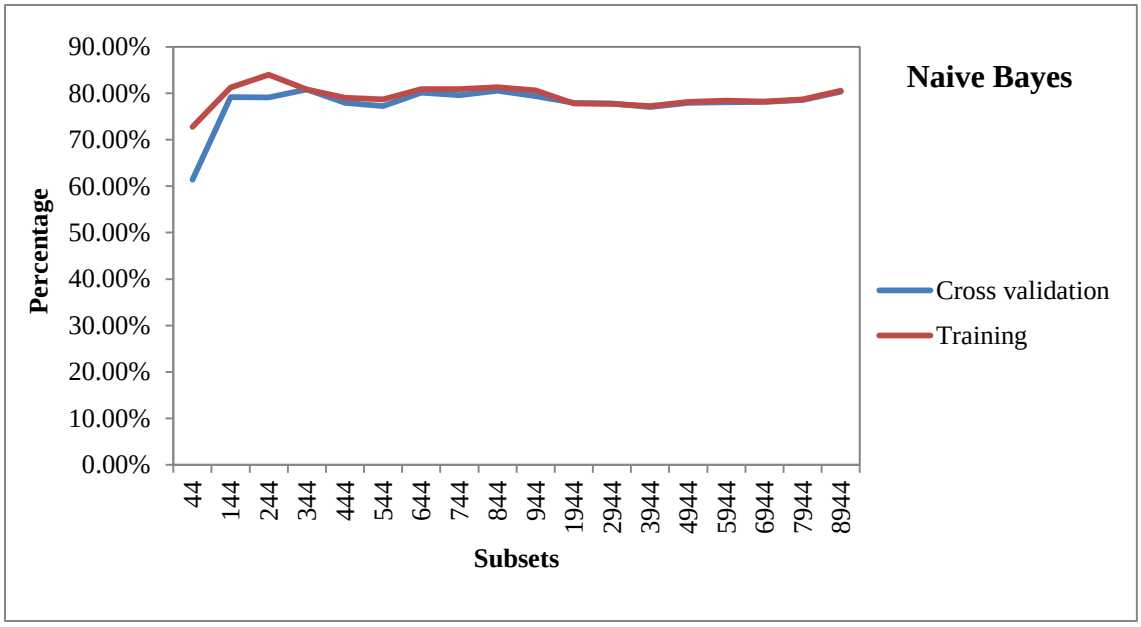
5.4.4. LEARNING CURVES FOR VARIOUS SUPERVISED ALGORITHMS

The performance of cross validation sets was plotted against training sets for each of the algorithms is shown below:

5.4.4.1.NAÏVE BAYES

Figure 28 shows the learning curve for NB. Although, NB assumes that the attributes are conditionally independent given the class value, however in real world scenario, this assumption may not be valid and can degrade the accuracy of NB. The training performance and testing performance are very close, indicating that there is almost no "overfitting" even when 2000 training examples are used. The performance slightly changes when the training sample size changes from 2000 to 3000. This probably shows that about 2000 training samples would be sufficient to train NB.

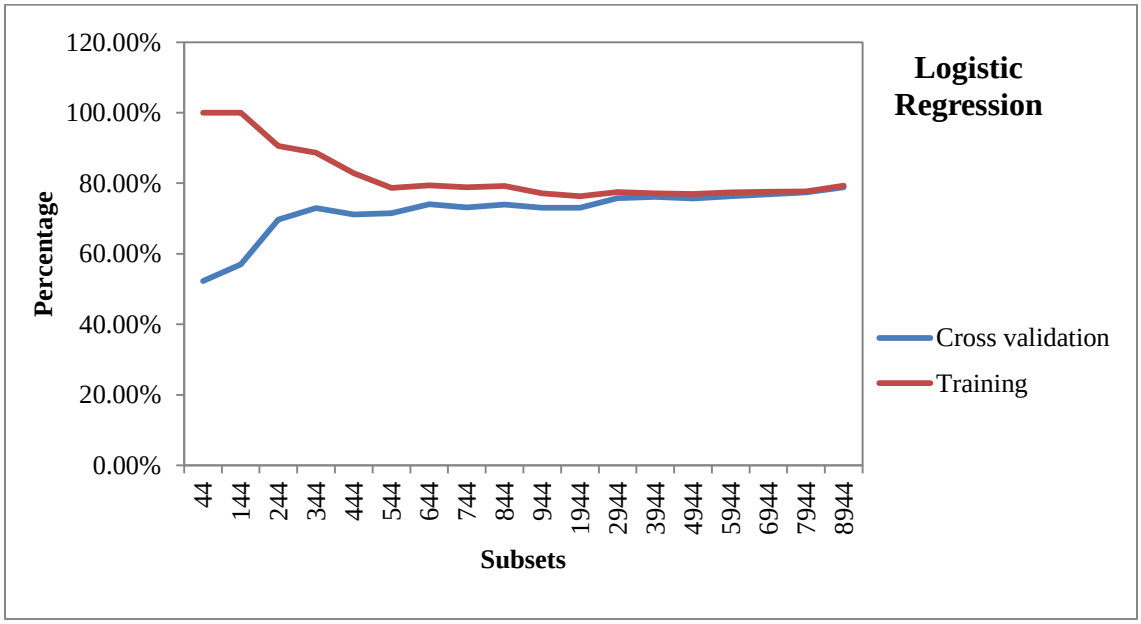
Figure 28: Learning Curve for Naive Bayes in MOC1



5.4.4.2. LOGISTIC REGRESSION

The results of the learning curve are shown in figure 29 for logistic regression. The training performance and testing performance are very close, indicating that there is almost no "overfitting" when 4000 training examples are used.

Figure 29: Learning Curve for Logistic Regression in MOC1

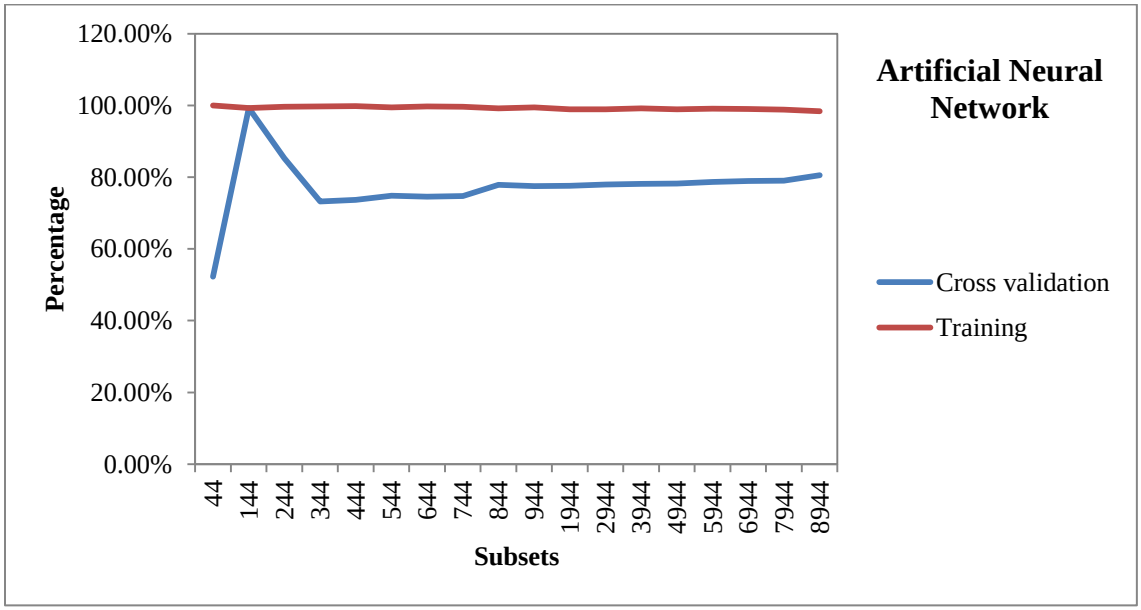


5.4.4.3.ARTIFICIAL NEURAL NETWORK

The results of the learning curve are shown in figure 27 for artificial neural network. The neural network models the relationship between outcome variables (class) and attributes using nonlinear functions and hence tends to fit any function. The training performance and testing performance are not close, indicating that there is an "overfitting". The findings show that the training performance is constant with a very high accuracy. Although the sample increases, the performance for training sample remains constant, thus representing a low bias and high

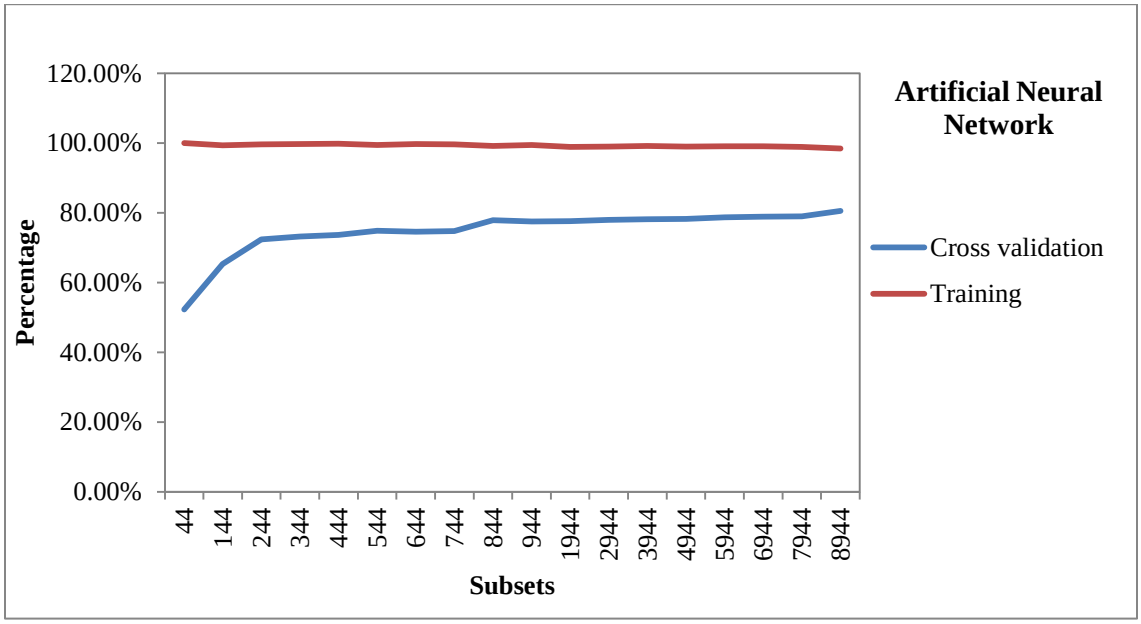
variance. On the other hand the cross validation set shows that as the sample size increases the accuracy increases, however the sample size for neural network may not seem to be sufficient.

Figure 30: Learning Curve for Artificial Neural Network in MOC1



Due to randomness of the instances, there is a possibility of differences in performances at various sample sizes. Such difference was seen for accuracy of ANN for cross validation set with a sample size 144 reaching an accuracy of 99.31% and sample size 244 reaching accuracy of 85.24%. To overcome such randomness, four different datasets of 144 and 244 samples, respectively were again trained and cross validated and their respective accuracy was averaged.

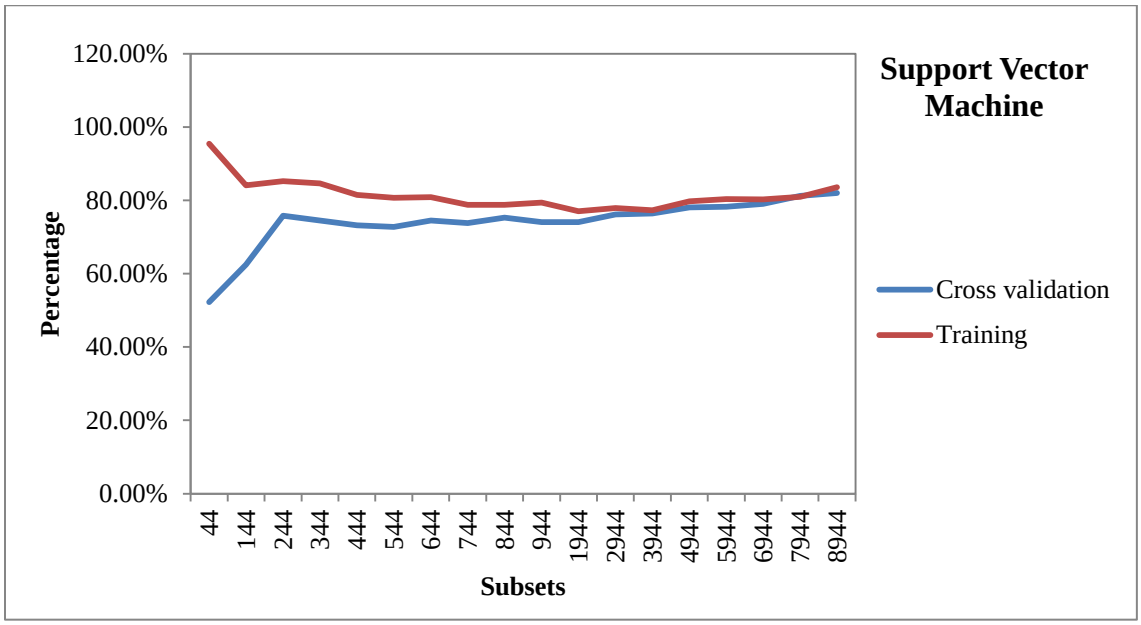
Figure 31 shows the learning curve for ANN after averaging the results of training and cross validation sets with sample size 144 and 244.



5.4.4.4 SUPPORT VECTOR MACHINE

The results of the learning curve are shown in figure 32 for support vector machine. As the default kernel function used in SVM was linear, this algorithm assumes relationship between class and the attribute as linear. As a result of this, SVM has similar assumption as that of LR and thus restricts the ability to fit the training data. The figure shows that as the sample increases the accuracy of training gradually decreases and the accuracy of the cross validation set increases. This case shows an example of high bias and low variance.

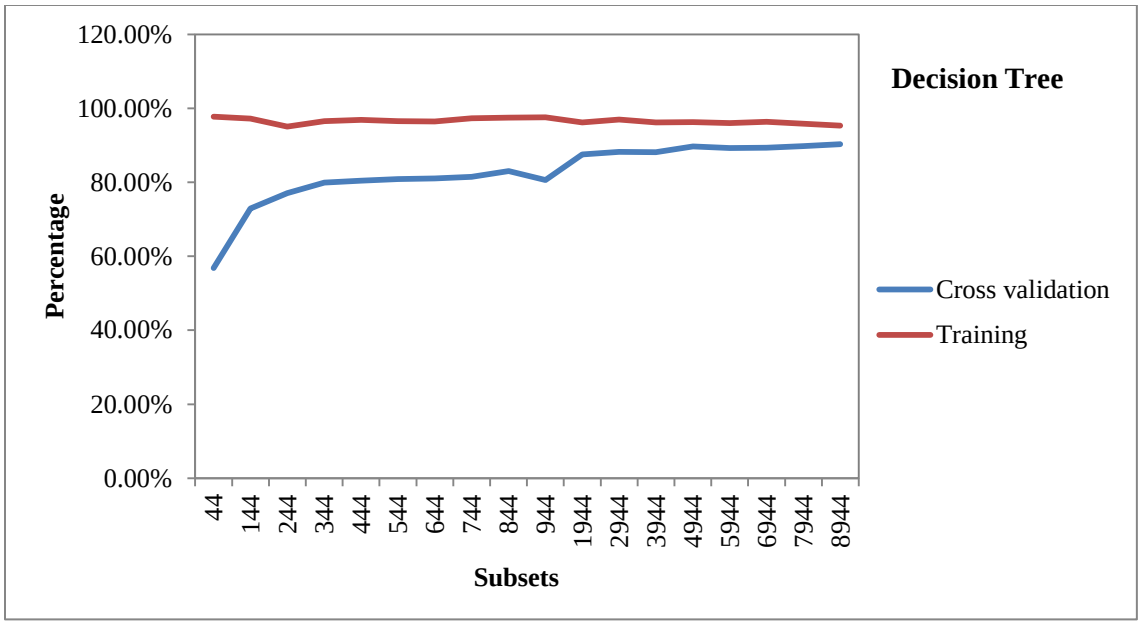
Figure 32: Learning Curve for Support Vector Machine in MOC1



5.4.4.4. DECISION TREE

The results of the learning curve are shown in figure 33 for decision tree. Decision tree is able to model nonlinear functions and works through segmentation process by splitting data into segments. The results of the learning show that DT has a low variance and low bias. The training performance remains constant while the DT cross validation set shows a slow learning

Figure 33: Learning Curve for Decision Tree in MOC1



5.4.5 COMPARISON OF DECISION TREE AND ENSEMBLES

The results of the performance of the ensembles are shown in table 8

Ensemble employing bagging on DT outperformed all the other models. The AUC was highest for Voting algorithm followed by B-DT and DT. B- DT indicated higher sensitivity and specificity as compared to voting and DT. Overall accuracy and precision was similar for Voting and B-ANN. In terms of MCC, DT outperformed B-DT and Voting. The AUC of B-DT was highest, followed by DT and voting.

Table 8: Results of proposed experiment for ensemble methods

Method	ML	Accuracy	Precision	Recall	Specificity	F-measure	AUC	MCC
Bagging (B)	B-NB	79.42%	79.30%	79.40%	78.29%	0.793	0.862	0.575
	B-LR	78.60%	78.50%	78.60%	77.33%	0.784	0.857	0.559
	B-SVM	78.74%	78.80%	78.70%	79.19%	0.785	0.810	0.561
	B-ANN	83.97%	84.00%	84.00%	82.65%	0.838	0.911	0.670
	B-DT	91.96%	92.00%	92.00%	92.15%	0.919	0.966	0.835
Voting		84.92%	85.00%	84.90%	84.12%	0.932	0.690	0.845
DT		90.19%	90.20%	90.20%	90.20%	0.896	0.799	0.899

5.4.5. VALIDATION OF RESULTANT MODELS BY AN EVALUATION SET ON IMBALANCED DATASETS

Table 9: Performance of Predictive Modeling for MOC 1 on imbalanced dataset by external evaluation set

ML	Accuracy	Precision	Recall	Specificity	F-measure	AUC	MCC
NB	80.07 %	80.80%	80.10%	79.10%	0.800	0.883	0.609
LR	78.80%	79.60%	78.80%	77.12%	0.787	0.868	0.584
SVM	78.08%	79.20%	78.10%	78.81%	0.779	0.781	0.573
ANN	84.60%	84.80%	84.60%	84.60%	0.846	0.923	0.920
DT	90.31%	90.50%	90.30%	90.30%	0.903	0.896	0.808

The results of validation show that DT could analyze the true positive and true negative cases better than NB, LR, SVM and ANN. The AUC and MCC were highest for ANN followed by DT, NB, SVM and LR.

5.4.7. SUMMARY

This dataset contained 11,048 instances with 190 variables. Of these 11,048, 10% (1,104) were used for external validation of the resultant model for imbalanced dataset. Other 9,944 instances were used for 10 fold cross validation. Based on the feature selection, the orange color (56 attributes) teeth surfaces along with attributes such as duration of diabetes, height and weight, Medicare status, total cholesterol, LDL, frequency of flossing were first eliminated. After second elimination, the dataset was left with 80 variables. Four different experiments were performed on this dataset. The results of balanced and imbalanced dataset showed that

imbalanced dataset outperformed balanced dataset regarding all performance metrics. Ensemble method showed that B-DT outperformed voting and DT, however F-measure was highest for voting and AUC was highest for B-ANN when compared to other algorithms. Of all the learners evaluated, the learning curves in NB, LR and SVM using linear kernel reached a plateau much earlier than other methods, suggesting that perhaps the low variance is a consequence of achieving consistent performance and further suggests that additional data does not improve results. Based on the results of this experiment, a progressively larger sample size is required for ANN to yield a higher accuracy with low variance and low bias.

The results of 10 fold cross validation for the original DT of MOC 1 containing 190 variables, yielded 317 leaves and 614 as the size of tree (number of nodes). With a confidence factor of 0.25 and minNumObj of 2, the root node was mesiolingual surface of tooth number 31. The interproximal surfaces were ranked near the top of the tree representing the important factors for predicting PD risk. This was followed by the (internal node) medical variable 'random blood glucose level'. At the internal node of random blood glucose level, the test condition showed >111 mg/dl as 'high risk' and $\leq$ 111 mg/dl was traversed to mesiolingual surfaces of tooth 28. Notably, the occurrence of random blood glucose levels after the interproximal surfaces represents the bidirectional association of PD and diabetes. Moreover, blood glucose levels with more than 111 mg/dl indicating a 'high risk' aligns with the blood glucose range indicating prediabetes. **Figure 34** shows the pruned MOC 1 decision tree. The dataset for MOC 1 containing 190 variables was pruned for the purpose of readability by lowering the confidence factor from 0.25 to 0.20 and increasing the minNumObj in WEKA from 2 to 15. This yielded a total number of 34 leaves and the size of tree was 66. All the nodes in the tree represent tooth

surfaces of PPD except for the patient reported tobacco use status which occurs at the 16th branch. The misclassification ratio for mesiolingual surface of tooth number 4 was highest in the tree (5744/362).

Figure 34 shows the pruned MOC 1 decision tree (J4.8) to identify the most important parameters that would influence the PD risk generated in WEKA.

```
MesialLingual31 <= 4
| DistalBuccal14 <= 4
| | MesialLingual2 <= 4
| | | MesialLingual18 <= 4
| | | | MesialLingual29 <= 4
| | | | | MesialBuccal13 <= 4
| | | | | | MesialLingual15 <= 4
| | | | | | | MesialLingual19 <= 4
| | | | | | | | MesialLingual3 <= 4
| | | | | | | | | MesialBuccal22 <= 4
| | | | | | | | | | MesialLingual30 <= 4
| | | | | | | | | | | MesialBuccal18 <= 4
| | | | | | | | | | | | MesialBuccal5 <= 4
| | | | | | | | | | | | | MesialBuccal2 <= 4
| | | | | | | | | | | | | | DistalBuccal25 <= 4
| | | | | | | | | | | | | | | DistalBuccal15 <= 4
| | | | | | | | | | | | | | | | MesialLingual14 <= 4
| | | | | | | | | | | | | | | | | DistalBuccal18 <= 4
| | | | | | | | | | | | | | | | | | DistalBuccal26 <= 4
| | | | | | | | | | | | | | | | | | | DistalBuccal2 <= 4
| | | | | | | | | | | | | | | | | | | | MesialLingual4 <= 4: LOW (5744.0/362.0)
| | | | | | | | | | | | | | | | | | | | MesialLingual4 > 4: HIGH (34.0/9.0)
| | | | | | | | | | | | | | | | | | | DistalBuccal2 > 4: HIGH (26.0/9.0)
| | | | | | | | | | | | | | | | | | DistalBuccal26 > 4: HIGH (35.0/5.0)
| | | | | | | | | | | | | | | | | DistalBuccal18 > 4
| | | | | | | | | | | | | | | | | | DistalBuccal31 <= 4
| | | | | | | | | | | | | | | | | | | Lingual10 <= 2
| | | | | | | | | | | | | | | | | | | | Tobacco use status = Never : HIGH (21.0/6.0)
| | | | | | | | | | | | | | | | | | | | Tobacco use status = Former : HIGH (25.0/12.0)
| | | | | | | | | | | | | | | | | | | | Tobacco use status = Current: LOW (20.0/4.0)
| | | | | | | | | | | | | | | | | | | Lingual10 > 2: HIGH (15.0/3.0)
| | | | | | | | | | | | | | | | | | DistalBuccal31 > 4: HIGH (17.0)
| | | | | | | | | | | | | | | | | MesialLingual14 > 4: HIGH (48.0/14.0)
| | | | | | | | | | | | | | | | DistalBuccal15 > 4: HIGH (38.0/10.0)
| | | | | | | | | | | | | | | DistalBuccal25 > 4: HIGH (30.0/1.0)
| | | | | | | | | | | | | | MesialBuccal2 > 4: HIGH (43.0/10.0)
```

| | | | | | | | | | | MesialBuccal5 > 4: HIGH (36.0/5.0)
| | | | | | | | | | | MesialBuccal18 > 4: HIGH (54.0/11.0)
| | | | | | | | | | MesialLingual30 > 4: HIGH (90.0/19.0)
| | | | | | | | | MesialBuccal22 > 4: HIGH (83.0/7.0)
| | | | | | | | MesialLingual3 > 4: HIGH (131.0/24.0)
| | | | | | | MesialLingual19 > 4: HIGH (80.0/9.0)
| | | | | | MesialLingual15 > 4: HIGH (163.0/23.0)
| | | | | MesialBuccal13 > 4: HIGH (68.0/3.0)
| | | | MesialLingual29 > 4: HIGH (173.0/13.0)
| | | MesialLingual18 > 4: HIGH (320.0/26.0)
| | MesialLingual2 > 4
| | | MesialLingual3 <= 4
| | | | MesialLingual15 <= 4
| | | | | MesialLingual18 <= 4
| | | | | | MesialBuccal27 <= 3
| | | | | | | MesialLingual5 <= 3
| | | | | | | | MesialLingual20 <= 3
| | | | | | | | | DistalBuccal3 <= 3
| | | | | | | | | | DistalBuccal31 <= 3: LOW (30.0/7.0)
| | | | | | | | | | DistalBuccal31 > 3: HIGH (16.0/3.0)
| | | | | | | | | DistalBuccal3 > 3: HIGH (57.0/9.0)
| | | | | | | | MesialLingual20 > 3: HIGH (37.0/3.0)
| | | | | | | MesialLingual5 > 3: HIGH (60.0/3.0)
| | | | | | MesialBuccal27 > 3: HIGH (75.0/2.0)
| | | | | MesialLingual18 > 4: HIGH (46.0)
| | | | MesialLingual15 > 4: HIGH (72.0)
| | | MesialLingual3 > 4: HIGH (130.0)
| DistalBuccal14 > 4: HIGH (546.0/15.0)
MesialLingual31 > 4: HIGH (1581.0/55.0)

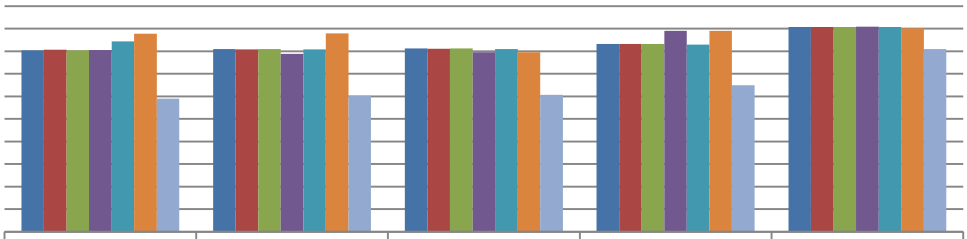
5.5 MODEL OF CARE 2: DENTAL ONLY

5.5.1 PATIENT CHARACTERISTICS

The overall mean age of the patients was 43.99 ± 16.70, with 64% (9643/15092) of patients being female. Of these, 12,418 were Medicaid patients and 4878 were Medicare patients. Mean brushing frequency was 1.6 ± 0.59 while flossing frequency was 0.67 ± 0.79. Patients with poor, fair, good and excellent oral hygiene was: 2360; 6952; 5357 and 269, respectively. The mean dental calculus determination for patients was 1.5 ± 0.69. Of these 15,092 patients, 4926 were current tobacco smokers, 4817 were former smoker and 5349 never smoked. The mean number of missing teeth was 3± 5.01. More than half of the patients showed a normal blood pressure levels. Approximately, 7390 patients were obese with a BMI of more than 30, whereas about 3880 patients were overweight with a BMI between 25 and 29.99.

5.5.2 RESULTS OF APPLICATION OF THE FIVE ALGORITHMS ON ALL THE VARIABLES

Figure 35: Results of performance metrics of application of five algorithms on MOC 2.



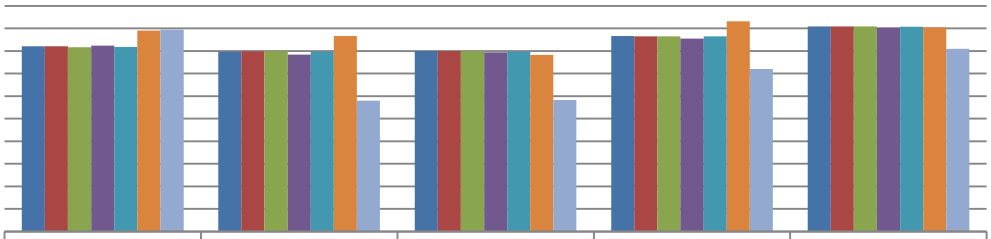
	NB	LR	SVM	ANN	DT
Accuracy	80.40%	80.92%	81.16%	83.18%	90.80%
Precision	80.60%	80.80%	81.10%	83.20%	90.80%
Recall	80.40%	80.90%	81.20%	83.20%	90.80%
Specificity	80.45%	78.80%	79.50%	89.10%	90.82%
F-measure	0.844	0.808	0.809	0.830	0.907
AUC	0.877	0.879	0.795	0.891	0.902
MCC	0.59	0.602	0.606	0.649	0.809

The accuracies for the classifiers: NB, LR, SVM, ANN and DT were 80.40%, 80.92%, 81.16%, 83.18% and 90.80%, respectively. The F-measure for LR and SVM were almost equal. The various performance metrics show that DT outperformed all the other classifiers. The sensitivity and specificity for DT were 90.80% (95% CI=90.22% to 91.41%) and 90.82% (95% CI 89.94% to 91.44%), respectively. The time taken to build the ANN model was 3,684 seconds. The AUC for DT was 0.902 followed by ANN, LR, NB and SVM. A paired t-test showed that there was no significant difference in terms of total accuracy between NB and LR , however DT was significantly higher in terms of accuracy as compared to other algorithms ($p<0.001$). The AUC for NB and LR was almost similar.

5.5.3 RESULTS OF FEATURE SELECTION

Variables including insurance status such as Medicare and Medicaid, frequency of flossing, systolic blood pressure and BMI were eliminated after feature selection method. The results of performance metrics after feature selection are shown in **Figure 36**

Figure 36: The results of performance metrics after feature selection for MOC 2



	NB	LR	SVM	ANN	DT
Accuracy	82.04%	79.69%	80.02%	86.55%	90.86%
Precision	82.10%	79.80%	80.00%	86.50%	90.90%
Recall	81.69%	79.90%	80.00%	86.50%	90.93%
Specificity	82.41%	78.35%	79.38%	85.43%	90.25%
F-measure	0.818	0.797	0.797	0.865	0.908
AUC	0.891	0.866	0.783	0.932	0.905
MCC	0.894	0.579	0.582	0.720	0.810

Overall, a slight increase in total accuracy was seen in NB, ANN and DT after feature selection. There was a decrease in value of MCC for SVM and LR, however showed a marked increase in ANN and NB. The F-measure remained constant for DT before and after the feature selection. DT outperformed other algorithms in terms of sensitivity 90.93% (95% CI 89.32 to 91.51) and specificity 90.25% (95% CI 89.47-91.00), however the ANN outperformed other algorithms in terms of AUC. There was an increase in the F-measure of ANN from 0.830 to 0.865. Similarly, the F-measure for DT increased slightly by 0.01. Precision for SVM and NB decreased marginally, whereas there was increase in precision for all other algorithms. Correspondingly, the total accuracy for all the algorithms increased slightly except for LR and SVM where a decrease in total accuracy was seen. The ranking performed by two tailed t-test ($\alpha = 0.05$) in WEKA experimental ranked the algorithms in a descending order of their performance as DT>ANN> NB>SVM> LR. A cross-validated t-tests (alpha <0.05) on AUC, total accuracy, precision, recall, F-measure and MCC showed that DT and ANN outperformed NB, LR and SVM. The perio chart with the representative surfaces of teeth are shown in **Figure 37.**

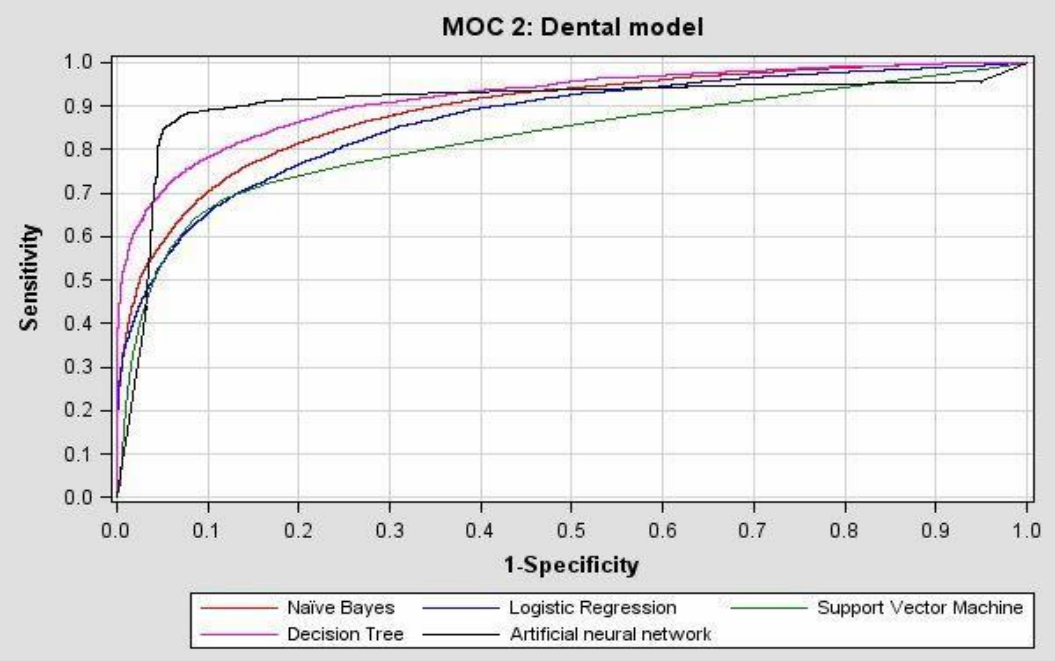
Figure 37: Periodontal chart and tooth surfaces to show the significant tooth surfaces after feature selection application to MOC 2

Figure 37: Periodontal chart and tooth surfaces to show the significant tooth surfaces after feature selection application to MOC 2															
	ML	ML	ML	ML	ML	ML	ML		ML	ML	DL	ML	ML	DB	ML	
	MB	DB	DL	MB	MB	MB	DL		DL	MB	DB	DB	DB	DL	MB	
	DL	DL	DB	DL	DB	DL	DB		DB	DB	ML	DL	DL	ML	DL	
	DB	MB	MB	DB	DL	DB	MB		MB	DL	MB	MB	MB	MB	DB	
	L	L	L	L	L	L	L		L	L	L	L	L	L	L	
	B	B	B	B	B	B	B		B	B	B	B	B	B	B	
1	2	3	4	5	6	7	8		9	10	11	12	13	14	15	16
32	31	30	29	28	27	26	25		24	23	22	21	20	19	18	17
	B	B	B	B	B	L	L		L	B	B	B	B	B	B	
	L	L	L	L	L	B	B		B	L	L	L	L	L	L	
	MB	MB	MB	DB	DL	ML	ML		ML	ML	ML	MB	MB	MB	MB	
	DB	DB	DL	DL	DB	MB	DL		MB	MB	DL	DL	DB	DB	DB	
	DL	DL	DB	MB	ML	DB	DB		DL	DL	DB	DB	DL	DL	DL	
	ML	ML	ML	ML	MB	DL	MB		DB	DB	MB	ML	ML	ML	ML	

The elimination in feature selection resulted in exclusion of the blue color (100 attributes) i.e. all the lingual and buccal surfaces of the teeth (56 attributes), interproximal surfaces including ML, DL and DB surfaces of all mandibular central incisors along with MB of tooth 25. The interproximal surfaces including MB, DB and DL of maxillary central and lateral incisors were eliminated after application of feature selection. Similarly the MB of left maxillary canine and DL of right maxillary canine were removed, as seen in MOC 1 and 3.

The ROC curve performance of each classifier for MOC 2 is shown in **Figure 38**.

Figure 38: Receiver Operating Characteristics curve for algorithms in MOC2



The ROC for ANN is higher than DT, NB, LR and SVM. The ROC of ANN begins at (0, 0) and eventually bends towards the right at (0.02, 0.5) and then runs vertical to (0.05, 0.85) indicating more true positives than false positives and correspondingly signaling a greater noise. The ROC curves of algorithms in descending order of their performance are ANN>DT>NB> LR>SVM.

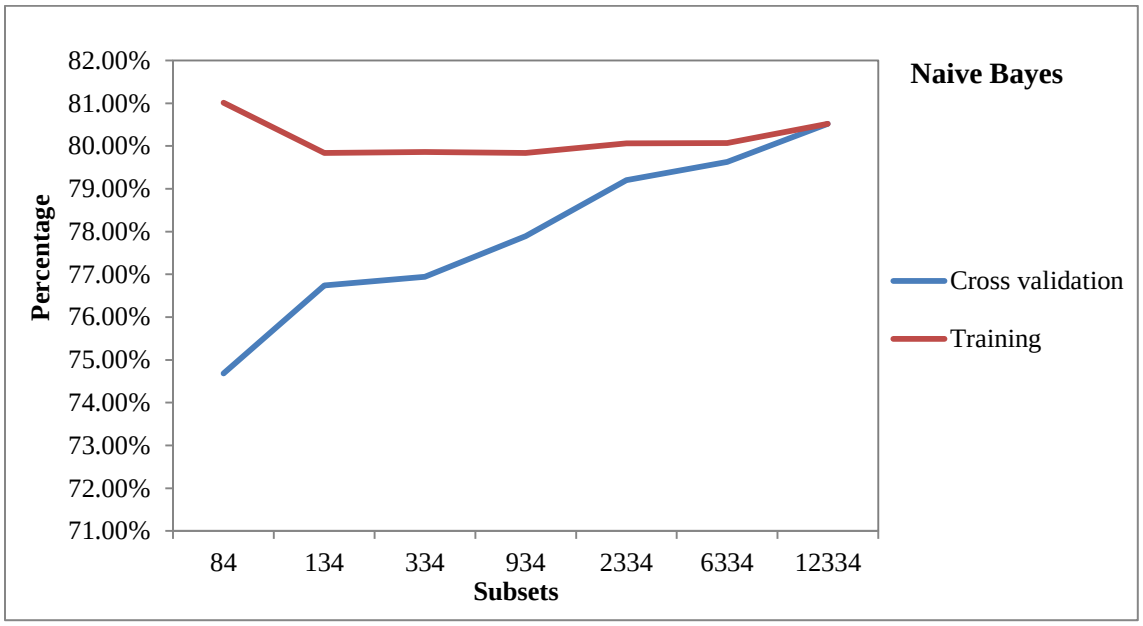
5.5.4 LEARNING CURVES FOR VARIOUS ALGORITHMS

The performance of cross validation sets was plotted against training sets for each of the algorithms is shown below:

5.5.4.1. NAÏVE BAYES

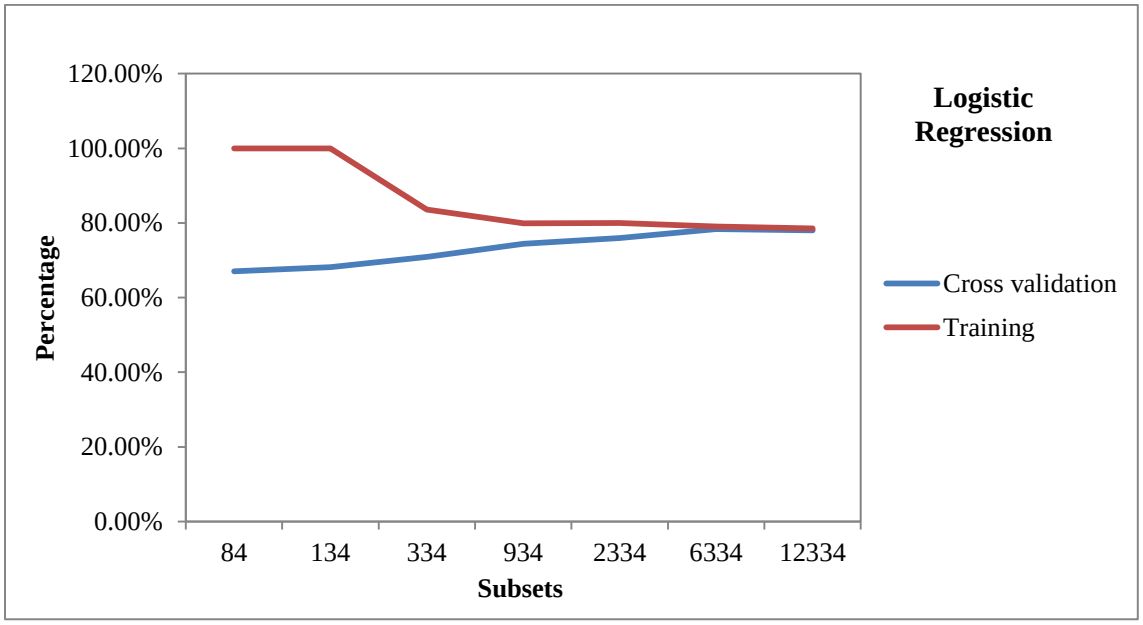
The training performance and testing performance are very close, indicating that there is almost no "overfitting" at a training sample size of approximately 12,000.

Figure 39: Learning curve for Naive Bayes in MOC2



5.5.4.2. LOGISTIC REGRESSION

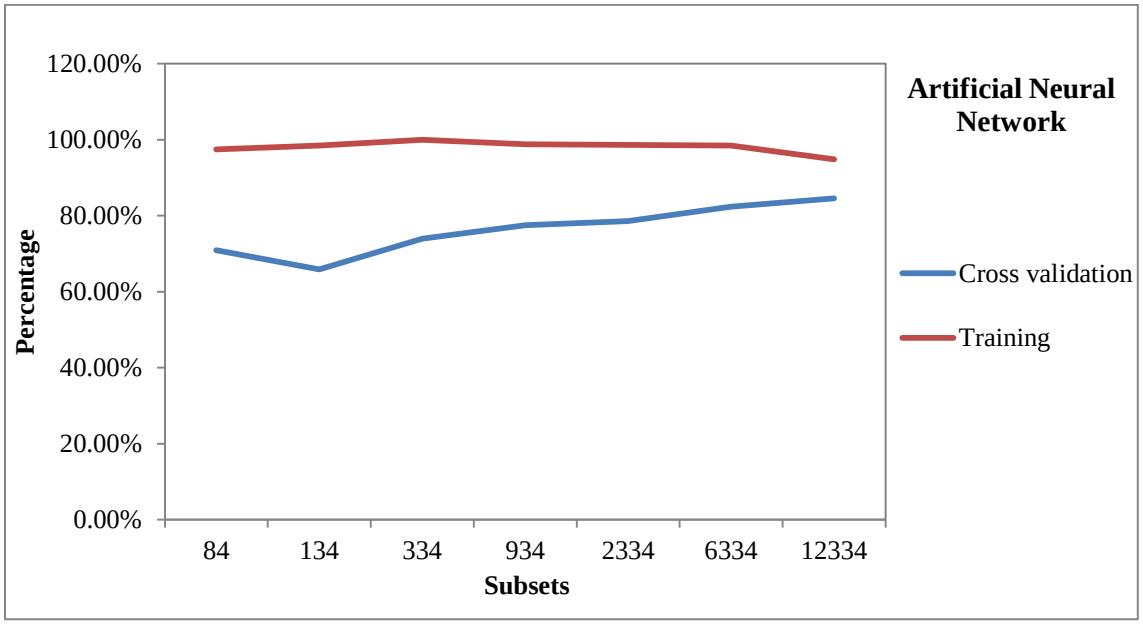
Figure 40: Learning curve for Logistic Regression in MOC2



The training curve for cross validation set for LR shows that the algorithm slowly learns while the training set shows a marked decrease in performance when the sample size increased from 334 to 934.

5.5.4.3. ARTIFICIAL NEURAL NETWORK

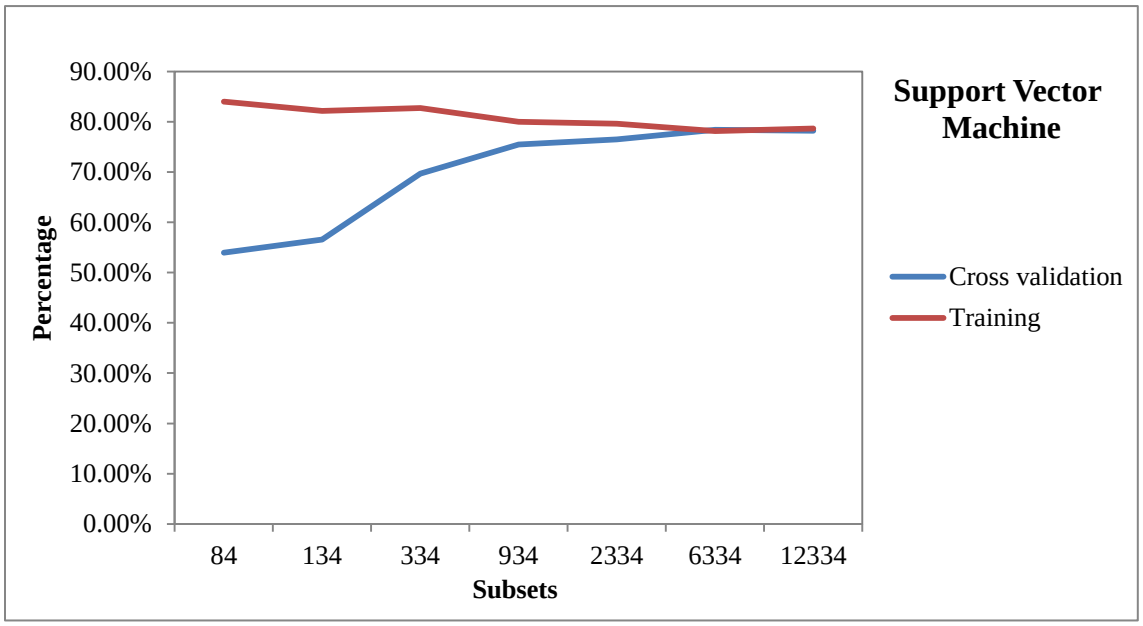
Figure 41: Learning curve for Artificial Neural Network in MOC2\



The results of the learning curve are shown in Figure 37 for ANN. Similar to MOC 1, the training performance and testing performance are not close, indicating that there is an "overfitting". The findings show that the training performance is constant with a very high accuracy. Although the sample increases, the performance for training sample remains constant, thus representing a low bias and high variance. On the other hand the neural network shows that as the sample size increases the accuracy increases, however the sample size for neural network may not seem to be sufficient.

5.5.4.4. SUPPORT VECTOR MACHINE

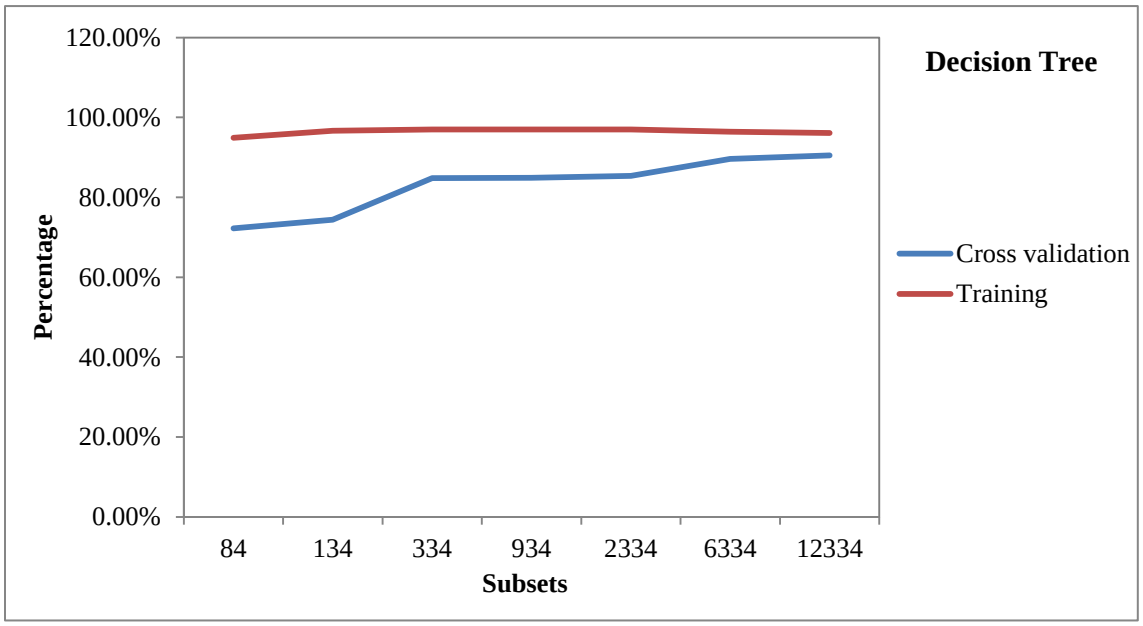
Figure 42: Learning curve for Support Vector Machine in MOC2



The results of the learning curve are shown in Figure 38 for support vector machine. Due to the default kernel function being linear, the learning curve for LR shows that as the sample increases the accuracy of training gradually decreases and the accuracy of the cross validation set increases. This case shows an example of high bias and low variance.

5.5.4.5. DECISION TREE

Figure 43: Learning curve for Decision Tree in MOC2



The results of the learning curve are shown in Figure 39 for decision tree. Decision tree is able to model nonlinear functions and works through segmentation process by splitting data into segments. The results of the learning show that DT has a low variance and low bias. The training performance remains constant while the DT cross validation set shows a slow learning.

5.5.5 VALIDATION OF RESULTANT MODELS BY AN EVALUATION SET

The results of validation by an external evaluation set shows that DT and ANN perform well in determining the recall and specificity. Similarly, the results of ANN show that the model D2 model after feature selection is able to detect the false negative and true positive cases. The MCC was highest for DT followed by ANN, NB, SVM and LR.

Table 10: Performance of Predictive Modeling for MOC 2 by external evaluation set

ML Algorithm	Accuracy	Precision	Recall	Specificity	F-measure	AUC	MCC
NB	80.31%	81.90%	80.30%	79.70%	0.801	0.909	0.622
LR	75.35%	76.90%	75.40%	74.30%	75.00%	0.852	0.523
SVM	76.07%	78.10%	76.10%	74.90%	0.756	0.761	0.541
ANN	85.98%	86.30%	86.00%	85.87%	0.859	0.931	0.723
DT	88.72%	89.00%	88.70%	88.30%	0.887	0.895	0.777

5.5.6 SUMMARY

This dataset contained 16,768 instances with 181 variables. Of these 16,768, 10% (1,676) were used for internal validation of the resultant model for the dataset. Other 15,092 instances were used for 10 fold cross validation. Feature selection eliminated about 100 teeth surfaces and 5 clinical variables resulting into 80 variables. An overall performance shows that DT outperformed other algorithms. The AUC for ANN and DT reached 0.932 and 0.905, respectively after feature selection, representing a better predicting performance.

The results of 10 fold cross validation for the original DT of MOC 2 containing 181 variables, yielded 468 leaves and 896 as the size of tree (number of nodes). With a confidence

factor of 0.25 and minNumObj of 2, the root node was mesiolingual surface of tooth number 31. The interproximal surfaces were ranked near the top of the tree representing the important factors for predicting PD risk. This was followed by the (internal node) dental variable 'calculus' and then by the medical variable 'BMI' and 'Age'. At the internal node of BMI=healthy, the test condition showed BMI= healthy as 'low risk' and BMI=obese was traversed to Age. Further the test condition at the internal node of age showed low risk for patients with age $\leq$ 56 and traversed to dental variables with age >56 years. Notably, the occurrence of data variable 'BMI=obese' after the interproximal surfaces represents that BMI more than 30 (obese) is an important predictive factor for PD risk in MOC 2. **Figure 34** shows the pruned MOC 2 decision tree. The dataset for MOC 2 containing 185 variables was pruned for the purpose of readability by lowering the confidence factor from 0.25 to 0.20 and increasing the minNumObj in WEKA from 2 to 20. This yielded a total number of 39 leaves and the size of tree was 75. All the nodes in the tree represent tooth surfaces of PPD except dental variable 'calculus' and medical variables 'BMI= obese', 'BMI=over weight', 'BMI=underweight' and 'gender=Female'. The misclassification ratio for distobuccal surface of tooth number 2 was highest in the tree (8657.0/474.0).

Figure 44: The pruned MOC 2 decision tree (J4.8) to identify the most important parameters that would influence the PD risk generated in WEKA

```
MesialLingual31 <= 4
| DistalBuccal14 <= 4
| | MesialLingual18 <= 4
| | | MesialLingual2 <= 4
| | | | MesialLingual19 <= 4
| | | | | MesialLingual15 <= 4
| | | | | | DistalBuccal23 <= 4
| | | | | | | MesialBuccal18 <= 4
| | | | | | | | MesialLingual14 <= 4
| | | | | | | | | MesialLingual30 <= 4
| | | | | | | | | | DistalBuccal12 <= 4
| | | | | | | | | | | MesialBuccal2 <= 4
| | | | | | | | | | | | MesialBuccal5 <= 4
| | | | | | | | | | | | | MesialLingual3 <= 4
| | | | | | | | | | | | | | MesialBuccal8 <= 4
| | | | | | | | | | | | | | | DistalBuccal15 <= 4
| | | | | | | | | | | | | | | | DistalBuccal26 <= 4
| | | | | | | | | | | | | | | | | DistalBuccal18 <= 4
| | | | | | | | | | | | | | | | | | MesialLingual29 <= 4
| | | | | | | | | | | | | | | | | | | DistalBuccal20 <= 4
| | | | | | | | | | | | | | | | | | | | MesialLingual22 <= 3
| | | | | | | | | | | | | | | | | | | | | DistalBuccal2 <= 4: LOW (8657.0/474.0)
| | | | | | | | | | | | | | | | | | | | | DistalBuccal2 > 4: HIGH (30.0/9.0)
| | | | | | | | | | | | | | | | | | | | MesialLingual22 > 3
| | | | | | | | | | | | | | | | | | | | | MesialLingual21 <= 4: LOW (316.0/71.0)
| | | | | | | | | | | | | | | | | | | | | MesialLingual21 > 4: HIGH (19.0)
| | | | | | | | | | | | | | | | | | | | DistalBuccal20 > 4: HIGH (27.0/8.0)
| | | | | | | | | | | | | | | | | | | MesialLingual29 > 4: HIGH (71.0/16.0)
| | | | | | | | | | | | | | | | | | DistalBuccal18 > 4
| | | | | | | | | | | | | | | | | | DistalBuccal31 <= 4
| | | | | | | | | | | | | | | | | | | MesialBuccal18 <= 3: LOW (61.0/17.0)
| | | | | | | | | | | | | | | | | | | MesialBuccal18 > 3
| | | | | | | | | | | | | | | | | | | | Calculus <= 1: LOW (2.0)
| | | | | | | | | | | | | | | | | | | | Calculus > 1: HIGH (6.0/1.0)
| | | | | | | | | | | | | | | | | | | | MesialBuccal3 <= 2: HIGH (15.0/1.0)
| | | | | | | | | | | | | | | | | | | | MesialBuccal3 > 2
| | | | | | | | | | | | | | | | | | | | | Buccal30 <= 2: HIGH (33.0/10.0)
| | | | | | | | | | | | | | | | | | | | | Buccal30 > 2: LOW (17.0/4.0)
| | | | | | | | | | | | | | | | | | | | Lingual22 > 2: HIGH (18.0/1.0)
| | | | | | | | | | | | | | | | | | | DistalBuccal31 > 4: HIGH (30.0)
```

| | | | | | | | | | | | | | | | | DistalBuccal26 > 4: HIGH (74.0/7.0)
| | | | | | | | | | | | | | | | DistalBuccal15 > 4
| | | | | | | | | | | | | | | | | DistalBuccal11 <= 3
| | | | | | | | | | | | | | | | | | MesialLingual27 <= 2: LOW (23.0/9.0)
| | | | | | | | | | | | | | | | | | MesialLingual27 > 2: HIGH (32.0/7.0)
| | | | | | | | | | | | | | | | | DistalBuccal11 > 3: HIGH (15.0)
| | | | | | | | | | | | | | | | MesialBuccal8 > 4: HIGH (27.0)
| | | | | | | | | | | | | | | | MesialLingual3 > 4
| | | | | | | | | | | | | | | | MesialLingual6 <= 3
| | | | | | | | | | | | | | | | | BMI = Over weight: HIGH (10.0/2.0)
| | | | | | | | | | | | | | | | | BMI = Obese
| | | | | | | | | | | | | | | | | | MesialLingual4 <= 3: LOW (19.0/2.0)
| | | | | | | | | | | | | | | | | | MesialLingual4 > 3: HIGH (16.0/6.0)
| | | | | | | | | | | | | | | | | BMI = Underweight: HIGH (2.0)
| | | | | | | | | | | | | | | | GENDER = M: HIGH (37.0/6.0)
| | | | | | | | | | | | | | | | MesialLingual6 > 3: HIGH (27.0/1.0)
| | | | | | | | | | | | | | | MesialBuccal5 > 4: HIGH (48.0/5.0)
| | | | | | | | | | | | | | MesialBuccal2 > 4
| | | | | | | | | | | | | | | DistalBuccal3 <= 4
| | | | | | | | | | | | | | | DistalBuccal23 <= 3: LOW (17.0/7.0)
| | | | | | | | | | | | | | | DistalBuccal23 > 3: HIGH (16.0/4.0)
| | | | | | | | | | | | | | | DistalBuccal3 > 4: HIGH (36.0)
| | | | | | | | | | | | | | DistalBuccal12 > 4: HIGH (52.0/4.0)
| | | | | | | | | | | | | MesialLingual30 > 4: HIGH (146.0/32.0)
| | | | | | | | | | | | MesialLingual14 > 4: HIGH (110.0/17.0)
| | | | | | | | | | | MesialBuccal18 > 4: HIGH (92.0/12.0)
| | | | | | | | | | DistalBuccal23 > 4: HIGH (179.0/5.0)
| | | | | | | | | MesialLingual15 > 4: HIGH (228.0/30.0)
| | | | | | | | MesialLingual19 > 4: HIGH (163.0/15.0)
| | | | | | | MesialLingual2 > 4: HIGH (575.0/60.0)
| | | | | | MesialLingual18 > 4: HIGH (619.0/30.0)
| | | | | DistalBuccal14 > 4: HIGH (846.0/27.0)
MesialLingual31 > 4: HIGH (2372.0/79.0)

5.6 MODEL OF CARE 3: DENTAL WITH PATIENT REPORTED MEDICAL

5.6.1 PATIENT CHARACTERISTICS

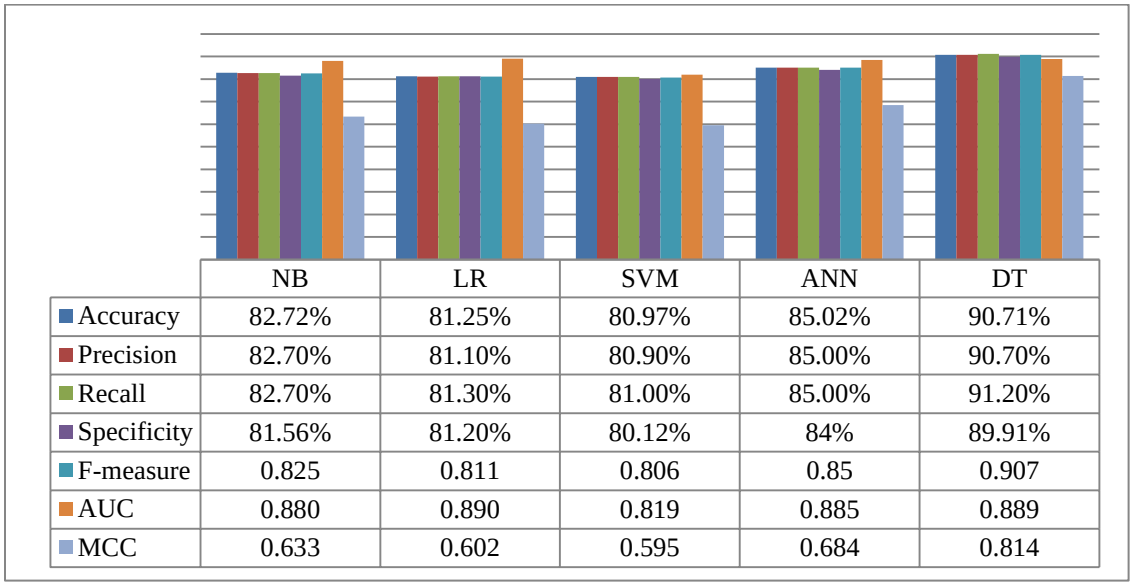
The overall mean age of patients was 41.19 $\pm$ 16.01, with 67% of patients being female. Of these 9,382 were Medicaid patients and 2,892 were Medicare patients. Mean brushing frequency was 1.6 $\pm$ 0.59. The distribution of the patients for presence or absence of diabetes was: 1,874 'Type 2 Diabetes and 10,299 'No diabetes'. The mean duration of diabetes was 1.2 $\pm$ 3.8 years. Of the 12,173 patients, 4,163 (34%) were current smokers, 3,640 (29.90%) were former smokers and 4,370(35.8%) never smoked tobacco. A majority of patients [72% (8,727/12,173)] had a diastolic blood pressure of less than 80 mm of Hg. About 52% (6,318/12,173) had a systolic blood pressure less than 120 mm of Hg. The mean height of the patients was 167.74 $\pm$ 9.5 cms and weight was 85.61 $\pm$ 23 kg. Approximately, 7,506 patients were obese with a BMI of more than 30, whereas about 3,872 patients were overweight with a BMI between 25 and 29.99. More than half patients had a healthy range of HDL and an optimal level of LDL. Similarly, most of the patients had a desirable level of total cholesterol and normal triglyceride levels.

5.6.2 RESULTS OF THE APPLICATION OF THE FIVE ALGORITHMS

The various performance metrics show that DT outperformed all the other classifiers. The sensitivity for DT was [90.60% (95%CI 89.7-92.51)], while specificity was [90.50% (95%CI: 89.85-92.87)]. MCC for DT was highest followed by ANN, SVM, LR and NB. F-measure for LR and SVM was almost equal, however lower than DT and ANN and slightly higher than NB. Precision was lowest for NB.

The results of the application of the five algorithms are shown in Figure 40

Figure 45: Results of application of ML algorithms to MOC 3

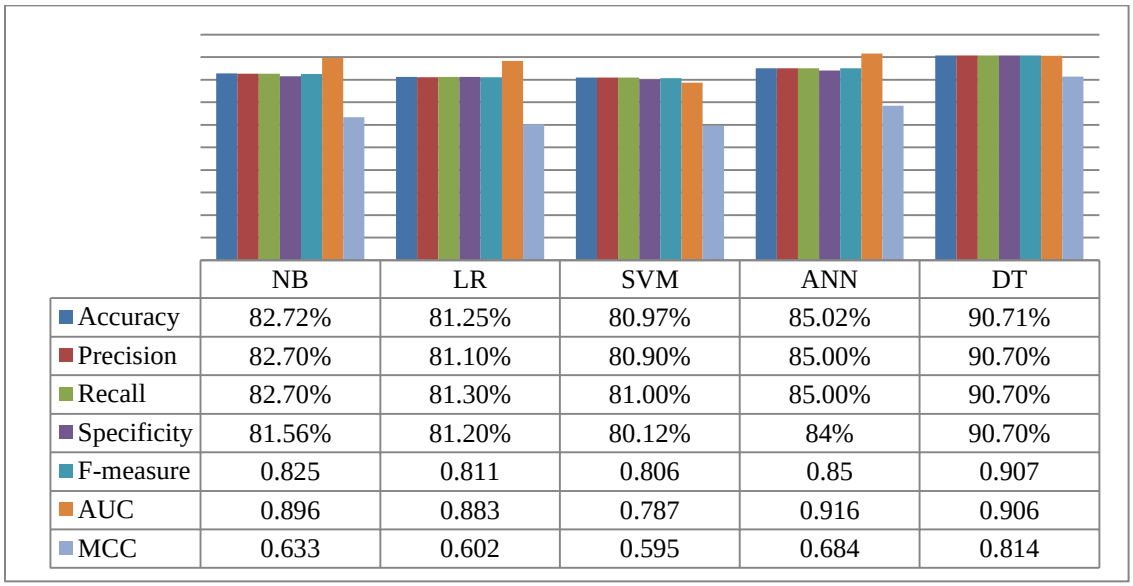


	NB	LR	SVM	ANN	DT
Accuracy	82.72%	81.25%	80.97%	85.02%	90.71%
Precision	82.70%	81.10%	80.90%	85.00%	90.70%
Recall	82.70%	81.30%	81.00%	85.00%	91.20%
Specificity	81.56%	81.20%	80.12%	84%	89.91%
F-measure	0.825	0.811	0.806	0.85	0.907
AUC	0.880	0.890	0.819	0.885	0.889
MCC	0.633	0.602	0.595	0.684	0.814

5.6.3 RESULTS OF FEATURE SELECTION

Variables including duration of diabetes, Medicare Status, frequency of flossing, height and weight were eliminated after feature selection method. The results of performance metrics after features selection are shown in **Figure 46**

Figure 46: The results of performance metrics after features selection for MOC 3



	NB	LR	SVM	ANN	DT
Accuracy	82.72%	81.25%	80.97%	85.02%	90.71%
Precision	82.70%	81.10%	80.90%	85.00%	90.70%
Recall	82.70%	81.30%	81.00%	85.00%	90.70%
Specificity	81.56%	81.20%	80.12%	84%	90.70%
F-measure	0.825	0.811	0.806	0.85	0.907
AUC	0.896	0.883	0.787	0.916	0.906
MCC	0.633	0.602	0.595	0.684	0.814

Overall, a slight increase in total accuracy was seen in NB, ANN and DT after feature selection. There was a decrease in value of MCC for DT, SVM and LR, however showed an increase in ANN and NB. The F-measure remained constant for DT before and after the feature selection. DT outperformed other algorithms in terms of sensitivity 91.20% (95% CI 90.54-91.83) and specificity 89.91% (95% CI 89.00-90.77). There was an increase in the F-measure of ANN and NB from 0.831 to 0.850 and 0.809 to 0.825, respectively. Similarly, the F-measure for DT increased slightly by 0.02. The MCC and AUC markedly increased from 0.645 to 0.684 and from 0.885 to 0.916, respectively for ANN. Precision for SVM decreased marginally, whereas there was increase in precision for all other algorithms. Correspondingly, the total accuracy for all the algorithms increased slightly except for SVM where a decrease in total accuracy was seen. The perio chart with the representative surfaces of teeth are shown in **Figure 47**

Figure 47: Periodontal chart and tooth surfaces to show the significant tooth surfaces after feature selection application to MOC3

Figure 47: Periodontal chart and tooth surfaces to show the significant tooth surfaces after feature selection application to MOC3

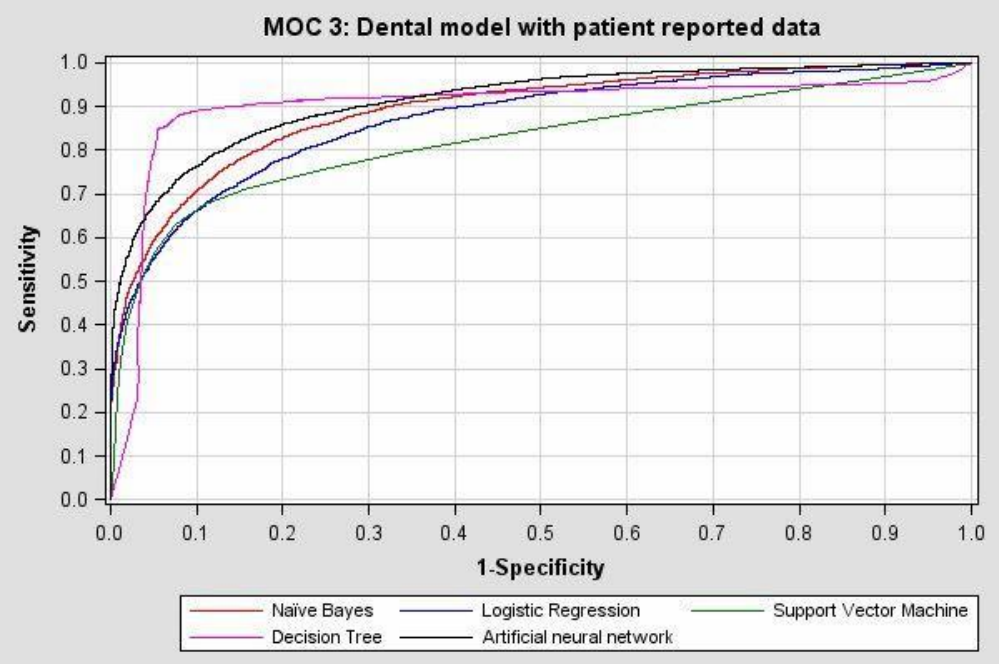
	ML	ML	ML	ML	ML	ML	ML		ML	ML	DL	ML	ML	DB	ML	
	MB	DB	DL	MB	MB	MB	DL		DL	MB	DB	MB	DB	DL	MB	
	DB	DL	DB	DL	DB	DL	DB		DB	DB	ML	DL	DL	ML	DL	
	DL	MB	MB	DB	DL	DB	MB		MB	DL	MB	DB	MB	MB	DB	
	L	L	L	L	L	L	L		L	L	L	L	L	L	L	
	B	B	B	B	B	B	B		B	B	B	B	B	B	B	
1	2	3	4	5	6	7	8		9	10	11	12	13	14	15	16
32	31	30	29	28	27	26	25		24	23	22	21	20	19	18	17
	B	B	B	B	B	L	L		L	B	B	B	B	B	B	
	L	L	L	L	L	B	B		B	L	L	L	L	L	L	
	MB	MB	MB	DB	DL	ML	ML		ML	ML	ML	MB	MB	MB	MB	
	DB	DB	DL	DL	DB	MB	DL		MB	MB	DB	DL	DB	DB	DB	
	DL	DL	DB	MB	ML	DL	DB		DB	DL	DL	DB	DL	DL	DL	
	ML	ML	ML	ML	MB	DB	MB		DL	DB	MB	ML	ML	ML	ML	

The elimination in feature selection resulted in exclusion of the blue color (99 attributes) i.e. all the lingual and buccal surfaces of the teeth (56 attributes), interproximal surfaces including ML,

DL and DB surfaces of all mandibular central incisors and mesiobuccal of tooth number 26. The interproximal surfaces including MB, DB and DL of maxillary central and lateral incisors were eliminated after application of feature selection. Similarly the MB of left maxillary canine and DL of right maxillary canine were removed, as seen in MOC 1.

The findings of this experiment shows that the AUC obtained by LR and NB were equal and higher than ANN, DT and SVM. The ROC curve performance of each classifier for MOC 3 is shown in Figure 48.

Figure 48: Receiver Operating Characteristic curves for algorithms in MOC 3



The ROC for DT is higher than ANN, NB, LR and SVM. The ROC of DT begins at (0, 0) and eventually bends towards the right at (0.02, 0.2) and then runs vertical to (0.05, 0.85) indicating more true positives than false positives and correspondingly signaling a greater noise. The ROC curves of algorithms in descending order of their performance are DT>ANN>NB> SVM.

5.6.4 VALIDATION OF RESULTANT MODELS BY AN EXTERNAL EVALUATION SET

The results of validation showed that NB, ANN and LR had a better analyzing capacity for true positive and true negative cases as compared to DT and SVM. The MCC was highest for NB followed by LR, ANN, SVM and DT. NB outperformed other algorithms in terms of precision and total accuracy.

The results of the evaluation of external set to MOC 3 are shown in Table 10

Table 11: Performance of predictive modelling for MOC 3 by external evaluation set

ML	Accuracy	Precision	Recall	Specificity	F-measure	AUC	MCC
NB	76.78%	78.00%	76.80%	75.67%	0.765	0.850	0.547
LR	75.35%	76.90%	75.40%	75.60%	0.750	0.852	0.523
SVM	74.56%	75.50%	74.60%	73.78%	0.743	0.824	0.500
ANN	75.00%	76.40%	75.00%	74.90%	0.747	0.750	0.514
DT	71.89%	77.00%	71.90%	74.67%	0.705	0.576	0.486

5.6.5 SUMMARY

This dataset contained 13,525 instances with 185 variables. This model of care incorporated patient reported medical data including presence or absence of diabetes and duration of diabetes along with oral hygiene behavior including frequency of tooth brushing, flossing and tobacco use. Of these 13,525, 10% (1,352) were used for internal validation of the resultant model for imbalanced dataset. Other 12,173 instances were used for 10 fold cross validation. Feature selection eliminated about 99 teeth surfaces and 5 clinical variables. An overall performance shows that DT outperformed other algorithms; however the results of validation showed a better analyzing capacity for NB, LR and ANN in terms of total accuracy as compared to DT and

SVM. The AUC for DT and ANN reached 0.916 and 0.906, respectively representing a better predicting performance.

The results of 10 fold cross validation for the original DT of MOC 3 containing 185 variables, yielded 340 leaves and 657 as the size of tree (number of nodes). With a confidence factor of 0.25 and minNumObj of 2, the root node was mesiolingual surface of tooth number 31. The interproximal surfaces were ranked near the top of the tree representing the important factors for predicting PD risk. This was followed by the (internal node) by the medical variable 'systolic blood pressure'. At the internal node of 'systolic blood pressure <120", the test condition showed systolic blood pressure <120 and between the range of 140-159 mm of Hg as 'high risk'. Further the test condition at the internal node of age showed 'low risk' for patients with age $\leq$ 44 and traversed to dental variables with age >44 years. **Figure 34** shows the pruned MOC 2 decision tree. The dataset for MOC 2 containing 185 variables was pruned for the purpose of readability by lowering the confidence factor from 0.25 to 0.20 and increasing the minNumObj in WEKA from 2 to 20. This yielded a total number of 38 leaves and the size of tree was 73. All the nodes in the tree represent tooth surfaces of PPD except medical variables 'systolic blood pressure'. The misclassification ratio for distobuccal surface of tooth number 19 was highest in the tree (7153.0/382.0).

Figure 49 The pruned MOC 3 decision tree (J4.8) to identify the most important parameters that would influence the PD risk generated in WEKA

```
MesialLingual31 <= 4
| DistalBuccal14 <= 4
| | MesialLingual18 <= 4
| | | MesialLingual2 <= 4
| | | | MesialLingual14 <= 4
| | | | | MesialLingual19 <= 4
| | | | | | MesialLingual15 <= 4
| | | | | | | MesialBuccal18 <= 4
| | | | | | | | MesialLingual3 <= 4
| | | | | | | | | MesialLingual30 <= 4
| | | | | | | | | | DistalBuccal18 <= 4
| | | | | | | | | | | DistalBuccal23 <= 4
| | | | | | | | | | | | DistalBuccal12 <= 4
| | | | | | | | | | | | | MesialBuccal2 <= 4
| | | | | | | | | | | | | | MesialBuccal5 <= 4
| | | | | | | | | | | | | | | MesialLingual4 <= 4
| | | | | | | | | | | | | | | | DistalBuccal2 <= 4
| | | | | | | | | | | | | | | | | MesialLingual27 <= 3
| | | | | | | | | | | | | | | | | | DistalBuccal19 <= 4: LOW (7153.0/382.0)
| | | | | | | | | | | | | | | | | | DistalBuccal19 > 4
| | | | | | | | | | | | | | | | | | | MesialLingual20 <= 3: LOW (16.0/6.0)
| | | | | | | | | | | | | | | | | | | MesialLingual20 > 3: HIGH (15.0/5.0)
| | | | | | | | | | | | | | | | | MesialLingual27 > 3
| | | | | | | | | | | | | | | | | | DistalBuccal26 <= 4
| | | | | | | | | | | | | | | | | | | MesialLingual29 <= 4
| | | | | | | | | | | | | | | | | | | | Lingual27 <= 3
| | | | | | | | | | | | | | | | | | | | | MesialBuccal27 <= 4: LOW (263.0/55.0)
| | | | | | | | | | | | | | | | | | | | | MesialBuccal27 > 4: HIGH (16.0/7.0)
| | | | | | | | | | | | | | | | | | | | Lingual27 > 3: HIGH (21.0/6.0)
| | | | | | | | | | | | | | | | | | | MesialLingual29 > 4: HIGH (18.0/2.0)
| | | | | | | | | | | | | | | | | | DistalBuccal26 > 4: HIGH (33.0/2.0)
| | | | | | | | | | | | | | | | | DistalBuccal2 > 4: HIGH (37.0/8.0)
| | | | | | | | | | | | | | | | MesialLingual4 > 4: HIGH (37.0/6.0)
| | | | | | | | | | | | | | | MesialBuccal5 > 4: HIGH (30.0/3.0)
| | | | | | | | | | | | | | MesialBuccal2 > 4: HIGH (51.0/10.0)
| | | | | | | | | | | | | DistalBuccal12 > 4: HIGH (30.0/2.0)
| | | | | | | | | | | | DistalBuccal23 > 4: HIGH (104.0/5.0)
| | | | | | | | | | | DistalBuccal18 > 4
| | | | | | | | | | | | DistalBuccal31 <= 4
| | | | | | | | | | | | | Lingual26 <= 2
| | | | | | | | | | | | | | DistalBuccal14 <= 2: HIGH (28.0/6.0)
```

| | | | | | | | | | | | | DistalBuccal14 > 2
| | | | | | | | | | | | | | DistalBuccal23 <= 3
| | | | | | | | | | | | | | | DistalBuccal14 <= 3: LOW (30.0/4.0)
| | | | | | | | | | | | | | | DistalBuccal14 > 3
| | | | | | | | | | | | | | | | MesialBuccal31 <= 3: HIGH (23.0/9.0)
| | | | | | | | | | | | | | | | MesialBuccal31 > 3: LOW (16.0/5.0)
| | | | | | | | | | | | | | DistalBuccal23 > 3: HIGH (19.0/5.0)
| | | | | | | | | | | | | Lingual26 > 2: HIGH (22.0/1.0)
| | | | | | | | | | | | DistalBuccal31 > 4: HIGH (35.0)
| | | | | | | | | MesialLingual30 > 4
| | | | | | | | | | MesialBuccal29 <= 2: LOW (17.0/6.0)
| | | | | | | | | | MesialBuccal29 > 2: HIGH (104.0/21.0)
| | | | | | | | MesialLingual3 > 4
| | | | | | | | | DistalBuccal4 <= 3
| | | | | | | | | | MesialLingual4 <= 3
| | | | | | | | | | | DistalBuccal4 <= 2: HIGH (26.0/8.0)
| | | | | | | | | | | DistalBuccal4 > 2: LOW (21.0/7.0)
| | | | | | | | | | MesialLingual4 > 3: HIGH (32.0/6.0)
| | | | | | | | | DistalBuccal4 > 3: HIGH (56.0/1.0)
| | | | | | | MesialBuccal18 > 4: HIGH (76.0/7.0)
| | | | | | MesialLingual15 > 4
| | | | | | | Systolic blood pressure = <120: HIGH (60.0/15.0)
| | | | | | | Systolic blood pressure = 140-159: LOW (4.0)
| | | | | | | Systolic blood pressure = 120-139: HIGH (75.0/5.0)
| | | | | | | Systolic blood pressure = >160: LOW (1.0)
| | | | | MesialLingual19 > 4: HIGH (114.0/13.0)
| | | | MesialLingual14 > 4: HIGH (128.0/14.0)
| | | MesialLingual2 > 4: HIGH (428.0/50.0)
| | MesialLingual18 > 4: HIGH (468.0/30.0)
| DistalBuccal14 > 4: HIGH (677.0/19.0)
MesialLingual31 > 4: HIGH (1889.0/66.0)

5.7 MODEL OF CARE 4: MEDICAL MODEL WITH PATIENT REPORTED DENTAL DATA

5.7.1 PATIENT CHARACTERISTICS

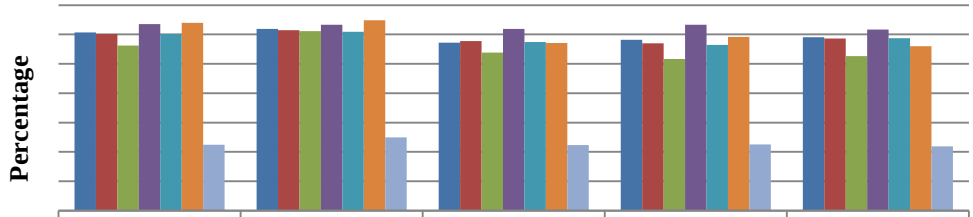
The overall mean age of patients was 47.36 $\pm$ 16.55, with 64% of patients being female. Of these 10,436 were Medicaid patients and 5,116 were Medicare patients. Mean brushing frequency was 1.6 $\pm$ 0.60. The distribution of the patients for presence or absence of diabetes was: 2,347 'Type 2 Diabetes, 3,301'Pre-diabetes' and 7,076 'No diabetes'. A majority of patients (1540/2347; 65.6%) diagnosed with diabetes had a duration of diabetes <1 year. Of the 14,135 patients, 4,241 (30%) were current smokers, 4,891 were former smokers (34.6%) and 5,002 (35.3%) never smoked tobacco. The mean height of the patients was 167.55 $\pm$ 9.9 cms and weight was 89.73 $\pm$ 24.9 kg. The mean random blood glucose level for this cohort was 112 $\pm$ 46.00. A majority of patients [66% (9,369/14,135)] had a diastolic blood pressure of less than 80 mm of Hg. About 87% (12,272/14,135) had a systolic blood pressure between the range of 120 and 139 mm of Hg. Approximately, 7,506 patients were obese with a BMI of more than 30, whereas about 3,872 patients were overweight with a BMI between 25 and 29.99. More than half patients had a healthy range of HDL and an optimal level of LDL. Similarly, most of the patients had a desirable level of total cholesterol and normal triglyceride levels.

5.7.2 RESULTS OF FIVE ALGORITHMS ON ALL THE VARIABLES

The various performance metrics show that LR outperformed all the other classifiers. Sensitivity was highest in LR [61.14% (95%CI 59.7-62.51)], while specificity was highest in SVM [61.86% (95%CI: 60.85-62.87)]. MCC for NB and ANN was almost equal, however lower than LR and

slightly higher than DT and SVM. Recall was lowest for ANN. **Figure 50** shows the results of application of ML algorithms to MOC 4

Figure 50: The results of application of ML algorithms to MOC 4

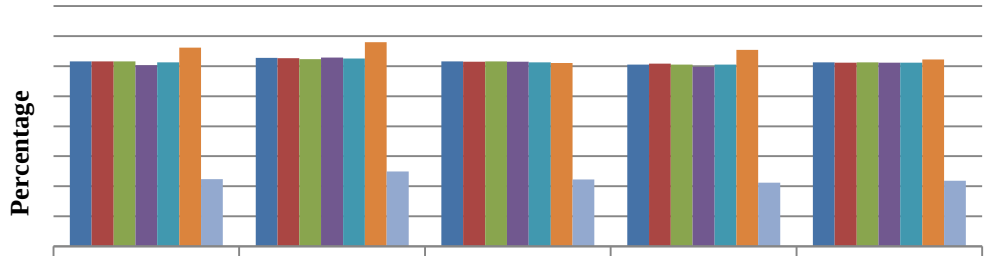


	NB	LR	SVM	ANN	DT
Accuracy	60.72%	61.86%	57.26%	58.18%	59.07%
Precision	60.20%	61.50%	57.70%	57.00%	58.60%
Recall	56.22%	61.14%	53.80%	51.69%	52.67%
Specificity	63.49%	63.31%	61.86%	63.28%	61.69%
F-measure	0.602	0.609	0.574	0.564	0.587
AUC	0.64	0.648	0.571	0.592	0.56
MCC	0.224	0.249	0.223	0.225	0.219

5.7.3 RESULTS OF FEATURE SELECTION

Variables including duration of diabetes, LDL, flossing, height and weight and Medicaid status were eliminated after feature selection method. The results of features selection are shown in **Figure 51**

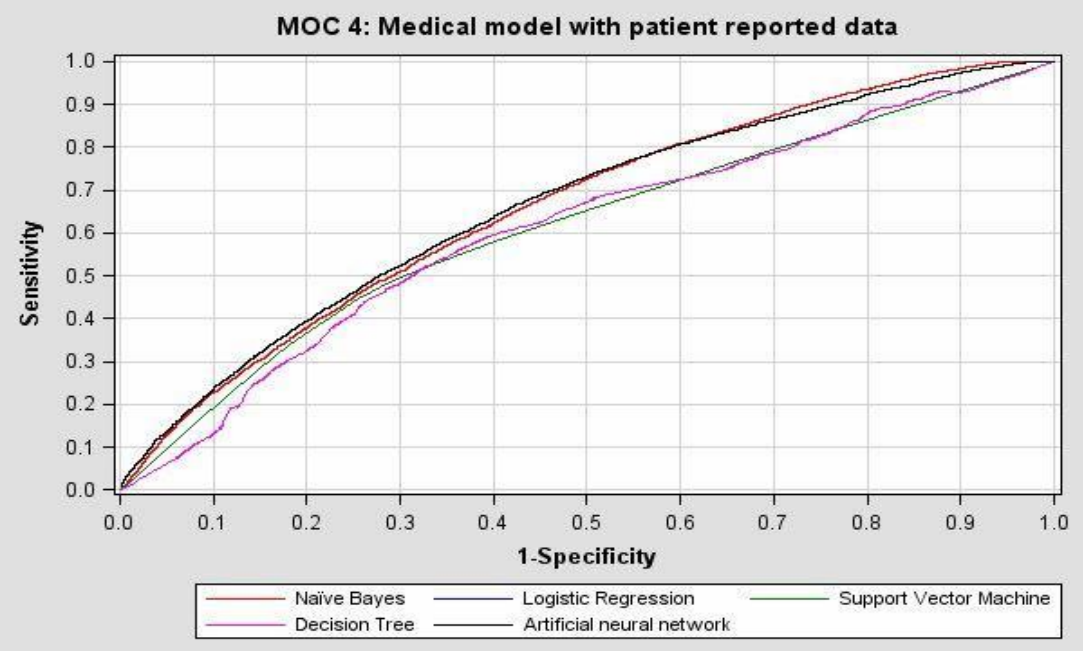
Figure 51: Results of performance metrics after feature selection in MOC 4



	NB	LR	SVM	ANN	DT
Accuracy	61.56%	62.74%	61.50%	60.46%	61.22%
Precision	61.50%	62.60%	61.40%	60.80%	61.10%
Recall	61.60%	62.27%	61.50%	60.50%	61.20%
Specificity	60.24%	62.78%	61.40%	59.78%	61.10%
F-measure	0.612	0.625	0.612	0.605	0.611
AUC	0.661	0.679	0.61	0.654	0.622
MCC	0.224	0.249	0.223	0.212	0.219

Overall, a slight increase in specificity is seen for all algorithms after feature selection. There was a slight decrease in specificity in LR after application of feature selection. LR outperformed other algorithms in terms of sensitivity 62.27% (95% CI 61.72-63.56) and specificity 62.78% (95% CI 60.89-63.75). There was an increase in the AUC of ANN from 0.592 to 0.654 after feature selection. Similarly, the AUC of LR increased by 0.15 from 0.648 to 0.679. The F-measure markedly increased from 0.564 to 0.605 for ANN, however there was a decreased in MCC of ANN from 0.225 to 0.212 after feature selection. **Figure 52** shows the ROC analysis displaying 5 ROC curves that are representing different levels of performance of the classifiers for MOC 4.

Figure 52: ROC analysis displaying 5 ROC curves that are representing different levels of performance of the classifiers for MOC 4.



The findings of this experiment shows that the AUC obtained by LR and NB were equal and higher than ANN, DT and SVM. The ROC curve performance of each classifier for MOC 4 is shown in **Figure 52.**

The ROC curves in the figure run from point (0, 0) and end at (1, 1). LR, ANN and NB show slightly symmetric curve which represent a lower performing model while DT and SVM are along the diagonal representing a random performance. The ROC curve of DT crosses the diagonal when the false positive rate (1-specificity) is 0.8 and true positive rate (sensitivity) is at 0.7 representing a worse than random performance. ROC curve for LR and NB performs better than DT, ANN and SVM.

5.7.4 VALIDATION OF RESULTANT MODELS BY AN EVALUATION SET

The results of the performance metrics with an external evaluation set are shown in Table 11. The results of validation show that SVM and DT had a better analyzing capacity for true positive and true negative cases as compared to NB, LR and ANN. The MCC was negative for NB, LR and ANN representing a negative relationship between the variables and the outcome.

Table 12: Performance metrics of MOC 4 with an external evaluation set

ML	Accuracy	Precision	Recall	Specificity	F-measure	AUC	MCC
NB	31.55%	30.40%	31.60%	30.62%	0.306	0.250	-0.380
LR	38.48%	38.30%	38.50%	36.78%	0.382	0.328	-0.232
SVM	53.49%	53.60%	53.60%	52.31%	0.533	0.535	0.070
ANN	33.95%	34.00%	33.90%	32.15%	0.338	0.347	-0.320
DT	51.00%	51.20%	51.00%	48.00%	0.493	0.469	0.022

5.7.5 SUMMARY

This dataset contained 15,705 instances with 15 variables. This model of care incorporated patient reported oral hygiene behavior such as frequency of tooth brushing, frequency of tooth flossing among others. Of these 15,705, 10% (1,570) were used for internal validation of the resultant model for imbalanced dataset. Other 14,135 instances were used for 10 fold cross validation. Feature selection eliminated 6 clinical variables. An overall performance shows that LR performed better than other algorithms; however the results of validation showed a better analyzing capacity of NB, SVM and DT. The AUC for LR reached 0.608 and ROC represented a random performance.

Table 13 shows the weights applied to each data variable in LR in form of coefficients.

Table 13: Weights applied to each data variable in logistic regression in form of their coefficients

Variable	HIGH
Random blood glucose	0.0007
Age	0.0271
Gender=M	0.5294
MCARE status	-0.1843
Tooth brushing	-0.0561
Type II Diabetes	0.0004
No diabetes	-0.0363
Pre-Diabetic	0.0467
Tobacco use status=Current	0.3524
Tobacco use status=Never	-0.1935
Tobacco use status=Former	-0.1315
Number of teeth present	0.0513
Blood Pressure Diastolic=<80	-0.0604
Blood Pressure Diastolic=90-99	0.1412
Blood Pressure Diastolic=80-89	0.0147
Blood Pressure Diastolic=>100	0.2353
Blood Pressure Systolic=120-139	-0.0317
Blood Pressure Systolic=140-159	0.0781
Blood Pressure Systolic=>160	-0.1815
BMI=Over Weight	0.0223
BMI=Obese	0.0033
BMI=Healthy	-0.0392
BMI=Under Weight	0.066
HDL=Low	0.0644
HDL=Healthy	0.0231
HDL= High	-0.103
LDL=Optimal	-0.073
LDL=Near Optimal	-0.0154
LDL= Borderline High	0.1258
LDL=High	0.1054
Total cholesterol=Borderline High	-0.0186
Total cholesterol=Desirable	0.0188
Total Cholesterol=very high	-0.0109
Triglyceride=Very High	0.4848
Triglyceride=High	0.0488
Triglyceride=Borderline High	-0.0058
Triglyceride=Normal	-0.0277

Intercept	-2.8231

5.8. MOC 5: MEDICAL ONLY

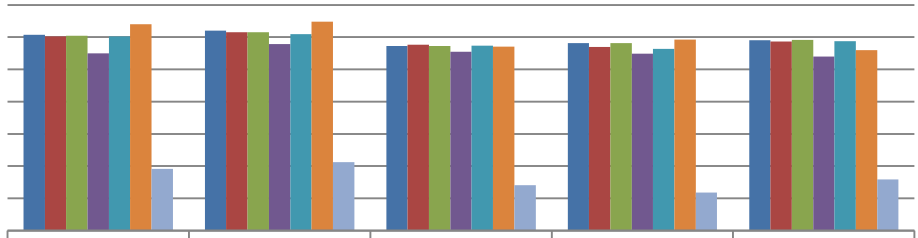
5.8.1 PATIENT CHARACTERISTICS

The overall mean age of patients was 47.36 $\pm$ 16.63, with 63% of patients being female. Of the 19,972 total patients in the cohort, 14,606 were Medicaid patients and 7,339 were Medicare patients. The distribution of the patients for presence or absence of diabetes was: 3,390 'Type 2 Diabetes, 5,090 'Pre-diabetes' and 11,492 'No diabetes'. The mean duration of diabetes for this cohort was 36.5 $\pm$ 4.16 years. Of the 19,972 patients, 6,559(32.80%) were current smokers, 7,060 were former smokers (35.35%) and 6,353 (31.85%) never smoked tobacco. The mean random blood glucose level for this cohort was 112 $\pm$ 46.31. A majority of patients (66%) had a diastolic blood pressure of less than 80 mm of Hg. About 45% had a systolic blood pressure between the range of 120 and 139 mm of Hg. Approximately, 52% of patients were obese with a BMI of more than 30, whereas about 27% patients were overweight with a BMI between 25 and 29.99. More than half patients had a healthy range of HDL and an optimal level of LDL. Similarly, most of the patients had a desirable level of total cholesterol and normal triglyceride levels.

5.8.2 RESULTS OF FIVE ALGORITHMS ON ALL THE VARIABLES

The various performance metrics show that LR outperformed all the other classifiers. Sensitivity was highest in LR [61.14% (95%CI 59.7-62.51)], while specificity was highest in SVM [61.86% (95%CI: 60.85-62.87)]. MCC for NB and ANN was almost equal, however lower than LR and slightly higher than DT and SVM. Recall was lowest for ANN. The results of the application of the five algorithms are shown in **Figure 53**

Figure 53: Results of application of ML algorithms to MOC 5

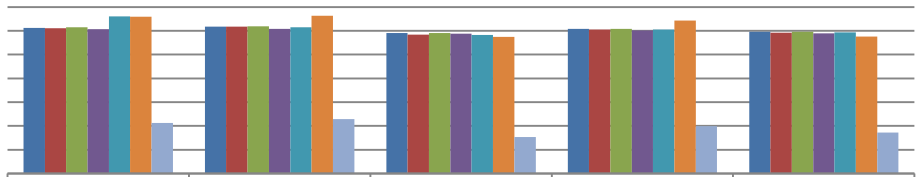


	NB	LR	SVM	ANN	DT
Accuracy	60.72%	61.99%	57.26%	58.18%	59.07%
Precision	60.20%	61.50%	57.70%	57.00%	58.60%
Recall	60.46%	61.57%	57.30%	58.20%	59.10%
Specificity	54.94%	57.86%	55.48%	54.84%	53.97%
F-measure	0.602	0.609	0.574	0.564	0.587
AUC	0.64	0.648	0.571	0.592	0.56
MCC	0.191	0.212	0.141	0.118	0.159

5.8.3. RESULTS OF FEATURE SELECTION

Variables including duration of diabetes, LDL and Medicare Status were eliminated after feature selection method. The results of features selection are shown in **Figure 54**

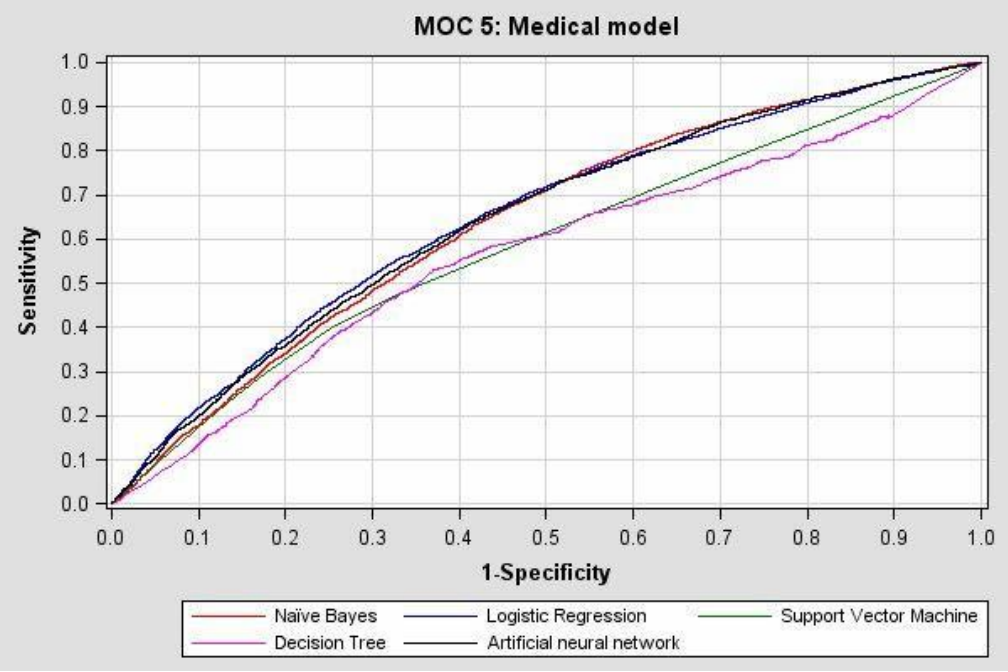
Figure 54: Results of performance metrics after feature selection in MOC 5



	Naïve Bayes (NB)	Logistic Regression (LR)	Support Vector Machine (SVM)	Artificial Neural Network (ANN)	Decision Tree (DT)
Accuracy	61.18%	61.80%	59.04%	60.79%	59.60%
Precision	61.10%	61.70%	58.40%	60.60%	59.20%
Recall	61.48%	61.90%	59.10%	60.80%	59.60%
Specificity	60.71%	60.82%	58.80%	60.17%	58.87%
F-measure	0.66	0.615	0.583	0.606	0.593
AUC	0.659	0.663	0.574	0.643	0.576
MCC	0.212	0.229	0.154	0.199	0.172

Overall, a slight increase in specificity is seen for all algorithms after feature selection. LR outperformed other algorithms in terms of sensitivity 61.90% (95% CI 61.52-63.36) and specificity 60.82% (95% CI 59.69-61.95). There was an increase in the AUC of ANN from 0.592 to 0.643 after feature selection. Similarly, the AUC for LR increased by 0.15; from 0.648 to 0.663. The F-measure and MCC was markedly increased from 0.546 to 0.606 and from 0.118 to 0.199 for ANN respectively. Precision remained the same for all the algorithms except for a slight increase in ANN precision. Correspondingly, the total accuracy for all the algorithms increased slightly. **Figure 55** displays 5 ROC curves that are representing different levels of performance of the classifiers for MOC 5.

Figure 55: ROC curve analysis displaying five ROC curves that are representing different levels of performance of the classifiers for MOC 5



The ROC curves in the figure begin at point (0, 0) and end at (1, 1). LR, ANN and NB show slightly symmetric curve towards the left of the diagonal as compared to DT and SVM. The ROC curve of DT crosses the diagonal when the false positive rate (1-specificity) is 0.8 and true

positive rate (sensitivity) is at 0.8 representing a worse than random performance. LR and NB perform better that DT, ANN and SVM.

5.8.4 VALIDATION OF RESULTANT MODELS BY AN EVALUATION SET

Results of performance of predictive modelling of MOC 5 with an external evaluation set.

The results of validation show that SVM and DT could analyze the true positive and true negative cases better than NB, LR and ANN. The MCC was negative for NB, LR and ANN representing a negative relationship between the variables and the outcome. The results of the performance of predictive modelling of MOC 5 with an external evaluation set is shown in table 12

Table 14: Performance of Predictive Modeling of MOC 5 with an external evaluation set

ML	Accuracy	Precision	Recall	Specificity	F-measure	AUC	MCC
NB	62.60%	68.00%	62.60%	62.12%	0.596	0.730	0.301
LR	62.60%	66.70%	62.60%	61.30%	0.602	0.715	0.290
SVM	61.40%	63.10%	61.40%	60.45%	0.601	0.614	0.245
ANN	59.35%	69.40%	59.30%	60.12%	0.533	0.760	0.269
DT	81.30%	82.50%	81.30%	80.23%	0.811	0.877	0.638

5.8.5 SUMMARY

This dataset contained 19,972 instances with 20 variables. Of these 19,972, 10% (1,997) were used for internal validation of the resultant model for imbalanced dataset. Feature selection method was applied to this set. Other 17,975 instances were used for 10 fold cross validation. Feature selection eliminated 3 clinical variables. An overall performance shows that LR performed better than other algorithms; however the results of validation showed a better analyzing capacity of NB, SVM and DT. The AUC for LR reached 0.608 and ROC represented a random performance.

Table 15: Weights applied to each data variable in logistic regression in form of their coefficients for MOC 5

	Class
Variable	LOW
Random blood glucose	-0.0008
Age	-0.015
GENDER=M	-0.5358
MCAID status	0.1168
No diabetes	0.0386
Type II Diabetes	0.0263
Pre-Diabetic	-0.0712
Tobacco use status : former	0.1024
Tobacco use status=Current	-0.299
Tobacco use status=Never	0.1854
Blood Pressure Diastolic=<80	0.089
Blood Pressure Diastolic=80-89	-0.0325
Blood Pressure Diastolic=90-99	-0.1865
Blood Pressure Diastolic=>100	-0.2305
Blood Pressure Systolic=120-139	-0.0239
Blood Pressure Systolic=<120	0.0212
Blood Pressure Systolic=140-159	-0.0136
Blood Pressure Systolic=>160	0.1022
BMI=Obese	-0.0134
BMI=Healthy	0.0452

BMI=Over Weight	-0.0181
BMI=Under Weight	-0.002
HDL=Healthy	0.0163
HDL=Low	-0.1101
HDL=High	0.1116
Total cholesterol=Desirable	-0.4581
Total cholesterol=High	4.3563
Total cholesterol=Borderline High	-0.2813
Triglyceride= Very High	-0.7534
Triglyceride=Borderline High	0.2069
Triglyceride=Normal	0.2097
Intercept	1.2347

5.9 MOC 6: MEDICAL WITHOUT PATIENT REPORTED DATA

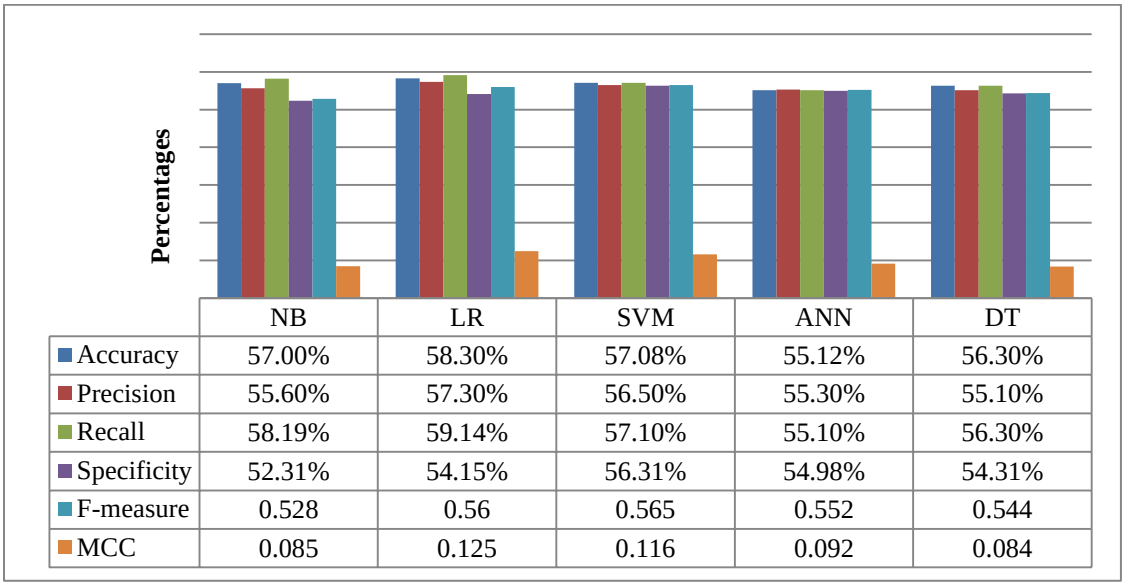
5.9.1 PATIENT CHARACTERISTICS

The distribution of the patients for presence or absence of diabetes was: 3458 'Type 2 Diabetes, 5178 'Pre-diabetes' and 13449 'No diabetes'. Approximately 774/3458 (22.3%) diagnosed with diabetes had a duration of diabetes <2 years. A majority of patients (14829/22085; 67%) had a diastolic blood pressure of less than 80 mm of Hg. About 45% (10110/22085) had a systolic blood pressure between the range of 120 and 139 mm of Hg. Approximately, 11394 patients were obese with a BMI of more than 30, whereas about 6179 patients were overweight with a BMI between 25 and 29.99. The mean random blood glucose level for this cohort was 111 $\pm$ 45.39. More than half patients had a healthy range of HDL and an optimal level of LDL. Similarly, most of the patients had a desirable level of total cholesterol and normal triglyceride levels.

5.9.2 RESULTS OF FIVE ALGORITHMS ON ALL THE VARIABLES

The various performance metrics show that LR outperformed all the other classifiers.

Figure 56: Results of application of five algorithms on MOC 6 dataset.

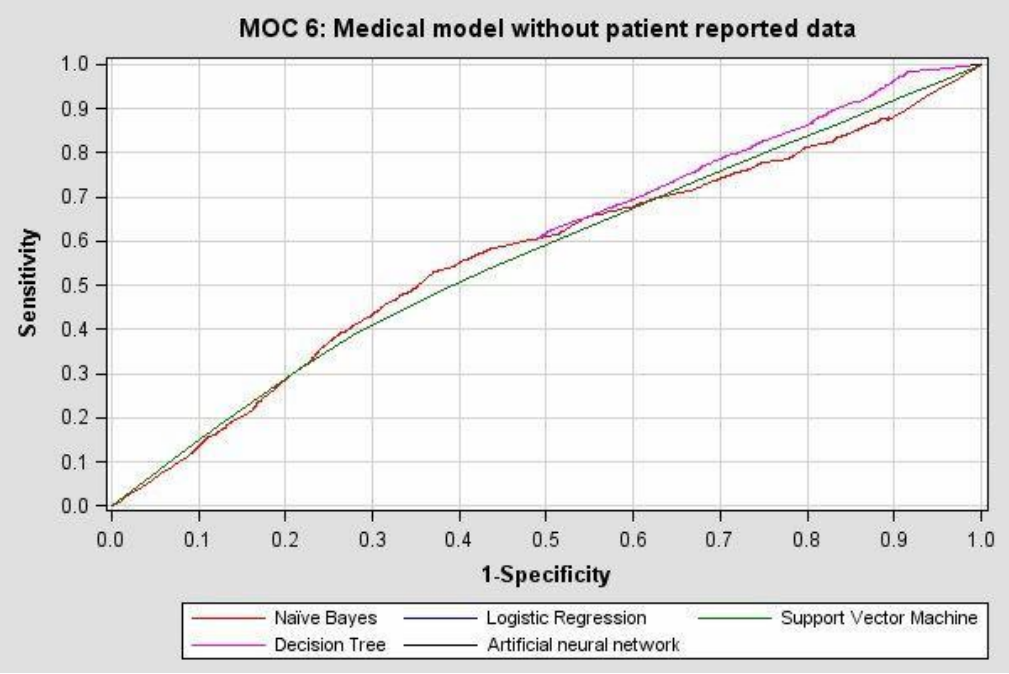


	NB	LR	SVM	ANN	DT
Accuracy	57.00%	58.30%	57.08%	55.12%	56.30%
Precision	55.60%	57.30%	56.50%	55.30%	55.10%
Recall	58.19%	59.14%	57.10%	55.10%	56.30%
Specificity	52.31%	54.15%	56.31%	54.98%	54.31%
F-measure	0.528	0.56	0.565	0.552	0.544
MCC	0.085	0.125	0.116	0.092	0.084

The F-measure for all the algorithms shows a similar range. The results of performance measures in terms of accuracy, precision, recall, F-measures, specificity and MCC showed a similar trend for NB, SVM, DT and ANN. The total accuracy for NB and SVM was almost equal. The recall and specificity for LR were 59.14% (95%CI 58.37%-55.42%) and 54.15%, (95% CI 52.87%-59.91%) respectively. The MCC of NB, DT and ANN were lower as compared to LR and SVM. Figure 51 shows the ROC analysis displaying 5 ROC curves that are representing different levels of performance of the classifiers for MOC 6.

The findings of this experiment shows that the AUC obtained by LR and NB were 0.608, and 0.597 respectively. ANN, DT and SVM showed 0.579, 0.571 and 0.556, respectively.

Figure 57: ROC curve analysis displaying five ROC curves that are representing different levels of performance of the classifiers for MOC 6



The beginning point for all the ROCs is (0, 0) and ending point is (1, 1). SVM, ANN and LR curves overlap with each other. The ROC curves for all the algorithms are very close to each

other and along the imaginary diagonal line connecting (0, 0) and (1, 1) representing a random performance. NB, ANN and LR performs better than DT and SVM.

5.9.3 SUMMARY

This dataset contained 22,085 instances with 10 variables. Feature selection was not performed on this dataset. An overall performance shows that LR performed better than other algorithms. The AUC for LR reached 0.608 and ROC represented a random performance.

Figure 16 Weights applied to each data variable in logistic regression in form of their coefficients

	Class
Variable	LOW
Random blood glucose	-0.0013
Type II Diabetes	-0.0881
No diabetes	0.1668
Pre-Diabetic	-0.1564
Blood Pressure Diastolic=80-89	-1.1705
Blood Pressure Diastolic=<80	-1.3543
Blood Pressure Diastolic=90-99	16.5951
Blood Pressure Diastolic=>100	1.9564
Blood Pressure Systolic=120-139	-0.1134
Blood Pressure Systolic=<120	0.1735
Blood Pressure Systolic=140-159	-0.1657
Blood Pressure Systolic=>160	-0.0004
BMI=Obese	0.059
BMI=Healthy	-0.0025
BMI=Over Weight	-0.0655
BMI=Under Weight	-0.123
HDL=Low	-0.1699
HDL=Healthy	0.0218
HDL= High	0.1785
LDL=Optimal	0.0756
LDL=Near Optimal	0.0288
LDL= Borderline High	-0.1045
LDL=High	-0.1957

Total cholesterol=Desirable	-0.0042
Total cholesterol=Borderline High	-0.0197
Total Cholesterol=very high	0.0669
Triglyceride=Normal	0.0273
Triglyceride=High	-0.0526
Triglyceride=Borderline High	0.009
Triglyceride=Very High	-0.222
Duration of diabetes	0.0004
Intercept	1.5145

5.10 MOC 7: MEDICAL MODEL WITH LIMITED DENTAL DATA

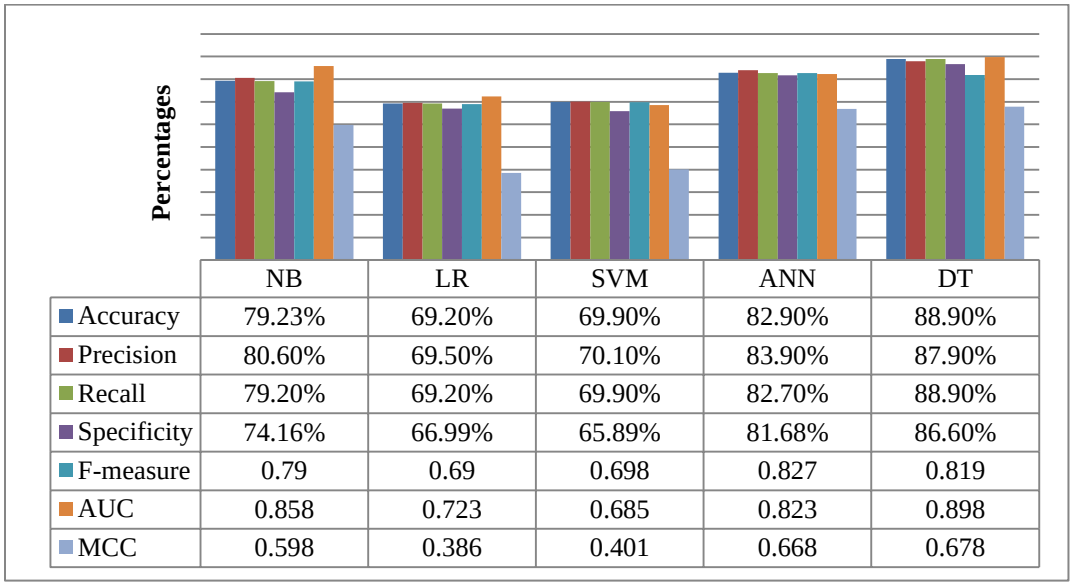
5.10.1 PATIENT CHARACTERISTICS

Of the 4,000 randomly selected medical and dental patients, 59.3% (2,372/4000) were female. The overall mean age of the patients was 49.97 years (SD 16.36 years). The distribution of the patients for presence or absence of diabetes was: 1089 'Type 2 Diabetes, 1183 'Pre-diabetes' and 1728 'No diabetes'. A majority of patients (621/991; 62.7%) diagnosed with diabetes fell under the category3 (duration of diabetes <5 years). Approximately 130 (3%) of patients had documentation of poor oral hygiene while the rest were categorized as good, fair and excellent. Of the 4,000 patients, 1128 (28%) were current smokers, 1464 were former smokers (36.7%) and 1411 (35%) never smoked tobacco. The total number of patients with Medicaid was 2,867 and with Medicare were 1,670. About 3,863 patients regularly brushed their teeth. The mean frequency of tooth brushing was 1.5$\pm$ 0.6 and that of flossing was 0.5 $\pm$ 0.6. The range of body mass index for the MOC patient cohort was between 16.92 and 79.53.

5.10.2 RESULTS OF FIVE ALGORIHTMS ON ALL THE VARIABLES

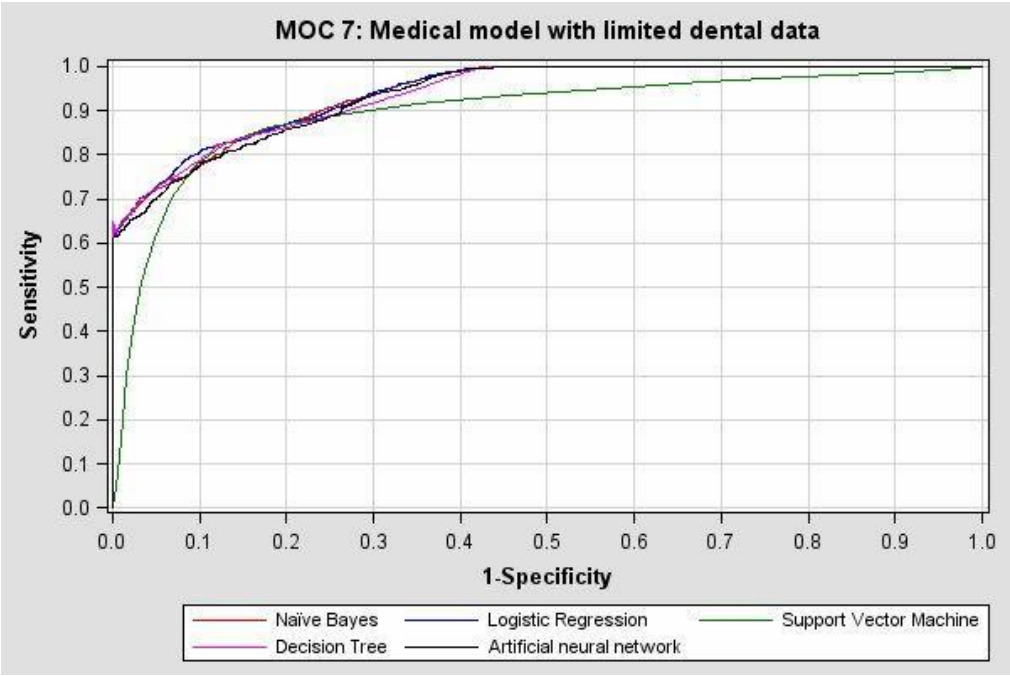
The various performance metrics show that DT outperformed all the other classifiers. Sensitivity was highest in DT [86.21% (95%CI 84.56 to 87.75)], while specificity was highest in SVM [61.86% (95%CI: 60.85-62.87)]. MCC and recall was lowest for LR. **Figure 58** show the results of application of ML algorithms to MOC 7

Figure 58: Results of application of ML algorithms to MOC 7



	NB	LR	SVM	ANN	DT
Accuracy	79.23%	69.20%	69.90%	82.90%	88.90%
Precision	80.60%	69.50%	70.10%	83.90%	87.90%
Recall	79.20%	69.20%	69.90%	82.70%	88.90%
Specificity	74.16%	66.99%	65.89%	81.68%	86.60%
F-measure	0.79	0.69	0.698	0.827	0.819
AUC	0.858	0.723	0.685	0.823	0.898
MCC	0.598	0.386	0.401	0.668	0.678

Figure 59 ROC curve analysis displaying five ROC curves that are representing different levels of performance of the classifiers for MOC 7



The ROC curves in the figure begin at point (0, 0) and end at (1, 1). DT, NB and ANN show higher curves as compared to LR and SVM. The ROC curves of algorithms in descending order of their performance are DT>NB>ANN> LR>SVM.

5.10.3 SUMMARY

This dataset contained 4,000 instances with 22 variables. This model of care incorporated patient reported medical data including presence or absence of diabetes and duration of diabetes along with oral hygiene behavior including frequency of tooth brushing, flossing and tobacco and variables such as number of teeth present, presence or absence of calculus, oral hygiene status, lipid panels including HDL, LDL, triglyceride and total cholesterol. Overall DT outperformed all the other algorithms. The ROC curve for DT, ANN and NB were almost symmetrical followed by LR and SVM.

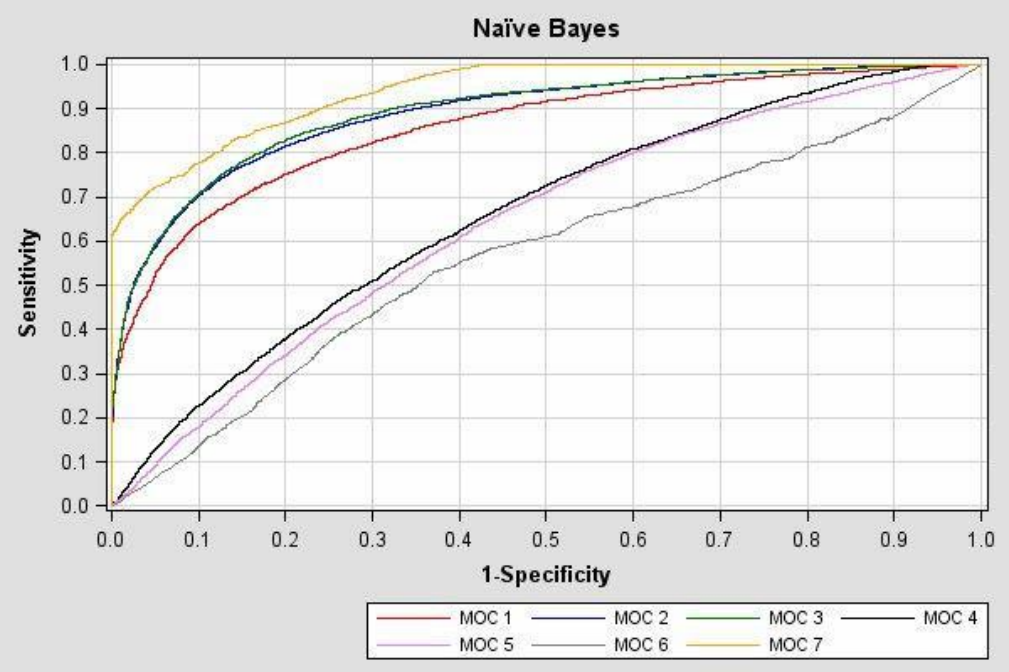
5.11.PERFORMANCE METRICS FOR INDIVIDUAL ALGORITHMS

The ROC curve plot of test sensitivity (true positive rate) versus 1-specificity (false positive rate) was also used to evaluate the performance of diagnostic tests for each of the algorithm on the MOC.

5.11.1. NAÏVE BAYES

5.11.1.1. ROC CURVE FOR NAÏVE BAYES

Figure 60: ROC curve displaying seven ROC curves that are representing different levels of performance of the MOCs in Naïve Bayes.

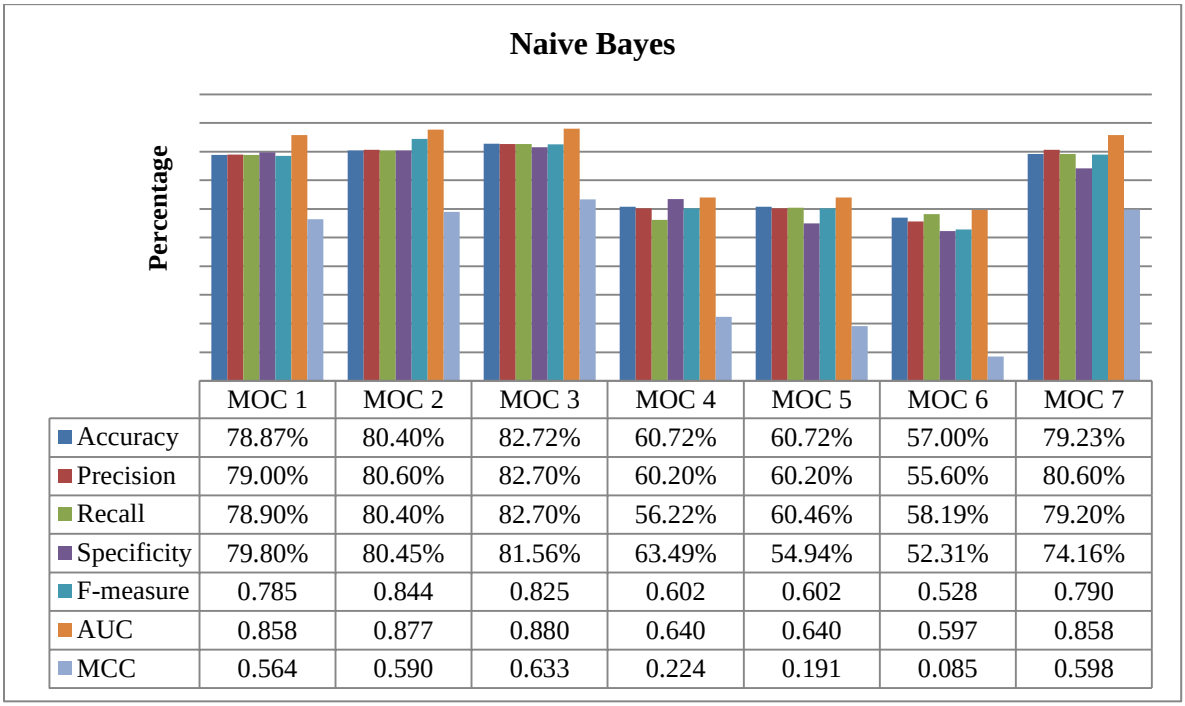


The ROC curves for MOC 1, MOC 2, MOC 3 and MOC 7 outperforms MOC 4, MOC 5 and MOC 6. The ROC curve for MOC 7 dominates the other MOCs and appears to be the best. The ROC curves for MOC 4, MOC 5 and MOC 6 reflects the performance of a diagnostic test that is no better than chance level, wherein the test yields the positive or negative results unrelated to the high PD risk status. The performance of ROC curves for MOCs in descending order: MOC 7> MOC 3> MOC 2> MOC 1> MOC 4> MOC 5> MOC 6.

5.11.1.2. PERFORMANCE METRICS WITH ALL DATA VARIABLES

The results of the performance metrics for NB show that NB performs well in MOC 3 followed by MOC 2. Precision values for MOC 7 and MOC 2 are almost similar; while accuracy, recall and specificity of MOC 7 is similar to MOC 1. Figure 61 shows the results of performance metrics of NB.

Figure 61: Results of performance metrics of Naïve Bayes in all seven models of care

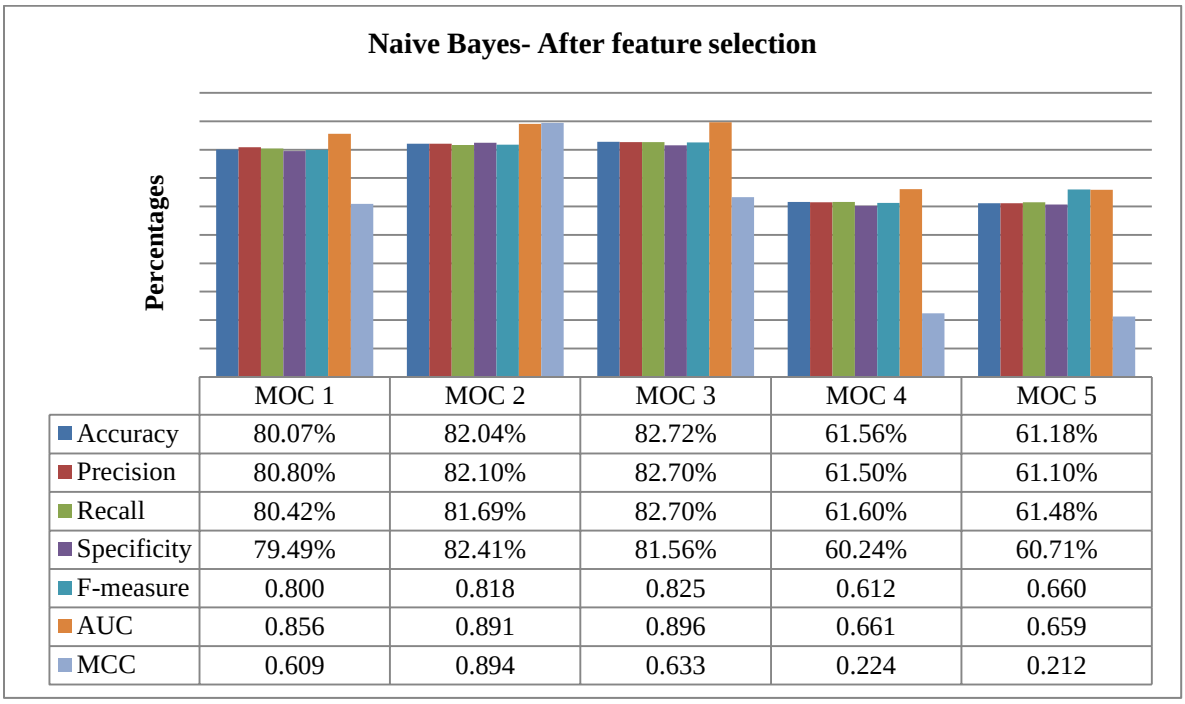


	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5	MOC 6	MOC 7
Accuracy	78.87%	80.40%	82.72%	60.72%	60.72%	57.00%	79.23%
Precision	79.00%	80.60%	82.70%	60.20%	60.20%	55.60%	80.60%
Recall	78.90%	80.40%	82.70%	56.22%	60.46%	58.19%	79.20%
Specificity	79.80%	80.45%	81.56%	63.49%	54.94%	52.31%	74.16%
F-measure	0.785	0.844	0.825	0.602	0.602	0.528	0.790
AUC	0.858	0.877	0.880	0.640	0.640	0.597	0.858
MCC	0.564	0.590	0.633	0.224	0.191	0.085	0.598

5.11.1.3. PERFORMANCE METRICS AFTER FEATURE SELECTION

Figure 62 shows the results of performance metrics of NB after feature selection

Figure 62: Results of the performance metrics of Naïve Bayes after feature selection



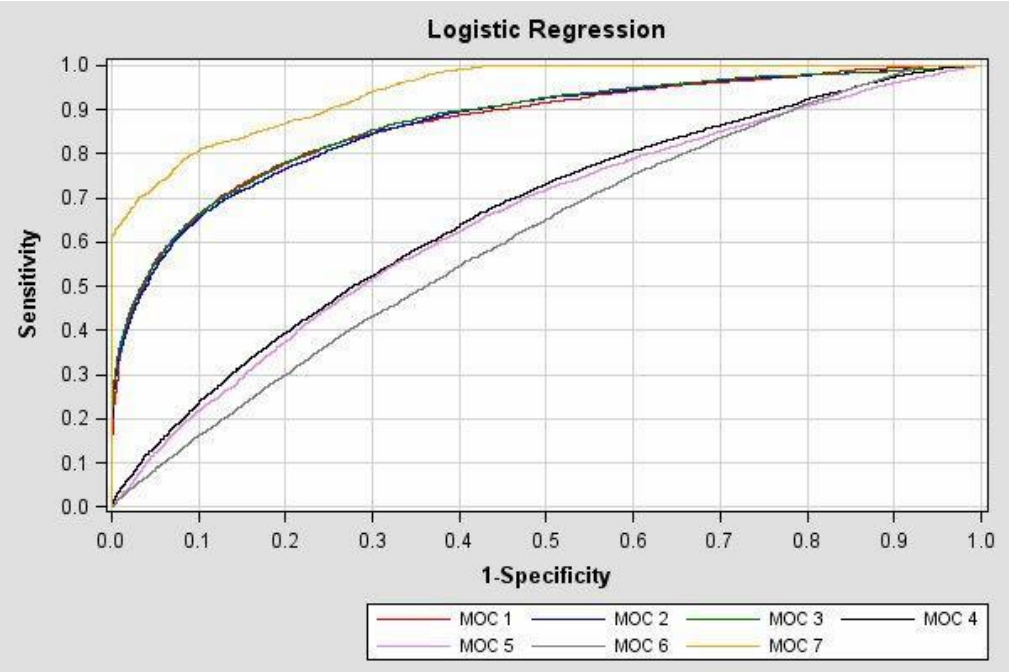
	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5
Accuracy	80.07%	82.04%	82.72%	61.56%	61.18%
Precision	80.80%	82.10%	82.70%	61.50%	61.10%
Recall	80.42%	81.69%	82.70%	61.60%	61.48%
Specificity	79.49%	82.41%	81.56%	60.24%	60.71%
F-measure	0.800	0.818	0.825	0.612	0.660
AUC	0.856	0.891	0.896	0.661	0.659
MCC	0.609	0.894	0.633	0.224	0.212

The results of feature selection show that MOC 3 and MOC 2 perform well in terms of accuracy, precision, recall and specificity as compared to other MOCs. The performance metrics for MOC 1 is slightly lower than MOC 2 and MOC 3, however more than MOC 4 and MOC 5.

5.11.2. LOGISTIC REGRESSION

5.11.2.1 ROC CURVE FOR LOGISTIC REGRESSION

Figure 63: ROC curve displaying seven ROC curves that are representing different levels of performance of the MOCs in Logistic Regression



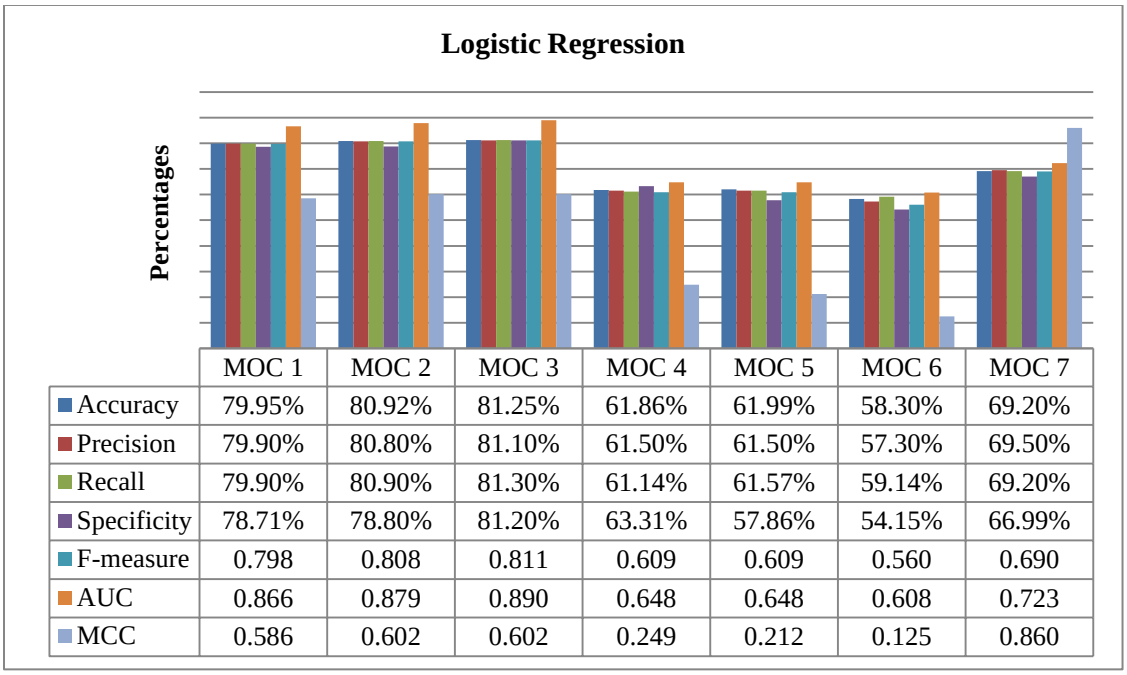
The ROC curves of MOC 7, 1, 2 and 3 are higher than MOC 4, MOC 5 and MOC 6.

The ROC curves of MOC 4, 5 and 6 are very close to each other and along the imaginary diagonal line connecting (0, 0) and (1, 1) representing a random performance. The ROC curves for MOC 7, 1, 2 and 3 represent a better performing model as compared to MOC 4, 5 and 6. The performance of ROC curves for MOCs in descending order: MOC 7> MOC 2> MOC 3> MOC 1> MOC 4> MOC 5> MOC 6.

5.11.2.2. PERFORMANCE METRICS WITH ALL DATA VARIABLES

Figure 64 shows the performance metrics of LR

Figure 64: Results of performance metrics of Logistic Regression in all seven models of care



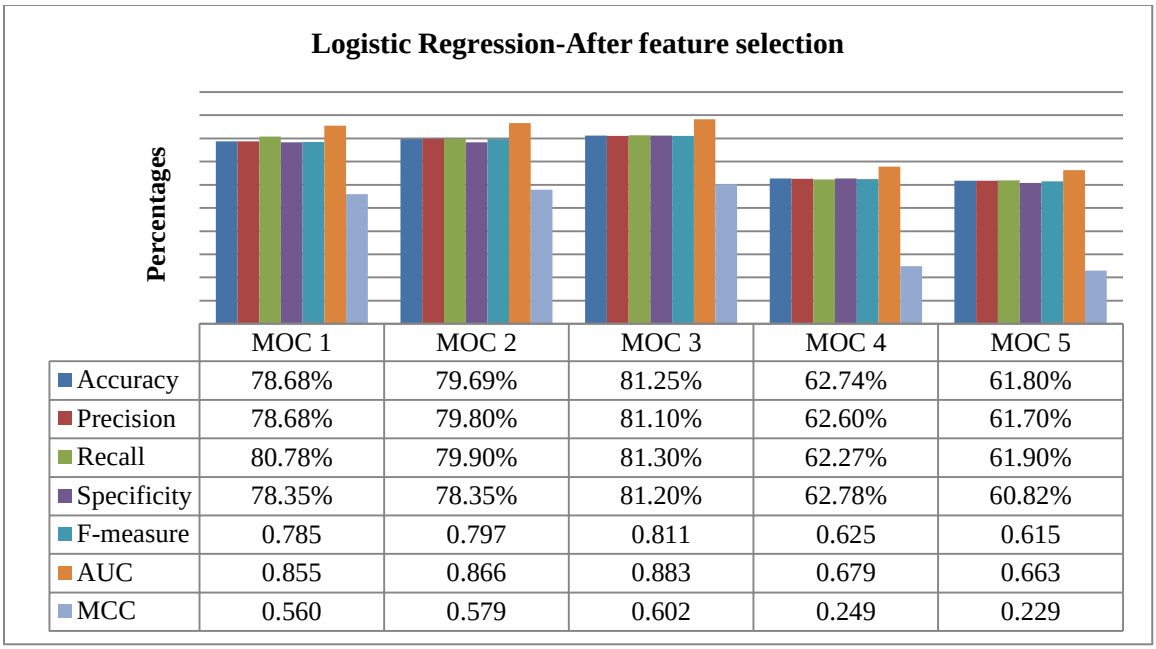
	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5	MOC 6	MOC 7
Accuracy	79.95%	80.92%	81.25%	61.86%	61.99%	58.30%	69.20%
Precision	79.90%	80.80%	81.10%	61.50%	61.50%	57.30%	69.50%
Recall	79.90%	80.90%	81.30%	61.14%	61.57%	59.14%	69.20%
Specificity	78.71%	78.80%	81.20%	63.31%	57.86%	54.15%	66.99%
F-measure	0.798	0.808	0.811	0.609	0.609	0.560	0.690
AUC	0.866	0.879	0.890	0.648	0.648	0.608	0.723
MCC	0.586	0.602	0.602	0.249	0.212	0.125	0.860

The results of the performance metrics for LR show that LR performs well in MOC 3 followed by MOC 2 and MOC 1. The MCC of MOC 7 outperforms all the MCC values of other MOCs. The AUC of MOC 3 outperforms other MOCs.

5.11.2.4. PERFORMANCE METRICS AFTER FEATURE SELECTION

Figure 65 shows the performance metrics of LR after feature selection

Figure 65: Results of the performance metrics of Logistic Regression after feature selection



	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5
Accuracy	78.68%	79.69%	81.25%	62.74%	61.80%
Precision	78.68%	79.80%	81.10%	62.60%	61.70%
Recall	80.78%	79.90%	81.30%	62.27%	61.90%
Specificity	78.35%	78.35%	81.20%	62.78%	60.82%
F-measure	0.785	0.797	0.811	0.625	0.615
AUC	0.855	0.866	0.883	0.679	0.663
MCC	0.560	0.579	0.602	0.249	0.229

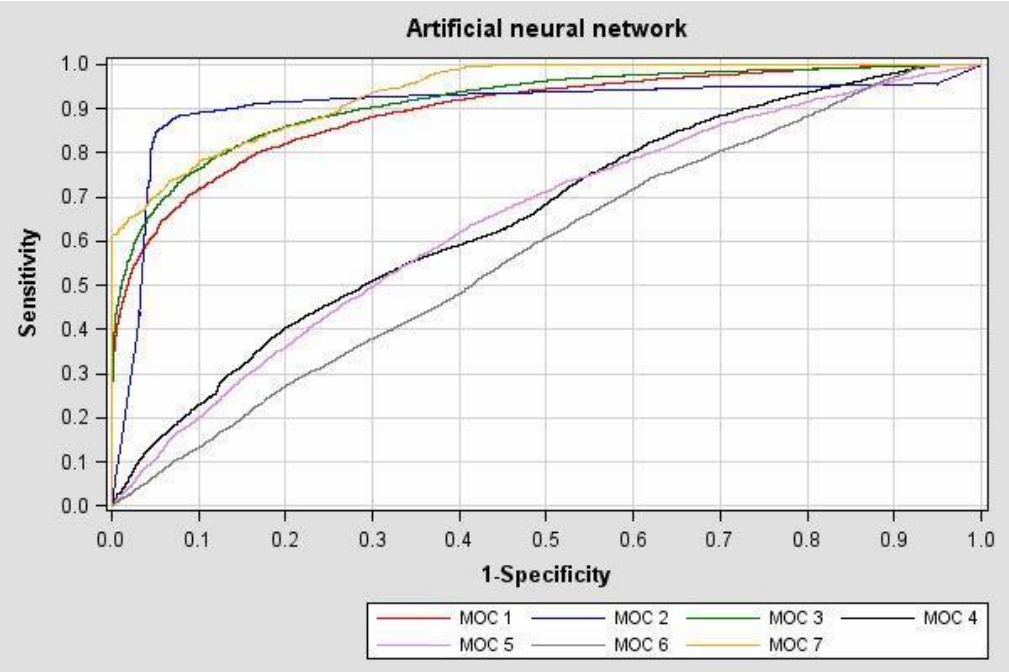
The results of feature selection show that MOC 3 outperforms other MOCs. The recall of MOC 1 is higher than MOC 2 and slightly lower than MOC 3.

5.11.3 ARTIFICIAL NEURAL NETWORK

5.11.3.1. ROC CURVE FOR ARTIFICIAL NEURAL NETWORK

Figure 66 shows the ROC curves for Artificial Neural Network for various models.

Figure 66: ROC curve displaying seven ROC curves that are representing different levels of performance of the MOCs in Artificial Neural Network

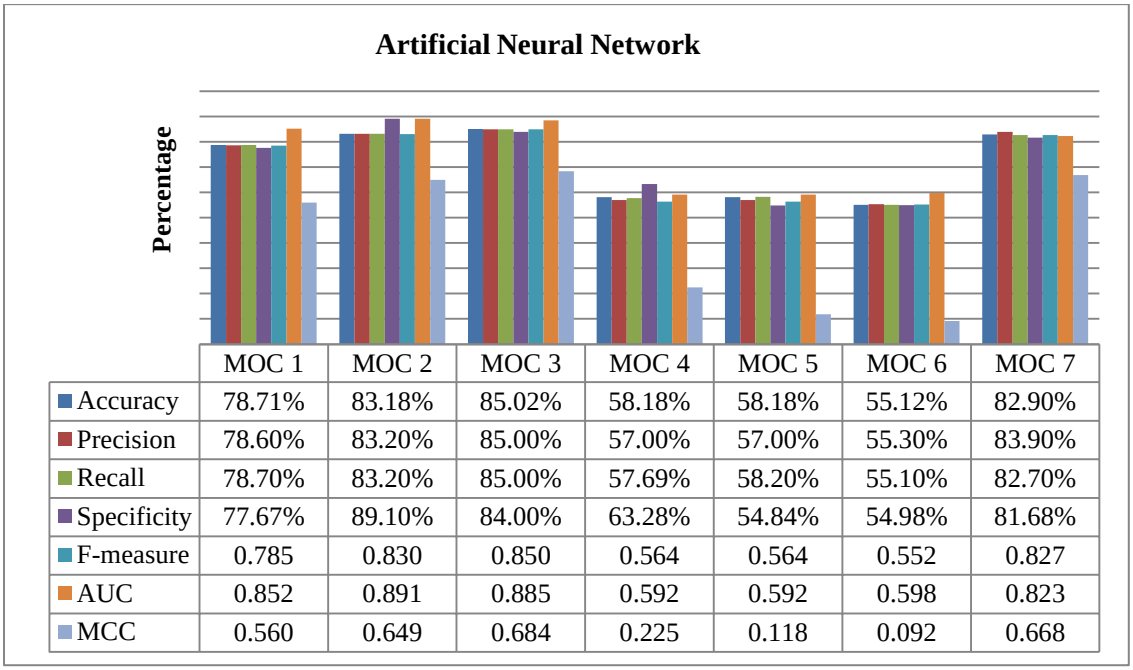


The ROC curves shows that MOC 2 had increasing discriminating power and accuracy when predicting PD risk as compared to MOC 7, 1,3,4,5 and 6. MOC 2, MOC 3 and MOC 1 outperformed MOC 4, MOC 5 and MOC 6. The ROC curve of MOC 6 lies along the diagonal representing a random performance. The ROC of MOC 2 begins at (0, 0) and eventually bends towards the right at (0.02, 0.1) and then runs vertical to (0.05, 0.85) indicating more true positives than false positives and correspondingly signaling a greater noise. The performance of ROC curves for MOCs in descending order: MOC 2> MOC 7> MOC 3> MOC 1> MOC 4> MOC 5> MOC 6

5.11.3.2.PERFORMANCE METRICS WITH ALL DATA VARIABLES

The results of the performance metrics for ANN show that ANN performs well in MOC 3 followed by MOC 2 and then MOC 7. The AUC is highest for MOC 2 followed by MOC 3, MOC 7 and MOC1. MOC 5 and MOC 6 perform poorly as compared to other MOCs. **Figure 67** shows the performance metrics of ANN

Figure 67: Results of performance metrics of Artificial Neural Network in all seven models of care

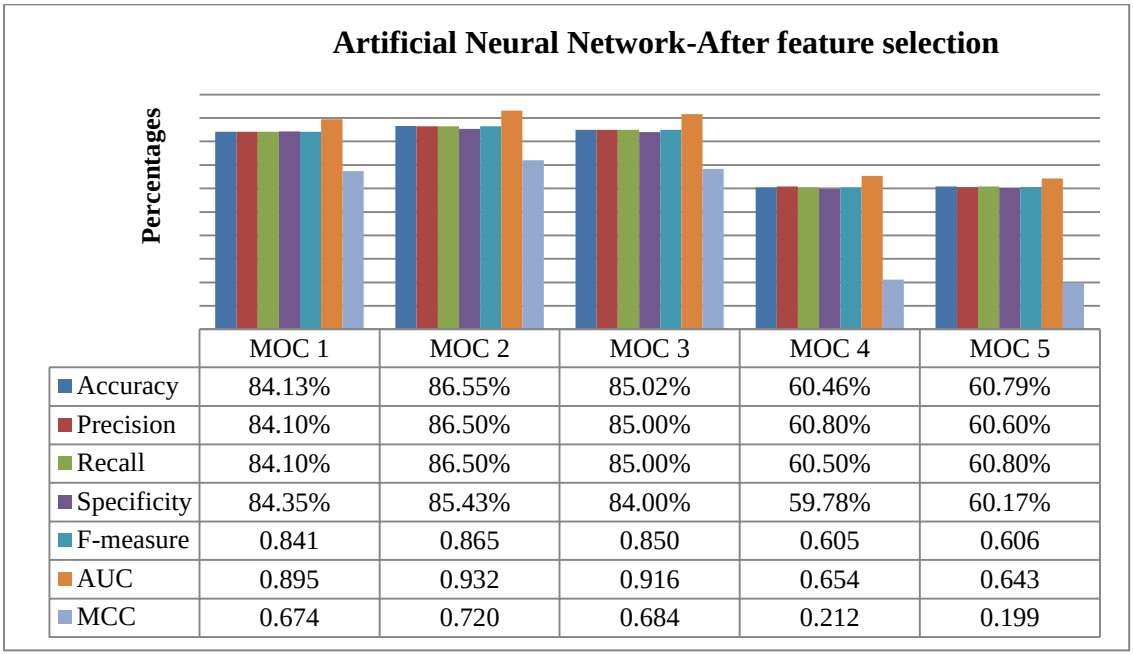


	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5	MOC 6	MOC 7
Accuracy	78.71%	83.18%	85.02%	58.18%	58.18%	55.12%	82.90%
Precision	78.60%	83.20%	85.00%	57.00%	57.00%	55.30%	83.90%
Recall	78.70%	83.20%	85.00%	57.69%	58.20%	55.10%	82.70%
Specificity	77.67%	89.10%	84.00%	63.28%	54.84%	54.98%	81.68%
F-measure	0.785	0.830	0.850	0.564	0.564	0.552	0.827
AUC	0.852	0.891	0.885	0.592	0.592	0.598	0.823
MCC	0.560	0.649	0.684	0.225	0.118	0.092	0.668

5.11.3.3.PERFORMANCE METRICS AFTER FEATURE SELECTION

Figure 68 shows the performance metrics of ANN after feature selection.

Figure 68: Results of the performance metrics of Artificial Neural Network after feature selection



	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5
Accuracy	84.13%	86.55%	85.02%	60.46%	60.79%
Precision	84.10%	86.50%	85.00%	60.80%	60.60%
Recall	84.10%	86.50%	85.00%	60.50%	60.80%
Specificity	84.35%	85.43%	84.00%	59.78%	60.17%
F-measure	0.841	0.865	0.850	0.605	0.606
AUC	0.895	0.932	0.916	0.654	0.643
MCC	0.674	0.720	0.684	0.212	0.199

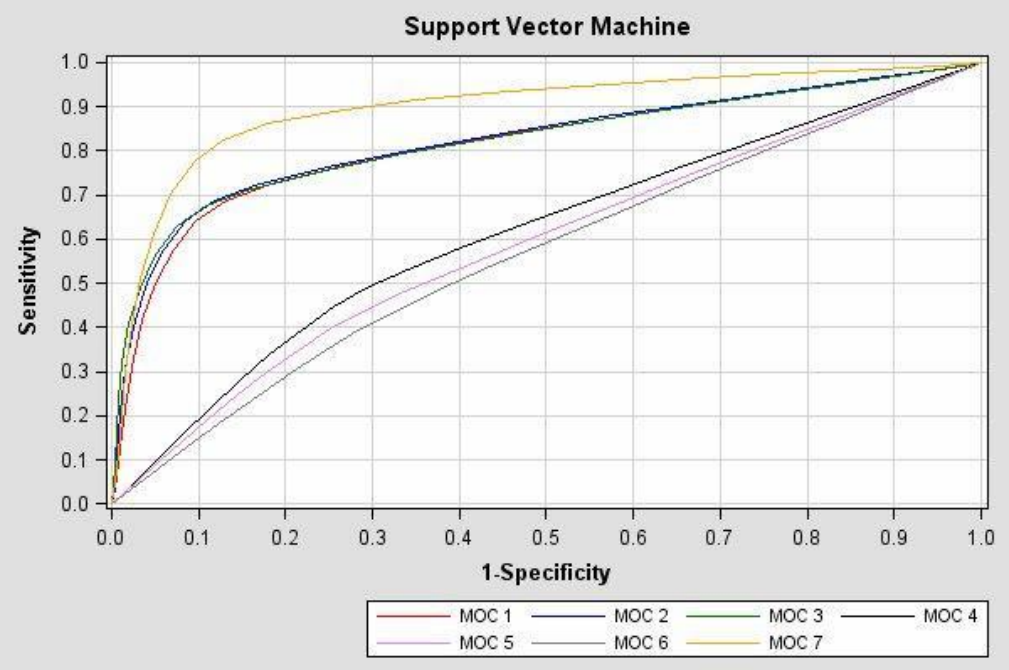
The results of feature selection show that MOC 2 outperforms other MOCs with respect to all performance metric values.

5.11.4. SUPPORT VECTOR MACHINE

5.11.4.1. ROC CURVE FOR SUPPORT VECTOR MACHINE

Figure 69 show the ROC curves for Support Vector Machines.

Figure 69: ROC curve displaying seven ROC curves that are representing different levels of performance of the MOCs in Support Vector Machine

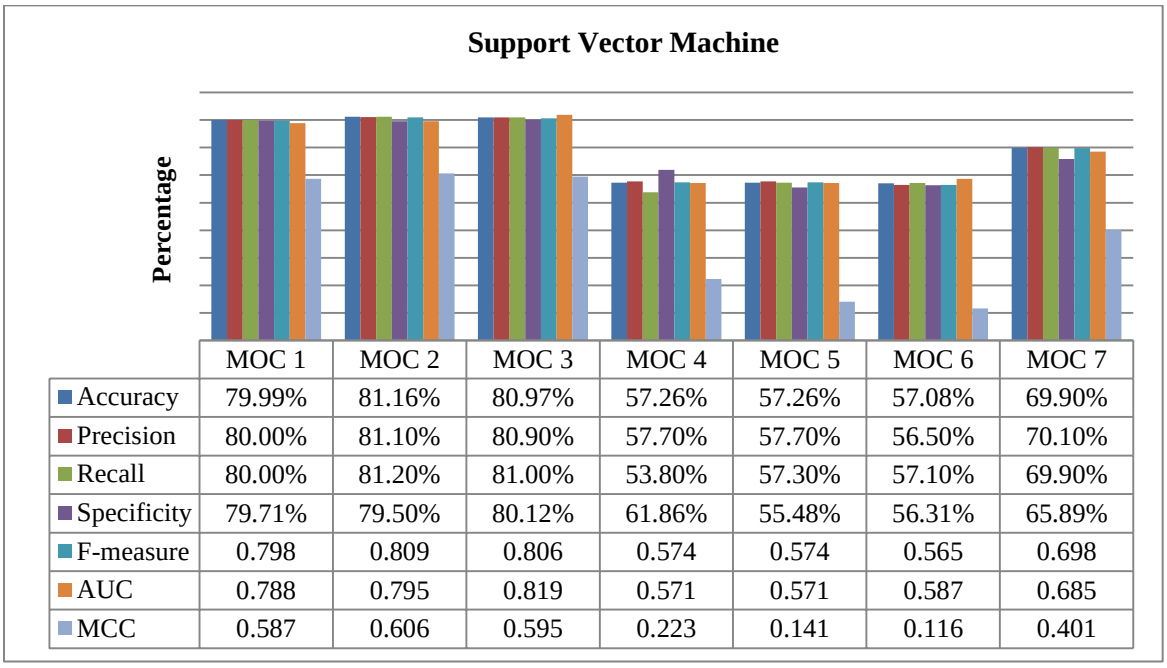


The ROC curves shows that MOC 7 outperformed MOC 1, 2, 3, 4, 5 and 6. MOC 4, MOC 5 and MOC 6 show symmetric curves that lie along the diagonal representing a poor performance. The ROC curve of MOC 6 lies along the diagonal representing a random performance. The ROC of MOC 1, 2 and 3 starts overlapping each other from (0.2, 0.74) and eventually ends at (1, 1). The performance of ROC curves for MOCs in descending order: MOC 7> MOC 3> MOC 2> MOC 1> MOC 4> MOC 5> MOC 6

5.11.4.2.PERFORMANCE METRICS OF ALL DATA VARIABLES

Figure 70 shows the performance of SVM for all MOCs

Figure 70: Results of performance metrics of Support Vector Machine in all seven models of care



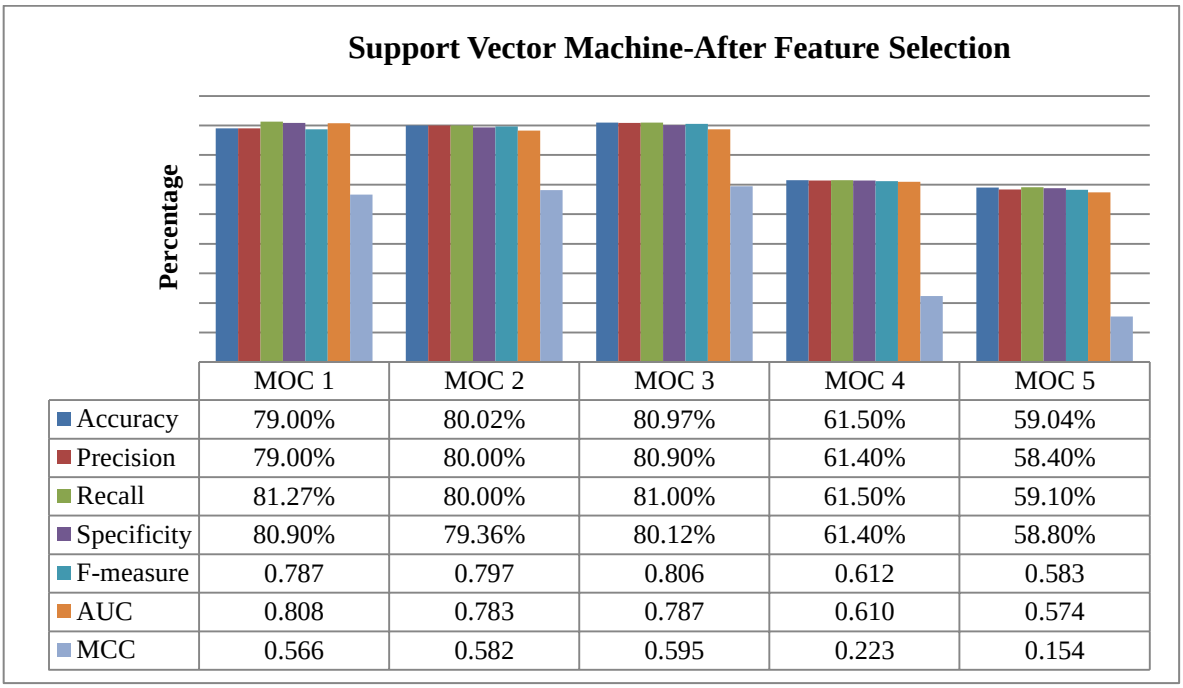
	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5	MOC 6	MOC 7
Accuracy	79.99%	81.16%	80.97%	57.26%	57.26%	57.08%	69.90%
Precision	80.00%	81.10%	80.90%	57.70%	57.70%	56.50%	70.10%
Recall	80.00%	81.20%	81.00%	53.80%	57.30%	57.10%	69.90%
Specificity	79.71%	79.50%	80.12%	61.86%	55.48%	56.31%	65.89%
F-measure	0.798	0.809	0.806	0.574	0.574	0.565	0.698
AUC	0.788	0.795	0.819	0.571	0.571	0.587	0.685
MCC	0.587	0.606	0.595	0.223	0.141	0.116	0.401

The results of the performance metrics for SVM show that SVM performs well in MOC 2 followed by MOC 3 and then MOC 1. The AUC is highest for MOC 3 followed by MOC 2, and MOC1. MOC 5, MOC 6 and MOC 7 perform poorly as compared to other MOCs.

5.11.4.3. PERFORMANCE METRICS AFTER FEATURE SELECTION

Figure 71 shows the performance of SVM after feature selection

Figure 71: Results of the performance metrics of Support Vector Machine after feature selection



	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5
Accuracy	79.00%	80.02%	80.97%	61.50%	59.04%
Precision	79.00%	80.00%	80.90%	61.40%	58.40%
Recall	81.27%	80.00%	81.00%	61.50%	59.10%
Specificity	80.90%	79.36%	80.12%	61.40%	58.80%
F-measure	0.787	0.797	0.806	0.612	0.583
AUC	0.808	0.783	0.787	0.610	0.574
MCC	0.566	0.582	0.595	0.223	0.154

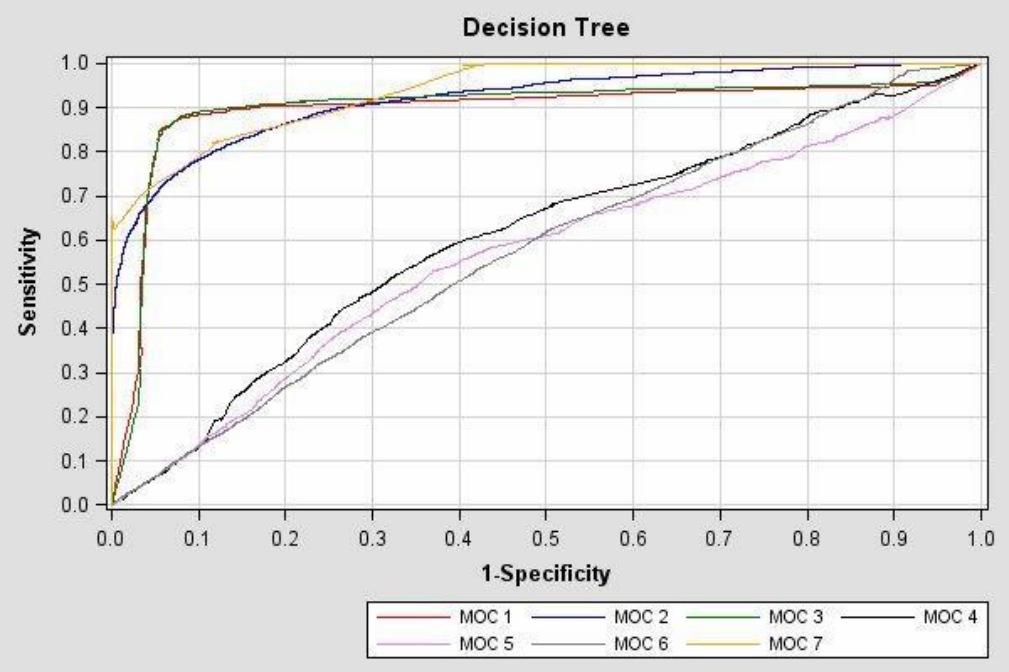
The results of feature selection show that MOC 2 and MOC 3 perform well in terms of accuracy and precision. The recall of MOC 1 is higher than other recall of other MOCs

5.11.5. DECISION TREE

5.11.5.1. ROC CURVE FOR DECISION TREE

Figure 72 show the ROC curves for Decision Tree.

Figure 72: ROC curve displaying seven ROC curves that are representing different levels of performance of the MOCs in Decision Tree

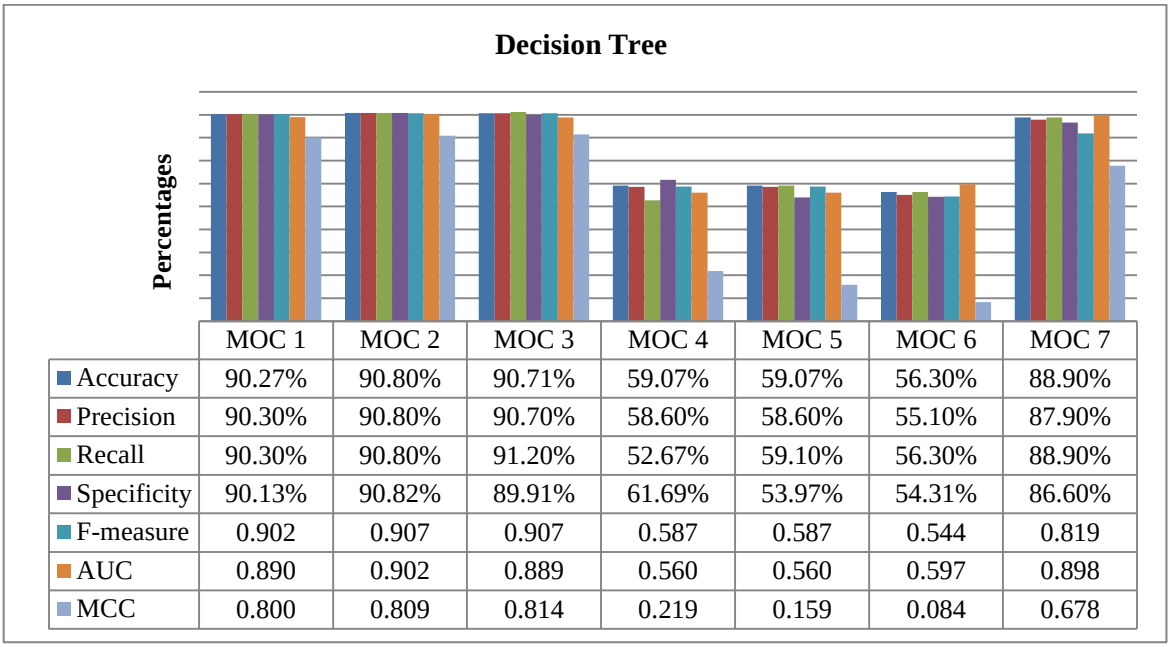


The ROC curves shows that MOC 1 and MOC 3 had increasing discriminating power and accuracy when predicting PD risk as compared to MOC 7, 2, 4, 5 and 6. The curve of MOC 1 begins at (0,0) and slightly bends to the right and then runs vertically where it overlaps the ROC curve of MOC 3 at (0.05, 0.85) and later runs along with ROC curve of MOC 3 to end at (1,1). The ROC curve of MOC 5 lies along the diagonal and crosses the diagonal at (0.75, 0.75) indicating worse than random performance. The ROC of MOC 4 and 6 run close to the diagonal indicating a random performance. The performance of ROC curves for MOCs in descending order: MOC 1> MOC 3> MOC 7> MOC 2> MOC 4> MOC 6> MOC 5.

5.11.5.2.PERFORMANCE METRICS FOR ALL THE DATA VARIABLES

Figure 73 shows the performance metrics of DT

Figure 73: Results of performance metrics of Decision Tree in all seven models of care



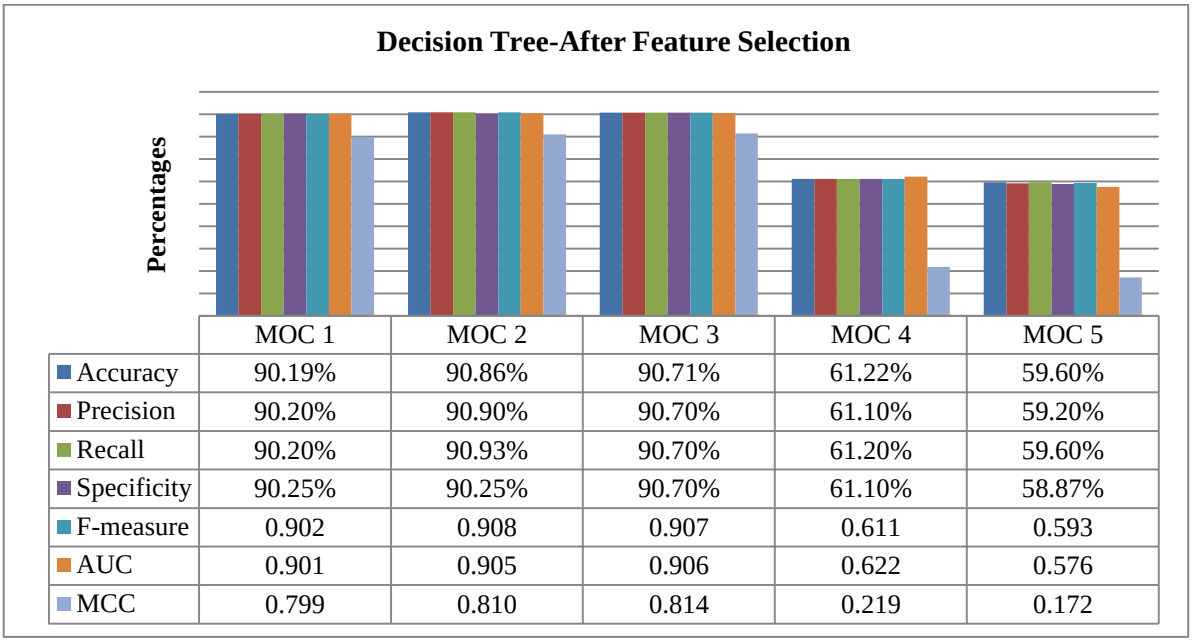
	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5	MOC 6	MOC 7
Accuracy	90.27%	90.80%	90.71%	59.07%	59.07%	56.30%	88.90%
Precision	90.30%	90.80%	90.70%	58.60%	58.60%	55.10%	87.90%
Recall	90.30%	90.80%	91.20%	52.67%	59.10%	56.30%	88.90%
Specificity	90.13%	90.82%	89.91%	61.69%	53.97%	54.31%	86.60%
F-measure	0.902	0.907	0.907	0.587	0.587	0.544	0.819
AUC	0.890	0.902	0.889	0.560	0.560	0.597	0.898
MCC	0.800	0.809	0.814	0.219	0.159	0.084	0.678

The results of the performance metrics for DT show that DT performs well in MOC 2 followed by MOC 3 and then MOC 1. The AUC of MOC 7 and MOC 3 are similar. MOC 5, MOC 6 and MOC 7 perform poorly as compared to other MOCs

5.11.5.3.PERFORMANCE METRICS AFTER FEATURE SELECTION

Figure 74 shows the performance metrics of DT after feature selection

Figure 74: Results of the performance metrics of Decision Tree after feature selection



	MOC 1	MOC 2	MOC 3	MOC 4	MOC 5
Accuracy	90.19%	90.86%	90.71%	61.22%	59.60%
Precision	90.20%	90.90%	90.70%	61.10%	59.20%
Recall	90.20%	90.93%	90.70%	61.20%	59.60%
Specificity	90.25%	90.25%	90.70%	61.10%	58.87%
F-measure	0.902	0.908	0.907	0.611	0.593
AUC	0.901	0.905	0.906	0.622	0.576
MCC	0.799	0.810	0.814	0.219	0.172

The results of feature selection show that there is slight difference in terms of all performance metrics between MOC 1, MOC 2 and MOC 3. The F-measure is almost similar for MOC 2 and MOC 3.

Chapter 6

DISCUSSION

6.1. OVERALL DISCUSSION

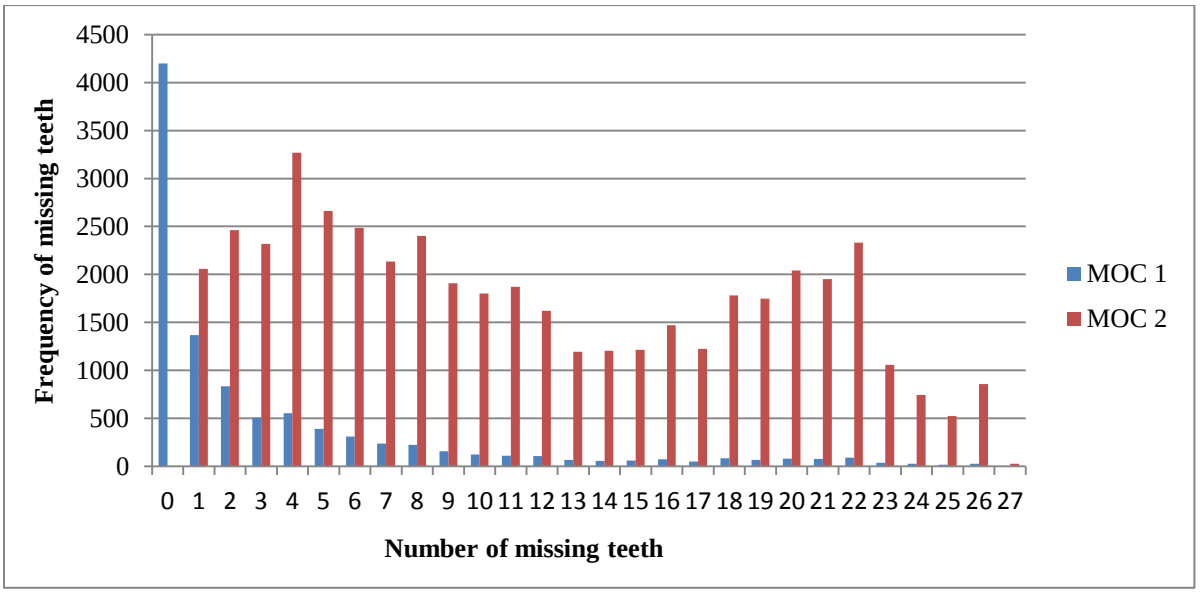
Scientific studies continue to produce evidence of oral-systemic associations, thereby promoting a focus on oral health screenings. Based on the general principles of oral health screening in medical and dental settings, this study presents a rationale for a preventive approach involving determination of a patients' risk for developing PD by assessment via seven models of care (MOC1 to MOC7) in an interdisciplinary setting. This retrospective study compared five predictive machine learning algorithms: NB, LR, SVM, ANN and DT, with no missing data, which has not been reported previously to identify patients at high risk for PD in interdisciplinary settings.

6.2. MODEL OF CARE

Overall, MOC 7, MOC 1, MOC 2 and MOC 3 exhibited a promising model for assessing PD risk as compared to MOC 4, MOC 5 and MOC 6. Based on the results of MOC 1, 2, 3 which included a subset of the comprehensive periodontal examination, including PPD, the latter variables carried a significant weight in determining PD risk. It is posited that the inclusion of only medical variables without any dental variables in MOC 4, 5 and 6 contributed to poor performance. An increase in total accuracy was seen to some extent for MOC 4 which incorporated patient reported dental data that was limited to oral hygiene habits such as frequency of tooth brushing, flossing and historical data surrounding tobacco use status.

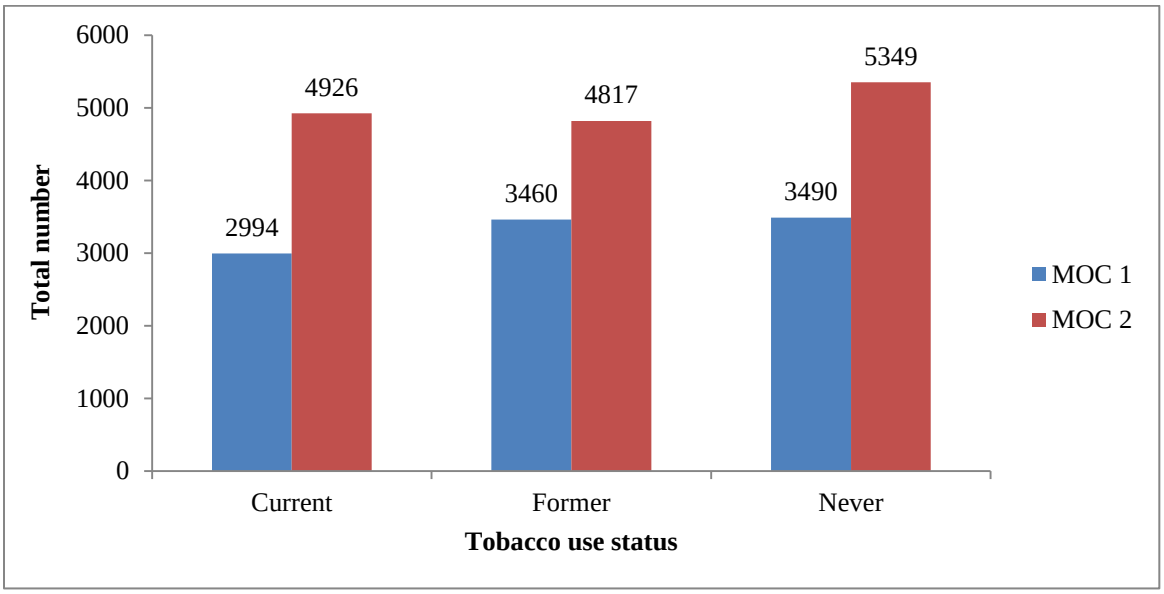
MOC 1 and MOC 7 explicitly identified the various broad determinants of PD risk, collected in an integrated care setting. Similarly, MOC-2 showed a slightly increased accuracy compared to MOC 1 and MOC 3 and displayed a favorable model that can be used in dental settings. It is posited that the increased accuracy of MOC 2 over MOC 1 may have been attributable to the sample size (n=9,944 in MOC 1 and n=15,092 in MOC 2) utilized in cross validation of the five algorithms in their respective MOCs. Tooth loss is considered as marker of long-term cumulative PD [136]. Closely looking at the MOC 1 and MOC 2 dataset, the frequency of number of missing teeth was more in MOC 2 and compared the MOC 1. Figure x shows the number of missing teeth in MOC 1 and MOC 2. The results of higher accuracy in MOC 2 as compared to MOC 1 could be attributable for the increasing number of missing teeth in MOC 2. Figure 75 shows the number of missing teeth in dataset of MOC 1 and MOC 2

Figure 75: Frequency of number of missing teeth in datasets of MOC 1 and MOC 2



Similarly, tobacco use has been associated with increase in incidence of tooth loss [136]. Figure Y shows the tobacco use status distribution in MOC 1 and MOC 2. Figure 76 shows the number of current smokers and former smokers are more in MOC 2 as compared to MOC 1

Figure 76: Total number of current smokers and former smokers are more in MOC 2 as compared to MOC 1



It is also posited that the increased accuracy of MOC 2 over MOC 3 may be due to inclusion patient-reported information such as presence or absence of diabetes, duration of diabetes, height, weight, tobacco use status, frequency of tooth brushing and flossing.

6.3.ALGORITHMS

Performance measure assessment showed that DT demonstrated higher analytical accuracy for disease risk classification than all other ML algorithms. The overall accuracy for DT in MOC 1, MOC 2 and MOC 3 was 90.31%, 90.86% and 90.71%, respectively. The collateral results after the empirical validation of the resultant models of MOC 1, MOC 2 and MOC 3 with external evaluation sets showed that the ability to identify true positive and false negative cases by all the DT in MOC1 and MOC 2 was higher as compared to MOC 3.

Following the performance of the DT in MOC 1, ANN emerged as a reliable algorithm to assess PD risk in interdisciplinary and dental settings. The present study utilized ANN applying a backpropagation method with a nonlinear sigmoid function in the hidden layer. However, the time required for cross validation for ANN ranged between 386 seconds and 7600 seconds. Although ANN is considered to be successful in tackling a wide range of problems, the run time was higher than that of other modeling approaches. Testing application of the algorithm to this cohort by modifying the sigmoid function to a hyperbolic tangent function (HTAN) and including an adaptive normalization routine as conducted in one of the studies [137] would be worthwhile. The results of HTAN showed that the run time was markedly decreased with least error. Results in the current study showed that ensemble methods exhibited improved performance compared to individual algorithms. Ambivalence associated with bagging could be due to use of the plurality vote, which sometimes results in two or more classes tied in a vote. The results of this technique showed that the high rate of overall identification (sensitivity and low rate of false negative identification), is important to informing preventive measures to reduce missing diagnosis of periodontal disease.

6.4.FEATURE SELECTION

Finding a minimum set of attributes not only enhances the classification accuracy but also the learning runtime [138]. For optimizing the model, this study employed information gain with ranker method for MOC 1, 2 and 3, and CFS with best search method for MOC 4 and MOC 5. The benefit of using multivariate filter such as CFS evaluated the individual predictive ability of each attribute and the degree of redundancy between the attributes, while a univariate filter such as the information gain ranked the features according to the information gain. For example,

this approach identified and eliminated the surfaces of teeth that are least significant for PPD at POC. In this study, these approaches (in MOC 1, 2 and 3) led to the novel observation that measurement of probing depths at interproximal tooth surfaces significantly outperformed measures taken at the buccal and lingual surfaces. The study posits that superiority of interproximal measurement may be attributable to the interproximal bone which is more coronal in position than the labial or lingual/palatal bone. A slight deepening of the pocket in the interproximal areas could more easily impact the bone. This observation challenges current standard of care with respect to commonly applied indices used in clinical dentistry (e.g. Silness and Loe), where buccal and lingual surfaces are the main focus for measuring probe depth [5-9]. The progression of PD involves furcation areas of the multi-rooted teeth in maxillary (upper jaw) and mandibular (lower jaw). Tooth surfaces such as mesiolingual, mesiobuccal, distobuccal and distolingual of maxillary and mandibular molars were identified as significant determinants during feature selection. For calculating clinical attachment loss, location of cemento-enamel junction (CEJ) is necessary. Interestingly, the sites identified by feature selection are also used as reference lines for determining relative clinical attachment (RCAL) loss when it is difficult to locate CEJ [32]. Moreover, these sites were also consistent with outcomes of a study that investigated the deepest crevice points in the mouth to provide the practitioners with minimum number of sites to probe [139].

6.5.DATA VARIABLES

Using multivariate and univariate filtering in various MOCs, the study recognized that random blood glucose, dental calculus, missing teeth, lipid panels including triglyceride levels and HDL, diastolic blood pressure, body mass index, oral hygiene status determined by the

dental provider, frequency of tooth brushing, diabetes status, tobacco use status, age, gender and PPD displayed highest performance in determining PD risk across all the MOCs. Although, evidence suggests that duration of diabetes and flossing of teeth are significant factors in PD severity, from the present work, by virtue of feature selection, factors such as patient self-reported frequency of flossing and duration of diabetes were eliminated from all the MOCs [140][141][142]. Studies have shown that the patient compliance with daily dental flossing was low due to difficulty in flossing [143][144]. Moreover it is also noted that the oral health literacy amongst the adult patients is at a low level that may also interfere with the ability to understand oral health information[145]

On closer examination of the data for duration of diabetes, only 313 patients out of 1709 had duration of Type 2 diabetes history with more than one year, while others were diagnosed within one year of their first visit to dental setting. Modelling and incorporating data with Type 2 diabetic patients with more than one year of history would help in understanding if the factor 'duration of diabetes' is significant or not. Oral hygiene status, which is clinically determined by the dental provider, takes into account the overall conditions present in the oral cavity, was also retained in the MOC 1, 2, 3 by feature selection. Notably, the dental calculus variable was located at the top of the decision tree following the most significant teeth surfaces as shown in Figure 77

Figure 77: MOC 2 decision tree (J4.8) showing the location of dental calculus variable at the top of the decision tree following the most significant teeth surfaces.

```
Calculus <= 1
|   Age <= 43: LOW (4697.0/955.0)
|   Age > 43
|   |   Missing Teeth <= 18
|   |   |   GENDER <= 0
|   |   |   |   Tobacco use status <= 2
```

```
|  |  |  |  |  Calculus <= 0: LOW (51.0/13.0)
|  |  |  |  |  Calculus > 0
|  |  |  |  |  |  Oral hygiene status <= 3
|  |  |  |  |  |  |  Oral hygiene status <= 1
|  |  |  |  |  |  |  |  Age <= 74
|  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: LOW (21.0/7.0)
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 9: HIGH (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 9: LOW (2.0)
|  |  |  |  |  |  |  |  |  |  |  brushing > 1: LOW (2.0)
|  |  |  |  |  |  |  |  |  Tobacco use status > 1: HIGH (18.0/7.0)
|  |  |  |  |  |  |  |  Age > 74: HIGH (11.0/1.0)
|  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  Age <= 47: LOW (178.0/31.0)
|  |  |  |  |  |  |  |  Age > 47
|  |  |  |  |  |  |  |  |  Tobacco use status <= 1: LOW (563.0/175.0)
|  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  brushing <= 1: LOW (94.0/29.0)
|  |  |  |  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 6
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 2
|  |  |  |  |  |  |  |  |  |  |  |  |  brushing <= 2: LOW (209.0/69.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  brushing > 2: HIGH (9.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 2
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 78
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (47.0/19.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 61: HIGH (10.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Age > 61: LOW (14.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Age > 78: LOW (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2: HIGH (4.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 6: LOW (53.0/12.0)
|  |  |  |  |  |  Oral hygiene status > 3
|  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  Missing Teeth <= 0: LOW (238.0/78.0)
|  |  |  |  |  |  |  |  Missing Teeth > 0
|  |  |  |  |  |  |  |  |  Missing Teeth <= 10
|  |  |  |  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1: LOW (67.0/26.0)
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 7: LOW (42.0/15.0)
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 7: HIGH (11.0/1.0)
|  |  |  |  |  |  |  |  |  |  brushing > 1
```

```
|  |  |  |  |  |  |  |  |  |  |  Age <= 51
|  |  |  |  |  |  |  |  |  |  |  |  Age <= 45: HIGH (9.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  Age > 45: LOW (49.0/14.0)
|  |  |  |  |  |  |  |  |  |  |  Age > 51: HIGH (207.0/89.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 10: LOW (73.0/20.0)
|  |  |  |  |  |  |  Diastolic Blood Pressure > 2
|  |  |  |  |  |  |  |  Missing Teeth <= 9
|  |  |  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1: HIGH (6.0/1.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 4: LOW (8.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 4: HIGH (3.0/1.0)
|  |  |  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  |  |  brushing <= 2: LOW (23.0/4.0)
|  |  |  |  |  |  |  |  |  |  brushing > 2: HIGH (2.0)
|  |  |  |  |  |  |  |  Missing Teeth > 9: LOW (10.0)
|  |  |  |  Tobacco use status > 2
|  |  |  |  |  Age <= 54
|  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: LOW (175.0/63.0)
|  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  brushing <= 1: LOW (24.0/8.0)
|  |  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  |  Missing Teeth <= 1: LOW (23.0/6.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 1
|  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 3: HIGH (14.0/5.0)
|  |  |  |  |  |  |  |  |  |  Oral hygiene status > 3
|  |  |  |  |  |  |  |  |  |  |  Age <= 47: LOW (3.0)
|  |  |  |  |  |  |  |  |  |  |  Age > 47: HIGH (7.0)
|  |  |  |  |  |  Diastolic Blood Pressure > 2
|  |  |  |  |  |  |  Missing Teeth <= 5
|  |  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 3: LOW (3.0)
|  |  |  |  |  |  |  |  |  Oral hygiene status > 3: HIGH (4.0/1.0)
|  |  |  |  |  |  |  |  brushing > 1: HIGH (15.0/4.0)
|  |  |  |  |  |  |  Missing Teeth > 5: LOW (6.0/1.0)
|  |  |  |  |  Age > 54
|  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  Age <= 70
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 3
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  |  |  |  |  Age <= 67: LOW (9.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  Age > 67: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1: HIGH (6.0/1.0)
|  |  |  |  |  |  |  |  |  Oral hygiene status > 3
```

```
| | | | | | | | | | Diastolic Blood Pressure <= 1
| | | | | | | | | | | Age <= 62: HIGH (10.0/4.0)
| | | | | | | | | | | Age > 62: LOW (2.0)
| | | | | | | | | | Diastolic Blood Pressure > 1: LOW (4.0)
| | | | | | | | Age > 70: HIGH (4.0)
| | | | | | | Diastolic Blood Pressure > 2: LOW (3.0)
| | | | | | brushing > 1
| | | | | | | Oral hygiene status <= 3: HIGH (63.0/28.0)
| | | | | | | Oral hygiene status > 3
| | | | | | | | Age <= 57: HIGH (7.0)
| | | | | | | | Age > 57
| | | | | | | | | Diastolic Blood Pressure <= 2
| | | | | | | | | | Diastolic Blood Pressure <= 1
| | | | | | | | | | | Missing Teeth <= 9
| | | | | | | | | | | | Missing Teeth <= 1
| | | | | | | | | | | | | Age <= 64
| | | | | | | | | | | | | | Age <= 58: LOW (3.0/1.0)
| | | | | | | | | | | | | | Age > 58: HIGH (2.0)
| | | | | | | | | | | | | Age > 64: LOW (3.0)
| | | | | | | | | | | | Missing Teeth > 1: HIGH (7.0)
| | | | | | | | | | | Missing Teeth > 9: LOW (4.0)
| | | | | | | | | | Diastolic Blood Pressure > 1: LOW (4.0/1.0)
| | | | | | | | | Diastolic Blood Pressure > 2: HIGH (2.0)
| | | GENDER > 0
| | | | Oral hygiene status <= 3
| | | | | Diastolic Blood Pressure <= 2
| | | | | | Tobacco use status <= 2
| | | | | | | Oral hygiene status <= 1
| | | | | | | | Calculus <= 0: HIGH (3.0)
| | | | | | | | Calculus > 0
| | | | | | | | | brushing <= 1: LOW (35.0/7.0)
| | | | | | | | | brushing > 1
| | | | | | | | | | brushing <= 2
| | | | | | | | | | | Missing Teeth <= 2: LOW (8.0/1.0)
| | | | | | | | | | | Missing Teeth > 2: HIGH (11.0/4.0)
| | | | | | | | | | brushing > 2: HIGH (2.0)
| | | | | | | Oral hygiene status > 1: LOW (525.0/215.0)
| | | | | | Tobacco use status > 2
| | | | | | | brushing <= 1
| | | | | | | | Oral hygiene status <= 1
| | | | | | | | | Diastolic Blood Pressure <= 1
| | | | | | | | | | Age <= 52: HIGH (7.0/1.0)
| | | | | | | | | | Age > 52: LOW (5.0/1.0)
| | | | | | | | | Diastolic Blood Pressure > 1
| | | | | | | | | | Missing Teeth <= 8
| | | | | | | | | | | Age <= 56: HIGH (2.0)
```

```
|  |  |  |  |  |  |  |  |  |  |  Age > 56: LOW (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  Missing Teeth > 8: LOW (6.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: LOW (25.0/10.0)
|  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1: HIGH (8.0/2.0)
|  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  Oral hygiene status <= 1: HIGH (6.0/1.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (35.0/13.0)
|  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1: LOW (19.0/7.0)
|  |  |  |  |  Diastolic Blood Pressure > 2
|  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  Diastolic Blood Pressure <= 3
|  |  |  |  |  |  |  |  Missing Teeth <= 9
|  |  |  |  |  |  |  |  |  Tobacco use status <= 1: LOW (7.0/1.0)
|  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 1: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  |  |  Age <= 62: LOW (9.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  Age > 62: HIGH (3.0)
|  |  |  |  |  |  |  |  Missing Teeth > 9: HIGH (3.0)
|  |  |  |  |  |  |  Diastolic Blood Pressure > 3: HIGH (3.0)
|  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  Missing Teeth <= 7: HIGH (55.0/16.0)
|  |  |  |  |  |  |  Missing Teeth > 7: LOW (5.0/1.0)
|  |  |  |  Oral hygiene status > 3
|  |  |  |  Missing Teeth <= 9
|  |  |  |  |  brushing <= 1
|  |  |  |  |  |  Missing Teeth <= 1
|  |  |  |  |  |  |  Missing Teeth <= 0
|  |  |  |  |  |  |  |  brushing <= 0
|  |  |  |  |  |  |  |  |  Tobacco use status <= 2: HIGH (3.0)
|  |  |  |  |  |  |  |  |  Tobacco use status > 2: LOW (3.0/1.0)
|  |  |  |  |  |  |  |  brushing > 0: LOW (95.0/44.0)
|  |  |  |  |  |  |  Missing Teeth > 0
|  |  |  |  |  |  |  |  Tobacco use status <= 2
|  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  Age <= 46: LOW (3.0)
|  |  |  |  |  |  |  |  |  |  |  Age > 46
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 49: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 49: LOW (5.0/1.0)
|  |  |  |  |  |  |  |  |  Tobacco use status > 1: LOW (19.0/8.0)
|  |  |  |  |  |  |  |  Tobacco use status > 2: HIGH (10.0/2.0)
|  |  |  |  |  |  Missing Teeth > 1
```

```
|  |  |  |  |  |  |  |  brushing <= 0
|  |  |  |  |  |  |  |  |  Age <= 70: LOW (5.0)
|  |  |  |  |  |  |  |  |  Age > 70: HIGH (2.0)
|  |  |  |  |  |  |  |  brushing > 0
|  |  |  |  |  |  |  |  |  Missing Teeth <= 3: HIGH (45.0/10.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 3
|  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 7: LOW (10.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 7: HIGH (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  |  |  |  |  Tobacco use status <= 2
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (22.0/9.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 67
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 5: HIGH (6.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 5: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Age > 67: HIGH (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  Tobacco use status > 2
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 56: HIGH (11.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 56: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2: HIGH (7.0/1.0)
|  |  |  |  |  |  brushing > 1: HIGH (278.0/113.0)
|  |  |  |  |  Missing Teeth > 9
|  |  |  |  |  |  Tobacco use status <= 2
|  |  |  |  |  |  |  brushing <= 1: LOW (38.0/8.0)
|  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  Tobacco use status <= 1: HIGH (11.0/4.0)
|  |  |  |  |  |  |  |  Tobacco use status > 1: LOW (21.0/7.0)
|  |  |  |  |  |  Tobacco use status > 2
|  |  |  |  |  |  |  Age <= 50: LOW (9.0/2.0)
|  |  |  |  |  |  |  Age > 50: HIGH (15.0/4.0)
|  |  Missing Teeth > 18: LOW (251.0/33.0)
Calculus > 1
|  Calculus <= 2
|  |  Age <= 35
|  |  |  GENDER <= 0: LOW (1178.0/407.0)
|  |  |  GENDER > 0
|  |  |  |  Oral hygiene status <= 1
|  |  |  |  |  brushing <= 0
|  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  Missing Teeth <= 3
|  |  |  |  |  |  |  |  Missing Teeth <= 0
|  |  |  |  |  |  |  |  |  Age <= 28: HIGH (12.0/3.0)
|  |  |  |  |  |  |  |  |  Age > 28: LOW (3.0)
|  |  |  |  |  |  |  |  Missing Teeth > 0: LOW (8.0)
```

```
| | | | | | | Missing Teeth > 3: HIGH (4.0)
| | | | | | Diastolic Blood Pressure > 2: LOW (2.0)
| | | | | brushing > 0
| | | | | | brushing <= 2
| | | | | | | brushing <= 1
| | | | | | | | Tobacco use status <= 2
| | | | | | | | | Diastolic Blood Pressure <= 1
| | | | | | | | | | Tobacco use status <= 1
| | | | | | | | | | | Age <= 30: HIGH (14.0/3.0)
| | | | | | | | | | | Age > 30: LOW (3.0)
| | | | | | | | | | Tobacco use status > 1
| | | | | | | | | | | Missing Teeth <= 0: LOW (7.0/1.0)
| | | | | | | | | | | Missing Teeth > 0
| | | | | | | | | | | | Age <= 22: LOW (2.0)
| | | | | | | | | | | | Age > 22: HIGH (8.0/1.0)
| | | | | | | | | | Diastolic Blood Pressure > 1: HIGH (18.0/6.0)
| | | | | | | | | Tobacco use status > 2: HIGH (51.0/13.0)
| | | | | | | | brushing > 1
| | | | | | | | | Age <= 26: LOW (20.0/6.0)
| | | | | | | | | Age > 26: HIGH (36.0/11.0)
| | | | | | brushing > 2: HIGH (9.0/1.0)
| | | | Oral hygiene status > 1
| | | | | Oral hygiene status <= 3
| | | | | | Missing Teeth <= 3
| | | | | | | brushing <= 1
| | | | | | | | Age <= 21: HIGH (3.0)
| | | | | | | | Age > 21: LOW (30.0/10.0)
| | | | | | | brushing > 1: LOW (47.0/13.0)
| | | | | | Missing Teeth > 3: HIGH (6.0/1.0)
| | | | | Oral hygiene status > 3
| | | | | | Age <= 22: LOW (32.0/8.0)
| | | | | | Age > 22
| | | | | | | brushing <= 1: LOW (148.0/71.0)
| | | | | | | brushing > 1
| | | | | | | | Diastolic Blood Pressure <= 1: HIGH (89.0/41.0)
| | | | | | | | Diastolic Blood Pressure > 1
| | | | | | | | | Diastolic Blood Pressure <= 2
| | | | | | | | | | Missing Teeth <= 3: HIGH (32.0/12.0)
| | | | | | | | | | Missing Teeth > 3: LOW (7.0)
| | | | | | | | | Diastolic Blood Pressure > 2: HIGH (5.0/1.0)
| | Age > 35
| | | Missing Teeth <= 18
| | | | Tobacco use status <= 2
| | | | | GENDER <= 0
| | | | | | brushing <= 2
| | | | | | | Age <= 60
```

```
|  |  |  |  |  |  |  |  Oral hygiene status <= 3
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 1
|  |  |  |  |  |  |  |  |  |  brushing <= 0: LOW (8.0/1.0)
|  |  |  |  |  |  |  |  |  |  brushing > 0
|  |  |  |  |  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (27.0/10.0)
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1: LOW (7.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 52: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Age > 52: HIGH (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 37: LOW (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 37: HIGH (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 0: HIGH (8.0/1.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 9: HIGH (14.0/4.0)
|  |  |  |  |  |  |  |  |  |  Missing Teeth > 9: LOW (2.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  Missing Teeth <= 0: LOW (96.0/27.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 0
|  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  |  |  Missing Teeth <= 3
|  |  |  |  |  |  |  |  |  |  Missing Teeth <= 2
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1: LOW (2.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1: HIGH (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1: HIGH (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  Missing Teeth > 2: LOW (3.0)
|  |  |  |  |  |  |  |  |  |  Missing Teeth > 3: HIGH (6.0)
|  |  |  |  |  |  |  |  |  |  brushing > 1: LOW (79.0/33.0)
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2: LOW (8.0/1.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 3
|  |  |  |  |  |  |  |  |  Missing Teeth <= 5
|  |  |  |  |  |  |  |  |  Missing Teeth <= 3
|  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1: LOW (151.0/72.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  Age <= 39: LOW (21.0/6.0)
|  |  |  |  |  |  |  |  |  |  |  Age > 39: HIGH (104.0/44.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 3: HIGH (49.0/15.0)
|  |  |  |  |  |  |  |  Missing Teeth > 5: LOW (54.0/17.0)
```

```
|  |  |  |  |  |  |  Age > 60
|  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  |  |  Missing Teeth <= 10
|  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Age <= 82
|  |  |  |  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 3
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 2: HIGH (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 2
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 5: LOW (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 5: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Oral hygiene status > 1: LOW (11.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Oral hygiene status > 3: HIGH (25.0/9.0)
|  |  |  |  |  |  |  |  |  |  |  |  Age > 82: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 3
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 80
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 76: HIGH (16.0/6.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  Age > 76: LOW (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 80: HIGH (6.0)
|  |  |  |  |  |  |  |  |  |  |  |  Oral hygiene status > 3
|  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 4: HIGH (31.0/14.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 4: LOW (9.0/1.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 1: LOW (35.0/14.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 1: HIGH (53.0/15.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 10: LOW (33.0/9.0)
|  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  Age <= 79: HIGH (91.0/19.0)
|  |  |  |  |  |  |  |  |  Age > 79
|  |  |  |  |  |  |  |  |  |  Missing Teeth <= 9: LOW (12.0/2.0)
|  |  |  |  |  |  |  |  |  |  Missing Teeth > 9: HIGH (3.0)
|  |  |  |  |  |  brushing > 2
|  |  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  |  Age <= 50: HIGH (12.0)
|  |  |  |  |  |  |  |  Age > 50
|  |  |  |  |  |  |  |  |  Missing Teeth <= 1: LOW (6.0/1.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 1: HIGH (11.0/2.0)
|  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  Missing Teeth <= 7: HIGH (11.0/3.0)
|  |  |  |  |  |  |  |  Missing Teeth > 7: LOW (2.0)
|  |  |  |  |  GENDER > 0
|  |  |  |  |  |  brushing <= 0
|  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  Age <= 47: HIGH (11.0/2.0)
```

```
|  |  |  |  |  |  |  |  Age > 47: LOW (20.0/6.0)
|  |  |  |  |  |  |  Diastolic Blood Pressure > 2: HIGH (2.0)
|  |  |  |  |  |  brushing > 0
|  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  Oral hygiene status <= 3
|  |  |  |  |  |  |  |  |  Age <= 43
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 38: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 38: HIGH (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 0: LOW (9.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 0: HIGH (6.0/1.0)
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2: LOW (3.0)
|  |  |  |  |  |  |  |  |  Age > 43
|  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1: HIGH (55.0/13.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (22.0/8.0)
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 4: HIGH (10.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 4: LOW (10.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 66: HIGH (5.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 66: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 51: LOW (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 51: HIGH (12.0/1.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 3: HIGH (252.0/77.0)
|  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  Oral hygiene status <= 1: HIGH (46.0/11.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 2
|  |  |  |  |  |  |  |  |  |  Missing Teeth <= 2
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Age <= 48: LOW (12.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  Age > 48
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 51: HIGH (7.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 51: LOW (8.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (10.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1: LOW (10.0/2.0)
```

```
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 0
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1: LOW (8.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1: HIGH (8.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 2
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 8: HIGH (19.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 8: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  Oral hygiene status > 2
|  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 6: HIGH (95.0/36.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 6: LOW (7.0/1.0)
|  |  |  |  |  |  |  |  |  |  Tobacco use status > 1: HIGH (134.0/43.0)
|  |  |  |  Tobacco use status > 2
|  |  |  |  |  GENDER <= 0
|  |  |  |  |  |  Age <= 39
|  |  |  |  |  |  |  brushing <= 1: LOW (30.0/8.0)
|  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 1: LOW (3.0)
|  |  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  |  Age <= 37: LOW (17.0/6.0)
|  |  |  |  |  |  |  |  |  |  Age > 37
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 0
|  |  |  |  |  |  |  |  |  |  |  |  Age <= 38: HIGH (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  Age > 38: LOW (6.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 0: HIGH (7.0/1.0)
|  |  |  |  |  |  |  |  Diastolic Blood Pressure > 1: HIGH (12.0/3.0)
|  |  |  |  |  |  Age > 39: HIGH (401.0/124.0)
|  |  |  |  |  GENDER > 0
|  |  |  |  |  |  Missing Teeth <= 7
|  |  |  |  |  |  |  brushing <= 0
|  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  |  Missing Teeth <= 0: LOW (9.0/2.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 0: HIGH (20.0/5.0)
|  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2: HIGH (3.0)
|  |  |  |  |  |  |  brushing > 0
|  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2: HIGH (273.0/47.0)
|  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2
|  |  |  |  |  |  |  |  |  Age <= 43
|  |  |  |  |  |  |  |  |  |  Age <= 40: HIGH (7.0/1.0)
|  |  |  |  |  |  |  |  |  |  Age > 40: LOW (11.0/1.0)
|  |  |  |  |  |  |  |  |  Age > 43: HIGH (24.0/1.0)
|  |  |  |  |  |  Missing Teeth > 7
|  |  |  |  |  |  |  Oral hygiene status <= 1: HIGH (44.0/12.0)
|  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  brushing <= 0: LOW (3.0)
|  |  |  |  |  |  |  |  brushing > 0
```

```
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 3: LOW (10.0/3.0)
|  |  |  |  |  |  |  |  |  Oral hygiene status > 3: HIGH (40.0/13.0)
|  |  |  Missing Teeth > 18: LOW (146.0/37.0)
|  Calculus > 2
|  |  Missing Teeth <= 16
|  |  |  GENDER <= 0
|  |  |  |  Age <= 39
|  |  |  |  |  Tobacco use status <= 1: HIGH (108.0/29.0)
|  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  Tobacco use status <= 2
|  |  |  |  |  |  |  |  Missing Teeth <= 1: HIGH (43.0/13.0)
|  |  |  |  |  |  |  |  Missing Teeth > 1
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 1: LOW (7.0)
|  |  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  |  Age <= 25: HIGH (3.0)
|  |  |  |  |  |  |  |  |  |  Age > 25
|  |  |  |  |  |  |  |  |  |  |  Age <= 26: LOW (2.0)
|  |  |  |  |  |  |  |  |  |  |  Age > 26: HIGH (5.0/1.0)
|  |  |  |  |  |  |  Tobacco use status > 2
|  |  |  |  |  |  |  |  Oral hygiene status <= 1: HIGH (81.0/27.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  Missing Teeth <= 3: HIGH (59.0/28.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 3
|  |  |  |  |  |  |  |  |  |  Age <= 26: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  Age > 26: LOW (8.0)
|  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  Diastolic Blood Pressure <= 2: LOW (44.0/18.0)
|  |  |  |  |  |  |  Diastolic Blood Pressure > 2
|  |  |  |  |  |  |  |  Age <= 33: HIGH (6.0/1.0)
|  |  |  |  |  |  |  |  Age > 33: LOW (6.0/2.0)
|  |  |  |  Age > 39
|  |  |  |  |  Tobacco use status <= 2
|  |  |  |  |  |  Diastolic Blood Pressure <= 1
|  |  |  |  |  |  |  brushing <= 1
|  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  Missing Teeth <= 12
|  |  |  |  |  |  |  |  |  |  Age <= 84: HIGH (21.0/1.0)
|  |  |  |  |  |  |  |  |  |  Age > 84: LOW (2.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 12: LOW (2.0)
|  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  Age <= 44: LOW (4.0)
|  |  |  |  |  |  |  |  |  Age > 44
|  |  |  |  |  |  |  |  |  |  Missing Teeth <= 5
|  |  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 3: HIGH (10.0/3.0)
|  |  |  |  |  |  |  |  |  |  |  Oral hygiene status > 3: LOW (3.0)
```

```
|  |  |  |  |  |  |  |  |  |  Missing Teeth > 5: HIGH (9.0)
|  |  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  |  Oral hygiene status <= 3
|  |  |  |  |  |  |  |  |  |  Oral hygiene status <= 1: HIGH (26.0/6.0)
|  |  |  |  |  |  |  |  |  |  Oral hygiene status > 1
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 2: LOW (4.0)
|  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 2
|  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 5: HIGH (3.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 5: LOW (2.0)
|  |  |  |  |  |  |  |  |  |  |  Tobacco use status > 1: HIGH (6.0/2.0)
|  |  |  |  |  |  |  |  |  Oral hygiene status > 3: HIGH (45.0/8.0)
|  |  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  |  Age <= 49: LOW (9.0/2.0)
|  |  |  |  |  |  |  |  |  Age > 49
|  |  |  |  |  |  |  |  |  |  Missing Teeth <= 9: HIGH (22.0/5.0)
|  |  |  |  |  |  |  |  |  |  Missing Teeth > 9: LOW (4.0/1.0)
|  |  |  |  |  |  |  |  Tobacco use status > 1
|  |  |  |  |  |  |  |  |  Missing Teeth <= 3: HIGH (14.0/2.0)
|  |  |  |  |  |  |  |  |  Missing Teeth > 3
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure <= 2
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth <= 4: LOW (5.0)
|  |  |  |  |  |  |  |  |  |  |  Missing Teeth > 4
|  |  |  |  |  |  |  |  |  |  |  |  brushing <= 1: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  brushing > 1
|  |  |  |  |  |  |  |  |  |  |  |  |  Age <= 53: HIGH (2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  Age > 53: LOW (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  Diastolic Blood Pressure > 2: HIGH (4.0/1.0)
|  |  |  |  |  Tobacco use status > 2
|  |  |  |  |  |  Diastolic Blood Pressure <= 1: HIGH (90.0/12.0)
|  |  |  |  |  |  Diastolic Blood Pressure > 1
|  |  |  |  |  |  |  Age <= 49: HIGH (18.0/1.0)
|  |  |  |  |  |  |  Age > 49
|  |  |  |  |  |  |  |  Oral hygiene status <= 1: HIGH (18.0/2.0)
|  |  |  |  |  |  |  |  Oral hygiene status > 1: LOW (13.0/3.0)
|  |  |  GENDER > 0
|  |  |  |  Age <= 25
|  |  |  |  |  Oral hygiene status <= 1
|  |  |  |  |  |  Tobacco use status <= 2
|  |  |  |  |  |  |  Tobacco use status <= 1
|  |  |  |  |  |  |  |  Age <= 24: LOW (3.0/1.0)
|  |  |  |  |  |  |  |  Age > 24: HIGH (2.0)
|  |  |  |  |  |  |  Tobacco use status > 1: LOW (2.0)
|  |  |  |  |  |  Tobacco use status > 2: HIGH (22.0/5.0)
|  |  |  |  |  Oral hygiene status > 1
```

| | | | | | Diastolic Blood Pressure <= 1: HIGH (10.0/4.0)
| | | | | | Diastolic Blood Pressure > 1: LOW (5.0)
| | | | Age > 25: HIGH (701.0/107.0)
| | Missing Teeth > 16
| | | Oral hygiene status <= 2
| | | | brushing <= 1
| | | | | Tobacco use status <= 2: LOW (21.0/4.0)
| | | | | Tobacco use status > 2
| | | | | | GENDER <= 0
| | | | | | | Missing Teeth <= 21
| | | | | | | | Age <= 54: LOW (9.0/2.0)
| | | | | | | | Age > 54: HIGH (2.0)
| | | | | | | Missing Teeth > 21: LOW (5.0)
| | | | | | GENDER > 0
| | | | | | | Diastolic Blood Pressure <= 1
| | | | | | | | Missing Teeth <= 17: HIGH (2.0)
| | | | | | | | Missing Teeth > 17: LOW (7.0/2.0)
| | | | | | | Diastolic Blood Pressure > 1: HIGH (9.0/3.0)
| | | | brushing > 1
| | | | | Missing Teeth <= 23
| | | | | | Diastolic Blood Pressure <= 1: HIGH (5.0)
| | | | | | Diastolic Blood Pressure > 1
| | | | | | | Tobacco use status <= 1: HIGH (2.0)
| | | | | | | Tobacco use status > 1: LOW (7.0/2.0)
| | | | | Missing Teeth > 23: LOW (3.0)
| | | Oral hygiene status > 2: LOW (35.0/10.0)

6.6.BODY MASS INDEX

Prevailing evidence suggests that obesity is positively associated with increased triglyceride levels and blood pressure [115]. It is possible that presence of measures defining either of these attributes (e.g. BMI or triglyceride levels) within the identified representative set could act as a confounding factor. Since the multivariate filtering process estimates the correlation between the subset of the attributes and class as well as the inter-correlation, this study assumed that there was no confounding factor in the representative subset after the CFS application. BMI is calculated with height and weight. Notably, these filter methods eliminated height and weight when BMI was present.

6.7.BLOOD GLUCOSE LEVELS

In the present study, random blood glucose (RBG) level was identified as an important representative factor in determining PD risk. Hemoglobin A1C (HbA1C), represents a measure of the average blood glucose levels across a temporal trajectory of three months, and has been routinely used in monitoring glycemic control [146]. Studies that have investigated the mathematical relationship between HbA1C and glucose have shown that HbA1C values and mean continuous measures of glucose levels are equivalent and interchangeable [147]. Exploring the relationship between random blood glucose levels and HbA1C in similar type of data set is worth considering. Further, the present study supports a relationship of tobacco use status and PD risk for all the MOCs [148]. Overall, a significant difference in the rate of never-smokers was observed in controls (low PD risk) compared to cases (high PD risk) ($p<0.0001$). Similarly the number of former smokers among control groups significantly exceeded than those detected among cases ($p<0.0001$). These significant differences may account towards potential for selection bias within the cohort.

6.8. AREA UNDER THE CURVE

According to the definition of AUC, an AUC of a classifier is equivalent to the probability that the classifier can rank a randomly selected positive instance higher than a randomly selected negative instance. The scale suggested by Allaire et al for interpretation of AUC value: AUC (0.5 to 0.6) as ‘poor’; AUC (0.6 to 07.) as ‘fair’; AUC (0.7 to 0.8) as ‘good’; AUC (0.8 to 0.9) as ‘very good’ and AUC (0.9 to 1.0) as ‘excellent’ was applied to all the algorithms in the MOCs for interpretation. Based on the scale, the study results suggests that the

overall performance of DT and ANN was excellent; performance of NB, LR, SVM was very good in MOC 1, 2, and 3; while it was fair in MOC 4, 5, and 6 [149].

6.9. F-MEASURE

It is important to note that in this study, the F-measure, which represents a harmonic mean of precision and recall, holds a distinctive importance by virtue of its evaluation of the relationships between high PD risk instances within the data and those given by the classifiers. Despite the association of high type 1 error with cross-validated t-tests, evidence suggests that cross-validated t-tests are powerful in determining whether a learning algorithm outperforms another on a particular learning task [150]. To determine the real difference between algorithms (seen in type II errors), this study statistically analyzed the results of the algorithms by conducting a 10-fold cross validated t-test on F-measures. MOC 1, 2 and 3 shows the MCC measures of all the algorithms, presenting evidence that ANN and BDT outperformed NB, LR and SVM in this study. This finding supports the excellent predictability of ANN and BDT in assessing PD risk. This difference could be partially attributable to the sample size used in this study for various MOCs.

6.10. IMBALANCED DATASETS

Although imbalanced datasets are thought to decrease the accuracy, the present study showed a slightly higher accuracy for the imbalanced dataset compared to the balanced dataset. The performance metrics in terms of total accuracy, precision, recall, AUC, MCC for NB and DT were almost similar for balanced and imbalanced data set supporting the statement that both the algorithms lack sensitivity to stratification [151]. Similarly, the findings of this experiment show that ANN, when applied to the balanced dataset performed better than the imbalanced

dataset. Studies have shown the use of cost sensitive multilayer perceptron (CSMLP) improves performance leading to smooth decision boundaries, reducing noisy data and overfitting [152]. Exploring the application of CSMLP in a similar type of data is worth considering with the aim of producing classification models that are not biased towards the overrepresented class which, in the present study, was represented by the controls including the low PD risk dataset. Studies have shown that SVM does not perform well in cases of balanced data, and this was also observed in the present study [153]. Overall, the findings of this study showed that undersampling the class (in case of balanced dataset) had a slight effect on the predictive performance of learned classifiers compared to performance with the imbalanced class. Based on previous studies, a common understanding is that dataset should represent the proportionately the prevalence within the population for a certain disease. Notably, in this study, the distribution of cases and controls represented the prevalence of low and high risk for PD.

6.11. EVALUATION BY EXTERNAL VALIDATION SET

This study validated the tool by creating a dataset of comparable population by creating a new subpopulation drawn from the population seen at the healthcare organization. The results of the validation show that MOC 1, 2 and 3 can assess sensitivity and specificity within 85 $\pm$ 5% as compared to MOC 4, 5 and 6. Evaluating another set on these models resulting from this experiment will be worth considering. The results of this experiment have some additional implications for assessing PD risk. Most notably, when the variable random blood glucose was added to the dataset, the total accuracy of the resultant models for all the classifiers remained about the same; however the accuracy dramatically increased in validation set supporting the bi-directional association of diabetes and PD.

6.12. MEDICAID AND MEDICARE STATUS

Marshfield Clinic Health System (MCHS) service area extends across largely rural communities in northern, western and central Wisconsin counties. Residents of many of these counties disproportionately exceed the State's average population statistics for persons that meet definitions ranking them in the lower socio-economic strata and also numbers of individuals >65 years of age. Variables such as status of Medicare and Medicaid were incorporated into the model to explore the relationship between insurance status and PD risk. Interestingly, the study results demonstrated correlation between insurance status and PD risk.

6.13. CLINICAL IMPLICATIONS-DENTAL CALCULUS EXAMINATION IN PRIMARY SETTINGS

Currently, medical providers lack training and are not equipped to conduct comprehensive periodontal examination in medical settings. To test the performance of the five algorithms using dental variables including dental calculus, number of teeth present and patient reported data including tobacco use status, frequency of tooth brushing and flossing was tested in the absence of PPD data on a dataset of 4,000 patients. Notably, the results of this attempt, showed excellent predictability in assessing PD risk as compared to other MOCs. This difference of predictability of MOC 7 could be partially attributable to the small sample size used in the MOC 7. A larger sample size for MOC 7 could help in validating the findings and thus the application of MOC 7 at POC. The results indicate that across the spectrum of oral examinations that could be screened by the medical providers, dental plaque /calculus and the number of teeth present in the mouth would need minimal training and education along with optimal conditions such as adequate light, mouth mirror and explorer [154]. A common concern is lack of time amongst the providers. Collaborative effort and using team based workflow planning have shown to be able to incorporate additional services and screening by other healthcare professional (such as medical assistants) without a significant impact on providers time [155]. Future studies to further the knowledge base could include emerging technologies such as dental endoscopy, intra-oral camera using fluorescence system [156][157]. Similarly, a ML approach could be utilized to analyze and quantify the presence of supra-gingival calculus on teeth by taking multiple photographs.

6.14. LIMITATIONS

The study acknowledges some limitations. The study data was collected at a single healthcare system and thus extrapolation of the predictive model developed in this study to another health care center would need to be evaluated. This raises the potential for selection bias within the healthcare system. Lack of racial/ethnic diversity in our population necessitated delimiting the study dataset to White/ Caucasian race and non-Hispanic/ Latino ethnicity, therefore limiting use of the algorithm in racial or ethnic minority populations within MCHS without prior evaluation of these algorithms in a representative population cohort. Ideally, based on previous studies, a comprehensive periodontal examination including pocket depth for all the teeth, radiographic findings, and clinical attachment loss among other oral characteristics carries a significant weight for PD risk assessment [62]. This study was limited to clinical variables and hence did not utilize radiographic findings. Moreover the data on clinical attachment loss, furcation involvement, presence of bacterial plaque, tooth mobility and bleeding on probing was insufficient (missing) and hence was not included in the study. Attributes including waist circumference were eliminated due to incomplete and missing data. Addition of variables, genetic markers, and/or laboratory values for surrogate biological markers of systemic inflammation, such as high sensitivity C-reactive protein, may be useful to further improve the PD risk prognostication in medical settings. Further, incomplete documentation resulting in missing structured data surrounding oral manifestations such as bleeding gums on tooth brushing, swollen gums among, others might have adversely affected the classification results. A potential useful approach could be engaging natural language processing for application to unstructured resources such as oral/ dental complaints of patients and clinical notes to identify or

test additional useful parameters to support the current PD risk model or capture data to supplement data incompletely captured in structured data fields.

Nonetheless, this study which was performed in a well-characterized interdisciplinary research database with various MOCs and exhibited additional features including robust sample size, and no missing data, shows a potential for translation into interdisciplinary practice including medical and dental settings.

Chapter 7

Informal Study

When testing the various MOC for this thesis, items with missing data were excluded. This exclusion might raise a concern related to the applicability of the results to natural settings, where missing data can be expected. To gain insight regarding these concerns, an informal study was conducted. A small dataset of 200 patients (1:1 ratio of case and control) were randomly selected from 10% of the external evaluation set of MOC 1. Missing data was artificially created by randomly removing [10% (1620/16200)] the data variables. This data set was then evaluated on the resultant model of MOC 1 to check for the clinical viability of the model with missing data. The results of evaluating the artificially created dataset on the resultant DT model showed a sensitivity and specificity of 89.34% (95%CI 82.47% to 94.20%) and 96.15% (95% CI 89.17% to 99.20%). As compared to the results of external evaluation set on MOC 1, the results with missing data showed a higher sensitivity and slightly lower specificity. However, the data was limited to just 200 patients as compared to 1104 patients in the MOC 1 external evaluation set. Based on this, it could be posited that the resultant model may perform well in actual clinical settings with 10% of missing data. More investigation is required to deduce a strong conclusion by evaluating a large data set with 10% of missing data on the resultant models.

Chapter 8

CONCLUSIONS

A summary of findings presented in this study are highly concordant with the premise that ML methods are effective when applied to improving patient care through early detection or preventive approaches by assisting healthcare professionals in evaluating risk of developing PD based on evaluation of patient data in the light of historical and current status. Although many factors affect individual variability in developing PD risk of a patient, this study considered a wide variety of predictive factors including oral factors such as dental calculus and number of teeth present among others, which are also evaluable in an interdisciplinary setting. MOC settings serving as a source of data from seven MOCs were used to determine relative PD risk and conducted application of multiple ML approaches to identify those with highest potential for translation into clinical care to assist healthcare providers in making effective and knowledge-driven decisions.

Datasets targeted for risk modeling consisted of factors previously established as candidates contributing risk in association with PD and several novel factors including patient's status for: Medicare and Medicaid, dental calculus, diabetes, oral hygiene, lipid profiles and blood pressure; number of teeth present, tooth brushing and flossing frequency, periodontal pocket depth (PPD) for all the present teeth and body mass index (BMI). Variables such as duration of diabetes, height, weight and total cholesterol did not contribute to PD risk prediction in any models, whereas random blood sugar levels, number of missing teeth, presence or absence of dental calculus and PPD contributed to model accuracy. PPD at specific tooth surfaces were identified as significant determinants during feature selection.

This study reinforced a role for interdisciplinary environment to promote development of new best practices for patient referrals and support mitigation of chronic disease onset or program. Further studies are needed to explore additional ways for advancing the MOCs to improve more integrated oral-systemic health care delivery. Future steps include incorporation of such models into the EHR and validating model performance in a clinical setting.

REFERENCES

[1] W. J. Loesche and N. S. Grossman, "Periodontal disease as a specific, albeit chronic, infection: diagnosis and treatment," *Clin. Microbiol. Rev.*, vol. 14, no. 4, pp. 727–752, table of contents, 2001.

[2] S. S. Socransky and a D. Haffajee, "The bacterial etiology of destructive periodontal disease: current concepts.," *J. Periodontol.*, vol. 63, no. 4, pp. 322–331, 1992.

[3] U.S. Department of Health and Human Services, "Oral Health in America: A Report of the Surgeon General," *Rockville, MD U.S. Dep. Heal. Hum. Serv. Natl. Inst. Dent. Craniofacial Res. Natl. Institutes Heal.* , 2000.

[4] P. I. Eke, B. a. Dye, L. Wei, G. O. Thornton-Evans, and R. J. Genco, "Prevalence of Periodontitis in Adults in the United States: 2009 and 2010," *J. Dent. Res.*, vol. 91, pp. 914–920, 2012.

[5] P. M. Preshaw and S. M. Bissett, "Periodontitis. Oral Complication of Diabetes.," *Endocrinology and Metabolism Clinics of North America*, vol. 42, no. 4. pp. 849–867, 2013.

[6] P. E. Petersen, D. Bourgeois, H. Ogawa, S. Estupinan-day, and C. Ndiaye, "Policy and Practice The global burden of oral diseases and risks to oral health," *Bull. World Health Organ.*, vol. 83, no. 05, pp. 661–669, 2005.

[7] RanftLeslie reviewed by Nordland P, "Gum Disease Treatments | What are Your Options?" [Online]. Available: http://www.yourdentistryguide.com/gum-disease-treatments/. [Accessed: 21-Apr-2017].

[8] K. Nasseh, M. Vujicic, and M. Glick, "The Relationship between Periodontal Interventions and Healthcare Costs and Utilization. Evidence from an Integrated Dental, Medical, and Pharmacy Commercial Claims Database," *Health Economics (United Kingdom)*, 2016.

[9] N. Shimpi, D. Schroeder, J. Kilsdonk, C. Ph, I. Glurich, and A. Acharya, "Assessment of Dental Providers ' Knowledge , Behavior and Attitude towards Incorporating Chairside Screening for Medical Conditions : A Pilot Study," vol. 2, no. 1, pp. 1–7, 2016.

[10] A. Acharya, "Marshfield Clinic Health System: Integrated Care Case Study.," *J. Calif. Dent. Assoc.*, vol. 44, no. 3, pp. 177–81, Mar. 2016.

[11] B. Bell and K. Thornton, "From promise to reality: Achieving the value of an EHR," *Healthc. Financ. Manag.*, vol. 65, no. 2, pp. 50–6, 2011.

[12] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care.," *Nat. Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.

[13] S. P. Goetzel Ron Z.,Staley Paula , Ogden Lydia, M. ; Jared Fox, PhD, ;, M. Jason Spangler, MD, M. ; Maryam Tabrizi, M. ; Meghan Beckowski, P. ; Niranjana Kowlessar,

;, P. Russell E. Glasgow, M. ; Martina V. Taylor, and ; Richards Chesley, “A Framework for Patient-Centered Health Risk Assessments Providing Health Promotion and Disease Prevention Ser vices to Medicare Beneficiaries,” *Centers Dis. Control Prev. Atlanta, GA*.

[14] R. I. Garcia, E. A. Krall, and P. S. Vokonas, “Periodontal disease and mortality from all causes in the VA Dental Longitudinal Study,” *Ann Periodontol*, vol. 3, no. 1, pp. 339–349, 1998.

[15] N. Shimpi, D. Schroeder, J. Kilsdonk, P.-H. Chyou, I. Glurich, E. Penniman, and A. Acharya, “Medical Providers’ Oral Health Knowledgeability, Attitudes, and Practice Behaviors: An Opportunity for Interprofessional Collaboration,” *J. Evid. Based. Dent. Pract.*, vol. 16, no. 1, pp. 19–29, 2016.

[16] W. Kye, R. Davidson, J. Martin, and S. Engebretson, “Current status of periodontal risk assessment,” *J. Evid. Based. Dent. Pract.*, vol. 12, no. 3 SUPPL., pp. 2–11, 2012.

[17] P. Glassman, M. Helgeson, and J. Kattlove, “Using telehealth technologies to improve oral health for vulnerable and underserved populations.,” *J. Calif. Dent. Assoc.*, vol. 40, no. 7, pp. 579–585, 2012.

[18] J. S. Schiller, J. W. Lucas, B. W. Ward, and J. a. Peregoy, “Summary health statistics for u.s. Adults: national health interview survey, 2012.,” *Natl. Cent. Heal. Stat. Vital Heal. Stat 10*, no. 252, pp. 1–171, 2012.

[19] F. R. Vogenberg, “Predictive and prognostic models: implications for healthcare decision-making in a modern recession.,” *Am. Heal. drug benefits*, vol. 2, no. 6, pp. 218–22, Sep. 2009.

[20] I. B. Lamster and K. Eaves, “A model for dental practice in the 21st century,” *Am. J. Public Health*, vol. 101, no. 10, pp. 1825–1830, 2011.

[21] R. P. Newhouse and B. Spring, “Interdisciplinary evidence-based practice: moving from silos to synergy.,” *Nurs. Outlook*, vol. 58, no. 6, pp. 309–17, 2010.

[22] “Locomotion | Define Locomotion at Dictionary.com.” [Online]. Available: http://www.dictionary.com/browse/periodontal. [Accessed: 21-Apr-2017].

[23] J. M. Albandar, L. J. Brown, J. A. Brunelle, and H. Löe, “Gingival state and dental calculus in early-onset periodontitis.,” *J. Periodontol.*, vol. 67, no. 10, pp. 953–9, 1996.

[24] S. F. G, “Risk Factors for the Periodontal Diseases 2014 AAPA,” 2014.

[25] R. J. Genco and W. S. Borgnakke, “Risk factors for periodontal disease.,” *Periodontol. 2000*, vol. 62, pp. 59–94, 2013.

[26] P. N. Papapanou, “Risk assessments in the diagnosis and treatment of periodontal diseases.,” *J. Dent. Educ.*, vol. 62, no. 10, pp. 822–39, Oct. 1998.

[27] I. Kroes, P. W. Lepp, and D. A. Relman, “Bacterial diversity within the human subgingival crevice.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 25, pp. 14547–52, 1999.

[28] L. a Ximénez-Fyvie, a D. Haffajee, S. Som, M. Thompson, G. Torresyap, and S. S. Socransky, "The effect of repeated professional supragingival plaque removal on the composition of the supra- and subgingival microbiota.," *J. Clin. Periodontol.*, vol. 27, no. October 1999, pp. 637–647, 2000.

[29] A. Tanner, M. F. Maiden, P. J. Macuch, L. L. Murray, and R. L. Kent, "Microbiota of health, gingivitis, and initial periodontitis.," *J. Clin. Periodontol.*, vol. 25, no. 2, pp. 85–98, 1998.

[30] A. Tanner, R. Kent, M. F. J. Maiden, and M. A. Taubman, "Clinical, microbiological and immunological profile of healthy, gingivitis and putative active periodontal subjects," *J. Periodontal Res.*, vol. 31, no. 3, pp. 195–204, 1996.

[31] S. S. Socransky and A. D. Haffajee, "Evidence of bacterial etiology: a historical perspective," *Periodontol. 2000*, vol. 5, no. 1, pp. 7–25, 1994.

[32] M. G. Newman, H. H. Takei, P. R. Klokkevold, and F. A. Carranza, *Carranza's Clinical Periodontology 11th Ed*, vol. XXXIII, no. 2. 2012.

[33] R. C. Page, "The role of inflammatory mediators in the pathogenesis of periodontal disease.," *J. Periodontal Res.*, vol. 26, no. 3 Pt 2, pp. 230–242, 1991.

[34] N. P. Lang, B. R. Cumming, and H. Löe, "Toothbrushing Frequency as It Relates to Plaque Development and Gingival Health," *J. Periodontol.*, vol. 44, no. 7, pp. 396–405, Jul. 1973.

[35] N. Wake, Y. Asahi, Y. Noiri, M. Hayashi, D. Motooka, S. Nakamura, K. Gotoh, J. Miura, H. Machi, T. Iida, and S. Ebisu, "Temporal dynamics of bacterial microbiota in the human oral cavity determined using an in situ model of dental biofilms," *npj Biofilms Microbiomes*, vol. 2, p. 16018, Aug. 2016.

[36] R. M. Kelner, B. R. Wohl, M. J. Deasy, and A. J. Formicola, "Gingival Inflammation as Related to Frequency of Plaque Removal," *J. Periodontol.*, vol. 45, no. 5.1, pp. 303–307, May 1974.

[37] B. Y. Hong, M. V. F. Araujo, L. D. Strausbaugh, E. Terzi, E. Ioannidou, and P. I. Diaz, "Microbiome profiles in periodontitis in relation to host and disease characteristics," *PLoS One*, vol. 10, no. 5, 2015.

[38] F. O. Ozden, O. Ozgonenel, B. Ozden, and A. Aydogdu, "Diagnosis of periodontal diseases using different classification algorithms: a preliminary study," *Niger J Clin Pr.*, vol. 18, no. 3, pp. 416–421, 2015.

[39] M. Matei, Madalina, Earar,Kamel, Jurja, Sanda, Rusu, "CORRELATIONS BETWEEN THE CLINICAL ASPECTS OF THE PERIODONTAL DISEASE AND TH...: Discovery Service for Marshfield Clinic," *Rom. J. Child Adolesc. Psychiatry*, vol. 2, no. 2, pp. 3–6, 2014.

[40] A. C. Solis, R. F. Lotufo, C. M. Pannuti, E. C. Brunheiro, A. H. Marques, and F. Lotufo-Neto, "Association of periodontal disease to anxiety and depression symptoms, and

psychosocial stress factors," *J Clin Periodontol*, vol. 31, no. 8, pp. 633–638, 2004.

[41] A. R. Kamer, R. G. Craig, A. P. Dasanayake, M. Brys, L. Glodzik-Sobanska, and M. J. de Leon, "Inflammation and Alzheimer's disease: possible role of periodontal diseases.," *Alzheimers. Dement.*, vol. 4, no. 4, pp. 242–50, 2008.

[42] A. Khocht, T. Yaskell, M. Janal, B. F. Turner, T. E. Rams, A. D. Haffajee, and S. S. Socransky, "Subgingival microbiota in adult Down syndrome periodontitis," *J. Periodontal Res.*, vol. 47, no. 4, pp. 500–507, 2012.

[43] T. E. Van Dyke and D. Sheilesh, "Risk factors for periodontitis.," *J. Int. Acad. Periodontol.*, vol. 7, no. 1, pp. 3–7, Jan. 2005.

[44] B. L. Mealey, "Periodontal disease and diabetes. A two-way street," *J. Am. Dent. Assoc.*, vol. 137 Suppl, no. October, p. 26S–31S, 2006.

[45] Wild Sarah, Roglic Gojka, Green Anders, Sicree Richard, and K. Hilary, "Global Prevalence of Diabetes: Estimates for the year 2000 and projection for 2030," *Diabetes Care*, vol. 27, no. 5, pp. 1047–1053, 2004.

[46] J. Kim and S. Amar, "Periodontal disease and systemic conditions: A bidirectional relationship," *Odontology*, vol. 94, no. 1. pp. 10–21, 2006.

[47] Nhlbi, "Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC7)," 2004.

[48] S. Engstrom, L. Gahnberg, H. Hogberg, and K. Svardsudd, "Association between high blood pressure and deep periodontal pockets: a nested case-referent study.," *Ups. J. Med. Sci.*, vol. 112, no. 1, pp. 95–103, 2007.

[49] U. J. Jung and M.-S. Choi, "Obesity and its metabolic complications: The role of adipokines and the relationship between obesity, inflammation, insulin resistance, dyslipidemia and nonalcoholic fatty liver disease.," *Int. J. Mol. Sci.*, vol. 15, no. 4, pp. 6184–6223, 2014.

[50] Guay Albert H, "Access to dental care Solving the problem for underserved populations," *J. Am. Dent. Assoc.*, vol. 135, pp. 1599–1605, 2004.

[51] B. Sanders, "DENTAL CRISIS IN AMERICA," 2012.

[52] A. Snyder, J. Antonishak, E. Potler, L. Grange, J. L. Breakell, C. Uriona, M. Mariani, V. L. Doggett, M. Maynard, N. Dueffert, K. Huh, A. Katzel, L. Lambert, M. Lyons, B. Maas, M. Mijic, M. F. Shaw, N. Augustine, B. Hill, N. Kallay, R. King, M. Mabanta, L. Norris, K. Patterson, A. Russell, F. Schecker, and S. Turner—for, "State Dental Policies Fail One in Five Children The Cost of Delay," 2010.

[53] "Oral Health Efforts Under Way to Improve Children's Access to Dental Services, but Sustained Attention Needed to Address Ongoing Concerns," 2010.

[54] K. Nasseh, M. Vujicic, and C. Yarbrough, "A Ten-Year, State-by-State, Analysis of

Medicaid Fee-for-Service Reimbursement Rates for Dental Care Services," 2014.

[55] National Institute of Dental & Craniofacial Research, "Healthy People 2010 Oral Health Toolkit : A Field Guide to Health Planning," 2010.

[56] Department of Health and Human Services Centers for Medicare & Medicaid Services, PCG, and Dpipd, "DEPARTMENT OF HEALTH AND HUMAN SERVICES Centers for Medicare & Medicaid Services Federally Qualified Health Center RURAL HEALTH SERIES FQHC BACKGROUND," 2017. [Online]. Available: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/fqhcfactsheet.pdf. [Accessed: 21-Apr-2017].

[57] C. H. Colla, C. Stachowski, S. Kundu, B. Harris, G. Kennedy, and M. Vujicic, "Dental Care Within Accountable Care Organizations: Challenges and Opportunities," Dartmouth, 2016.

[58] P. M. Preshaw, "Detection and diagnosis of periodontal conditions amenable to prevention.," *BMC Oral Health*, vol. 15 Suppl 1, no. Suppl 1, p. S5, 2015.

[59] World Health Organization, "WHO | Risk factors," *WHO*, 2014. [Online]. Available: http://www.who.int/topics/risk_factors/en/. [Accessed: 21-Apr-2017].

[60] American Academy of Periodontology, "Parameters on Comprehensive Periodontal Examination," *J. Periodontol.*, vol. 71, no. 5, 2000.

[61] G. R. Persson, L. A. Mancl, J. Martin, and R. C. Page, "Assessing periodontal disease risk: a comparison of clinicians' assessment versus a computerized tool.," *J. Am. Dent. Assoc.*, vol. 134, no. 5, pp. 575–82, May 2003.

[62] R. C. Page, E. A. Krall, J. Martin, L. Mancl, and R. I. Garcia, "Validity and accuracy of a risk calculator in predicting periodontal disease.," *J. Am. Dent. Assoc.*, vol. 133, no. 5, pp. 569–76, May 2002.

[63] N. P. Lang, / Maurizio, and S. Tonetti, "Periodontal Risk Assessment (PRA) for Patients in Supportive Periodontal Therapy (SPT)," *Oral Heal. Prev. Dent.*, vol. 1, pp. 7–16, 2003.

[64] R. C. Page, J. A. Martin, and C. F. Loeb, "The Oral Health Information Suite (OHIS): its use in the management of periodontal disease.," *J. Dent. Educ.*, vol. 69, no. 5, pp. 509–20, May 2005.

[65] H. C. H. Sandberg and U. G. H. Fors, "The HIDEP model--a straightforward dental health care model for prevention-based practice management.," *Swed. Dent. J.*, vol. 31, no. 4, pp. 171–9, 2007.

[66] R. V. Chandra, "Evaluation of a novel periodontal risk assessment model in patients presenting for dental care.," *Oral Health Prev. Dent.*, vol. 5, no. 1, pp. 39–48, 2007.

[67] P. Eickholz, J. Kaltschmitt, J. Berbig, P. Reitmeir, and B. Pretzl, "Tooth loss after active periodontal therapy. 1: patient-related factors for risk, prognosis, and quality of outcome," *J. Clin. Periodontol.*, vol. 35, no. 2, pp. 165–174, Jan. 2008.

[68] H. Jansson and O. Norderyd, “Evaluation of a periodontal risk assessment model in subjects with severe periodontitis. A 5-year retrospective study.,” *Swed. Dent. J.*, vol. 32, no. 1, pp. 1–7, 2008.

[69] L. Trombelli, R. Farina, S. Ferrari, P. Pasetti, and G. Calura, “Comparison between two methods for periodontal risk assessment.,” *Minerva Stomatol.*, vol. 58, no. 6, pp. 277–87, Jun. 2009.

[70] M. Leininger, H. Tenenbaum, and J.-L. Davideau, “Modified periodontal risk assessment score: long-term predictive value of treatment outcomes. A retrospective study,” *J. Clin. Periodontol.*, vol. 37, no. 5, pp. 427–435, May 2010.

[71] S. Lindskog, J. Blomlöf, I. Persson, A. Niklason, A. Hedin, L. Ericsson, M. Ericsson, B. Järncrantz, U. Palo, G. Tellefsen, O. Zetterström, and L. Blomlöf, “Validation of an Algorithm for Chronic Periodontitis Risk Assessment and Prognostication: Risk Predictors, Explanatory Values, Measures of Quality, and Clinical Use,” *J. Periodontol.*, vol. 81, no. 4, pp. 584–593, Apr. 2010.

[72] R. Shankarapillai, L. K. Mathur, M. A. Nair, N. Rai, and A. Mathur, “Periodontitis Risk Assessment using two artificial Neural Networks-A Pilot Study,” *Int. J. Dent. Clin. ©INTERNATIONAL J. Dent. Clin.*, vol. 2, no. 2, pp. 36–40, 2010.

[73] F. O. Costa, L. O. Miranda Cota, E. J. Pereira Lages, A. P. Lima Oliveira, S. C. Cortelli, J. R. Cortelli, T. C. Medeiros Lorentz, and J. E. Costa, “Periodontal Risk Assessment Model in a Sample of Regular and Irregular Compliers Under Maintenance Therapy: A 3-Year Prospective Study,” *J. Periodontol.*, vol. 83, no. 3, pp. 292–300, Mar. 2012.

[74] S. T. Teich, “Risk Assessment-Based Individualized Treatment (RABIT): a comprehensive approach to dental patient recall.,” *J. Dent. Educ.*, vol. 77, no. 4, pp. 448–57, Apr. 2013.

[75] D. Lü, H. Meng, L. Xu, R. Lu, L. Zhang, Z. Chen, X. Feng, D. Shi, Y. Tian, and X. Wang, “New Attempts to Modify Periodontal Risk Assessment for Generalized Aggressive Periodontitis: A Retrospective Study,” *J. Periodontol.*, pp. 1–14, Jan. 2013.

[76] M. Busby, E. Chapple, R. Matthews, and I. L. C. Chapple, “Practitioner evaluation of a novel online integrated oral health and risk assessment tool: a practice pilot,” *BDJ*, vol. 215, no. 3, pp. 115–120, Aug. 2013.

[77] R. Pivovarov and N. Elhadad, “Automated methods for the summarization of electronic health records,” *J. Am. Med. Informatics Assoc.*, vol. 22, no. 5, pp. 938–947, 2015.

[78] K. Liu, A. Acharya, S. Alai, and T. K. Schleyer, “Using Electronic Dental Record Data for Research,” *J. Dent. Res.*, vol. 92, no. 7_suppl, pp. S90–S96, Jul. 2013.

[79] T. K. Schleyer, A. Ruttenberg, W. Duncan, M. Haendel, C. Torniai, A. Acharya, M. Song, T. P. Thyvalikakath, K. Liu, and P. Hernandez, “An ontology-based method for secondary use of electronic dental record data.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2013, pp. 234–8, 2013.

[80] A. Acharya, J. J. VanWormer, S. C. Waring, A. W. Miller, J. T. Fuehrer, and G. R. Nycz, "Regional epidemiologic assessment of prevalent periodontitis using an electronic health record system.," *Am. J. Epidemiol.*, vol. 177, no. 7, pp. 700–7, Apr. 2013.

[81] S. I. Hay, D. B. George, C. L. Moyes, and J. S. Brownstein, "Big Data Opportunities for Global Infectious Disease Surveillance," *PLoS Med.*, vol. 10, no. 4, 2013.

[82] D. Weatherall, B. Greenwood, H. L. Chee, and P. Wasi, *Science and Technology for Disease Control: Past, Present, and Future*. 2006.

[83] A. Rostami, Reihaneh, Hegde , Harshad, Shimpi, Neel, Pack, Gary, Olson, Brent, Acharya, "2016 AADR/CADR Annual Meeting & Exhibition - Session Details," in *Disparities, Health Literacy, and Oral Cancer*, 2016.

[84] A. Koehler, Krista, Shimpi, Neel, Hegde, Harshad, Pack, Gary, Chyou, Po-Huang, Acharya, "Development of Prototypical Design of Oral Cancer Risk Assessment Tool," in *IADR/AADR/CADR General Session & Exhibition*, 2015.

[85] C. B. Wiebe and E. E. Putnins, "The periodontal disease classification system of the American Academy of Periodontology - An update," *Journal of the Canadian Dental Association*, vol. 66, no. 11. pp. 594–597, 2000.

[86] S. R. Barber, M. J. Davies, K. Khunti, and L. J. Gray, "Risk assessment tools for detecting those with pre-diabetes: A systematic review," *Diabetes Research and Clinical Practice*, vol. 105, no. 1. pp. 1–13, 2014.

[87] K. J. M. Janssen, I. Siccama, Y. Vergouwe, H. Koffijberg, T. P. A. Debray, M. Keijzer, D. E. Grobbee, and K. G. M. Moons, "Development and validation of clinical prediction models: Marginal differences between logistic regression, penalized maximum likelihood estimation, and genetic programming," *J. Clin. Epidemiol.*, vol. 65, no. 4, pp. 404–412, 2012.

[88] I. H. Witten, E. Frank, and M. a. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, vol. 54, no. 2. 2011.

[89] P. Wang, "The limitation of Bayesianism," *Artif. Intell.*, vol. 158, no. 1, pp. 97–106, 2004.

[90] S. Ogino and R. B. Wilson, "Bayesian Analysis and Risk Assessment in Genetic Counseling and Testing," *J. Mol. Diagnostics*, vol. 6, no. 1, pp. 1–9, 2004.

[91] E. Miranda, E. Irwansyah, A. Y. Amelga, M. M. Maribondang, and M. Salim, "Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier.," *Healthc. Inform. Res.*, vol. 22, no. 3, pp. 196–205, Jul. 2016.

[92] M. Langarizadeh and F. Moghbeli, "Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review.," *Acta Inform. Med.*, vol. 24, no. 5, pp. 364–369, Oct. 2016.

[93] I. Rish, J. Hellerstein, and T. Jayram, "An analysis of data characteristics that affect naive Bayes performance," *Tec. Rep. RC21993, IBM Watson …*, 2001.

[94] L. Cochon, J. Esin, and A. A. Baez, “Bayesian comparative model of CT scan and ultrasonography in the assessment of acute appendicitis: results from the Acute Care Diagnostic Collaboration project,” *Am. J. Emerg. Med.*, vol. 34, no. 11, pp. 2070–2073, 2016.

[95] M. Kayri and Murat, “Predictive Abilities of Bayesian Regularization and Levenberg–Marquardt Algorithms in Artificial Neural Networks: A Comparative Empirical Study on Social Data,” *Math. Comput. Appl.*, vol. 21, no. 2, p. 20, May 2016.

[96] F. Burden and D. Winkler, “Bayesian regularization of neural networks,” *Methods Mol. Biol.*, vol. 458, pp. 25–44, 2008.

[97] J. P. Hilbert, S. Zasadil, D. J. Keyser, and P. B. Peele, “Using Decision Trees to Manage Hospital Readmission Risk for Acute Myocardial Infarction, Heart Failure, and Pneumonia,” *Appl. Health Econ. Health Policy*, vol. 12, no. 6, pp. 573–585, 2014.

[98] S. Chaganti, A. J. Plassard, L. Wilson, M. A. Smith, M. B. Patel, and B. A. Landman, “A Bayesian Framework for Early Risk Prediction in Traumatic Brain Injury.,” *Proc. SPIE--the Int. Soc. Opt. Eng.*, vol. 9784, 2016.

[99] M. Mitchell and T, “Machine learning,” *MIT Press*, p. 414, 1997.

[100] M. Moon and S.-K. Lee, “Applying of Decision Tree Analysis to Risk Factors Associated with Pressure Ulcers in Long-Term Care Facilities,” *Healthc. Inform. Res.*, vol. 23, no. 1, p. 43, Jan. 2017.

[101] M. Mohri, A. Rostamizadeh, and A. Talwalkar, “Foundations of Machine Learning,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

[102] P. J. Lisboa and A. F. G. Taktak, “The use of artificial neural networks in decision support in cancer: A systematic review.”

[103] M. F. Møller, “A scaled conjugate gradient algorithm for fast supervised learning,” *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[104] H. Zhao, A. P. Sinha, and G. Bansal, “An extended tuning method for cost-sensitive regression and forecasting,” *Decis. Support Syst.*, vol. 51, no. 3, pp. 372–383, 2011.

[105] L. A. Zadeh, “Fuzzy algorithms,” *Inf. Control*, vol. 12, no. 2, pp. 94–102, Feb. 1968.

[106] Y.-L. Huang, K.-L. Wang, and D.-R. Chen, “Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines,” 2005.

[107] P. S. Hiremath and J. R. Tegnoor, “Follicle Detection and Ovarian Classification in Digital Ultrasound Images of Ovaries.”

[108] T. Razzaghi, O. Roderick, I. Safro, and N. Marko, “Multilevel weighted support vector machine for classification on healthcare data with missing values,” *PLoS One*, vol. 11, no. 5, 2016.

[109] G. A. Brooks, A. J. Kansagra, S. R. Rao, J. I. Weitzman, E. A. Linden, and J. O. Jacobson,

"A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy.," *JAMA Oncol.*, vol. 1, no. 4, pp. 441–7, 2015.

[110] R. L. Schaefer, "Bias correction in maximum likelihood logistic regression.," *Stat. Med.*, vol. 2, no. October 1981, pp. 71–8, 1983.

[111] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S.-F. Chen, "Lung cancer cell identification based on artificial neural network ensembles," *Artif. Intell. Med.*, vol. 24, no. 1, pp. 25–36, 2002.

[112] W. J. Lin and J. J. Chen, "Class-imbalanced classifiers for high-dimensional data," *Briefings in Bioinformatics*, vol. 14, no. 1. pp. 13–26, 2013.

[113] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Learning from positive examples when the negative class is undetermined--microRNA gene identification.," *Algorithms Mol. Biol.*, vol. 3, p. 2, 2008.

[114] Q. Wei and R. L. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS One*, vol. 8, no. 7, 2013.

[115] V. Hjellvik, S. Sakshaug, and H. Strøm, "Body mass index, triglycerides, glucose, and blood pressure as predictors of type 2 diabetes in a middle-aged Norwegian cohort of men and women," *Clin. Epidemiol.*, vol. 4, no. 1, pp. 213–224, 2012.

[116] W. Yin, Y. Yi, X. Guan, L. Zhou, J. Wang, D. Li, and X. Zuo, "Preprocedural Prediction Model for Contrast-Induced Nephropathy Patients," *J. Am. Heart Assoc.*, vol. 6, no. 2, 2017.

[117] J. L. Bruse, M. A. Zuluaga, A. Khushnood, K. McLeod, H. N. Ntsinjana, T.-Y. Hsia, M. Sermesant, X. Pennec, A. M. Taylor, and S. Schievano, "Detecting Clinically Meaningful Shape Clusters in Medical Image Data: Metrics Analysis for Hierarchical Clustering applied to Healthy and Pathological Aortic Arches," *IEEE Trans. Biomed. Eng.*, pp. 1–1, Feb. 2017.

[118] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest.," *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, p. 51, 2011.

[119] R. Batuwita and V. Palade, "Class Imbalance Learning Methods for Support Vector," *Imbalanced Learn. Found. Algorithms, Appl.*, pp. 83–100, 2013.

[120] T. R. Hoens and N. V. Chawla, "Generating diverse ensembles to counter the problem of class imbalance," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6119 LNAI, no. PART 2, pp. 488–499.

[121] C. X. Ling and V. S. Sheng, "Cost-sensitive learning and the class imbalance problem," *Encycl. Mach. Learn.*, pp. 231–235, 2008.

[122] National Institute of Dental and Craniofacial Research, "Periodontal Disease in Adults

(Age 20 to 64)," *National Institute of Health*. [Online]. Available: https://www.nidcr.nih.gov/DataStatistics/FindDataByTopic/GumDisease/PeriodontaldiseaseAdults20to64.htm. [Accessed: 21-Apr-2017].

[123] C. Enders, "Traditional Methods for Dealing with Missing Data," in *Applied Missing Data Analysis*, Second., T. Little, Ed. New York: The Guild Press, 2010, p. 377.

[124] S. Woltering, I. Granic, C. Lamm, and M. D. Lewis, "Neural changes associated with treatment outcome in children with externalizing problems," *Biol. Psychiatry*, vol. 70, no. 9, pp. 873–879, 2011.

[125] "Citing SPSS within your thesis." [Online]. Available: http://libanswers.brenau.edu/faq/165512. [Accessed: 21-Apr-2017].

[126] J. W. Tukey, "Mathematics and the Picturing of Data*," in *Proceedings of the International Congress of Mathematicians*, 1974.

[127] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance.," *BMC Med. Inform. Decis. Mak.*, vol. 12, p. 8, 2012.

[128] "Sample Size Software | Power Analysis Software | PASS | NCSS.com," *NCSS Statistical Software*. [Online]. Available: https://www.ncss.com/software/pass/?gclid=Cj0KEQjw2-bHBRDEh6qk5b6yqKIBEiQAFUz29pS3KxOOeKbVmnQlFFvRt-JXJ34jFYI3_EPbbuSTflUaAtvS8P8HAQ. [Accessed: 21-Apr-2017].

[129] D. J. Hand and R. J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," in *Machine Learning*, vol. 45, no. 2, Kluwer Academic Publishers, 2001, pp. 171–186.

[130] L. Ladha and T. Deepa, "Feature selection methods and algorithms," *Int. J. ...*, vol. 3, no. 5, pp. 1787–1797, 2011.

[131] M. Hall, "Correlation-based Feature Selection for Machine Learning," *Methodology*, vol. 21i195–i20, no. April, pp. 1–5, 1999.

[132] F. Provost, D. Jensen, and T. Oates, "Efficient Progressive Sampling," *Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 23–32, 1999.

[133] B. High and T. Cholesterol, "ATP III At-A-Glance : Quick Desk Reference," *Hypertension*, pp. 1–6, 2009.

[134] "Understanding Blood Pressure Readings," *American Heart Association*. [Online]. Available: http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/KnowYourNumbers/Understanding-Blood-Pressure-Readings_UCM_301764_Article.jsp#.WPp8AfnytQI. [Accessed: 21-Apr-2017].

[135] I. Frank, Eibe, Hall , Mark, Witten, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java," 2016. [Online]. Available:

http://www.cs.waikato.ac.nz/ml/weka/citing.html. [Accessed: 21-Apr-2017].

[136] M. Desvarieux, R. T. Demmer, T. Rundek, B. Boden-Albala, D. R. Jacobs, P. N. Papapanou, R. L. Sacco, and R. L. Oral Infections and Vascular Disease Epidemiology Study (INVEST), "Relationship between periodontal disease, tooth loss, and carotid artery plaque: the Oral Infections and Vascular Disease Epidemiology Study (INVEST).," *Stroke*, vol. 34, no. 9, pp. 2120–5, Sep. 2003.

[137] V. I. Anireh and E. N. Osegi, "A Modified Activation Function with Improved Run-Times For Neural Networks," *Neural Evol. Comput.*, Jul. 2016.

[138] A. Janecek, W. N. W. Gansterer, M. Demel, and G. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy.," *Fsdm*, vol. 4, pp. 90–105, 2008.

[139] J. R. Pritchard and A. J. Laws, "Gingival crevice depth. I. Predictability of probing deepest points," *Aust. Dent. J.*, vol. 29, no. 6, pp. 404–410, Dec. 1984.

[140] N. S. Rajhans, R. M. Kohad, V. G. Chaudhari, and N. H. Mhaske, "A clinical study of the relationship between diabetes mellitus and periodontal disease.," *J. Indian Soc. Periodontol.*, vol. 15, no. 4, pp. 388–92, Oct. 2011.

[141] K. Han and J.-B. Park, "Association between oral health behavior and periodontal disease among Korean adults: The Korea national health and nutrition examination survey.," *Medicine (Baltimore).*, vol. 96, no. 7, p. e6176, Feb. 2017.

[142] D. Sambunjak, J. W. Nickerson, T. Poklepovic, T. M. Johnson, P. Imai, P. Tugwell, and H. V Worthington, "Flossing for the management of periodontal diseases and dental caries in adults," in *Cochrane Database of Systematic Reviews*, T. M. Johnson, Ed. Chichester, UK: John Wiley & Sons, Ltd, 2011.

[143] J. Asadoorian and D. Locker, "The impact of quality assurance programming: a comparison of two canadian dental hygienist programs.," *J. Dent. Educ.*, vol. 70, no. 9, pp. 965–71, Sep. 2006.

[144] B. Schüz, A. U. Wiedemann, N. Mallach, and U. Scholz, "Effects of a short behavioural intervention for dental flossing: randomized-controlled trial on planning when, where and how," *J. Clin. Periodontol.*, vol. 36, no. 6, pp. 498–505, Jun. 2009.

[145] M. Jones, J. Y. Lee, and R. G. Rozier, "Oral Health Literacy Among Adult Patients Seeking Dental Care," *J. Am. Dent. Assoc.*, vol. 138, no. 9, pp. 1199–1208, 2007.

[146] G. Lippi and G. Targher, "Glycated hemoglobin (HbA1c): old dogmas, a new perspective?," *Clin. Chem. Lab. Med.*, vol. 48, no. 5, pp. 609–614, Jan. 2010.

[147] D. M. Nathan, H. Turgeon, and S. Regan, "Relationship between glycated haemoglobin levels and mean glucose levels over time," *Diabetologia*, vol. 50, no. 11, pp. 2239–2244, Oct. 2007.

[148] G. K. Johnson and N. A. Slach, "Impact of tobacco use on periodontal status.," *J. Dent. Educ.*, vol. 65, no. 4, pp. 313–21, Apr. 2001.

[149] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[150] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[151] C. Drummond, R. C. Holte, N. V. Chawla, V. S. Sheng, B. Gu, W. Fang, and J. Wu, “Exploiting the cost (in)sensitivity of decision tree splitting criteria,” *Int. Conf. Mach. Learn.*, vol. 66, no. 1, pp. 239–246, 2003.

[152] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, J. Ye, and I. Alzheimer’s Disease Neuroimaging, “Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study,” *Neuroimage*, vol. 87, pp. 220–241, 2014.

[153] C. Kadie and C. Kadie, “Quantifying the Value of Constructive Induction, Knowledge, and Noise Filtering on Inductive Learning,” in *Proceeding of the 8th Machine Learning Workshop*, 1991, pp. 153–157.

[154] “Wisconsin Diabetes Mellitus Essential Care Guidelines 2012.”

[155] “Oral Health: An Essential Component of Primary Care,” 2015.

[156] B. B. Partido, A. A. Jones, D. L. English, C. A. Nguyen, and M. E. Jacks, “Calculus detection calibration among dental hygiene faculty members utilizing dental endoscopy: a pilot study.,” *J. Dent. Educ.*, vol. 79, no. 2, pp. 124–32, Feb. 2015.

[157] F. Shakibaie and L. J. Walsh, “Dental calculus detection using the VistaCam,” *Clin. Exp. Dent. Res.*, vol. 2, no. 3, pp. 226–229, Dec. 2016.

[158] US Census Bureau, “Census 2010,” *US Census Bureau*, 2010. [Online]. Available: http://quickfacts.census.gov/qfd/states/13/13135.html.

[159] MEDICARE, “Medicare. gov,” *01/05/2013*, 2013. [Online]. Available: http://es.medicare.gov/HospitalCompare/(X(1)S(kutnv0q5nm2qmbfa3sfseq0x))/About/HOSInfo/Hospital-Info.aspx?AspxAutoDetectCookieSupport=1.

[160] Centers for Disease Control and Prevention, “Body Mass Index: Considerations for Practitioners.” [Online]. Available: https://www.cdc.gov/obesity/downloads/bmiforpactitioners.pdf. [Accessed: 21-Apr-2017].

[161] “Measuring waist circumference: The importance of waist circumference: Cut the Waist.” [Online]. Available: http://www.cutthewaist.com/measuring.html. [Accessed: 21-Apr-2017].

[162] and B. I. National Heart, Lung, “Description of High Blood Pressure - NHLBI, NIH,” *U.S.Department of Health and Human Services*. [Online]. Available: https://www.nhlbi.nih.gov/health/health-topics/topics/hbp. [Accessed: 21-Apr-2017].

[163] P. Drouin, J. F. Blickle, B. Charbonnel, E. Eschwege, P. J. Guillausseau, P. F. Plouin, J. M. Daninos, N. Balarac, J. P. Sauvanet, and D. O. F. Diabetes, “Diagnosis and

classification of diabetes mellitus," *Diabetes Care*, vol. 32, no. SUPPL. 1, pp. S62–S67, 2009.

[164] "DENTAL NUMBERING SYSTEMS," *Justi Educational Department*, 2003. [Online]. Available: http://www.americantooth.com/downloads/instructions/Dental_Sys_Permanent_Teeth.pdf. [Accessed: 21-Apr-2017].

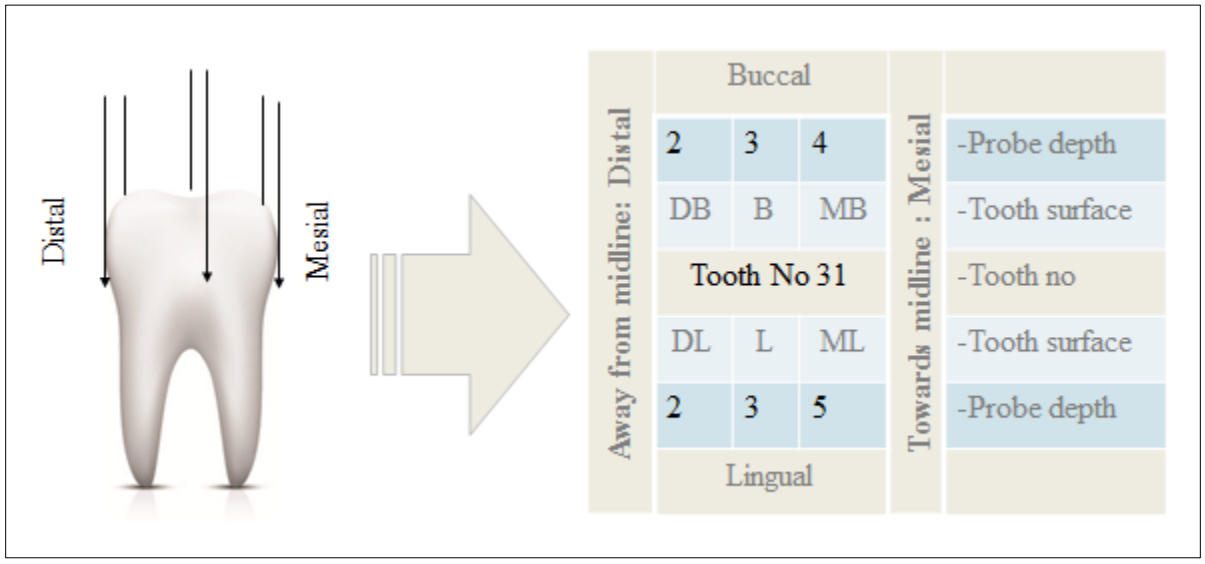
APPENDIX A: DATA DICTIONARY

Demographic information	Characteristics/Definition
Age	Patients between 18 years and 89 years. Age was defined as the age of the patient at the first dental visit.
Gender	Male, Female
Race	According to the Federal Office of Management and Budget (OMB) and United States Census Bureau, define the concept of race, " as social and cultural characteristics as well as ancestry" [158] White/Caucasians, American Indian or Alaska Native Native Hawaiian or Other Pacific Islander Black or African American Asian
Ethnicity	According to the Federal Office of Management and Budget (OMB) and United States Census Bureau, all respondents are categorized by membership in one of the two ethnic categories as following: [158] Not Hispanic/Latino Hispanic or Latino
Insurance information	
Medicaid status	Medicare is, " A Federal and a State program that provides health coverage to patients who have very low income" [159]
Medicare status	Medicaid is, "A Federal program that provides health coverage to patients who are 65 years or older or have a severe disability, irrespective of their income" [159].
Social history information	
Tobacco use	A patient who smokes tobacco. Due to limited data on smokeless tobacco, this study did not use the smokeless tobacco use variable.
Tobacco use status	Tobacco use status was defined as the latest patient self-reported tobacco use behavior. It included the categories of current, former and never smoker.
Current smoker	Self-reported data by the patients who are current tobacco smokers
Former smoker	Self-reported data by the patients who were former smokers
Never smoker	Self-reported data by the patients who have never smoked tobacco product.
Clinical observations	

Height	Height of the patient in meters
Weight	Weight of the patient in kilograms
Body Mass Index (BMI)	According to Centers for Disease Control and Prevention (CDC), BMI is defined as a person's weight in kilograms divided by the square of height in meters [160].
Waist Circumference	Is waist measurement to assess central fat distribution and degree of abdominal obesity [161].
Blood Pressure	According to National Heart, Lung, and Blood Institute (NHBI), Blood pressure is defined as, "the force of blood pushing against the walls of the arteries as the heart pumps blood" [162].
Systolic blood pressure	According to NHBI, Systolic blood pressure is blood pressure when the heart beats while pumping blood [162]
Diastolic blood pressure	According to NHBI, Diastolic blood pressure is blood pressure when the heart is at rest between beats [137]
Laboratory findings	
Random Blood Glucose	Is blood glucose measurement that is carried out at random, regardless of the time the patient eats his food.
High Density Lipids	Component of lipid panel laboratory test
Low Density Lipids	Component of lipid panel laboratory test
Total Cholesterol	Component of lipid panel laboratory test
Triglyceride Levels	Component of lipid panel laboratory test
Co-morbid conditions	
Type 1 Diabetes	Patient diagnosed with Type1 Diabetes (ICD9/10)
Type 2 Diabetes	Diagnosis of patient with Type 2 Diabetes (ICD9/10)
Prediabetes	As defined according to American Diabetes Association[163]
Dental variables	
(Dental-clinical)	Elements collected during comprehensive dental exam

Periodontal Pocket Depth (PPD)	Probing depth of gingival/periodontal pocket in mm
Missing teeth	Number of missing teeth
Present teeth	Number of teeth present in mouth excluding tooth root
Low PD risk	Patients who are at a low risk of developing PD
High PD risk	Patients who are at a high risk of developing PD
Type of Oral hygiene	Status of oral hygiene determined by the dental provider
Dental Calculus and stain	Hard deposit on tooth surface which is difficult to remove by mechanical cleansing such as tooth brushing
Bleeding on probing	Bleeding gums while measuring probing depth
Tooth mobility	Different grades of movements of teeth
Clinical Attachment Loss	Loss of periodontal attachment
Dental plaque	A thin biofilm on tooth surface
Furcation involvement	Loss of periodontal attachment in interproximal areas of tooth
(Dental-Oral hygiene)	
Tooth brushing	Frequency of brushing teeth in a day
Tooth flossing	Frequency of flossing teeth in a day
Tooth numbering	Used of Universal Numbering System [164]
Tooth surfaces	
Mesial	Towards the midline of dental arch
Distal	Away from the midline of dental arch
Buccal surface	Tooth surface that faces towards buccal mucosa/ cheek mucosa
Facial surface	Tooth surface that faces towards lips
Palatal surface	Tooth surface that faces towards palatal arch
Lingual surface	Tooth surface that faces towards tongue
Proximal surface	Mesial and Distal
Common Dental Terms	
Cemento enamel junction	Junction between enamel of tooth crown and cementum of tooth root
Gingiva	Gums
Periodontal ligament	Fibrous connective tissue that runs from tooth root to the alveolar bone (anchors the tooth) [127]
Alveolar bone	Bone surrounding the teeth

APPENDIX B: PERIODONTAL CHART SHOWING TOOTH SURFACES OF A MOLAR FOR MEASURING PROBING DEPTH



CURRICULUM VITAE

NEEL SHIMPI

EDUCATION

Master of Management with concentration in Healthcare Informatics **June 2012**
Cambridge College of Management, MA, USA

Post Graduate Diploma in Healthcare Management **May 2009**
Post Graduate Certificate in Clinical Research **May 2008**
Symbiosis Center of Health Sciences, Pune, India

Bachelor in Dental Surgery **August 2005**
Maharashtra University of Health Sciences, India**,** Graduated in top 10% of the state

Related Projects and core course work completed at University of Wisconsin-Milwaukee

- Datamining and Database Systems, Infrastructure for Information Systems, Information Technology Strategy and Management, Service-Oriented Analysis and Design, Data and Information Management, Ethics and Integrity in Science, Multivariate Techniques in Management Research and Biomedical and Health Care Terminology and Ontology
- Doctoral dissertation: "**Development and Evaluation of An Interdisciplinary Periodontal Risk Prediction Tool Using a Machine Learning Approach**".

WORK EXPERIENCE

Research Specialist at Institute for Oral and Systemic Health, **(June 2013 to present)**
Marshfield Clinic Research Foundation, Marshfield, WI, USA

- Involved in substantive dental and oral health research.
- Facilitated, coordinated and led activities of research projects, including responsibility for keeping deadlines.
- Involved in evaluation and development of clinical decision support systems for an interdisciplinary environment
- Assisted in writing abstracts, research reports for academic publication in scholarly journals and project publications.
- Developed computer based education modules for dental and medical providers
- Conducted various state-wide and nation-wide surveys.
- Engaged in writing grants and manuscripts.
- Developed patient education modules.
- Participated in research related activities within the organization
- Co-mentored summer interns and dental residents.

Research Assistant at Biomedical Informatics Department,
Marshfield Clinic Research Foundation, Marshfield, WI, USA **(Aug 2012- June 2013)**

- Facilitated with research projects and managed to perform independent research with minimal supervision.

- Engaged in substantive research and analysis for decision support systems based on artificial intelligence.
- Actively engaged as a project lead for developing a curriculum for a three day workshop. Demonstrated the ability to delegate and design core curriculum for the workshop.
- Demonstrated the ability of facilitating and coordinating activities of research projects and taking the responsibility for keeping deadlines and undertaking research activities as directed by the team leader.
- Assisted in writing manuscripts and reports.
- Involved in development of proposals for grants/funding.
- Actively worked on analysis of meaningful use stage 2 objectives.
- Demonstrated high level organizational competencies in customer satisfaction

Faculty member of Maharashtra University of Health Sciences, Pune, India (Nov 2011- Jan 2012)

- Conducted CME workshops on 'Introduction to Medical Informatics' and 'Research Methodology'
- Created theoretical and educational modules for medical students that simulated examinations and interviews of virtual patients; enabled students to order tests, therapies, and consults and provided feedback.
- Developed study modules that allowed the doctors and medical students to specify patient symptoms, view information and images and diagnosis.

Study Coordinator and Intern at S.K. Medical Center and Noble Hospital, Pune, India (2009-2010)

- Used e-tools that allowed the integration of the data collected from various investigative sites and laboratories.
- Monitored the Clinical Trial progress with Standard Operating Procedures.
- Responsible for Site monitoring, Site Management and registry Management for Clinical Studies.
- Conducted an independent study project on 'Infections in ICU'.

Dental Surgeon at Polyclinic and Dental Office, Aurangabad, India. **(2006-2007)**

- Created and used dental learning and study models for patients on computer.
- Constructed small templates having specific workflow and data for standardizing some of the basic components.
- Demonstrated great work coordination with the co. surgeons in team performances during major oral surgeries.
- Competently managed 60-70 patients per month and actively involved in all phases of setting up the own clinic.

Dental Resident at C.S.M.S.S Dental College and Hospital, Aurangabad, India. (2004-2005)

- Handled the responsibilities of assisting senior dentist in performing dental procedures.
- Served as a mentor for 1st and 2nd year dental students and was selected as head of student and sports committee.

Peer Reviewed Papers:

- Acharya A, **Shimpi N**, Mahnke A, Mathias R, Zhan Y. Medical care providers' perspective on dental information needs in electronic health records. J Am Dent Assoc. 2017; S0002-8177(17)30094-6. doi: 10.1016/j.adaj.2017.01.026

- **Shimpi N**, Bharatkumar A, Jethwani M, Chyou P, Glurich I, Blamer J, Acharya A. Knowledgeability, Attitude and Behavior of Primary Care Providers Towards Oral Cancer: a Pilot Study. J Cancer Educ. 2016. doi:10.1007/s13187-016-1084-4

- **Shimpi N**, Schroeder D, Kilsdonk J, Chyou P, Glurich I, Penniman E, Acharya A. Medical Providers' Oral Health Knowledgeability, Attitudes, and Practice Behaviors: an Opportunity for Interprofessional Collaboration. Journal of Evidence Based Dental Practice. 2016;16(1):19-29. doi: 10.1016/j.jebdp.2016.01.002

- **Shimpi N**, Schroeder D, Kilsdonk J, Chyou PH, Glurich I, et al. (2015) Assessment of Dental Providers' Knowledge, Behavior and Attitude towards Incorporating Chairside Screening for Medical Conditions: a pilot study. J Den Oral Care 2(1): 102

Peer Reviewed Abstracts:

- Survey of Primary Care Providers' Oral Health Knowledge/Attitude/Practices. **Shimpi N**, Glurich I, Schroeder D, Hegde H, Chyou P, Acharya A. J Dent Res Vol # 96 (Spec Iss A): 2445, 2017(www.iadr.org).

- Community Awareness Towards Association of Diabetes and Oral Health. **Shimpi N**, Hegde H, Glurich I, Schroeder D, Chyou P, Acharya A. J Dent Res Vol # 96 (Spec Iss A): 1798, 2017(www.iadr.org).

- Interdisciplinary Diabetes Management: Qualitative assessment of Medical/Dental Practitioners' Perspectives. Glurich I, Schwei K, Lindberg S, **Shimpi N**, Schroeder D, Acharya A. J Dent Res Vol # 96 (Spec Iss A):3078, 2017(www.iadr.org).

- Dental Quality Analytics: Dental Measures and Provider Performance Dashboard. Hegde H., Steinmetz A., Theisen J., Koralkar R., Finamore J., Halstead S., Leege S., O'Brien J., **Shimpi N,** Acharya A.. J Dent Res Vol # 96 (Spec Iss A): 0566, 2017(www.iadr.org).

- Awareness, Knowledge And Attitudes Of Patients Towards Oral Cancer. **Shimpi N**, Jethwani M, Bharatkumar A, Chyou P, Glurich I, Acharya A. J Dent Res Vol #95 (Spec Iss A): 2392532, 2016 (www.iadr.org).

- Oral Cancer Risk Assessment Using Machine Learning Algorithms. Rostami R., Hegde H., **Shimpi N**, Pack G., Olson B., Acharya A. J Dent Res Vol #95 (Spec Iss A): 1486 ,2016.(www.iadr.org).

- Smoking Status Classification Of Clinical Notes Using Natural Language Processing. Hegde H., **Shimpi N**, Pack G., Rostami R., Acharya A. J Dent Res Vol #95 (Spec Iss A): 2383016,2016 (www.iadr.org).

- Usability Heuristic Evaluation of a Web-based Case Simulator for Dentists. Schwei K, Thomas K, Mahnke A, **Shimpi N**, Thirumalai V, Enstad C, Johnson K, Godlevsky O, Johnson N, Rush B, Acharya A. J Dent Res Vol #95 (Spec Iss A): 1895,2016 (www.iadr.org).

- DentaSeal: A Web-Based Application For Wisconsin Seal-A-Smile Program. Ray W, Steinmetz A, Hegde H, Halstead S, Baker K, **Shimpi N**, Acharya A. J Dent Res Vol #95 (Spec Iss A): 2016 (www.iadr.org).

- User-Centered Approach For Developing Web-based Dental Sealant Registry Called 'DentaSeal'. Steinmetz A, Thomas K, Ray W, Hegde H, Halstead S, **Shimpi N**, Acharya A. J Dent Res Vol #95 (Spec Iss A): 2016 (www.iadr.org).

- Acharya A, Glurich I, Schwei K, **Shimpi N**, Jansen M, O'Brien J, Kleutsch T, Penniman E, Schroeder D, Nycz G. Developing Medical-Dental Integrated Care Models (ICM) to Manage Diabetes. WREN conference. 2015.

Scientific Presentations

- **Shimpi N**, Glurich I, Schroeder D, Hegde H, Chyou P, Acharya A, 'Survey of Primary Care Providers' Oral Health Knowledge/Attitude/Practices'. *Poster presentation* at the 95thIADR General Session & Exhibition, 46thAnnual Meeting of the AADR, 41stAnnual Meeting of the CADR, Boston, Massachusetts, March 22nd – 25th March, 2015.

- **Shimpi N,** Jethwani M, Bharatkumar A, Chyou PH, Glurich I, Acharya A, Awareness, Knowledge And Attitudes Of Patients Towards Oral Cancer. Oral presentation at the 94th AADR/CADR Annual Meeting, Los Angeles, CA, March 16-19, 2016.

- **Shimpi N**, Schroeder D, Kilsdonk G, Chyou P, Acharya A, Knowledge, Attitude and Behavior of Medical Providers Towards Oral Health. Oral presentation at the 93rd IADR General Session & Exhibition, 44th Annual Meeting of the AADR, 39th Annual Meeting of the CADR, Boston, Massachusetts, March 11th – March 14th, 2015.

- Koehler K, **Shimpi N**, Hegde H, Pack G, Chyou P, Acharya A, 'Development of Prototypical Design of Oral Cancer Risk Assessment Tool'. *Oral presentation* at the 93rd IADR General Session & Exhibition, 44th Annual Meeting of the AADR, 39th Annual Meeting of the CADR, Boston, Massachusetts, March 11th – March 14th, 2015.

- **Shimpi N,** Hegde H**,** Bohne J, Acharya A**,** 'Natural language processing (NLP) pipeline to facilitate clinical free text information extraction: A pilot effort', Third Coast Consortium Biomedical and Health Informatics, Milwaukee, Wisconsin, April 17th, 2015.

- Vesel M, **Shimpi N**, Acharya A**,** Developing a Reference Cross Mapping Between Different Dental Diagnostic Terminologies. Oral presentation at the 43rd AADR Annual Meeting & Exhibition, 38th Annual Meeting of the CADR, Charlotte, North Carolina, March 19th – March 22nd , 2014.

Intramural Presentations at Marshfield Clinic

- Shimpi N, 'Conducting Oral examination in Pediatric Patients and pediatric oral risk assessment tool' presentation to the pediatric residents, 2016.

- Shimpi N, 'Conducting Oral examination in Pediatric Patients' presentation to the pediatric residents, 2015.
- Shimpi N, 'Oral and Systemic Diseases', Journal Club Presentation. 2015
- Shimpi N, 'Association of Periodontal Disease and Alzheimer Disease', Journal Club Presentation. 2014
- Shimpi N, 'Big data analytics in Healthcare', Health Innovation Chat. 2014
- Shimpi N, Pathak R, 'Diabetes and Oral Health', Grand Rounds at Marshfield Clinic. 2013

Proficiencies:
- Machine learning and ontology: Waikato Environment for Knowledge Analysis (WEKA), PR-OWL
- SPSS, R- Amelia II for multiple imputation
- Working knowledge of SAS, Apache Spark, BayesOWL
- C, C++ certification. Operating systems: Windows XP, Vista, Linux, Mackintosh.
- PowerPoint, Access, Excel, Word, Outlook, Adobe photoshop.

Awards:
- Honored as one of the Best Outgoing Student in the Dental School.
- College awards for securing highest marks in Periodontics, Pedodontics , Orthodontics, Oral Medicine and Radiology

Professional Memberships

Member of the International Association for Dental Research (IADR), 2014-present

Extra-curricular activities: State Level Champion of Archery.(1998-1999), State Level Champion of Badminton (1992-1994), District Level Champion of Throw Ball(1995-1996), Medals and Trophies for Chess, Lawn Tennis and Table tennis(2000-2005), Sports Secretary, Dental College(2004-2005), Was interviewed on All India Radio on "Prevention of Oral and Dental Diseases" (Nov, 2006). Series of 12 articles on Dental related topics were published in leading newspapers in India (2005-2006).