DEVELOPING AND DEPLOYING DATA MINING TECHNIQUES IN HEALTHCARE

By

SAEED PIRI

Bachelor of Science in Industrial and System Engineering
Amirkabir University of Technology
Tehran, Iran
2008

Master of Science in Industrial Engineering
Sharif University of Technology
Tehran, Iran
2011

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2017

DEVELOPING AND DEPLOYING DATA MINING

TECHNIQUES IN HEALTHCARE

Dissertation Approved:

Dr. Tieming Liu

Dissertation Adviser

Dr. Sunderesh Heragu

Dr. Farzad Yousefian

Dr. William Paiva

Dr. Dursun Delen

ACKNOWLEDGEMENTS

It has been a long journey since January 2013 when I joined the IEM program at OSU. From the very beginning, I was surrounded with scholars who guided, inspired, and helped me reach this point.

First, I would like to thank Dr. Tieming Liu, my advisor for all his help and support in the past four years, I appreciate everything he did throughout my PhD program. Additionally, I am grateful for the advice, support, and guidance of my co-advisor Dr. Dursun Delen who was tremendously helpful and supportive during this process. Completing and earning my dissertation would not have been possible without his guidance and dedication.

I acknowledge and appreciate the financial support of CHSI and guidance of its executive director Dr. Willima Paiva for the past three years. He has been a pleasure to work with throughout these years. I am also grateful for the advice and support I received from my committee members, Dr. Sunderesh Heragu, Dr. Farzad Yousefian, and Dr. Arash Pourhabib. They were all very supportive during the process and were very helpful in getting me to the finish line.

I would also like to thank my parents. They have always been my greatest inspirations in all stages of my life and words cannot express how much gratitude I have for all that they have done for me to ensure I was able to get to the point I am now. I also have to thank all my family members and friends for their kind support.

Last but not least, I would like to thank the love of my life, Yasamin. Without her support, it would not have been possible for me to get thorough my PhD. I know I can do anything with her by my side.

Name: SAEED PIRI

Date of Degree: JULY, 2017

Title of Study: DEVELOPING AND DEPLOYING DATA MINING TECHNIQUES IN HEALTHCARE

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract: Improving healthcare is a top priority for all nations. US healthcare expenditure was $3 trillion in 2014. In the same year, the share of GDP assigned to healthcare expenditure was 17.5%. These statistics shows the importance of making improvement in healthcare delivery system. In this research, we developed several data mining methods and algorithms to address healthcare problems. These methods can also be applied to the problems in other domains.

The first part of this dissertation is about rare item problem in association analysis. This problem deals with the discovering rare rules, which include rare items. In this study, we introduced a novel assessment metric, called adjusted_support to address this problem. By applying this metric, we can retrieve rare rules without over-generating association rules. We applied this method to perform association analysis on complications of diabetes.

The second part of this dissertation is developing a clinical decision support system for predicting retinopathy. Retinopathy is the leading cause of vision loss among American adults. In this research, we analyzed data from more than 1.4 million diabetic patients and developed four sets of predictive models: basic, comorbid, over-sampled, and ensemble models. The results show that incorporating comorbidity data and oversampling improved the accuracy of prediction. In addition, we developed a novel "confidence margin" ensemble approach that outperformed the existing ensemble models. In ensemble models, we also addressed the issue of tie in voting-based ensemble models by comparing the confidence margins of the base predictors.

The third part of this dissertation addresses the problem of imbalanced data learning, which is a major challenge in machine learning. While a standard machine learning technique could have a good performance on balanced datasets, when applied to imbalanced datasets its performance deteriorates dramatically. This poor performance is rather troublesome especially in detecting the minority class that usually is the class of interest. In this study, we proposed a synthetic informative minority over-sampling (SIMO) algorithm embedded into support vector machine. We applied SIMO to 15 publicly available benchmark datasets and assessed its performance in comparison with seven existing approaches. The results showed that SIMO outperformed all existing approaches.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 1

Introduction and Motivation 1
Problem Statements 5
Research Objectives 8
Contributions 8
Organization of the Dissertation 10

II. LITERATURE REVIEW 11

Rare Rules Association Analysis: The Case of Diabetes Complications 11
CDSS for Diabetic Retinopathy 14
Imbalanced Data Learning Algorithms 17

III. DEVELOPMENT OF A NEW METRIC TO IDENTIFY RARE PATTERNS IN ASSOCIATION ANALYSIS: THE CASE OF ANALYZING DIABETES COMPLICATIONS 22

Methodology 23
Results 30
Discussion and Conclusion 42

IV. A DATA ANALYTICS APPROACH TO BUILDING A CLINICAL DECISION SUPPORT SYSTEM FOR DIABETIC RETINOPATHY: DEVELOPING AND DEPLOYING A MODEL ENSEMBLE 46

Methodology 46
Results 61
Discussion and Conclusion 73

V. DEVELOPING A SYNTHETIC INFORMATIVE MINORITY OVER-SAMPLING (SIMO) ALGORITHM EMBEDDED INTO SUPPORT VECTOR MACHINE TO LEARN FROM IMBALANCED DATASETS 77

Support Vector Machine 78
SIMO and W-SIMO Algorithms 82
Numerical Experiments 87
Discussion and Conclusion 99

Chapter Page

VI. CONCLUSION.....105

REFERENCES 109

LIST OF TABLES

Table Page

3.1-Hypothetical Dataset H....25
3.2- Generated rules from Dataset H....25
3.3-Diabetes complications count and percentage in the data....29
3.4-Comorbidity index value by race/ethnicity....31
3.5- Comorbidity index value in rural vs urban diabetic patients....32
3.6- Comorbidity index value by gender....32
3.7- Two-item association rules- general population....37
3.8- Two-item association rules in various demographic groups....39
3.9-Three-item association rules- general population....42
4.1- Demographic variables....62
4.2- Lab procedure variables....62
4.3- Comorbidity variables....63
4.4- Set 1 - Basic models' results....65
4.5- Set 2 - Comorbid models' results....66
4.6- Set 3 - Over-sampled models' results....67
4.7- Set 4 - Ensemble models' results....71
5.1-Notations for SIMO and W-SIMO algorithms....86
5.2 -Benchmark datasets characteristics....91
5.3- Performance of imbalanced data learning approaches (using G mean)....93
5.4- Performance of imbalanced data learning approaches (using AUC)....94
5.5-Average difference between our algorithm and other approaches....96
5.6-Overall ranking on linear SVM....96
5.7-Overall ranking- SVM-RBF kernel....96
5.8-Overall ranking on logistic regression....96
5.9- Overall ranking on decision tree....96
5.10-The performance of best approach in each machine learning technique....97
5.11- Imbalanced gap and average # of synthetically generated data points....98
5.12- Sensitivity analysis on SIMO parameters....100

LIST OF FIGURES

Figure Page

1.1- Three analytics levels-descriptive, predictive, and prescriptive....3
3.1- Final dataset structure....30
3.2-Comorbidity index value by race/ethnicity....31
3.3a- Comorbidity index value in rural vs urban diabetic patients....32
3.3b- Comorbidity index value by gender....32
3.4- Diabetes complications' prevalence by race/ethnicity....33
3.5-Diabetes complications' prevalence: rural versus urban....35
3.6-Diabetes complications' prevalence: females versus males....36
4.1- A simplified conceptual data model for Cerner Health Facts....48
4.2- Data preparation steps....49
4.3- Modeling procedure....52
4.4- Two-layer perceptron neural network....53
4.5- Predictive model sets....56
4.6- Synthetic minority over-sampling technique (SMOTE)....57
4.7- AUC comparison among modeling techniques and modeling sets....68
4.8 - ROC charts of modeling techniques in different sets....69
4.9- Variable importance ranking in detecting diabetic retinopathy....72
5.1- Linear SVM hyperplane....79
5.2- SIMO algorithm mechanism (simplified)....84
5.3- Confusion matrix....88
5.4- ROC chart....90
5.5- 4-fold cross validation mechanism....92

CHAPTER I

INTRODUCTION

1.1. Introduction and Motivation

Improving healthcare is a top priority for all nations. In 2014, US healthcare expenditure was $3 trillion, or $9,523 per person. In the same year, the share of GDP assigned to healthcare expenditure was 17.5% [1]. These statistics shows the importance of improving the healthcare delivery system.

Diabetes is one of the most serious and prevalent chronic conditions affecting approximately 415 million people worldwide, with this number is expected to grow to 642 million by 2040 [2]. The situation is particularly dire in the U.S., which has the highest prevalence of diabetes of all developed nations. Approximately 86 million adults aged 20 years and older (37%) were diagnosed as pre-diabetic between 2009 and 2012. By 2014, the estimated number of adults with diagnosed or undiagnosed diabetes topped 29 million, representing about 9% of the U.S. adult population (*National Diabetes Statistics Report, 2014)*. Minority racial/ethnic groups have higher rates of diabetes than non-Hispanic whites, with Native Americans having the highest rate at nearly 16%, followed by African American (13.2%), Hispanics (12.8%), Asian Americans (9%)

and Caucasians (7.6%). Each year, approximately 1.4 million people are diagnosed with either type 1 or type 2 diabetes, making it the nation's seventh leading cause of death in 2010. The cost of diabetes in America in 2012 was approximately $245 billion, which included about $69 billion in indirect costs related to impairment, job loss, and premature death.

In recent years, due to modern healthcare technology, a large amount of data from various sources, such as patient care, as well as compliance and regulatory requirements has become available [4]. The digitization of this data has been happening rapidly in the recent years [5]. The extensive availability of healthcare data, as well as advances in the area of data mining and machine learning, has generated the interesting field of healthcare analytics. The development of decision support systems by data analysts with the aid of clinical experts' knowledge has eased the burden on physicians and clinicians and smoothed clinical procedures. Analyzing healthcare data and applying machine learning techniques in this area have several benefits: patients can be stratified based on the severity of a particular disease or condition and, consequently, suitable treatments can be provided for each group; risk factors of different diseases can be identified, leading potentially to better health management; and diseases can be detected at early stages, allowing for appropriate interventions and treatments. For a comprehensive discussion about healthcare analytics, its promises and its potentials we refer readers to [5].

Similar to other domains, three types of analytics can be conducted in healthcare: descriptive, predictive, and prescriptive. Most of the analytics projects start with descriptive analytics [6]. Descriptive analytics tells us what has happened in the past and what is going on at the present. In descriptive analytics, hypotheses are tested, and trends are identified. Descriptive analytics could lead to discovering interesting patterns in the data. The next step in analytics is predictive analytics. Predictive analytics tells us what is going to happen in the future. In predictive analytics, by using statistical and machine learning models to analyze historical data, the relationship between the target and predictors can be detected [7]. The final step in an analytics

project is prescriptive analytics. In this step, analysts use optimization techniques to identify the best course of action. In Figure 1.1, different levels of analytics can be seen [8].

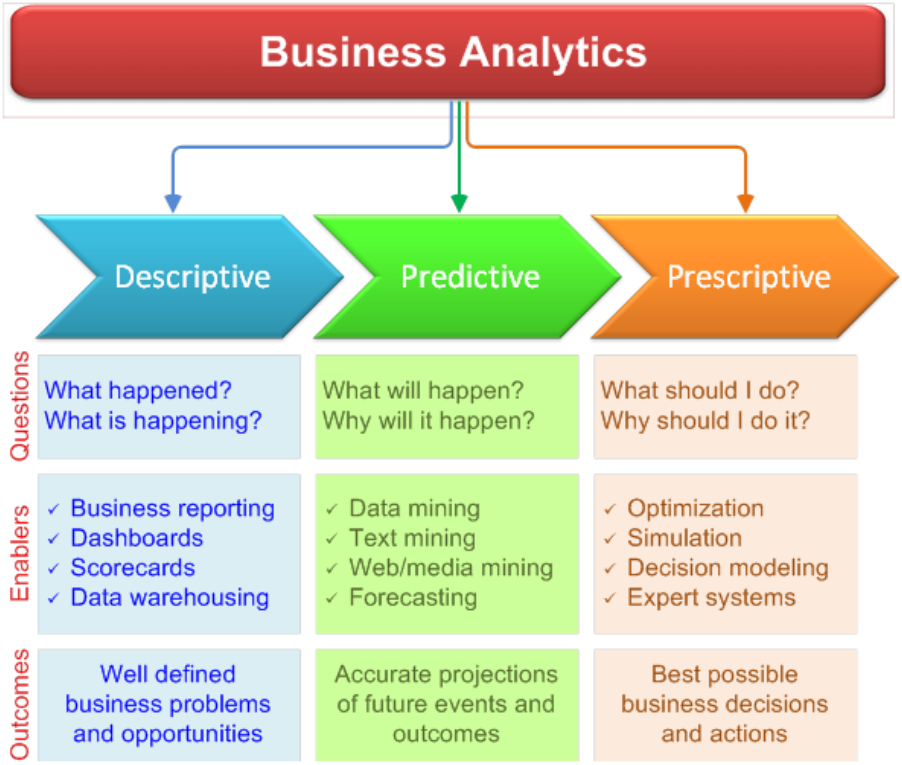


Figure 1.1- *Three analytics levels-descriptive, predictive, and prescriptive. Adopted from Real-World Data Mining: Applied Business Analytics and Decision Making, by D. Delen, 2015: FT Press (a Pearson Publishing Company). Adapted with permission*

Discovering the affinities and associations among various items has long been of interest to managers and data analysts. Association rule mining (or market basket analysis in the marketing and business literature) is a data mining method that aims to reveal the association/affinity patterns/rules among various items (objects or events) that occur together. We can enumerate several implications for association rule mining. In the retail industry, it facilitates finding solutions for assortment planning, coupon design, and product discounting [9]. In health care settings, association rule mining may help answer questions such as whether the presence of a particular health condition increases the probability of developing other conditions and which

preventive measures could best reduce risk of complications. These questions can be addressed through an understanding of the associations among different complications.

One critical challenge in healthcare domain is detecting the disease in early stages. One of the major complications of diabetes that has not received enough attention is diabetic retinopathy. This complication is the most common cause of vision loss among people with diabetes and a leading cause of blindness for American adults. According to the 2014 National Diabetes Statistics Report, between the years 2005 and 2008, 4.2 million of American diabetics aged 40 or older suffered from diabetic retinopathy. Among patients who have had diabetes for up to 20 years, almost all type I and more than 60% of type II diabetics develop retinopathy [10]. This complication is caused by damage to the blood vessels of retina, the light-sensitive tissue at the back of the eye. At its early stages, diabetic retinopathy may be asymptomatic or only show mild vision problems, but if it is not diagnosed and treated in time, it can eventually cause blindness.

Another challenge that data analysts face in the healthcare domain as in many other domains is dealing with imbalanced datasets. A dataset is called imbalanced when the distribution of different classes in the data is not similar. For instance, in the case of two-class data, there are many more examples of one class (negative examples) compared to the other class (positive examples). Let us call the class with fewer examples the minority class, and the class with more examples the majority class. The imbalanced datasets are very common in real-life problems, especially in pattern recognition problems. For example, if a sample of people were tested for a specific disease, only a small portion of them would actually have the disease. Another example is credit card fraud detection where only a few numbers of transactions in the whole sample of transactions are actually fraud [11].

In imbalanced datasets, the prediction accuracy, especially for the minority class, is a critical challenge. When the standard machine learning techniques are applied to the imbalanced data, the

result will be in favor of the majority class, i.e. a big portion of the minority class examples will be classified as the majority. In real world applications, the detection accuracy of the minority class is critically important because the minority class usually is the class of interest. Thus, misclassifying the minority class has much higher cost compared to misclassifying a majority class example. To make it clearer, compare the cost of misclassifying a cancerous patient as non-cancerous to the cost of misclassifying a non-cancerous as cancerous; in the former case, the misclassification may lead to death of a person but in the latter case, there will be some more tests and screenings.

1.2. Problem Statements

One of the most critical problems in association rule mining is the rare item problem. This problem emerges when there are items that occur rarely compared to items that are more frequent. For instance, in retail, purchases of items such as electronics or jewelry are likely rarer than grocery purchases. Rare items, however, may be equally as important as frequent items or even more so. In health care, some complications may not be as frequent as other complications, but they may be more critical or even fatal. Therefore, it is important to recognize and discover the association patterns among rare items as well as the association patterns between rare and frequent items. Any association rule that includes a rare item is called a rare rule. Discovering rare rules is a critical challenge in association analysis. In classical association rule mining, a minimum support is specified to extract the association rules. Support of a rule specifies the fraction of a population for which the rule is true. Setting a high threshold for support may lead to losing rare association rules. On the other hand, specifying a low threshold for support will lead to over generating association rules. Therefore, the problem statement is "how to retrieve and discover rare association rules without over-generating association rules?"

Diabetes typically leads to several complications, often presenting simultaneously. The American Diabetes Association (ADA) classifies these conditions as: skin complications, eye complications (retinopathy), neurological manifestations (neuropathy), foot complications, diabetic ketoacidosis (DKA) and ketones, renal manifestations (nephropathy), high blood pressure (hypertension), stroke, hyperosmolar hyperglycemic nonketotic syndrome (HHNS), gastroparesis, heart disease, stroke, and mental health. The existence of more than one distinct condition in a patient is defined as comorbidity [12]. Comorbidity is highly prevalent in diabetics. Research shows that between 1999 and 2004, only 14% of type 2 diabetic patients were not diagnosed with any additional comorbid conditions [13]. The benefit of considering comorbidities as opposed to studying different diseases in isolation has been showed by many researchers ([14], [15], [16], [17], and [18]). The high prevalence of comorbid conditions among diabetics and the benefits of studying comorbidities shown by other researchers motivated us to study the comorbidity in diabetes and conduct association analysis among its complications. The problem statement here is "is there any strong and interesting association among complications of diabetes?"

Although retinopathy is preventable and existing treatments can slow down the disease progress, vision loss that happens in the late stages of retinopathy cannot be restored. Thus, it is critical to diagnose this complication as early as possible. The current method for diagnosing diabetic retinopathy is a comprehensive eye examination in which after a patient's eye is dilated, an ophthalmologist examines the retina with an indirect ophthalmoscope and a special lens. Unfortunately, and despite the high prevalence of retinopathy, the annual diabetic retinopathy evaluation has one of the lowest rates of patient compliance for several reasons. First, many patients do not seek proper medical attention because this disease is asymptomatic at the early stages; second, availability of ophthalmologists is low or even nonexistent in many areas, especially in rural communities; and third, many patients find the necessary eye dilation unpleasant. Because of this low compliance rate, about 50% of patients with diabetic retinopathy

are undiagnosed (National Eye Institute report, 2015). Therefore, the rising prevalence of diabetes, coupled with barriers to ophthalmological screenings that lead to a high rate of undiagnosed diabetic retinopathy patients, create an urgent need for a tool to detect this complication. To be useful, this tool should be non-invasive, readily available to diabetic patients, validated on a large number of cases, and eliminate the need for specialized equipment that is not universally available. This study sets out to employ an analytics approach on data collected during routine primary care visits to fills this gap. Specifically, we build a clinical decision support system (CDSS) for prediction of diabetic retinopathy that satisfies the aforementioned requirements for diagnostic tools. Our problem statement in this research topic is "How to detect diabetic retinopathy at early stages when retina images are not available?"

To develop the CDSS for diabetic retinopathy, we have applied several machine learning techniques such as neural networks, logistic regression, decision tree and random forest. Ensemble models are learning approaches that combine multiple single classifiers and then make the final classification decision by an averaging or voting mechanism [19]. Use of ensemble models in several studies in the literature, including the studies that used lab and demographic data to predict diabetic retinopathy, points out to the complexity of this problem domain. Ensemble models have the benefit of being more robust than single models [20], and therefore, improve prediction accuracy. For this reason, we also employ an ensemble modeling approach in developing our CDSS. Specifically, we develop a heterogeneous confidence margin ensemble and illustrate how it outperforms the existing ensemble techniques. In this research topic, the problem statement is "how to improve the prediction accuracy of single classifiers through developing an efficient ensemble approach?"

There are various remedies for the imbalanced data learning problem. One of these remedies is to modify the imbalance ratio in the dataset. Decreasing the imbalanced ratio can be achieved through either under-sampling the majority class, i.e. removing some portion of data that belong

to the majority class; or over-sampling the minority class, i.e. generating synthetic data points that belong to the minority class. In this research topic, the problem statement is "how to address the problem of learning from imbalanced data, by developing an efficient over-sampling algorithm?"

1.3. Research Objectives

Our research objectives in this dissertation are as follows,

- Developing a new assessment metric to discover rare rules in association analysis
- Discovering the potential existing associations among complications of diabetes by applying association analysis
- Developing a CDSS to detect diabetic retinopathy using lab data collected during a routine diabetic primary care visit
- Developing a novel ensemble approach to further improve the prediction accuracy of single classifiers
- Developing an over-sampling algorithm embedded into support vector machine to enhance the performance of machine learning techniques when applied to imbalanced datasets

1.4. Contributions

In this research, we have three topics. First topic is "developing a new metric to identify rare patterns in association analysis: the case of analyzing diabetic complications". Second topic in this dissertation is "a data mining approach to build a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble". Finally, the third topic is "developing a synthetic informative minority over-sampling (SIMO) algorithm embedded into support vector machine to learn from imbalanced datasets."

In the first research topic, our contributions to the fields of decision support systems and medical informatics are twofold. At the methodological level, we introduced adjusted_support, a new assessment metric for association rule mining that addresses the rare item problem. By using adjusted_support, we will be able to address the problem of rare items by extracting rare rules without over-generating association rules. At the application level, we performed association rule mining on the complications of diabetes by applying the adjusted_support metric. Our findings in this research shed light on the association patterns among diabetes complications, which has not received enough attention in the literature.

Second research study contributes to the data mining and medical decision support literatures from three perspectives: methodology, data management, and application. In the methodological aspect, we develop and evaluate a novel approach in building ensemble models. This approach aggregates the predictions of individual models by calculating a weighted confidence margin across all models. However, in contrast to the existing weighted averaging ensembles that assign weights to individual models based on their overall prediction performance, our confidence margin ensemble assigns varying weights to the constituting models. These weights are calculated for each observation in the data and are based on the distance between the estimated probabilities of records and the decision cut off point. We show how this approach improves the accuracy of decisions made by our CDSS. From the data management perspective, we processed a very large transactional database of clinical encounters and aggregated the observations at the patient level. This enabled us to create a single data set containing comorbid conditions of patients to develop an accurate picture of their health status. Consequently, our CDSS is able to consider a larger number of risk factors and provides a more realistic depiction of the coexistence of chronic diseases. Finally, in the application aspect, we develop an accessible, easy-to-implement, and inexpensive solution to the currently high proportion of undiagnosed retinopahty

among diabetics. This CDSS reduces direct and indirect medical costs of the healthcare system in the US and more importantly, saves eyesight for a large number of citizens.

In the third research topic, we propose a novel over-sampling algorithm integrated with support vector machine (SVM). We can numerate several advantages for our proposed algorithm. First, it is embedded into a powerful classifier, i.e. SVM, and therefore better results are expected compared to other pre-processing approaches. Second, we conduct over-sampling rather than under-sampling that may lead to information loss due to discarding a fraction of data. Finally, we perform the over-sampling only on the informative minority examples. In this way, we generate the least amount of synthetic data points; therefore, the distribution of the training data will not change dramatically. In addition, because the amount of synthetic generated data is much less compared to other existing methods such as SMOTE, Borderline SMOTE, Safe-Level SMOTE, and Cluster-SMOTE, the computational cost of training machine learning techniques will be lower.

1.5. Organization of the Dissertation

The remainder of this dissertation is organized as follows. The second chapter contains the literature review on the three topics that are covered in this research. The third chapter presents the association analysis method and its implementation on diabetes complication. In Chapter 4, different stages of developing a CDSS for diabetic retinopathy is described and the results of the developed predictive and ensemble models are provided. Chapter 5 presents our developed over-sampling algorithm, SIMO. This algorithm is evaluated compared to other existing approaches and the results are provided in Chapter 5. Finally, in Chapter 6, we summarize and conclude the dissertation.

CHAPTER II

LITERATURE REVIEW

In this chapter, we review the related literatures for the research topics in this study. First, we provide the literature review for association analysis and the rare rule problem. We also review the studies related to the application of association analysis on diabetes complications. Then, we discuss the related studies to the diabetic retinopathy CDSSs and ensemble models. Finally, we review the studies in the domain of imbalanced data learning.

2.1. Rare Rules Association Analysis: The Case of Diabetes Complications

In recent years, the application of analytics and data mining in health care has received much attention. Although association analysis has been one of the popular data mining methods applied to health care data, its application to studies of diabetes complications, especially rare complications, is still limited. Agrawal, et al. [9] introduced association analysis for the first time. They studied associations among items in a large customer transaction dataset. Following their work, association analysis has been applied to many domains, such as bioinformatics ([21], [22], [23]), social domain ([24], [25], [26]), and earth science ([27], [28]).

One of the components of association analysis is assessing rules such that only the most useful ones are retained. Several assessment measures, termed interestingness metrics, exist for evaluating rules and filtering out the least useful. Two traditional measures, support and confidence, will be described in detail in Chapter 3. Several other interestingness metrics for association rules were introduced by researchers, such as h-confidence by Xiong, et al. [29], NConf by Liu, et al. [30] and relative-confidence by Yan, et al. [31]. Tan, et al. [32] extensively reviewed 21 metrics for association patterns and described their usefulness in different application areas. Various association analysis techniques and their properties were reviewed by Kotsiantis and Kanellopoulos [33].

Rare item problem is one of the important challenges in association analysis. Theoretically studying the rare item problem in association analysis started by Liu, et al. [34]. They proposed a multiple min-support approach in which every item in the data has its own min item support (MIS). MIS is specified by comparing a lowest allowable support and the support of the item times a parameter, β. In this way, rare items have lower min-support compared to frequent items, thus they will not be ignored in the rule generation procedure. There are two are main downsides of this approach. First, specifying MIS when the number of items in the dataset is large is a tedious job, and second, determining the optimal value of β is not easy. To address these issues, Yun, et al. [35] introduced relative support. Their formula does not include the parameter β, therefore they did not have the challenge of determining the optimal value for β. Nevertheless, their approach still requires specifying multiple suitable min-support for various itemsets.

Wang, et al. [36] also introduced a framework to address the rare item problem. Similar to the two previously mentioned studies, they assigned different min-support to various itemsets by tracking the dependency chain of itemsets in generating the itemset. While other multi min-support approaches focus on the frequency of items, the approach developed by Seno and Karypis [37] focuses on the length of the itemset, i.e., the itemsets with more items have lower min-

support while itemsets with fewer items have higher min-support. Any-confidence, all-confidence, and bond were introduced by Omiecinski [38] as alternatives for supports in rare item association analysis. Adjusted_support is similar to any-confidence in 2-item rules. These metrics are effective for rules that all items are rare, but they are not effective for rules containing both rare and frequent items. Kiran and Re [39] proposed an improved multiple min-support approach for extracting the rare association rules. Their approach requires specifying multiple minimum support, which is inconvenient compared to a single minimum adjusted_support in real-world application. For a comprehensive review of existing methods and metrics for rare item association analysis, we refer readers to [40].

Several researchers have applied association analysis to diabetes. Simon, et al. [41], Simon, et al. [42], Ramezankhani, et al. [43], and Kamalesh, et al. [44] applied association analysis to assess the risk of developing diabetes. Shin, et al. [45] analyzed the data of 5,022 patients diagnosed with essential hypertension and Valent, et al. [46] analyzed the data of 9,358 diabetic patients. By applying association analysis, they showed that essential hypertension and diabetes mellitus were strongly associated. This result is not surprising since about 71% of diabetic patients have hypertension [3]; thus, this association pattern cannot be considered very useful or interesting. Kim, et al. [47] analyzed the data of 20,314 diabetic patients in South Korea and assessed the associations among various diseases and type 2 diabetes. They found strong associations between diabetes and hypertension; diabetes, hypertension, and stroke; and diabetes, hypertension, and dyslipidemia. However, the relatively small number of patients as well as the limitation of race to Asians diminishes the reliability and generalizability of their results. Therefore, there is an urgent need for a comprehensive study of the associations among diabetes complications using a large dataset that represents the diversity of diabetic patients. Our study fills this gap in the literature by performing association analysis on complications associated with diabetes and introducing a new interestingness assessment metric that captures both rare and frequent association patterns.

2.2. CDSS for Diabetic Retinopathy

Even though CDSSs based on EMR data have been broadly used by practitioners in recent years, its implementation in the field of ophthalmology is still limited [48]. This dearth exists while several researchers have studied the relationship between diabetic retinopathy and different potential risk factors. For instance, Karma, et al. [49] studied the existence of diabetic retinopathy in 328 diabetic patients using ophthalmoscopy and wide field fundus photography and tried to identify the association between diabetes duration and other risk factors, such as nephropathy and coronary disease. In another study, Klein, et al. [50] measured the relationship between retinopathy and hyperglycemia by studying 1878 diabetics.

Most of the existing CDSSs for diabetic retinopathy use image processing algorithms. While these algorithms facilitate early detection of diabetic retinopathy, they require an image of the retina. Therefore, although they ease the burden of assessing the images of retina, they fail to address the evident barrier of patients' access to specialists. Examples of studies that belong to this category are (Kahai, et al. [51], Paunksnis, et al. [52], Marsolo, et al. [53], Tsai, et al. [54], Noronha, et al. [55] Bursell, et al. [56], Kumar and Madheswaran [57], and Xiao, et al. [58]). We refer the readers to Mookiah, et al. [59] for a comprehensive review of research in this category. The other category of CDSSs for diabetic retinopathy includes those matched with lenses or an ophthalmoscope that can be used on a smartphone. Prasanna, et al. [60] proposed a portable smartphone-based CDSS that requires attaching an ophthalmoscope to a smartphone to capture fundus images, and captured images will be processed by the algorithm installed on the smartphone. Bourouis, et al. [61] also proposed a smartphone-based algorithm integrated with microscopic lenses used to capture retinal images. Their CDSS uses a neural network model to analyze such images and provide the results. Despite all the benefits of these algorithms,

additional equipment is still required for retinal imaging, which, for many diabetics and primary care providers, may be cost-prohibitive or unavailable.

Many research projects have studied the association of retinopathy and different lab tests. For instance, the association of retinopathy and hemoglobin A1c has been shown in several studies ([62], [63], [64]). Researchers have also studied the relationship between cholesterol and retinopathy and have found the two to be related ([65], [66]). The Diabetes Control and Complications Trial (DCCT) and the U.K. Prospective Diabetes Study (UKPDS) have shown that controlling the glucose level could reduce the risk of retinopathy [67]. Other studies have shown that retinopathy and hypertension are associated [68]. Besides blood tests, some urine tests such as proteinuria are shown to be associated with retinopathy [69].

While these studies show the potential for developing tools that can detect or predict retinopathy using lab results, only a few studies have used lab and demographic data to detect diabetic retinopathy without requiring retinal imaging. Skevofilakas, et al. [70] developed a CDSS using data from 55 type I diabetic patients to predict the risk of diabetic retinopathy. They applied classification-based Rule Induction with C5.0, Hybrid Wavelet Neural Network (HWNN), Classification and Regression Tree (CART), and neural network, and merged their results using a voting mechanism. In another work, Balakrishnan, et al. [71] used data from 140 diabetic patients in Malaysia to build a diabetic retinopathy predictive system, which employed a voting mechanism to select the final outcome from the results of decision tree and case-based reasoning (CBR).

Although these two research projects did not use any retinal images to predict the risk of diabetic retinopathy, they are limited in a number of ways. First, they are based on small samples (55 in the first and 140 patients in the second study). Second, they consider a limited number of risk factors. These charactreistics not only contribute to lack of a comprehensive image of the

patients' health status, but also make the final results less generalizable. Additionally, while according to the NIH statistics, 95% of the diabetics are type II, the first study has only focused on type I diabetic patients. Research shows between 74.9% to 82.3% of type I diabetics have retinopathy [72]. Therefore, the baseline model for predicting retinopathy among type I diabetics will have an accuracy of about 80%. Moreover, almost all type I diabetic patients who have had the disease for 20 years develop this vision complication. Thus, despite the first study's high accuracy (98%) in predicting retinopathy among type I diabetic patients, it does not address the more important problem of detecting retinopathy in type II diabetics. This limitation is addressed in the second study, but with an overall accuracy of 85%, it leaves room for improvement. Therefore, another promise of the current effort is to develop a model that addresses the limitations of the extant literature, while improving upon their results.

Ensemble models for supervised learning were first introduced by Tukey [73] and have since been studied by many researchers. At a high level, there are two categories of ensemble models: homogeneous and heterogeneous ensembles. Homogeneous ensembles combine multiple variations of a single classifier technique. Ensembles in this category use such algorithms as Bagging and AdaBoost to manipulate the training dataset and to develop multiple training datasets. These training datasets will be used by a data mining technique such as decision tree, and at the end, a voting or averaging mechanism will be used to make the final prediction using the outputs of single classifiers [19]. One of the most famous ensemble models in this category is random forest. Heterogeneous ensembles, on the other hand, combine various single classifiers (that are built using different data mining techniques) on the same training dataset. Simple average, weighted average, and voting based ensembles belong to this category [74]. A comprehensive review of ensemble techniques can be found in Rokach [75]. In this study, we developed a novel heterogeneous ensemble approach that will be explained in Chapter 4.

2.3. Imbalanced Data Learning Algorithms

Studying the imbalanced data classification has received a considerable amount of attention in recent years. He and Garcia [76] classified the different approaches of analyzing imbalanced data into four main classes,

- Sampling methods
- Cost-sensitive methods
- Kernel-based methods and active learning methods
- Other methods such as, one-class learning, novelty detection, etc.

Sampling methods: The aim of the sampling methods is to reach some degree of balanced distribution in the dataset. These methods can be categorized into two major streams, those that under-sample the majority class and those that over-sample the minority class. In under-sampling methods, some parts of the majority examples are removed. As a result, the distribution of the classes will be more balanced. The simplest method in this category is the random under-sampling. There is not any specific mechanism for under-sampling in this approach and it functions merely randomly. Other under-sampling approaches such as BalancedCascade and EasyEnsemble presented by Liu, et al. [77] are called informed under-sampling. In EasyEnsemble, several samples of the majority class data are taken and combined with minority class data. Multiple models are built based on these datasets, and at the end an ensemble model makes the final decision. The main criticism of the under-sampling methods is that by removing some parts of the data, potential important information in the data can be lost.

Over-sampling on the other hand, is to re-sample or generate extra examples of the minority class. The most basic over-sampling method is random over-sampling in which minority examples in the data are randomly duplicated. The main downside of random over-sampling is over-fitting. Another major approach in over-sampling is synthetic data generation. SMOTE

(Synthetic Minority Over-Sampling Technique) is one of the most well-known methods in synthetic data generation. In this method, synthetic data points are generated on the line connecting the minority samples to their *k* nearest minority class neighbors [78]. The major drawback in SMOTE is that it may lead to over-generalization.

There are extensions to the SMOTE that tried to improve the performance of this technique. Han, et al. [79] proposed a synthetic over-sampling method named Borderline-SMOTE. In this method, only a subset of minority data points is over-sampled by SMOTE technique. Those minority data points are located near the border of two classes. Borderline minority data points are identified as minority examples that most of their nearest neighbors belong to the majority class. On the other hand, Bunkhumpornpat, et al. [80] introduced a method named Safe-Level SMOTE. This method calculates a parameter called safe-level. The greater that a safe-level is for a minority example shows that example is farther away from the borderline. After identifying the minority examples in safe regions, those data points will be over-sampled using SMOTE. Cieslak, et al. [81] introduced the cluster SMOTE method. This method first clusters the minority examples, and then over-samples data points within each cluster by applying SMOTE.

Barua, et al. [82] proposed a majority weighted minority oversampling technique that first identifies hard to learn minority examples by considering their distance from the majority neighbors, and then it over-samples those examples using a clustering approach. There are other studies in the area of synthetic data generation ([79], [83], and [84]). Generally speaking, synthetic oversampling significantly improves the classification accuracy, especially for the minority class. Another advantage is that by generating the synthetic minority data (not simply replicating existing minority data), the minority region is generalized and overfitting can be avoided [85]. For a more comprehensive review of the sampling methods, we refer readers to He and Garcia [76].

Cost-sensitive methods: Unlike sampling methods that alter the distribution of the data through either generating synthetic minority data points or removing some portion of majority data points, the idea of cost-sensitive methods is based on the different misclassification costs for different classes in the dataset. Usually the cost of misclassifying the minority class is much higher than the majority class misclassification [86]. To perform cost-sensitive methods, a matrix, called cost matrix is required. This matrix shows the misclassification cost for different classes in the dataset [87]. The main concern about cost-sensitive methods is that in most of the situations the exact misclassification cost related to various classes is unknown [88].

There are three major categories in cost sensitive approaches [76]. The first category includes techniques that assign various weights to the examples in the dataspace. Methods in this category are motivated by the AdaBoost algorithm [89]. AdaBoost is a meta-algorithm that begins with the original dataset and trains a model on this dataset. Incorrectly classified examples are identified, and in the next iteration more weight (higher error cost) will be assigned to them. In this way, more focus will be on the examples that are misclassified. This process repeats and the classifier performance improves. The second group encompasses approaches are those that use ensemble schemes integrated with cost-sensitive approaches. Many of the research studies in these two categories have combined various weighting and adaptive boosting techniques. For instance Sun, et al. [90] and Fan, et al. [91] proposed algorithms for updating the weights in AdaBoost in imbalanced data learning. Lee, et al. [92] used SVM to adjust the weights of the examples in AdaBoost to learn from imbalanced data. In the third category, cost-sensitive methods incorporate the misclassification costs directly into the classifiers. Cost-sensitive decision tree [93], cost-sensitive neural networks [94], and cost-sensitive SVM [95] are in this category.

Kernel-based methods: Kernel-based methods are mostly integrated with SVM. Many researchers have studied imbalanced data learning through support vector machine. Wu and Chang [96] developed a boundary-alignment algorithm, which makes a change in the kernel function to move

the boundary toward the negative instances. Akbani, et al. [97] proposed an algorithm by integrating the different error cost method [95] and the SMOTE over-sampling method, however they performed the SMOTE over-sampling independent from the SVM model. Wang and Japkowicz [98] applied boosting and asymmetric error cost for minority and majority classes. Mathew, et al. [99] proposed a kernel-based SMOTE for SVM. In their approach, the over-sampling through the SMOTE technique happens in kernel feature space. Yu, et al. [100] developed the SVM-OTHR algorithm. In this algorithm, they adjusted the decision threshold by moving the decision hyperplane toward the majority class data.

Tang and Zhang [101] proposed a granular SVM with repetitive under-sampling. They utilized SVM for under-sampling in a way that they repeatedly developed SVM models and each time discarded the negative (majority class) support vectors from the data. Even though they performed the under-sampling integrated with the SVM, the problem of losing potential important information by under-sampling still exists. As Akbani, et al. [97] showed in their paper, under-sampling the majority class may decrease the total error, but it usually deteriorates the performance of the SVM on the test data, because it fails to approximate the orientation of the ideal hyperplane. Batuwita and Palade [102] suggested an over-sampling method in which they selected the majority examples near the boundary as the informative negative data points, and then they randomly over-sampled the minority examples to have relatively balanced data. This work can be critiqued in two ways. First, they focused on the informative majority examples, while the primary interest in imbalanced datasets is on the minority examples, therefore the focus on the informative majority examples may lead to even more bias toward the majority class. Second, they simply applied random over-sampling that is not as powerful as synthetic data generation methods and may lead to over-fitting. The two former studies did not compare their model's performance with other exiting methods; therefore, it is not easy to comment on generalizability and efficiency of their model. [103] proposed a preprocessing approach using

SVM for imbalanced data. In their approach, they first trained SVM on the original data, and then replaced the actual target variable value by the SVM predicted value. They claimed that SVM will classify a portion of the majority examples as minority, and therefore the processed data will have a more balanced distribution. Their claim is questionable, because in imbalanced data learning most of the time there is poor accuracy on minority class and good accuracy on majority. This means that most of the minority examples are misclassified as majority not the other way around. They tested their approach only on one dataset; therefore, their results could be because of the characteristics of that special dataset.

CHAPTER III

DEVELOPMENT OF A NEW METRIC TO IDENTIFY RARE PATTERNS IN ASSOCIATION ANALYSIS: THE CASE OF ANALYZING DIABETES COMPLICATIONS

In this chapter, we present the methodology of association analysis and our developed metric, adjusted_support. Following that we provide the results of applying adjusted_support and association analysis to complications of diabetes. The number of co-existing complications among diabetic patients could be a meaningful index to evaluate their health status. In this study, we defined comorbidity index as the mean number of co-exiting complications in a diabetic patient. Besides association analysis, we also performed a comorbidity analysis on diabetic patients by calculating their comorbidity index and compared the comorbidity index of patients in different demographic groups. This analysis will provide insight on the comorbidity status of diabetics at a more granular level and can lead to better decision making by healthcare administrative professionals and clinicians. In addition, we studied the prevalence of diabetes complications among various demographic groups.

3.1. Methodology

In this section, we start with briefly describing association analysis, define its common parameters and metrics, and explain the algorithmic extent of our proposed rare item/patterns identification metric.

Let $I = \{i_1, i_2, \dots, i_n\}$ be the set of all items (in our study, items are various diabetes complications), and $R = \{r_1, r_2, \dots, r_N\}$ be the set of all records in our data (each record corresponds to each patient and contains the patient's complications). Let X be a subset of I, i.e., a subset of items (diabetes complications), then X is called an *itemset* ($X = \{i_1, i_2, \dots, i_k\}$). An itemset with k items is called *k-itemset*. Every record, r_i, includes a subset of items (complications) in I, thus each r_i is an itemset ($r_i \subseteq I$). Suppose itemset X contains the following complications: retinopathy, nephropathy, and neuropathy. The *support count* for X is the number of records that include items in X. The support count for X is denoted by $support\_count(X)$, and is calculated as in Equation 3.1.

$$support\_count(X) = |\{r_i | X \subseteq r_i, r_i \in R\}| \tag{3.1}$$

Where |.|, denotes the cardinality of a set.

An association rule is defined as $X \longrightarrow Y$, where X and Y are itemsets and their intersection is $\emptyset$. When the number of items (in our case, the number of complications) increases, the number of generated rules grows exponentially. As a result, generated association rules should be evaluated and useful rules identified. Two sets of assessment measures, *objective* and *subjective*, can be applied to selecting beneficial and interesting rules [104]. We describe these measures in the following sections.

Objective Measures

In this category, there are two major classical assessment metrics to evaluate an association rule's strength: *support* and *confidence*. The support of an association rule is the occurrence probability of the rule among all records, or, in our study, the proportion of patients for which the rule is true. It is calculated as in Equation 3.2.

$$Support\ (X \longrightarrow Y) = \frac{support\_count(X \cup Y)}{N} \tag{3.2}$$

In traditional association analysis, a minimum support is specified before generating rules. When rare (infrequent) items exist in the data, this approach is inefficient. A high min-support will lead to elimination of all rules containing any rare item, and a low min-support will lead to over-generating rules that may not be interesting enough. Therefore, a new assessment metric is required that can simultaneously solve the problem of eliminating rare items and over-generating uninteresting rules. In this study, we introduce a new assessment metric termed adjusted_support. It is calculated as shown in Equation 3.3.

$$Adjusted\_Support\ (X \longrightarrow Y) = \frac{support\_count(X \cup Y)}{Min\ \{support\_count(X),\ \ support\_count(Y)\}} \tag{3.3}$$

The calculation of adjusted_support begins by comparing the number of records with items in X (left hand side of rule) to the number of records with items in Y (right hand side of rule) and selecting the smallest group. Using the selected group, the proportion of records with items in both X and Y (patients that are diagnosed with the complications in both X and Y) is calculated. To calculate the adjusted_support, instead of considering the entirety of patients' records, we focus on a subset of records in order to capture the rare association rules. When a rare item (complication) exists in a rule, the denominator of the adjusted_support will be a small number, and, as a result, adjusted_support will be large enough to satisfy the minimum adjusted_support condition. For the frequent rules, the adjusted_support still will be large enough because these rules will have a large numerator. Therefore, by considering a single pre-specified minimum

adjusted_support, both rare and frequent association rules can be discovered and extracted from the data. Here we provide a simple example to describe the adjusted_support more clearly. Suppose we have a hypothetical dataset with the following characteristics (Table 3.1):

Table 3.1-*Hypothetical Dataset H (Total number of the records=5000)*

Itemset	Support count of the itemset
Retinopathy ($\boldsymbol{X_1}$)	1800
Nephropathy ($\boldsymbol{X_2}$)	2000
Gastroparesis ($\boldsymbol{X_3}$)	200
Retinopathy, Nephropathy	500
Retinopathy, Gastroparesis	80
Nephropathy, Gastroparesis	5

Consider the following association rules and their calculated support and adjusted_support in Table 3.2:

Table 3.2- *Generated rules from dataset H*

No.	Rule	Support	Adjusted_support
1	$X_1 \longrightarrow X_2$	$500/5000 = 0.10$	$500/Min(1800, 2000) = 0.28$
2	$X_1 \longrightarrow X_3$	$80/5000 = 0.016$	$80/Min(1800,200) = 0.40$
3	$X_2 \longrightarrow X_3$	$5/5000 = 0.001$	$5/Min(2000,200) = 0.025$

Suppose both minimum support and adjusted_support are specified as 5%. In this case, rule number 1 passes both criteria and rule number 3 passes neither of them. However, without considering adjusted_support, rule number 2 will not be selected, but obviously there is a strong association between X_1 and X_3. By considering adjusted_support as the assessment metric, this rule will be selected as a strong and interesting rule. Therefore, we can see that adjusted_support is effective in all cases, i.e., capturing both strong rare and frequent rules (rules 2 and 1) and removing weak rules (rule 3).

The next measure that we used is confidence. The confidence of a rule measures how often records include items in Y, given they include items in X (in our case, how often patients have the complications in Y (right hand side of rule) when they are diagnosed with complications in X (left hand side of rule)). For instance, a confidence of 60% means that 60% of the patients with

complications in X also have complications in Y. Thus, a rule with higher confidence is more dependable. The calculation of confidence is shown in Equation 3.4.

$$Confidence\ (X \longrightarrow Y) = \frac{support\_count(X \cup Y)}{support\_count(X)} \tag{3.4}$$

As can be seen in Equations 3.2 and 3.3, support and adjusted_support are commutative operations, i.e., $Support\ (X \longrightarrow Y) = Support\ (Y \longrightarrow X)$ and $Adjusted\_Support\ (X \longrightarrow Y) = Adjusted\_Support\ (Y \longrightarrow X)$. However, confidence is not a commutative operation, i.e., if the direction of a rule changes, its confidence will also change.

Only considering support, adjusted_support, and confidence for evaluating an association rule might be misleading. *Lift* is a metric that considers both confidence and support concepts at the same time. *Lift*, also called *improvement*, is calculated as in Equation 3.5.

$$Lift\ (X \longrightarrow Y) = \frac{Confidence\ (X \longrightarrow Y)}{Support(Y)} \tag{3.5}$$

Lift measures the usefulness of a rule compared to a random guess. Suppose we have a rule such as $diabetic\ neuropathy \longrightarrow diabetic\ retinopathy$. If the lift of this rule is 3, it means that the chance of a diabetic neuropathy patient having diabetic retinopathy is 3 times higher than a random diabetic patient. Rules with lift higher than 1 are considered as useful rules. Lift is also a commutative operation; we can mathematically show this property of the lift as follows,

$$Lift\ (X \longrightarrow Y) = \frac{Confidence\ (X \longrightarrow Y)}{Support(Y)} = \frac{\frac{support\_count(X \cup Y)}{support\_count(X)}}{\frac{support\_count(Y)}{N}} = \frac{support\_count(X \cup Y)N}{support\_count(X)support\_count(Y)} \tag{3.6}$$

$$Lift\ (Y \longrightarrow X) = \frac{Confidence\ (Y \longrightarrow X)}{Support(X)} = \frac{\frac{support\_count(Y \cup X)}{support\_count(Y)}}{\frac{support\_count(X)}{N}} = \frac{support\_count(Y \cup X)N}{support\_count(X)support\_count(Y)} \tag{3.7}$$

from (3.6), (3.7), and $support\_count(X \cup Y) = support\_count(Y \cup X)$

$$\Rightarrow Lift\ (X \longrightarrow Y) = Lift\ (Y \longrightarrow X) \quad (3.8)$$

Subjective Measures

Subjective measures are based on subjective reasoning. We call an association rule subjectively interesting, if it unveils interesting and unexpected patterns in the data. For instance, discovering the strong association between diabetic retinopathy and ophthalmic complications of diabetes is expected because they are both eye-related diseases. An association rule such as "$diabetic\ retinopathy \longrightarrow ophthalmic\ complication$" may have high support, adjusted_support, confidence, and lift, but this rule does not uncover any interesting and unexpected pattern in the data and therefore, does not help physicians in diagnosis and treatment. Evaluating an association rule via subjective measures requires domain knowledge, and a data analyst cannot assess an association rule subjectively by herself without consulting domain experts.

It is important to note that association rules should not be interpreted as cause and effect. These rules only illustrate associations, not causality [47]. For example, in this study associations represent the co-existence of different complications. Therefore, a strong association between two complications in a rule does not in and of itself indicate any causality, but it can point to the need for future research on the causal nature of related complications.

Data Preparation

Data for this study came from the Cerner Health Facts data warehouse, one of the largest commercial databases of electronic medical records (EMR) in the U.S. For research purposes, data are de-identified in accordance with HIPAA requirements and are linked through unique identifiers. Each admission has information recorded for patient demographics, admission source, diagnoses, procedures, drugs dispensed, laboratory test results, and billing and primary payer. At

the time of this study, Health Facts contained data for more than 58 million unique patients, about 84 million patient visits, over 320 million prescriptions, and about 2.4 billion clinical lab results that were collected since 2000 from 480 affiliated hospitals and hospital systems across the nation.

For this study, we extracted admission and diagnosis data for diabetes and its complications for patient visits between September 1999 and January 2016. The first dataset included 2,317,259 unique diabetic patients with various complications. Among them, 624,810 were only diagnosed with diabetes, and 1,086,005 had only essential hypertension co-existing with diabetes (here we need to note that all of the patients in our study were diagnosed with diabetes or one of its complications, and among them 1,502,946 patients had hypertension that could be co-existing with other diabetes complications). Because the number of these two conditions were extremely large compared to the other diabetes complications, and also more than 70% of diabetics are known to have hypertension, patients diagnosed with only diabetes and/or hypertension i.e., the diabetic patients without other diabetes complications were excluded from the association analysis. Diabetes and related complications were defined by International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and by the ICD, Tenth Revision (ICD-10-CM). In the data, among various diagnosis for diabetic patients, there were two diagnoses as “Other specified manifestations” and “Unspecified complication”. Since these two diagnosis (complications) did not have any specific and clear meaning, we removed the records for them for the data. The final dataset included 492,025 unique patients with diabetes and associated complications. The number of unique patient/complication in the dataset was 753,733. The size of our dataset compared to the existing literature in the field is much larger, thus the output of our analysis is expected to be more dependable. Table 3.3 illustrates the distribution of diabetes complications in our data. The complications used in this study were selected based on the ICD 9 and ICD 10 classification for diabetes complications. Although diabetic patients may

have other complications, these conditions were outside the scope of this study and therefore not included in the analysis.

Table 3.3-*Diabetes complications count and percentage in the data*

Complication	Frequency Count	Frequency Percentage
Neurological manifestations (Neuro)	202634	26.88
Renal manifestations (Renal)	122882	16.30
Stroke	79985	10.61
Ophthalmic manifestations (Ophthal)	74198	9.84
Retinopathy (Retino)	61046	8.10
Peripheral circulatory disorder (PCD)	53804	7.14
Ketoacidosis (Keto)	49661	6.59
Heart Disease (Heart)	43007	5.71
Gastroparesis (Gastro)	30032	3.98
Diabetes with hyperglycemia (Hyper-g)	14886	1.97
Hyperosmolarity (Hyperos)	14530	1.93
Other coma (Coma)	3213	0.43
Skin complications (Skin)	2154	0.29
Diabetes with hypoglycemia (Hypo-g)	1427	0.19
Diabetic Arthropathy (Arthro)	264	0.04
Oral complications (Oral)	10	0.00

ICD-10 diagnosis codes, which went into effect October 1, 2015, are more granular than ICD 9 codes. Because ICD-10 has been implemented for a relatively short time, there are limited numbers of records for the more granulated diagnosis. For instance, in ICD 9 the code 250.8 is described as “diabetes with other specified manifestations”; while in ICD 10, under the E10.6 (or E11.6) which is for diabetes with other specified manifestations, there are diabetic arthropathy, skin complications, oral complications, diabetes with hypoglycemia, and diabetes with hyperglycemia. Thus, as can be seen in Table 3.3, the number of patients diagnosed with diabetic arthropathy, skin complications, oral complications, diabetes with hypoglycemia, and diabetes with hyperglycemia are too small.

To prepare the data, we needed to perform several steps such as merging tables, creating new variables, and changing the structure of the data tables. In the final dataset, we needed a patient

identifier and a complication in each row. Therefore, patients with multiple co-existing complications had multiple records in the final dataset. Figure 3.1 shows the structure of the final dataset used for the association analysis.

Patient ID	Complication
XXX-XXXX	Retinopathy
XXX-XXXX	Renal manifestations
XXX-XXXX	Neurological manifestations
XXX-XXXX	Ophthalmic manifestations
XXX-XXXX	Ketoacidosis
YYY-YYYY	Retinopathy
YYY-YYYY	Neurological manifestations
YYY-YYYY	Stroke
ZZZ-ZZZZ	Gastroparesis
ZZZ-ZZZZ	Renal manifestations

Figure 3.1- *Final dataset structure*

3.2. Results

In this section, we first present the results of the comorbidity index analysis in different demographic groups of patients. Following that, we compare the prevalence of major diabetes complications in different demographic groups. Finally, we represent the results of the association analysis among diabetes complications.

Comorbidity Index Analysis

In this study, we calculated what we have termed comorbidity index, which is the mean number of complications. The overall index value for the study population was 1.53. Because of the large proportion of diabetic patients with hypertension, we excluded it from our analysis, therefore, the inclusion of hypertension and diabetes itself increases the index value by 2 points to 3.53. Table 3.4 shows index values and descriptive statistics for the racial/ethnic groups examined in this study.

Table 3.4-*Comorbidity index value by race/ethnicity*

Race/Ethnicity	Num. of Observations	Percentage	Mean	Maximum	Lower 95% CI	Upper 95% CI
Whole Population	**492025**	**100%**	**1.53**	**11**	**1.529**	**1.535**
Biracial	557	0.11%	1.76	9	1.667	1.859
African American	101582	20.65%	1.66	10	1.650	1.663
Hispanic	7434	1.51%	1.62	8	1.598	1.646
Native American	5394	1.10%	1.59	8	1.567	1.621
Caucasian	302801	61.54%	1.51	11	1.502	1.509
Asian	7145	1.45%	1.50	7	1.482	1.525
Other	14207	2.89%	1.47	9	1.451	1.481
Middle Eastern Indian	103	0.02%	1.45	6	1.281	1.612
Pacific Islander	443	0.09%	1.41	5	1.329	1.484
Asian/Pacific Islander	129	0.03%	1.32	3	1.211	1.424
Missing	52230	10.62%	1.45	8	1.439	1.454

As can be seen in Table 3.4 and Figure 3.2, biracial patients had highest index value with an average of 1.76 complications. African American, Hispanic, and Native American patients all had index values above the population average. Asian/Pacific Islander patients had the lowest index value at 1.32.

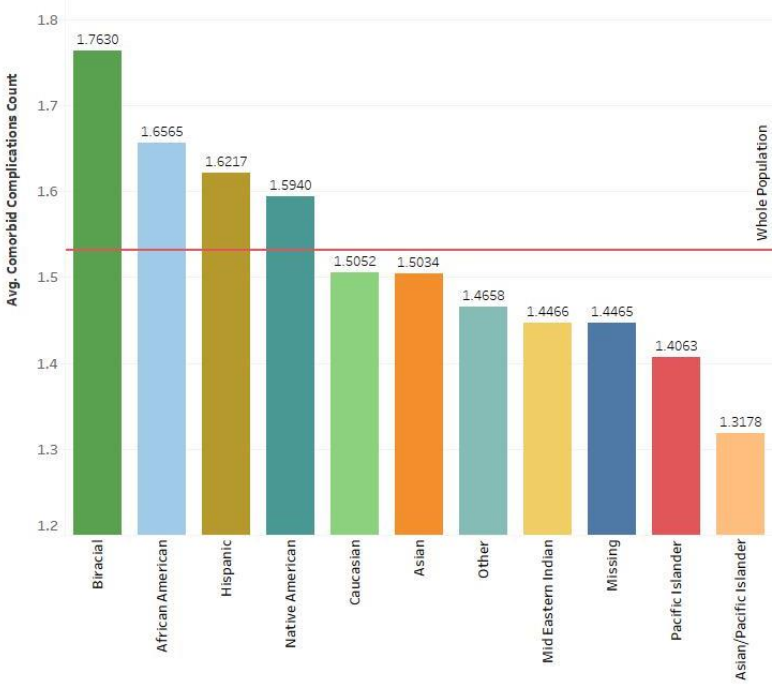


Figure 3.2- *Comorbidity index value by race/ethnicity*

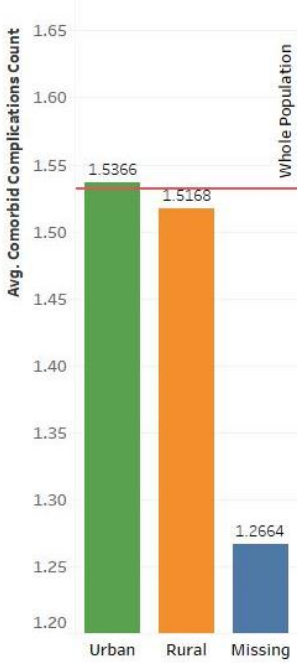
Table 3.5- *Comorbidity index value in rural vs urban diabetic patients*

Urban/Rural Status	Num. of Observations	Percentage	Mean	Maximum	Lower 95% CI	Upper 95% CI
Whole Population	**492025**	**100%**	**1.53**	**11**	**1.529**	**1.535**
Rural	99628	20.25%	1.52	11	1.511	1.523
Urban	391087	79.49%	1.54	10	1.534	1.540
Missing	1310	0.27%	1.27	5	1.241	1.292

Table 3.5 and Figure 3.3a depict the comorbidity index of rural and urban diabetics. Urban patients had slightly higher number of co-exiting complications and this difference was statistically meaningful at the level of 95%. Comorbidity index was not statistically different between males and females as is demonstrated in Table 3.6 and Figure 3.3b.

Table 3.6-*Comorbidity index value by gender*

Gender	Num. of Observations	Percentage	Mean	Maximum	Lower 95% CI	Upper 95% CI
Whole Population	**492025**	**100%**	**1.53**	**11**	**1.529**	**1.535**
Female	231924	47.14%	1.54	10	1.535	1.543
Male	233391	47.43%	1.53	11	1.528	1.536
Missing	26710	5.43%	1.47	8	1.460	1.481



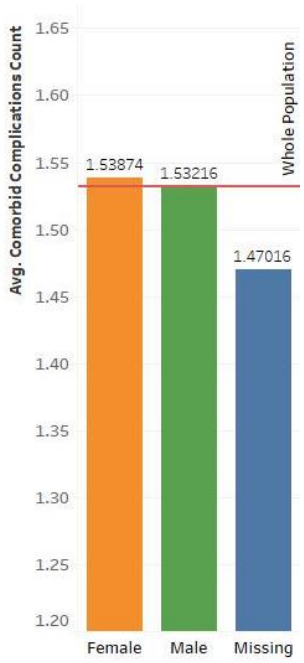


Figure 3.3a *Figure 3.3b*

Figure 3.3a- *Comorbidity index value in rural vs urban diabetic patients*

Figure 3.3b- *Comorbidity index value by gender*

Diabetes Complications Prevalence in Different Demographic Groups

In this section, we compare the prevalence of major diabetes complication by demographic groups. Figures 3.4a and 3.4b demonstrate the racial/ethnic prevalence of complications. Caucasians had the highest prevalence of neurological manifestations and heart disease, while renal manifestations were highest among Asians. Strokes were more prevalent among African Americans, and Native Americans suffered from the highest rates of ophthalmic manifestations, retinopathy, and hyperglycemia. Ketoacidosis and gastroparesis were most common in Hispanics, and hyperosmolarity was more prevalent in African American and Asian patients than other races.

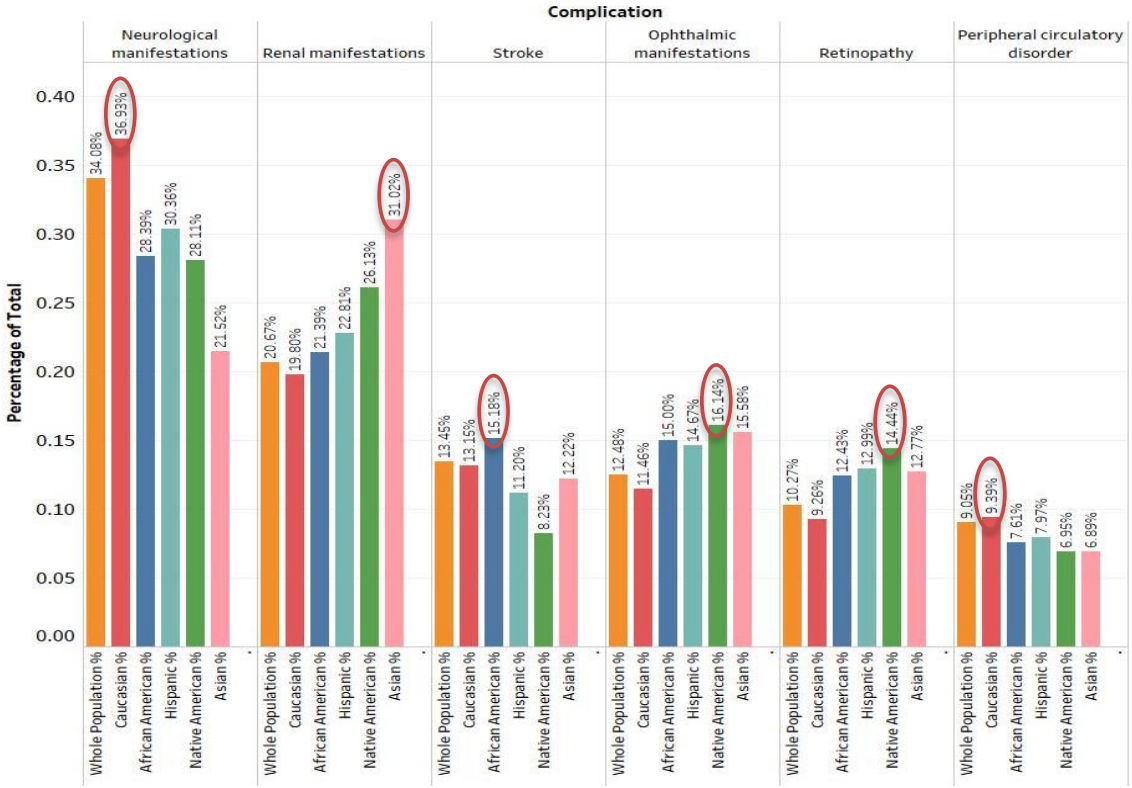


Figure 3.4a-*Diabetes complications' prevalence by race/ethnicity*

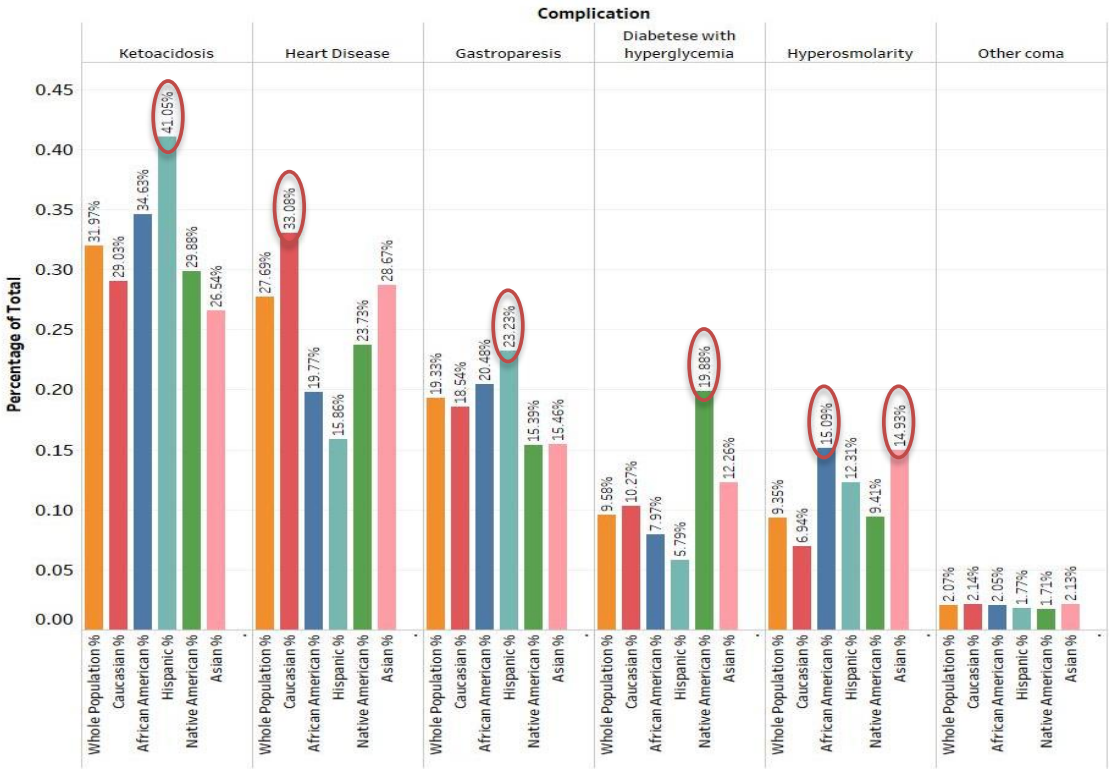


Figure 3.4b- *Diabetes complications' prevalence by race/ethnicity*

Figure 3.5 compares the prevalence of diabetes complications among rural versus urban patients. Based on our analysis, neurological manifestations, stroke, heart disease, and gastroparesis were more prevalent among patients in urban areas compared to rural areas. On the other hand, renal manifestations, ophthalmic manifestations, retinopathy, peripheral circularity disorder, and hyperglycemia had a higher prevalence in rural compared to urban areas. Rates of other complications were similar for both rural and urban areas.

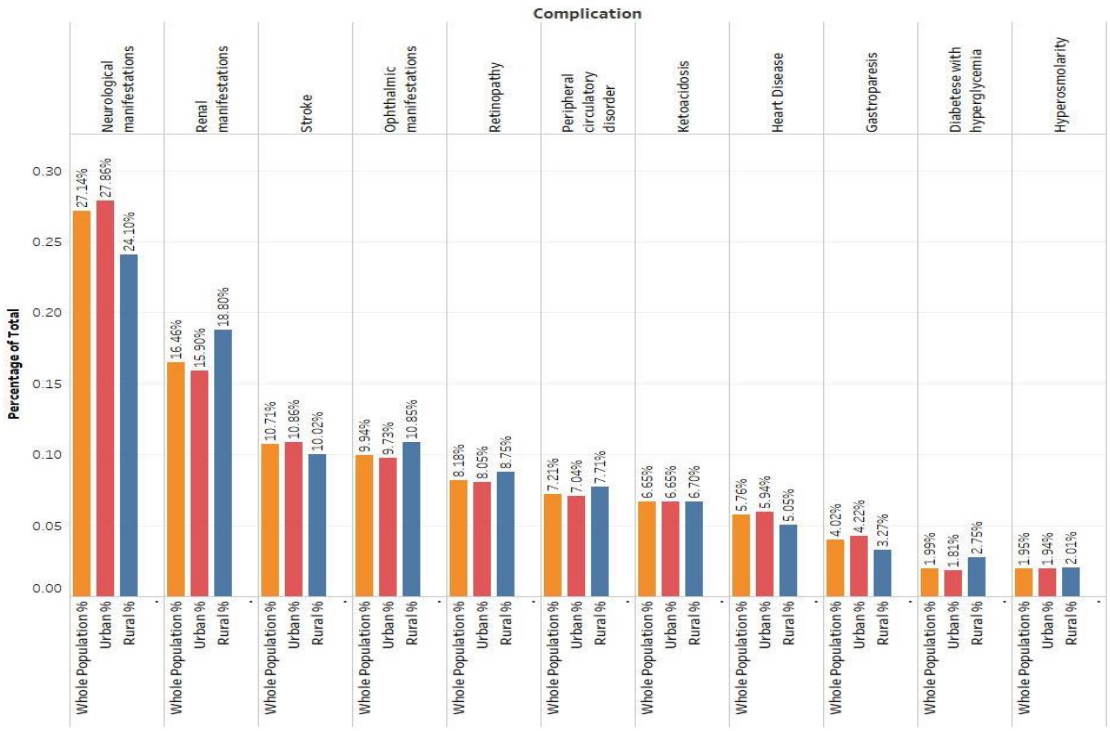


Figure 3.5-*Diabetes complications' prevalence: rural versus urban*

A comparison of complications between males and females is shown in Figure 3.6. Neurological manifestations, stroke, ophthalmic manifestations, retinopathy, and gastroparesis were more prevalent among women than men, while renal manifestations, peripheral circularity disorder, and heart disease were more common in men. Rates of other complications were similar for both groups.

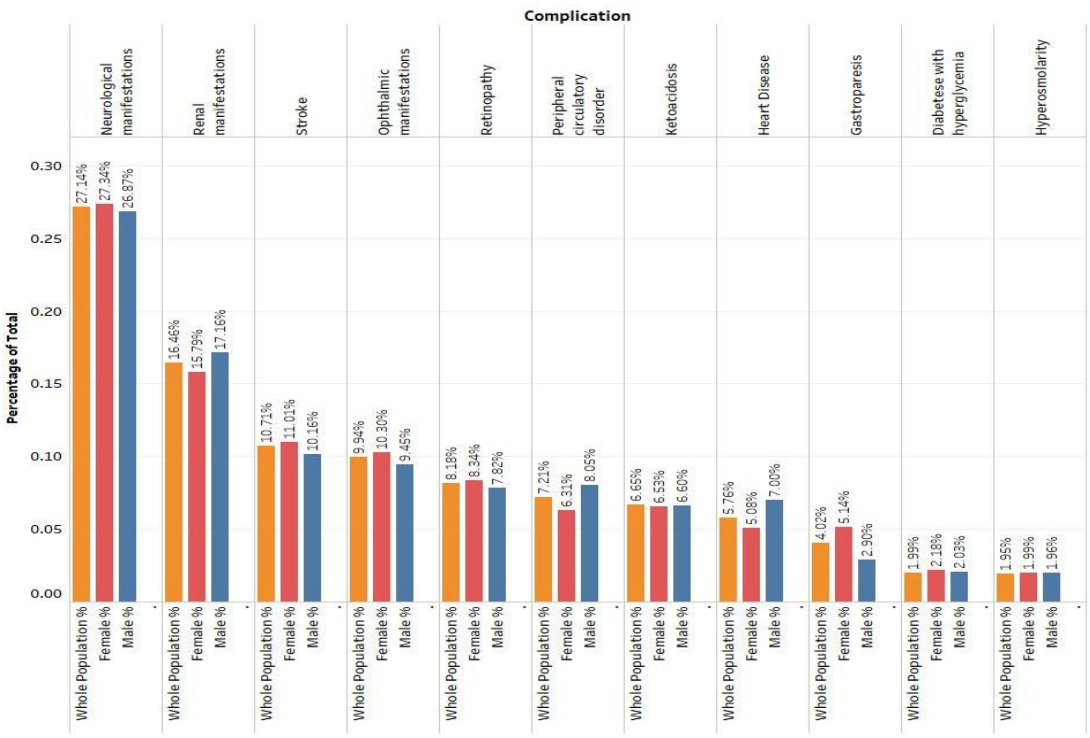


Figure 3.6-*Diabetes complications' prevalence: females versus males*

Association Analysis Results

The results of the association analysis among the major diabetes complications are provided in this section. We first performed association analysis on the entire study population followed by analyses for each demographic group. We first evaluated each set of generated rules based on objective measures (support, adjusted_support, and lift), and second, by consulting clinical experts, we assessed the rules subjectively. All rules shown in Tables 3.7, 3.8a, 3.8b, 3.8c, and 3.8d were strong two-item rules, which means they passed the thresholds for the objective measures. All of these rules had the following characteristics:

$$Adjusted\_Support \geq 10\%$$

$$Lift \geq 1$$

Since all of the presented rules met thresholds for adjusted_support and confidence, we compared rules using lift and sorted them in the following tables by lift values. As described in Section 3.1, any rule with lift greater than 1 is considered useful rule, which means that rule provides extra information that helps better decision making.

Table 3.7- *Two-item association rules- general population*

No.	Rule	Support (%)	Adjusted_Support (%)	Lift
1	Hypo_g ==> Hyper_g	0.06	20.95	6.93
2	Ophthal ==> Retino	11.38	91.71	6.08
3	Skin ==> Hyper_g	0.07	16.53	5.46
4	Hypo_g ==> Heart	0.07	24.95	2.86
5	Coma ==> Keto	0.16	23.90	2.37
6	Skin ==> PCD	0.11	24.14	2.21
7	Skin ==> Heart	0.07	16.62	1.90
8	Hyperos ==> Keto	0.52	17.63	1.75
9	Arthro ==> Neuro	0.04	70.83	1.72
10	Hyper_g ==> Heart	0.45	14.83	1.70
11	Neuro ==> Gastro	4.16	68.08	1.66
12	Renal ==> Retino	4.70	37.90	1.52
13	Gastro ==> Keto	0.88	14.37	1.43
14	Renal ==> Ophthal	5.33	35.37	1.42
15	Skin ==> Neuro	0.25	57.06	1.39
16	Gastro ==> Retino	1.04	16.98	1.37

Even though all of the rules in Table 3.7 were objectively strong, some were not subjectively interesting or meaningful. For example, rule “Hypo_g ==> Hyper_g” shows a strong association between “diabetes with hypoglycemia” and “diabetes with hyperglycemia” and values for adjusted_support, confidence, and lift exceeding threshold values. However, this rule is not very insightful as hospitalized diabetics may experience both high and low glycemic levels due to their medications. Another rule that appeared to objectively meaningful but was not subjectively interesting was “Ophthal ==> Retino.” Both items in the rule (ophthalmic manifestations and retinopathy) are eye-related complications; therefore, discovering a strong association between two closely related complications would be expected and, subsequently, of little use for clinicians.

As the results in Table 3.7 show, diabetic patients with skin complications had a high potential of also being diagnosed with hyperglycemia, peripheral circulatory disorder, heart disease, and neurological manifestations. Hypoglycemia was strongly associated with heart disease as demonstrated by rule 4, which shows diabetics with hyperglycemia were 2.86 times more likely to develop heart disease than the general population of diabetic patients. Rule 8 shows that diabetics with hyperosmolarity were 1.75 times more likely to have ketoacidosis. Based on rules 9 and 11, neurological manifestations were strongly associated with diabetic arthropathy and gastroparesis. Diabetics with renal manifestations (rules 12 and 14) were 1.52 and 1.42 times more likely to have retinopathy and ophthalmic manifestations compared to random diabetic patients.

According to the results presented in Table 3.8a, the association between neurological manifestations and gastroparesis was much stronger among African American diabetics compared to general diabetics' population. Besides that, African American diabetics diagnosed with other coma, were about two times more likely to have hyperosmolarity compared to a random African American diabetic, but we did not discover this association rule in general diabetics' population. In addition, there was a strong association between hypoglycemia and gastroparesis among African Americans while there was not such association among general population of diabetics. In Hispanic diabetic population, we could find a couple of differences compared to general population (Table 3.8b). For instance, the association between neurological manifestation and gastroparesis was stronger (lift of 1.90 versus 1.66). However, the association of other coma and ketoacidosis among Hispanics was weaker compared to general population, lift for the rule "Coma ==> Keto" was 1.45, while lift for the same rule in general population was 2.37.

We could discover some association rules among Native American diabetics that we did not observe in general population (Table 3.8c). The association rule "Gastro ==> PCD" was one of them that shows the association between gastroparesis and peripheral circularity disorder. The

other one was the association between neurological manifestations and peripheral circularity disorder (rule number 6 in Table 3.8c). Besides these two rules that did not exist in general population, there were other rules with significantly different degree of strength. For instance, rules "Skin ==> PCD", "Neuro ==> Gastro", and "Skin ==> Neuro" were stronger among Native Americans, but rule "Skin ==> Hyper_g" was weaker among them compared to the general population.

Two-item association rules in race/ethnicity groups that were different from general population

Table 3.8a- *African Americans*

No.	Rule	Support (%)	Adjusted_Support (%)	Lift
1	Neuro ==> Gastro	5.55	73.53	2.03
2	Coma ==> Hyperos	0.08	11.02	1.99
3	Hypo_g ==> Gastro	0.05	14.09	1.87

Table 3.8b- *Hispanics*

No.	Rule	Support (%)	Adjusted_Support (%)	Lift
1	Neuro ==> Gastro	6.42	76.82	1.90
2	Coma ==> Keto	0.14	20.34	1.45

Table 3.8c- *Native Americans*

No.	Rule	Support (%)	Adjusted_Support (%)	Lift
1	Skin ==> Hyper_g	0.19	27.45	4.12
2	Skin ==> PCD	0.21	29.41	3.67
3	Neuro ==> Gastro	3.59	74.29	2.15
4	Skin ==> Neuro	0.44	62.75	1.81
5	Gastro ==> PCD	0.55	11.43	1.43
6	Neuro ==> PCD	3.84	47.85	1.38

Table 3.8d- *Asians*

No.	Rule	Support (%)	Adjusted_Support (%)	Lift
1	Arthro ==> Hyperos	0.01	50.00	13.92
2	Hypo_g ==> Heart	0.09	41.18	5.63
3	Skin ==> Heart	0.05	28.57	3.91
4	Skin ==> PCD	0.05	28.57	3.29
5	Arthro ==> Retino	0.01	50.00	3.09
6	Arthro ==> Renal	0.03	100.00	2.61
7	Gastro ==> Neuro	2.47	65.98	2.44
8	Skin ==> Neuro	0.10	57.14	2.11
9	Skin ==> Ophthal	0.06	35.71	1.80
10	Hypo_g ==> PCD	0.03	11.76	1.35
11	Hypo_g ==> Neuro	0.08	35.29	1.30

Among Asian diabetics, we found several association rules that did not exist in general population (Table 3.8d). Diabetic arthropathy patients were extremely likely to have hyperosmolarity and that was about 14 times more likely compared to a random Asian diabetic. The same group of patients (Asians with diabetic arthropathy), were at a very high risk of having retinopathy and renal manifestations (rules number 6 and 7 in Table 3.8d). Among Asians, hypoglycemia was strongly associated with both peripheral circularity disorder and neurological manifestations, associations that we did not recognize in the general diabetics' population. Besides these new rules, there were multiple rules that were significantly stronger form their counterparts in general population. Rules number 2, 3, 4, 7, 8, and 9 in Table 3.8d were in that category.

Association analysis in rural and urban diabetic patients revealed some difference between these two groups. The strength of several association rules were significantly different comparing the rural and urban patients. The following rules were stronger among rural diabetic patients: "Hypo_g ==> Heart", "Arthro ==> Neuro", "Neuro ==> Gastro", "Skin ==> Neuro". On the other hand, these rules had greater lift among urban patients: "Skin ==> Hyper_g", "Coma ==> Keto", "Hyperos ==> Keto", "Hyper_g ==> Heart", "Renal ==> Retino", "Gastro ==> Keto, "Gastro ==> Retino". Other than these strength differences, there was an association between hypoglycemia and ketoacidosis among urban diabetics while we did not observe such an association among rural diabetic patients. Association patterns among female and male diabetics were different in some rules. The association rule "Hypo_g ==> Gastro" which was discovered among female diabetics, was not captured in male patients. Several rules had significantly different strength comparing female and male diabetics. For instance, rules "Skin ==> Heart" and "Skin ==> PCD" were stronger among females. And rules "Skin ==> Hyper_g", "Hyperos ==> Keto", "Arthro ==> Neuro", "Hyper_g ==> Heart", "Keto ==> Gastro", and "Gastro ==> Retino" had higher degree of strength among male diabetic patients.

Association rules can include more than two items. So far, what we showed in Tables 3.7, 3.8a, 3.8b, 3.8c, and 3.8d, all were two-item rules. But, in Table 3.9, we present three-item association rules among diabetes complications in general population of diabetics. As it is clear form a comparison between the results in Tables 3.7 and 3.9, the lift of the three-item rules was much greater than the lift of the two-item association rules. The reason is that more items in an association rules, conveys more information, and as a results we expect to have stronger rules. For instance, compare the rule number 3 in Table 3.7, to the rule number 5 in Table 3.9. These two rules had the following characteristics:

Skin ==> Hyper_g, lift= 5.46

Skin & PCD ==> Hyper-g, lift= 8.01

In the second association rule, we had more information and that was the knowledge about the existence of both PCD and skin complications, therefore the likelihood of having hyperglycemia was about 2.5 times higher with this extra information. Other rules in the Table 3.9 can be compared to their counterpart two-item rules in Table 3.7. Similar to previous rules in Tables 3.7, 3.8a, 3.8b, 3.8c, and 3.8d, three-item rules must be assessed with subjective measures as well as objective ones. For instance, rule number 3, because of the existence of hyperglycemia and hypoglycemia at two sides of the rule, and rule number 8, because of the existence of ophthalmic manifestations and retinopathy at two sides of the rule, failed to pass the subjective assessment. There were other rules that did not pass the subjective measure, thus are not shown in Table 3.9. The same analysis has been performed for higher number of items in association rules and also for different demographic groups, but because of the space limitation we did not provide the results in this manuscript, but they would be available upon request.

Based on what we observed in all of the tables representing the association rules, there were numerous association rules that would not be discovered if we did not used the adjusted_support.

The value of support in many of these rules was even less than 1%, but they had high enough adjusted_support to be selected, and the results showed those rules had high lift values that is indicator of their usefulness.

Table 3.9-*Three-item association rules- general population*

No.	Rule	Support (%)	Adjusted_Support (%)	Lift
1	PCD & Hyper-g ==> Skin	0.03	20.69	47.26
2	Skin ==> Neuro & Hyper-g	0.05	12.12	17.73
3	Neuro & Hypo-g ==> Hyper-g	0.03	31.21	10.32
4	Skin & Heart ==> Hyper-g	0.02	28.49	9.42
5	Skin & PCD ==> Hyper-g	0.03	24.23	8.01
6	Skin & Renal ==> Hyper-g	0.02	23.87	7.89
7	Skin ==> Neuro & Heart	0.05	11.37	7.36
8	Ophthal & Gastro ==> Retino	1.01	89.41	7.21
9	Skin & Neuro ==> Hyper-g	0.05	21.24	7.02
10	Hypo-g ==> Neuro & Heart	0.03	10.02	6.48
11	Gastro ==> Neuro & Keto	0.77	37.90	6.22
12	Coma & Gastro ==> Keto	0.03	57.20	5.68
13	Retino & Keto ==> Gastro	0.25	33.21	5.45
14	Ophthal & Keto ==> Gastro	0.27	31.47	5.16
15	Retino & Coma ==> Gastro	0.02	31.38	5.15
16	Hyperos & Gastro ==> Keto	0.11	50.61	5.02
17	Keto & Heart ==> Hyperos	0.02	14.48	4.91
18	Coma & Ophthal ==> Gastro	0.02	28.14	4.62
19	Coma & Neuro ==> Gastro	0.05	27.55	4.52
20	Renal & Keto ==> Gastro	0.31	27.49	4.51

3.3. Discussion and Conclusion

The application of data analytics in health care has led to discovering insightful and interesting information that can lead to advances in health care delivery [5]. The rapid advances in data science coupled with the growing amount of available data in all aspects of the health care industry make health care analytics even more efficient and beneficiary. In this study, we tried to add to the data mining literature by introducing the adjusted_support in rare item association analysis, and to the health care analytics literature by performing comorbidity and association analysis among major complications of diabetes.

Several research studies have shown the impact of comorbidity on the management of diabetes. Poor management of comorbidity may lead to ineffective control of the disease and subsequent increases in both mortality rates and treatment costs ([105], [106]). The comparison of comorbidity index values, showed that among racial/ethnic groups, Biracial, African Americans, Hispanics, and Native Americans had the highest number of comorbid complications. Moreover, Pacific Islanders, Middle Eastern Indians, Asians, and Caucasians had the lowest comorbidity index values, respectively. These results show a potential effective comorbidity management among Biracial, African Americans, Hispanics, and Native Americans and more attention should be paid to these races for a better disease control.

The next step in this study was taking a closer look at different diabetes complications prevalence among various demographic groups. Knowledge about the prevalence of complications among different groups of patients at more granular levels has several benefits. First, it could help policy makers to provide more effective high-level plans. For instance, if we know there is a higher rate of retinopathy among Native Americans compared to other races, it could be the indication of low level of ophthalmic care management among that race, thus necessary actions can be taken. Second, it would help the researchers to study relationships between genetic characteristics of people and different diseases. Third, it could help clinicians to provide targeted treatments and interventions for specific groups of patients. Based on the results of our study, neurological manifestations and heart disease were more prevalent among Caucasians, renal disease and hyperosmolarity were more prevalent among Asians, stroke and hyperosmolarity had the highest prevalence among African Americans, eye related diseases and hyperglycemia were more common among Native Americans, and Hispanics had the highest prevalence in ketoacidosis and gastroparesis compared to other races. By comparing diabetics in urban versus rural areas, we reached to these results: neurological manifestations, stroke, heart disease, and gastroparesis were more prevalent among urban patients; and renal manifestations, ophthalmic manifestations,

retinopathy, peripheral circularity disorder, and hyperglycemia had a higher prevalence among diabetics in rural areas. Between different genders, females had higher rates of diagnosis with neurological manifestations, stroke, ophthalmic manifestations, retinopathy, and gastroparesis compared to males; and males were diagnosed with renal manifestations, peripheral circularity disorder, and heart disease more often compared to females.

Our methodological contribution in this research topic was addressing the rare item problem. To address this problem, we proposed a new objective metric for association rules and called it adjusted_support. By considering adjusted_support instead of support that has been used in traditional association analysis, we could capture the rare association rules from the data without over generating the useless association rules. We performed association analysis both in general diabetics' population as well as various demographic groups for better understanding of the association patterns among complications in those demographic groups. The knowledge about the association among complications of diabetes can facilitate the diagnosing of different complication of diabetes, it also can be a hint to study the scientific reasons behind those associations, and last but not the least it could lead to better management of diabetes and its comorbid complications.

All of the generated rules in our analysis were assessed by both objective and subjective metrics. For objective assessment, we used our proposed metric, adjusted_support beside support and lift. In addition, for subjective assessment we consulted with our medical advisors. Based on our results, skin complication was strongly associated with hyperglycemia, Peripheral circulatory disorder, heart disease, and neurological manifestations. Hyperosmolarity co-existed with ketoacidosis very often. Neurological manifestations co-existed with diabetic arthropathy and gastroparesis very frequently. Diabetics with renal manifestations were highly potential of suffering from eye related disease such as retinopathy. Finally, gastroparesis was strongly

associated with both ketoacidosis and retinopathy. The results of the association analysis are provided in Section 3.2 in various demographic groups in more details.

Similar to any other research, we faced some limitations in this study. First, the scope of our research was limited to the complications of diabetes that are specified in ICD 9 and ICD 10. Therefore, other disease that patients may have been diagnosed with, were not considered in our study. Perhaps including those potential existing complications would lead to even more insightful findings. Another limitation was related to the nature of EHR data. Because these types of dataset are collected for reasons other than the purpose of this research, they may lack some degree of accuracy, for instance, some of the complications of a patient may not be recorded in her visit. However, the large amount of the data that was available in our study can compensate this limitation.

CHAPTER IV

A DATA ANALYTICS APPROACH TO BUILDING A CLINICAL DECISION SUPPORT SYSTEM FOR DIABETIC RETINOPATHY: DEVELOPING AND DEPLOYING A MODEL ENSEMBLE

In this chapter, we explain various steps of developing the CDSS for diabetic retinopathy. We also present our proposed ensemble approach, confidence margin and assess its performance in comparison with existing ensemble methods. We expect that the CDSS we develop in this effort will be able to detect diabetic retinopathy at its early stages with a high degree of accuracy. This CDSS, which relies exclusively on lab data, not only helps overcome one of the major barriers to the early diagnosis of diabetic retinopathy, but also provides a new standard of care that will improve quality and increase compliance in healthcare without raising costs.

4.1. Methodology

Data cleaning and preprocessing is a very important aspect in any comprehensive data analytics study. This is even more important in healthcare analytics, especially when real-world EMR data is involved—because the data are captured and stored in different clinical/hospital settings and

for reasons other than data analytics [107]. Hence, in this study, data preparation was taken very seriously.

Dara Preprocessing

The data used for this research was obtained from the Cerner Corporation's Health Facts data warehouse; a comprehensive, relational repository of real-world, de-identified, and HIPAA-compliant patient data. A simplified conceptual data diagram of the Cerner Health Facts data warehouse is presented in Figure 4.1.

Processing and analyzing large EMR datasets involves various challenges. Jagadish, et al. [108] classify these challenges into five categories: data acquisition; information extraction and cleaning; data integration, aggregation, and representation; modeling and analysis; and interpretation. Regarding the large number of variables in the data warehouse, we spent a significant amount of time to understand the purpose and relevance of each variable. An even more demanding step in preparing the dataset for final analysis was aggregrating the records at the patient level and integrating patients' comorbid conditions. Hence, information extraction and cleaning, together with data integration, aggregation, and representation constituted the majority of our data preprocessing efforts.

The nature of EMR data posed yet another difficulty to this study. Because EMR data is collected for purposes other than performing data analytics, it suffers from mutiple defficiencies. First, since EMR data is collected from several facilities around the country, it lacks integrity and consistency. For instance, different units or even naming might be used in different hospitals. Second, data missingness or incompleteness, which are endemic in EMR data, need to be addressed. And third, outliers and other data entry errors are prevalent in EMR data. We describe the approach we used in the data preparation step to address these challenges later in this section.

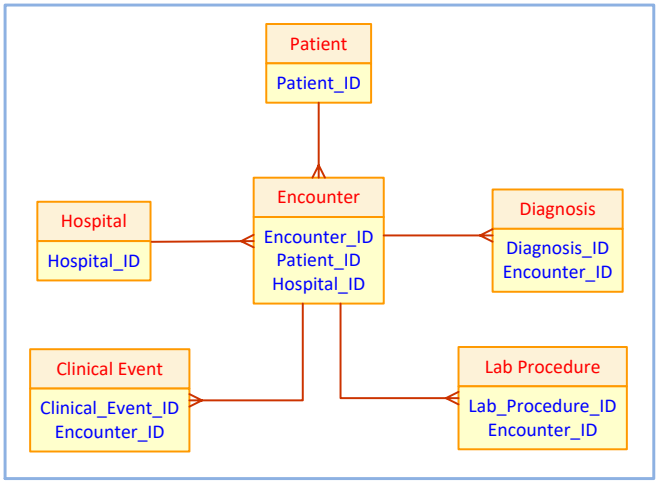


Figure 4.1- *A simplified conceptual data model for Cerner Health Facts data warehouse*

For the purpose of this research, we extracted data of more than 1.4 million unique diabetic patients from approximately 5.3 million visits. Since the number of variables collected from different data tables was rather large (300+), we needed to take many data selection, aggregation, and preparation steps.

First, data from all tables (e.g., encounter, patient, lab procedure, clinical event, etc.) for diabetes diagnosis and all associated complications, such as diabetic neuropathy, nephropathy, and retinopathy, were extracted. The table that included the lab procedure data was very important in this regard. The first dataset extracted from this table for diabetic patients included more than 800 different lab procedures. This primary dataset was very sparse, since not every patient had all of these lab results. We dropped those lab procedures that lacked sufficient results in the data. After taking several data cleaning steps, 88 lab procedures remained. Because EMR data are collected from hundreds of facilities across the United States, different names may be used for the same lab procedures. We consulted with clinical experts and merged identical lab procedures into one variable. As a result of this step, 58 lab procedures remained in our dataset.

The lab procedure table contained a column labeled “lab_procedure_name” that included all lab procedures for individual visits (encounters). We transposed this table so that each lab procedure

had its own column. This increased the number of columns in the lab procedure table from 35 to about 100. Since every patient at each visit (hospital stay) could have multiple results for the same lab procedure, we retained the last result as consultations with physicians and clinical experts suggested these values could be considered the stable condition for a patient. Moreover, because our focus was on developing a CDSS for the early detection of retinopathy, we selected each patient's first chronological visit to increase the validity and generalizability of our findings. In the next step, we used table keys (i.e., "Patient_ID," "Encounter_ID," and "Diagnosis_ID") to join data from multiple tables into a single table that included lab results, demographic data, and diagnosis data, with each record representing an individual diabetic patient. The resulting table included data from over 300,000 unique patients. Figure 4.2 depicts different data preparation steps in our study.

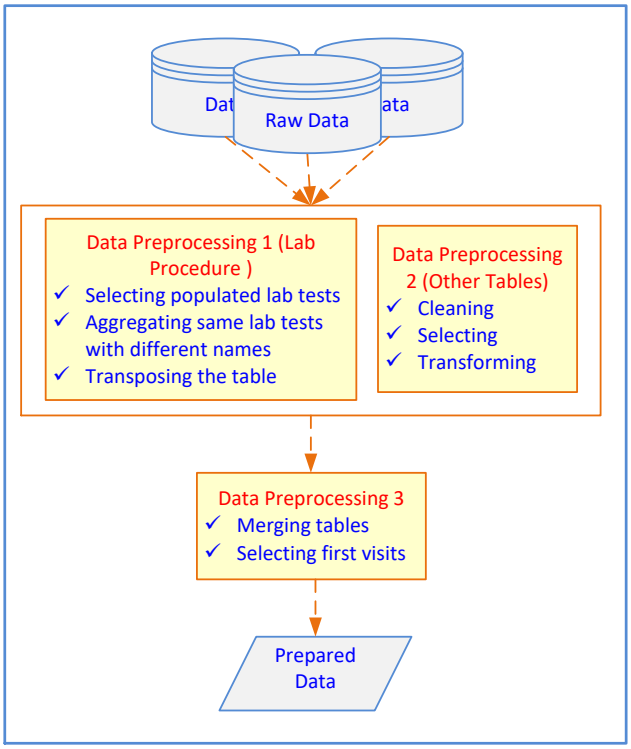


Figure 4.2- *Data preparation steps*

Modeling Procedure

To develop the predictive models, we employed logistic regression, decision tree, random forest, and artificial neural networks. Building each of these models was preceded by various data manipulation steps, including transforming variables to approximate a normal distribution, replacing and filtering extreme data points, and applying different imputation methods, and at the end, we compared the results. The modeling procedure is shown in Figure 4.3.

The target variable was a binary variable, where 0 denoted *no diabetic retinopathy diagnosis* and 1 denoted *diabetic retinopathy diagnosis*. Diabetic patients without a retinopathy diagnosis were included as the control group. The final dataset was largely imbalanced. In fact, we may face imbalanced data in several real world problems: fraud detection, oil-spill detection, and medical applications (Kubat, et al. [109], Rao, et al. [110], Chan, et al. [111]). The majority class in the dataset was diabetic patients without retinopathy (95%), and our class of interest, diabetics with retinopathy (5%), was the minority class. The main challenge in analyzing imbalanced datasets is that the performance of most standard machine learning techniques will be poor in terms of identifying the target variable [112]. Therefore, a balanced dataset is necessary to develop predictive models with high accuracy. Since there were a reasonable number of retinopathy patients in the minority class (about 15,000 patients), we created a balanced dataset by randomly under-sampling the majority class. The next step was to partition the data into training and validation datasets to objectively assess the different model types. In the following section, we provide a brief description of each of the modeling techniques used in this study.

Modeling Techniques

Logistic regression: Logistic regression is a classic statistical model. This method is capable of predicting and classifying categorical variables, but is mostly used for binary variables [113]. It is an extended version of linear regression, but instead of modeling a continuous value, binary

logistic regression models the log odds of the probability of an event, as opposed to its occurrence, as a linear function of the predictors.

Selection methods are often used to construct an optimal regression equation using a large number of predictors. Three statistical regression methods of variable selection are forward selection, backward elimination, and stepwise selection. The training in forward selection starts with an empty equation and adds predictors one at a time starting with the most significant predictor. Selection ends when all remaining predictors fail to meet the specified *F*-to-enter value. Backward elimination training starts with all predictors and removes, one at a time, the least significant predictors. Elimination ends when all remaining predictors fail to meet the specified *F*-to-remove value. The stepwise method is a variation of the above methods. It starts with an empty model, and after each step in which a predictor is added based on the *F*-to-enter value, it evaluates predictors in the model against the specified significance level. Those that fall below this level are removed. In this study, we applied the stepwise method. The binary logistic regression equation is shown in Equation 4.1 and Equation 4.2.

$$logit[P(x)] = \ln\left[\frac{P(x)}{1-P(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{4.1}$$

$$P(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} = \frac{1}{1+e^{-logit[p(x)]}} \tag{4.2}$$

$P(x)$ is the probability that the target variable belongs to a specific category (in our study, a patient has retinopathy) and β_i is the coefficient of the i^{th} predictor.

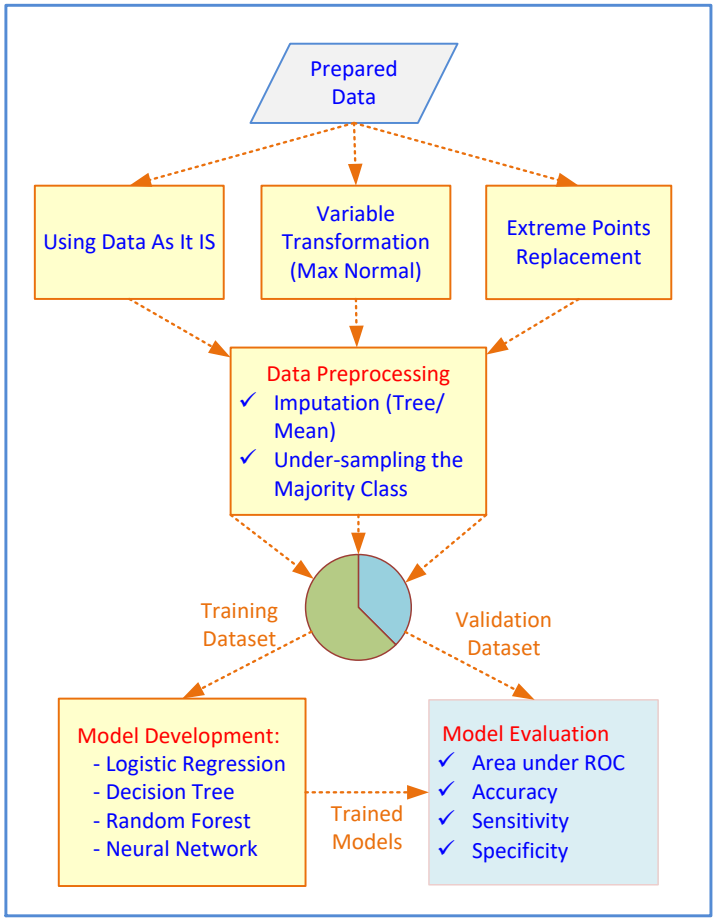


Figure 4.3- *Modeling procedure*

Artificial Neural Network: Artificial neural network (ANN) is a very popular model in healthcare analytics. ANN can be defined as "massively parallel processors, which tend to preserve experimental knowledge and enable their further use" [114]. One of the advantages of the neural network model is its ability in handling highly complex problem structures with non-linear relationships among variables. A limitation of this method, however, is its high sensitivity to model parameters (i.e., structure/architecture of the model, learning rate, number of layers and neurons in each layer, etc.) [115]. Figure 4.4 exhibits a simple two-layer perceptron network. In this example, there are three inputs and two neurons in the hidden layer. There is a transfer function for the output layer and for each neuron in the hidden layer. In this study, we used two-layer perceptron networks with hyperbolic tangent transfer functions in the hidden layer and a

soft-max transfer function in the output layer (see Equations 4.3, 4.4, and 4.5). We also used the conjugate-gradient optimization technique to optimize the network. For more details about the neural networks design, we refer the readers to Hagan, et al. [116].

$$a_1 = f_1(w_{11}x_1 + w_{21}x_2 + w_{31}x_3 + b_1) \tag{4.3}$$

$$a_2 = f_2(w_{12}x_1 + w_{22}x_2 + w_{32}x_3 + b_2) \tag{4.4}$$

$$y = f(w_1a_1 + w_2a_2 + b) \tag{4.5}$$

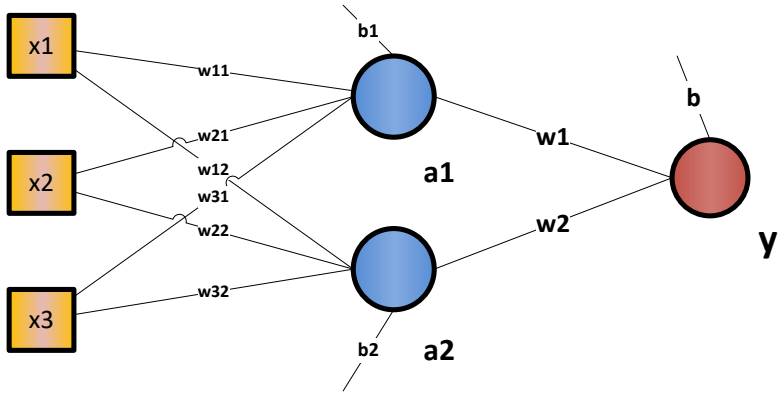


Figure 4.4- *Two-layer perceptron neural networks*

In these equations, $x_1, x_2\ and\ x_3$ are input variables; w_{ij} is the weight of the input i for neuron j; b_j and a_j are the bias and output of neuron j respectively; f_1 and f_2 are the transfer functions for the hidden layer; f is the transfer function of the output layer; and y is the output of the network. In this study, we developed neural network models in two settings. In the first setting, we fed all of the variables into the neural network models, but in the second setting, we only used the variables that were selected through the stepwise method in logistic regression.

Decision Tree: Decision tree is a method that recursively partitions the data based on a predictor [117]. The training process in this method starts at the root node (i.e., all the records and predictors). The tree is built by splitting the records at each stage (i.e., each node) according to the best cut-off value of a predictor. There are several criteria to select the best split. In this study,

we used Pearson's χ^2 p-value and the Gini index. Pearson's χ^2 p-value measures the level of separation achieved by the split. To calculate this measure, consider a 2×2 contingency table for the split. Columns represent the branch directions and rows specify the target variable (0 or 1). The χ^2 value is calculated as in Equation 4.6.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (4.6)$$

In this equation O_{ij} is the observed frequency in row i and column j, and E_{ij} is the expected frequency in row i and column j. The p-value of the χ^2 is then calculated. The smaller the p-value, the better the split or the higher the level of separation.

The Gini index shows the level of purity achieved by the split. Gini is the probability that two randomly selected members of a population are the same. For a pure population, this index would be 1. The calculation of the Gini index in each leaf of a split is as in Equation 4.7, where p_1 and p_2 are the proportions of each level.

$$Gini = {p_1}^2 + {p_2}^2 \quad (4.7)$$

Then, the Gini score of the split is calculated as in Equation 4.8, where w_{left} and w_{right} are the proportion of the records in each leaf.

$$Gini_{score} = w_{left} Gini_{left} + w_{right} Gini_{righ} \quad (4.8)$$

The higher the Gini score, the higher the level of purity achieved by the split. Although the decision tree method is easy to understand, especially for those without knowledge of theories underlying data mining methods, one of its major drawbacks is that data partitioning may result in one leaf comprised of few data points, precluding any useful information from that portion of the data [115].

Random Forest: Random forest could be considered an extension of decision tree. This method develops multiple smaller trees that classify each member of the sample data. The final predicted

class for a particular sample member is determined using a voting mechanism based on the prediction of all trees [118]. Each tree in the random forest uses a subset of records and variables. Random sampling with replacement is used for building each tree. In this study, after examining several scenarios developed by altering model characteristics, we used 60% of the training data and the square root of the number of variables to build each tree. Several advantages can be enumerated for random forest. Besides high accuracy, this method provides a variable importance metric that can be used for identifying important risk factors. Random forest can also handle datasets with a large number of variables [16].

Predictive Model Sets

In this research, four different sets of predictive models were developed (see Figure 4.5). The first set, called the basic models, encompassed models that were developed using lab procedures and demographic data of diabetic patients. In the second set, models were built on lab procedures, demographics, and comorbidity data. These models are called comorbid models. The third set, dubbed over-sampled models, consisted of models built using the over-sampled data by applying the synthetic minority over-sampling technique (SMOTE). And, the fourth set included ensemble models that were developed based on the outputs of individual classifiers.

Basic Models: In this set of models, we used the data compiled during the data preparation phase. We call this dataset "basic data" as it only included demographic and lab results of the diabetic patients.

Models Based on Comorbid Data: The second set of predictive models was based on the comorbidity information. To develop these models, comorbidity data were added to the basic data through several data preparation steps. In these models, we considered the existence of other diabetes-related complications to predict diabetic retinopathy. The following complications were included in our analyses: neuropathy, nephropathy, peripheral circularity, hyperosmolarity,

diabetes-related coma, and other specified diabetes-related conditions. To prepare the comorbid dataset, we performed several steps on the primary data table, which consisted of the list of patients, their complication (diagnosis code), and their demographic and lab data. Since each complication of a patient generated a different record in the database, we extracted all records in which the diagnosis was one of the aforementioned complications and saved them in separate tables. Next, we merged these tables by patient ID and added a binary variable for each complication. Therefore, for each patient, in addition to the demographic and lab data, we added information about their other co-existing complications. After taking these steps, the dataset became ready for the development of the predictive models.

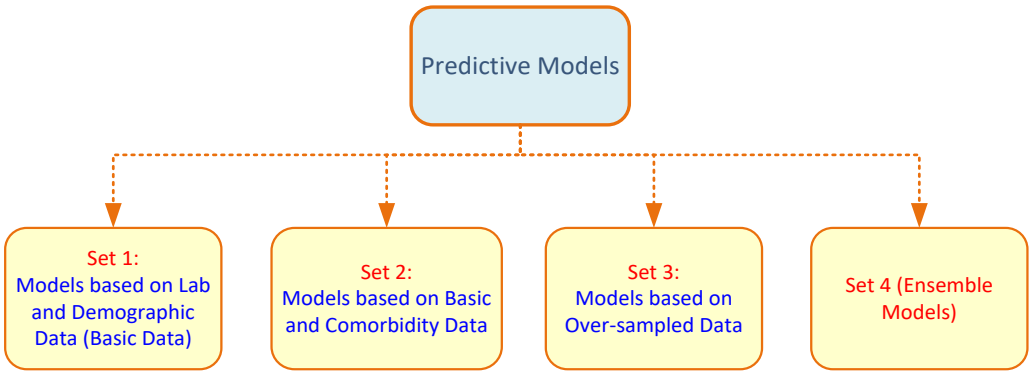


Figure 4.5- *Predictive model sets*

Models Based on Over-Sampled Data: In the previous two sets of models, we carried out random under-sampling for the majority class to create a balanced dataset. One obvious limitation of under-sampling is the possibility of losing important information about the majority class by removing some fractions of the data [76]. The other available approach to create a balanced dataset is to over-sample the minority class. Numerous over-sampling methods have been proposed in recent years, and among them synthetic data generation for the minority class has one of the best performances. By generating synthetic data, new examples of the minority class are generated using different techniques to reach some desired degree of balanced class distribution. SMOTE [78] is one of the most famous methods in this regard. In this method, synthetic data

points are generated on the line joining each minority sample and any/all of its k minority class nearest neighbors (minority class with the smallest Euclidean distance from the original sample). Consider x a minority class, and x_i one of its k minority class nearest neighbors. The new data will be generated as in Equation 4.9,

$$x_{new} = (1 - \delta)x + \delta x_i \tag{4.9}$$

where δ is a random number between $[0, 1]$. Figure 4.6 depicts the synthetic data generation process.

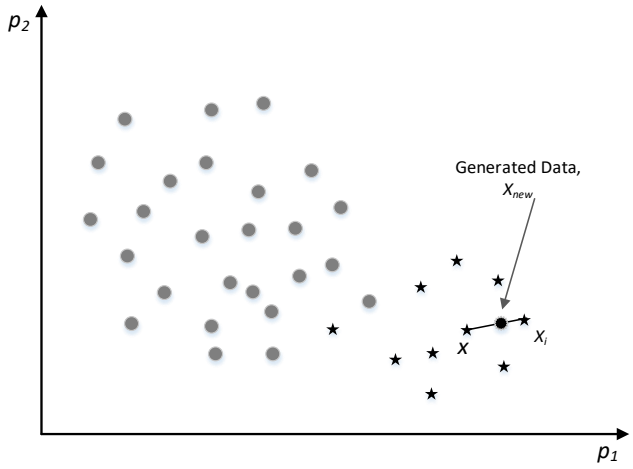


Figure 4.6 - *Synthetic Minority Over-Sampling Technique (SMOTE)*

The number of k nearest neighbors to be used depends on the amount of over-sampling required. For instance, if we want to increase the minority class by 300%, k would be equal to 3. We can enumerate several advantages for this method. First, it requires no information other than the dataset itself [119]. Further, since it is a preprocessing method, over-sampled data can be used in any classification technique with good performance on balanced data [120]. Finally, by generating synthetic minority data, as opposed to simply replicating existing minority data, the minority region can be generalized and overfitting, a limitation of replication, can be avoided [85].

In this study, we considered 5 neighbors to increase the size of the minority class by 10 times, which means we generated two synthetic data points on each line connecting the minority class and each of its five nearest neighbors. Rather than over-sampling the minority class up to the level of the majority class, we increased the size of the minority class to some extent and then under-sampled the majority class to reach a balanced dataset, which is consistent with Chawla, et al. [78] study that showed the combination of SMOTE and under-sampling of the majority class has a better result compared to plain under-sampling. For simplicity, from this point to the end of the this chapter, we call models in Set 1 "basic models", models in Set 2 "comorbid models", and models in Set 3 "over-sampled models".

Ensemble Models

In this study, we developed a new ensemble approach, called confidence margin ensemble. We assessed the performance of confidence margin in comparison to four other exiting ensemble techniques, which are simple average, weighted average, voting-based, and random forest. We explained the random forest model earlier in this section. A description of other approaches is presented next.

Simple average ensemble models – In this method, each model contributes an equally weighted output to compute an average. For instance, if there are 5 single models A, B, C, D and E with outputs of $0.66, 0.45, 0.76, 0.92,$ and 0.48 for a record, the average output would be 0.654. The final classification decision would be YES, even though two of the single models would classify the record as NO (assuming that the decision cut-off is set at 0.50).

Weighted average ensemble models – Rather than assuming all classifiers contribute equally; this method assigns different weights to single classifier outputs for calculating the final result. The weight of each single classifier corresponds to its accuracy. Again, suppose there are 5 single models A, B, C, D and E. The weights will be determined as in Equation 4.10.

$$W^A = \frac{Accuracy^A}{Accuracy^A + Accuracy^B + Accuracy^C + Accuracy^D + Accuracy^E} \quad (4.10)$$

So, the more accurate a model, the greater weight the model will have in calculating the weighted average. The final prediction for a record can be calculated as in Equation 4.11.

$$P = W^A P^A + W^B P^B + W^C P^C + W^D P^D + W^E P^E \quad (4.11)$$

Confidence margin ensemble models - To build these ensemble models, we define a metric, named confidence margin (c_m). Confidence margin is calculated for each record predicted by each single model. The cut-off point to make the decision is considered to be 0.5, hence the confidence margin for record i, predicted by model A is defined as in Equation 4.12.

$$c\_m_i^A = \begin{cases} P_i^A - 0.5 & if\ P_i^A > 0.5 \\ 0.5 - P_i^A & if\ P_i^A \leq 0.5 \end{cases} \quad (4.12)$$

where P_i^A is the prediction output of model A for record i.

Therefore, we say a model is more confident in predicting a record (or the confidence margin is greater) when its prediction is farther away from the cut-off. Final predictions are obtained after calculating confidence margins for all records in all models. Similar to the two previous ensemble models, the average of the single models' predictions is calculated, but different weights are used at both model and record levels. The weight of the prediction for record i by model A is calculated as in Equation 4.13.

$$W_i^A = \frac{c\_m_i^A}{c\_m_i^A + c\_m_i^B + c\_m_i^C + c\_m_i^D + c\_m_i^E} \quad (4.13)$$

Therefore, the final prediction for record i can be calculated as in Equation 4.14.

$$P_i = W_i^A P_i^A + W_i^B P_i^B + W_i^C P_i^C + W_i^D P_i^D + W_i^E P_i^E \quad (4.14)$$

Voting-based ensemble models - In these models, a voting procedure is utilized to make the final decision. When the number of single models is odd, the final decision is the majority vote. But when the number of single models is even, in case of equal votes between two classes, the final decision is made by comparing the sum of confidence margins for models that voted YES to the sum of confidence margins for models that voted NO (see Equation 4.15).

$$\begin{cases} if\ \sum_{j^+ \in Modles\ which\ voted\ YES} c\_m_i^{j^+} > \sum_{j^- \in Modles\ which\ voted\ NO} c\_m_i^{j^-} & \longrightarrow YES \\ if\ \sum_{j^+ \in Modles\ which\ voted\ YES} c\_m_i^{j^+} \leq \sum_{j^- \in Modles\ which\ voted\ NO} c\_m_i^{j^-} & \longrightarrow NO \end{cases} \quad (4.15)$$

Variable Importance Evaluation

One of the major benefits of analytics and data mining in healthcare is the identification of factors that have strong predictive power in detecting a disease. In this section, we elucidate the procedure used to evaluate the predictive power of different independent variables (predictors). To assess the predictive power (variable importance) of a variable, we used the Gini impurity reduction metric in the random forest models. As was mentioned earlier, random forest is a collection of multiple decision trees. Thus, to calculate the Gini impurity reduction for different variables in a random forest model, the average Gini impurity reduction for each variable in all decision trees needs to be computed. The calculation of the Gini purity index for a node in a decision tree was shown in Equation 4.7; therefore, the Gini impurity index, which is showed by $Gini_{imp}$, is calculated as in Equation 4.16,

$$Gini_{imp} = 1 - Gini = 1 - ({p_1}^2 + {p_2}^2) = p_1(1 - p_1) + p_2(1 - p_2) \quad (4.16)$$

where p_1 and p_2 are the proportions of each level of the target variable at the node. If a variable is used for a split in a decision tree, the Gini impurity reduction (GIR) for that variable is calculated as in Equation 4.17,

$$GIR = Gini_{imp_{parent}} - w_{left} Gini_{imp_{left}} - w_{right} Gini_{imp_{right}} \quad (4.17)$$

where w_{left} and w_{right} are the proportion of the records in each leaf. Now suppose T trees are generated in a random forest model; the GIR for a variable in that random forest will be given as in Equation 4.18,

$$GIR_{RF} = \frac{\sum_{t=1}^{T} GIR_t}{T} \tag{4.18}$$

The higher the GIR of a variable in the random forest, the more important that variable is for detecting the target variable. Since we have developed several random forest models in different data preparation settings, we develop the final ranking of important variables by following this procedure: first GIR for all variables is calculated in all random forest models in Sets 2 and 3; then, GIRs are normalized in each random forest model; and finally, the average of normalized GIRs for each variable in different random forest models is computed to obtain the variable importance metric for that variable. Since any manipulation in the data could change the models' output and, hence, the variable importance ranking, we believe our procedure for creating the list of important variables is more robust and more reliable.

4.2. Results

Variable Description

The independent variables (predictors) in our dataset can be divided into three categories: demographic, lab, and comorbidity variables. Tables 4.1, 4.2 and 4.3 provide a brief description of the variables (out of 68 variables in the data) that made it to the final models. Table 4.1 describes the four demographic variables: gender, race, marital status, and urban/rural status. There were slightly more females than males. Most of the patients were Caucasian, followed by African Americans and Hispanics. More than 37% of the patients were married and others were single, widowed, or legally separated.

Table 4.1- *Demographic variables*

Name	Levels	Mode
Gender	Female/Male/ Unknown	Female
Race	Caucasian/ African American/ Hispanic/ Asian/ Native American/ Pacific Islander/ Unknown/ Null	Caucasian
Marital Status	Married/ Single/ Widowed/ Legally Separated/ Life Partner/ Null	Married
Urban Rural Status	Urban (99%)/ Rural (1%)	Urban

Table 4.2 provides some descriptive statistics for lab procedures. In this table, mean, standard deviation (SD), skewness, kurtosis, and the missing value percentage for each variable are presented. A brief description of the comorbidity variables (other diabetes-related complications) is provided in Table 4.3.

Table 4.2- *Lab procedure variables*

Name	Description	Mean	StDev	Skewness	Kurtosis	Missing (%)
Alanine Aminotransferase (ALT)	This test assesses the level of ALT enzyme in the blood.	31.14	31.66	5.814	46.38	0.49
Albumin Serum	This test measures the level of albumin in the blood.	3.11	0.80	3.673	135.16	0.47
Anion Gap (Blood)	This test evaluates the electrically charged particles such as sodium, chloride, and bicarbonate in the blood.	9.61	3.68	0.868	3.49	0.39
Aspartate Aminotransferase (AST)	This test measures the level of AST enzyme in the blood. AST test usually ordered with ALT.	28.49	26.03	5.174	34.73	0.44
Blood Urea Nitrogen (BUN)	BUN measures the amount of nitrogen in the blood that comes from urea. This test evaluates the functionality of kidneys and liver.	20.70	14.59	2.786	11.94	0.22
Calcium Serum	This test evaluates the amount of the calcium in the blood.	8.48	1.11	-3.501	18.86	0.26
Chloride Serum	This test helps detecting the abnormal amounts of chloride in the blood.	102.52	4.66	-1.47	30.47	0.21
Creatinine Serum	This test evaluates kidney function. Creatinine is a waste product of muscle metabolism and eating meat.	0.77	1.20	4.088	22.67	0.20

Glucose Serum Plasma	This test assesses the blood glucose level, a major test to screen for pre-diabetes and diabetes.	155.19	80.50	2.39	10.47	0.29
Hematocrit	This test measures the percentage of red blood cells in the blood.	35.27	9.09	-2.104	6.32	0.30
Hemoglobin	This test measures the amount of hemoglobin in the blood.	12.27	4.45	15.217	297.57	0.28
Mean Corpuscular Hemoglobin Concentration	MCH measures the average concentration of hemoglobin per red blood cell.	33.23	1.39	-5.627	163.68	0.34
Mean Platelet Volume (MPV)	MPV is automated measurement of average size of platelets in the blood.	8.38	1.52	0.238	3.60	0.43
Potassium Serum	This test measures the level of potassium in the blood.	3.76	0.61	-0.003	2.55	0.20
Protein Total Serum	This test evaluates the amounts of albumin and globulin proteins in the blood.	6.13	1.16	-2.199	9.96	0.50
Red Blood Cell (RBC) Count	RBC measures the number of red blood cells in the blood and usually is ordered as a part of a complete blood cell test.	3.72	1.31	21.354	725.58	0.36
Sodium Serum	This test assesses the level of sodium and detects abnormal low/high sodium in the blood.	138.03	3.95	-6.576	226.32	0.22
White Blood Cell (WBC) Count	This test determines the number of WBC in the blood and helps to diagnose infections and other medical conditions.	8.16	4.77	11.475	281.13	0.30

Table 4.3- *Comorbidity variables*

Name	Description
Neuropathy	Type of nerve disorder; particularly prevalent in the feet and legs.
Nephropathy	Kidney disease
Peripheral Circulatory	Arterial blockage due to atherosclerosis; mostly affects lower extremities.
Ketoacidosis	High levels of ketone bodies, or blood acids, resulting from the breaking down of fat rather than glucose for energy; may lead to coma.
Hyperosmolarity	Extremely high blood sugar level in which excess blood sugar is passed into the urine; may lead to life-threatening dehydration.
Other Complications	Other specified complications of diabetes are in this category

Models' Output

The results of the different sets of models are presented in Tables 4.4, 4.5, and 4.6, which include area under the curve (AUC) of the receiver operating characteristic (ROC), accuracy, sensitivity, and specificity. As the results show, the accuracy, sensitivity, specificity, and AUC of the models improved from the basic to the comorbid, and from the comorbid to the over-sampled models. Finally, for each set of models, ensemble models outperformed single models.

In set 1 (basic models), the best performance was obtained from random forests without imputation of missing values, with AUCs standing at 81.90%, 81.80% and 81.50%. This was followed by neural networks, with AUCs of 80.00% and 79.80%. In this set, following the random forest and neural networks, logistic regression models ranked third, and decision trees had the poorest performance.

Table 4.4 – *Set 1 - Basic models' results*

Data Manipulation	Imputation Technique	Modeling Technique	AUC	Accuracy	Sensitivity	Specificity
No Manipulation	No Imputation	DT-Gini	77.30%	70.76%	63.95%	77.56%
		DT-Prob χ^2	77.30%	70.71%	65.04%	76.38%
		RF*	81.80%	73.78%	72.37%	75.19%
	Mean	DT-Gini	76.30%	70.98%	65.50%	76.47%
		DT-Prob χ^2	75.60%	71.12%	59.76%	82.48%
		RF	79.20%	71.87%	67.46%	76.29%
		LR	78.00%	71.26%	63.27%	79.24%
		ANN	79.40%	72.96%	69.69%	76.24%
		ANN-Reg#	79.80%	72.28%	67.64%	76.92%
	Tree	DT-Gini	75.20%	70.25%	56.62%	83.89%
		DT-Prob χ^2	75.30%	70.41%	60.13%	80.70%
		RF@	79.80%	72.33%	67.50%	77.15%
		LR	78.00%	71.92%	65.50%	78.33%
		ANN	79.60%	72.67%	70.82%	74.51%
		ANN-Reg#	80.00%	72.99%	68.50%	77.47%
Extreme Point Replacement	No Imputation	DT-Gini	77.30%	70.76%	63.95%	77.56%
		DT-Prob χ^2	77.30%	70.71%	65.04%	76.38%
		RF*	81.90%	73.78%	71.01%	76.56%
	Mean	DT-Gini	76.40%	71.01%	65.91%	76.10%
		DT-Prob χ^2	76.00%	71.10%	60.17%	82.02%
		RF	79.20%	72.05%	67.14%	76.97%
		LR	76.70%	69.71%	59.76%	79.65%
		ANN	78.50%	72.01%	66.36%	77.65%
		ANN-Reg	78.30%	72.14%	63.40%	80.88%
	Tree	DT-Gini	75.10%	70.35%	56.26%	84.53%
		DT-Prob χ^2	76.20%	70.89%	60.63%	81.16%
		RF	79.30%	71.94%	66.32%	77.56%
		LR	77.80%	71.48%	63.68%	79.29%
		ANN#	80.00%	73.26%	67.96%	78.56%
		ANN-Reg	79.40%	72.30%	67.73%	76.88%
Max Normal Variable Transformation	No Imputation	DT-Gini	77.20%	70.69%	64.00%	77.38%
		DT-Prob χ^2	77.30%	70.71%	65.04%	76.38%
		RF*	81.50%	73.33%	69.64%	77.01%
	Mean	DT-Gini	76.60%	71.01%	62.72%	79.29%
		DT-Prob χ^2	75.60%	71.17%	60.67%	81.66%
		RF@	79.70%	72.05%	67.46%	76.65%
		LR	78.10%	71.69%	66.77%	76.60%
		ANN	79.60%	72.60%	68.37%	76.83%
		ANN-Reg	79.30%	72.23%	66.95%	77.51%
	Tree	DT-Gini	75.30%	70.69%	61.00%	80.38%
		DT-Prob χ^2	76.60%	70.91%	61.77%	80.06%
		RF	79.50%	71.71%	66.50%	76.92%
		LR	78.20%	71.55%	67.60%	75.51%
		ANN@	79.70%	72.78%	68.32%	77.24%
		ANN-Reg	79.40%	72.46%	68.55%	76.38%

* Top three models # Second top three models @ Third top three models

Table 4.5 – *Set 2 - Comorbid models' results*

Data Manipulation	Imputation Technique	Modeling Technique	AUC	Accuracy	Sensitivity	Specificity
No Manipulation	No Imputation	DT-Gini	83.40%	78.65%	76.83%	80.47%
		DT-Prob χ^2	83.60%	79.06%	78.88%	79.24%
		RF*	88.40%	80.61%	75.97%	85.25%
	Mean	DT-Gini	84.20%	79.38%	79.88%	78.88%
		DT-Prob χ^2	83.80%	79.18%	78.88%	79.47%
		RF#	87.50%	79.81%	74.37%	85.25%
		LR	86.20%	79.09%	70.41%	87.76%
		ANN	87.00%	80.29%	74.42%	86.16%
		ANN-Reg	86.90%	79.97%	73.83%	86.12%
	Tree	DT-Gini	83.30%	78.81%	72.14%	85.48%
		DT-Prob χ^2	84.70%	78.77%	68.78%	88.76%
		RF#	87.70%	80.18%	75.92%	84.43%
		LR	86.50%	79.40%	71.19%	87.62%
		ANN@	87.40%	80.13%	75.47%	84.80%
		ANN-Reg#	87.40%	80.18%	75.24%	85.12%
Extreme Point Replacement	No Imputation	DT-Gini	83.40%	78.65%	76.83%	80.47%
		DT-Prob χ^2	83.60%	79.06%	78.88%	79.24%
		RF*	88.60%	80.38%	74.60%	86.16%
	Mean	DT-Gini	84.20%	79.38%	79.88%	78.88%
		DT-Prob χ^2	83.80%	79.18%	78.88%	79.47%
		RF@	87.40%	79.84%	74.28%	85.39%
		LR	85.90%	78.88%	69.78%	87.98%
		ANN	86.40%	80.13%	74.15%	86.12%
		ANN-Reg	86.30%	79.72%	73.51%	85.94%
	Tree	DT-Gini	84.20%	79.20%	74.51%	83.89%
		DT-Prob χ^2	84.40%	79.22%	74.51%	83.93%
		RF	87.40%	79.95%	74.37%	85.53%
		LR	85.90%	79.13%	70.28%	87.98%
		ANN	87.30%	79.97%	75.56%	84.39%
		ANN-Reg	86.10%	79.77%	72.05%	87.48%
Max Normal Variable Transformation	No Imputation	DT-Gini	83.40%	78.70%	76.74%	80.66%
		DT-Prob χ^2	83.60%	79.06%	78.88%	79.24%
		RF*	88.50%	80.41%	74.74%	86.07%
	Mean	DT-Gini	84.10%	79.27%	79.79%	78.74%
		DT-Prob χ^2	83.80%	79.18%	78.88%	79.47%
		RF@	87.40%	79.90%	73.78%	86.03%
		LR	86.30%	79.49%	72.64%	86.35%
		ANN	86.60%	79.72%	73.96%	85.48%
		ANN-Reg	86.50%	79.63%	73.69%	85.57%
	Tree	DT-Gini	84.30%	78.95%	71.83%	86.07%
		DT-Prob χ^2	85.00%	79.20%	73.60%	84.80%
		RF	87.40%	79.79%	73.46%	86.12%
		LR	85.70%	79.34%	71.46%	87.21%
		ANN	86.30%	79.40%	74.65%	84.16%
		ANN-Reg	86.10%	79.70%	71.73%	87.67%

* Top three models # Second top three models @ Third top three models

Table 4.6 – *Set 3 - Over-sampled models' results*

Data Manipulation	Imputation Technique	Modeling Technique	AUC	Accuracy	Sensitivity	Specificity
No Manipulation	No Imputation	DT-Gini	93.00%	89.13%	89.05%	89.20%
		DT-Prob χ^2	93.00%	89.14%	88.85%	89.42%
		RF*	97.90%	92.71%	90.00%	95.43%
	Mean	DT-Gini	89.90%	84.16%	89.03%	79.07%
		DT-Prob χ^2	90.10%	84.16%	89.18%	79.13%
		RF@	95.20%	88.03%	90.77%	85.28%
		LR	87.40%	78.33%	68.25%	88.42%
		ANN	92.20%	84.97%	86.95%	83.00%
		ANN-Reg	93.20%	85.96%	87.47%	84.46%
	Tree	DT-Gini	91.90%	87.17%	86.53%	87.81%
		DT-Prob χ^2	92.10%	87.18%	86.51%	87.84%
		RF#	96.00%	89.83%	90.04%	89.63%
		LR	87.50%	79.79%	72.72%	86.85%
		ANN	93.30%	86.91%	87.44%	86.37%
		ANN-Reg	91.40%	84.06%	87.57%	80.56%
Extreme Point Replacement	No Imputation	DT-Gini	93.00%	89.13%	89.05%	89.21%
		DT-Prob χ^2	93.00%	89.14%	89.08%	89.21%
		RF*	97.90%	92.76%	90.22%	95.30%
	Mean	DT-Gini	89.90%	84.14%	88.72%	79.57%
		DT-Prob χ^2	90.10%	84.14%	88.87%	79.42%
		RF@	95.20%	88.02%	90.75%	85.28%
		LR	83.40%	77.50%	67.38%	87.62%
		ANN	92.20%	85.06%	86.66%	83.46%
		ANN-Reg	91.70%	84.21%	88.08%	80.36%
	Tree	DT-Gini	92.20%	87.27%	86.74%	87.80%
		DT-Prob χ^2	92.20%	87.28%	86.76%	87.79%
		RF	91.00%	89.66%	60.11%	94.20%
		LR	87.80%	79.54%	72.13%	86.96%
		ANN	93.50%	87.03%	87.74%	86.33%
		ANN-Reg	91.50%	83.72%	86.72%	80.73%
Max Normal Variable Transformation	No Imputation	DT-Gini	93.00%	89.13%	89.05%	89.21%
		DT-Prob χ^2	93.00%	89.14%	89.08%	89.21%
		RF*	97.90%	92.71%	89.98%	95.43%
	Mean	DT-Gini	90.30%	84.36%	89.12%	79.61%
		DT-Prob χ^2	89.30%	84.31%	89.25%	79.37%
		RF#	95.20%	88.04%	90.69%	85.51%
		LR	87.20%	77.86%	68.21%	87.51%
		ANN	92.40%	85.17%	86.72%	83.63%
		ANN-Reg	93.40%	86.27%	87.64%	84.91%
	Tree	DT-Gini	92.20%	88.01%	88.45%	87.57%
		DT-Prob χ^2	92.20%	88.04%	88.65%	87.43%
		RF#	96.10%	89.94%	90.18%	89.69%
		LR	88.80%	80.66%	83.90%	77.43%
		ANN@	93.90%	87.57%	87.22%	87.92%
		ANN-Reg	92.00%	84.30%	86.60%	82.00%

* Top three models # Second top three models @ Third top three models

In set 2 (comorbid models), random forests once again had the best performance with AUCs of 88.60%, 88.50% and 88.40% for the top three models. Similar to the basic models, neural networks, logistic regressions, and decision trees ranked second, third, and fourth, respectively. In set 3 (over-sampled models), random forest models had the highest accuracy in detecting retinopathy among diabetic patients. The best models in set 3 had an AUC of 97.80%, which is remarkable. AUCs for other models in this set were significantly high, mostly over 92%. Neural networks were the second-best models in this set, but unlike sets 1 and 2, decision trees had the third rank in over-sampled models, and logistic regressions had the worst performance.

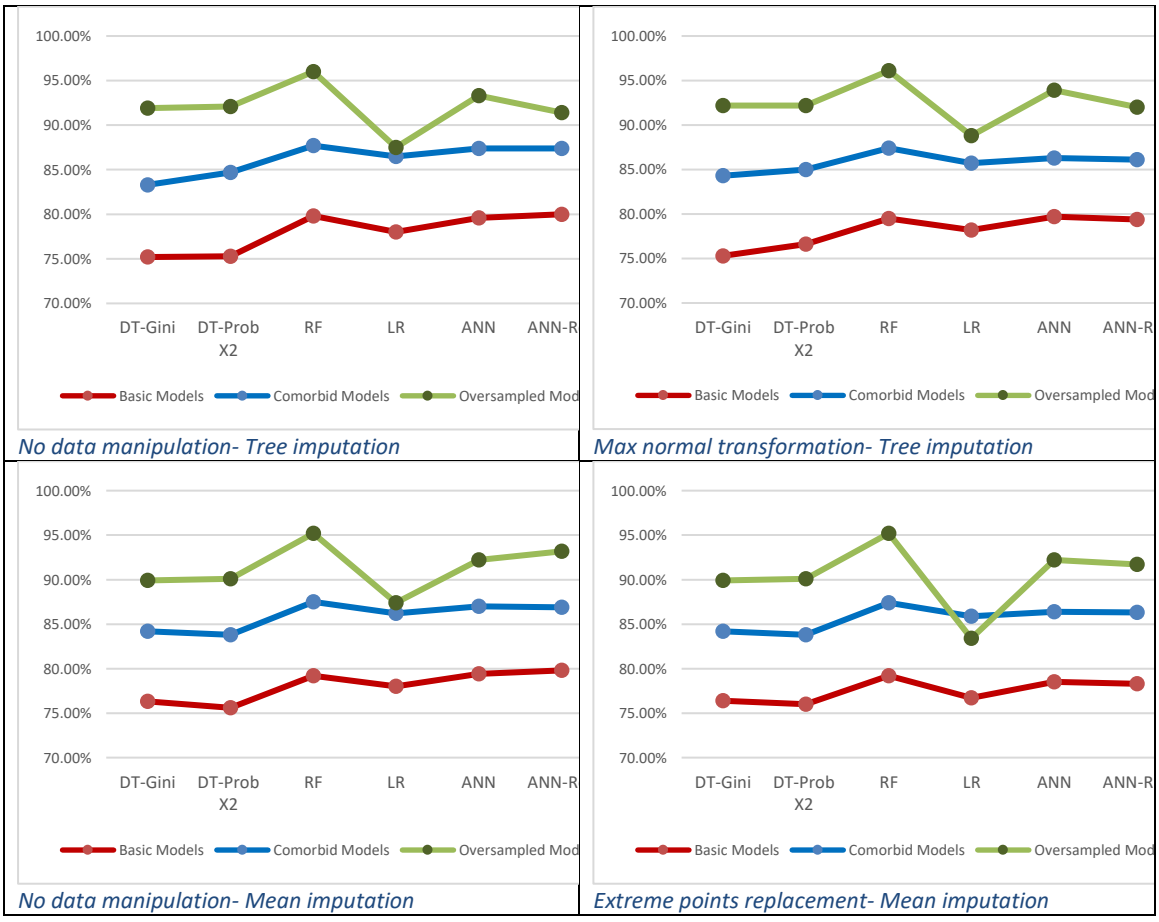
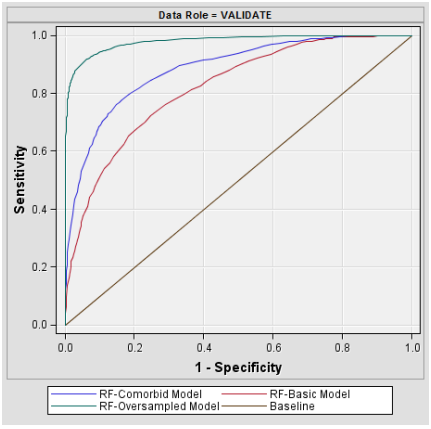


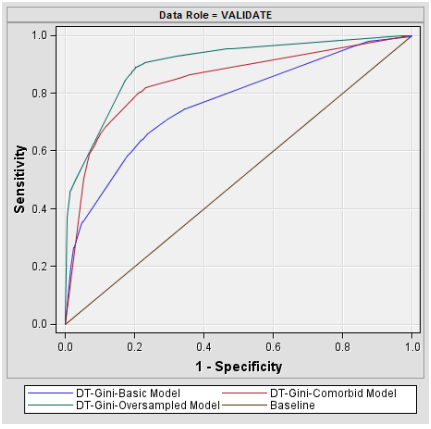
Figure 4.7 - *AUC comparison among modeling techniques and modeling sets*

Graphs in Figure 4.7 compare the modeling techniques within and between the three sets: basic, comorbid, and over-sampled. As expected, there clearly is a significant improvement (about 10%) from the basic to the comorbid models. Except for the logistic regression, the accuracy of the

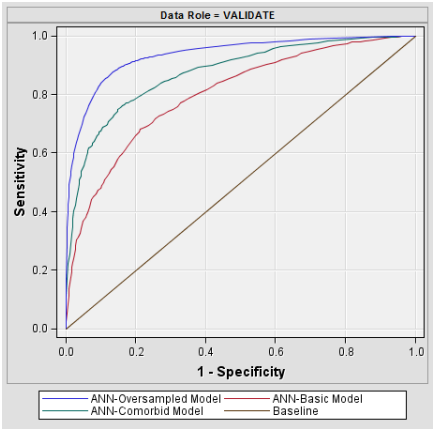
models built on the over-sampled sets are considerably higher than that of the comorbid sets for all the modeling techniques. Specifically, the AUCs of the over-sampled models (for random forest, neural networks, and decision tree techniques) were about 8% higher than the AUCs of their counterparts in the comorbid models. This improvement is expected, since in over-sampled models there are more data points available to train the models. The only modeling technique that did not improve by using over-sampled data was logistic regression, which was the only linear model used in this study. All other models (i.e., random forest, neural networks, and decision tree) are non-linear; so, unlike logistic regression, they can take advantage of richer and more complicated data, leading to higher accuracy with over-sampled data.



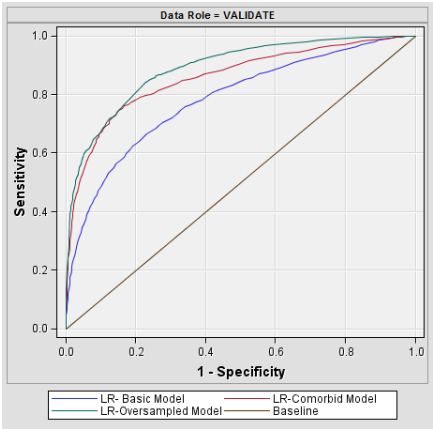
Random forest models (No data manipulation- no imputation)



Decision tree-Gini (Extreme point replacement- mean imputation)



Neural networks (Extreme point replacement- tree imputation)



Logistic regression (Max normal transformation- tree imputation)

Figure 4.8 - *ROC charts of modeling techniques in different sets*

The ROC charts of each one of the modeling techniques used in the three modeling sets is provided in Figure 4.8, which demonstrates the superiority of the over-sampled models over the comorbid and basic models, and the dominance of the comorbid models over the basic models. In addition to the AUC, we compared the basic, comorbid, and over-sampled models using the difference between sensitivity (true positive rate) and specificity (true negative rate) for each model in each set. Even though the overall accuracy is an important metric to compare models, a desired model is one that can classify both positive targets (in our case, patients with diabetic retinopathy) and negative targets (patients without diabetic retinopathy) at a high rate. Therefore, the lower the difference between sensitivity and specificity of a model, the better and more reliable that model. The average differences between sensitivity and specificity for models in set 1 (basic models), set 2 (comorbid models), and set 3 (over-sampled models) were 12.93%, 9.63%, and 6.06%, respectively. Thus, not only did the over-sampled models had the highest accuracy, but they were also more robust and reliable than the comorbid and basic models, with the comorbid models being more robust than the basic models.

Table 4.7 presents the accuracies of the best single model, existing ensemble approaches, and our proposed ensemble approach for the various datasets used in this study. As it can be seen, almost all ensemble models (except for 5 out of 90) have improved the prediction accuracy compared to the best single model. Those 5 simple average ensembles did not outperform their best constituting single models for a simple reason: not all models composing the ensemble had a good performance, however, they all were given an equal weight in determining the ensemble's final decision. Consequently, the overall accuracy of the ensembles created by simple averaging suffered in comparison to the accuracy of the best constituting single models. In fact, one of the most important requirements for building a good ensemble is to have comparably accurate and diverse single classifiers [121]. The results also show that the prediction accuracies improve from the simple average to the weighted average and culminate in the confidence margin ensemble

models. Reaching to a performance peak in the confidence margin ensembles reflects their more accurate and more reliable assignment of weights compared to the weighted-average ensemble models. Voting-based ensembles outperformed the best single models in all cases. In two cases, voting-based models had the best predictive performance among all ensemble models. Similarly, random forest models outperformed all other types of ensembles in two cases. Overall, our proposed ensemble approach, the confidence margin ensemble, had the best performance in this study. More specifically, it excelled in 14 out of 18 total different data settings. It deserves to mention that we used logistic regression, decision tree, and neural network models for developing the simple average, weighted average, voting based, and confidence margin ensemble models.

Table 4.7- *Set 4 - Ensemble models results*

Accuracy of Ensemble Models	Best Single Model	**Ensemble Model Type**				
		Existing Approaches				Our Approach
		Simple Average	Weighted Average	Voting Based	Random Forest	Confidence Margin
Ensemble of models in set 1						
No Manipulation /Tree Imputation	72.99%	73.31%	73.37%	73.36%	72.33%	**73.56%**
No Manipulation /Mean Imputation	72.96%	72.93%	72.99%	73.03%	71.87%	**73.45%**
Extreme Point Replacement /Tree Imputation	73.26%	73.28%	73.35%	**73.46%**	71.94%	73.42%
Extreme Point Replacement / Mean Imputation	72.14%	72.20%	72.34%	72.30%	72.05%	**73.56%**
Variable Transformation/Tree Imputation	72.78%	72.60%	72.89%	72.83%	71.71%	**73.25%**
Variable Transformation/Mean Imputation	72.60%	72.62%	72.75%	72.80%	72.05%	**73.85%**
Ensemble of models in set 2						
No Manipulation/Tree Imputation	80.18%	80.16%	80.60%	80.78%	80.18%	**80.64%**
No Manipulation /Mean Imputation	80.29%	80.57%	80.65%	80.62%	79.81%	**80.97%**
Extreme Point Replacement /Tree Imputation	79.97%	80.15%	80.29%	80.35%	79.95%	**80.80%**
Extreme Point Replacement / Mean Imputation	80.13%	80.25%	80.40%	**81.05%**	79.84%	80.93%
Variable Transformation/Tree Imputation	79.70%	80.11%	80.37%	80.40%	79.79%	**80.75%**
Variable Transformation/Mean Imputation	79.72%	80.20%	80.38%	80.40%	79.90%	**80.81%**
Ensemble of models in set 3						
No Manipulation/Tree Imputation	87.18%	88.23%	88.27%	88.03%	**89.83%**	88.38%
No Manipulation /Mean Imputation	85.96%	86.15%	86.92%	87.11%	**89.83%**	89.21%
Extreme Point Replacement /Tree Imputation	87.28%	87.01%	88.68%	88.60%	89.66%	**89.84%**
Extreme Point Replacement / Mean Imputation	85.06%	86.08%	86.98%	87.15%	88.02%	**88.42%**
Variable Transformation/Tree Imputation	88.04%	87.91%	88.58%	88.52%	89.94%	**90.12%**
Variable Transformation/Mean Imputation	86.27%	86.52%	87.05%	87.62%	88.04%	**88.25%**

Variable Importance

As we mentioned in previous sections, 68 independent variables were included in our analyses. Understanding the predictive power of each of these variables could be helpful for physicians in

better managing the course of the disease by controlling factors that are highly associated with retinopathy. As the results show, random forest models have the best performance among other modeling techniques; therefore, to specify the importance of variables according to their predictive powers, we applied the Gini reduction metric on the validation datasets using the output from random forest models. Figure 4.9 shows the variable importance in detecting retinopathy based on the Gini reduction score in multiple random forest models. Based on our findings, neuropathy; creatinine serum and blood urea nitrogen (both measures of kidney function); glucose serum plasma (used to screen for pre-diabetes and diabetes); and hematocrit (a measure of red blood cell concentration) were the most important variables for detecting diabetic retinopathy.

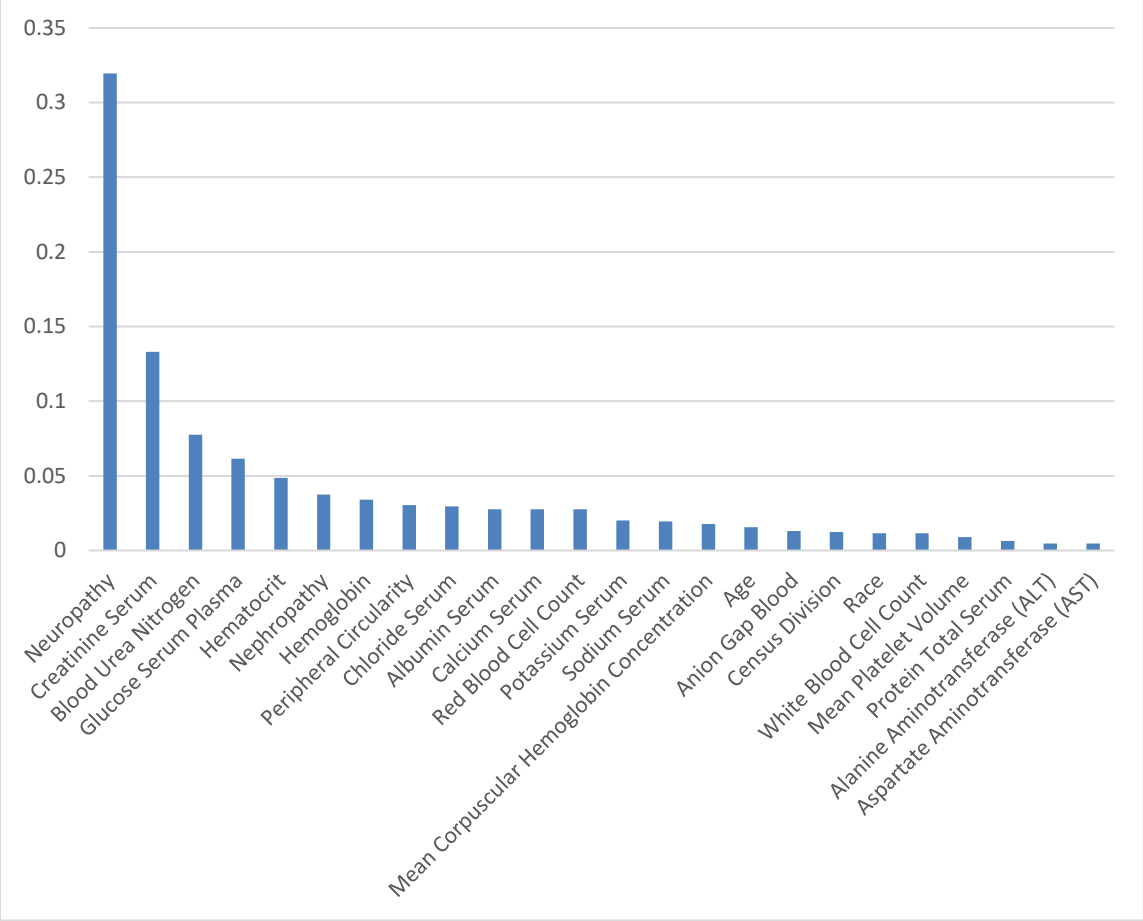


Figure 4.9- *Variable importance ranking in detecting diabetic retinopathy*
[*Y-axis represents the average of normalized Gini reduction*]

4.3. Discussion and Conclusion

In this study, we analyzed data from more than 1.4 million diabetic patients. Our objective was to develop a CDSS to detect diabetic retinopathy using demographic, lab procedure, and comorbidity data only. Several aspects distinguish this research study from other existing studies in this area. First, this study included a far greater number of patients and risk factors in the analysis, which contribute to the rigor and robustness of the findings. While other similar studies used data from several hundred patients, we employed data from more than 300,000 patients to develop the predictive models. Second, through a series of database operations and data preparation steps conducted in SAS, we incorporated patients' comorbidity data into our models, which significantly improved the accuracy of our predictive analytics-based CDSS. The results we obtained from the comorbid models were in line with other research that emphasizes the importance, and advantage, of analyzing concurrent conditions instead of studying each condition in isolation. Third, we over-sampled the rare event, which made it possible to preserve important features in the data that could have been lost through pure under-sampling. Models we built by utilizing over-sampled data were much more accurate and robust compared to comorbid and basic models, especially in non-linear modeling techniques (such as random forest, neural network, and decision tree) that can handle more complex data.

In this study, we developed and evaluated a novel ensemble approach, which we call the confidence margin ensemble. Our evaluations showed that confidence margin ensembles had better overall performance compared to the existing ensemble models. In confidence margin ensemble, we assign different weights in both model and record levels. To calculate these weights, first, the confidence margin metric needs to be computed for every model at each record. Confidence margin metric is the absolute difference between the model's output and decision cut-off value. The greater the confidence margin, the more confident the model in predicitng the record. For instance, if the decisin cut-off is 50%, a model with the output of 95% is more

confident compared to a model with the output of 60%. After computing all of the confidence margin metrics, the weight of the prediction for each record by every model is calculated by normalizing the confidence metrics of all models in each record. In this research we also addressed the issue of tie in voting-based ensemble models by comparing the confidence margins of the base predictors.

To make sure that the quality and veracity of our datasets were acceptable, we devoted a considerable amount of time and effort to data cleaning and preparation. We analyzed the distribution, missing value percentage, and potential outliers of each of the predictors, and included only those variables that were rich enough and had significant predictive powers. As a result, there should be little doubt about the quality, robustness, and accuracy of our models.

Based on our analyses, diabetic neuropathy had the strongest predictive power in detecting diabetic retinopathy, followed by creatinine serum, blood urea nitrogen, glucose serum plasma, and hematocrit. The results showed it is possible to predict diabetic retinopathy with 92.76% accuracy using only the results of a routine blood test. The accuracy of our CDSS may not be as high as the fundus image-based solutions (for example, Kumar and Madheswaran [57] reached an accuracy of 97%), but considering the ease-of-use and the cost-effectiveness of our approach, the resulting CDSS is not only competitive to the existing fundus image-based solutions, but also it can reduce the currently high rate of noncompliance with annual routine ophthalmologic exams. A conspicuous benefit of our CDSS, therefore, is a lower number of patients who would eventually develop retinopathy, which in turn will result in a more efficient healthcare delivery and management system.

Medical researchers and clinicians (e.g., Sabanayagam, et al. [122]) have noted the necessity of developing an accurate predictive model for diabetic retinopathy. We believe our CDSS, which has several practical and clinical applications, fills this gap. First, as mentioned earlier, it makes

up for the low compliance rate of annual comprehensive eye examination for retinopathy. This annual eye examination is necessary for every diabetic patient, but because of inconvenience of the procedure and unavailability of equipment and specialists, it has one of the lowest compliance rates in the U.S. healthcare. Second, the input to this CDSS is nothing but the results of a routine blood test that makes it easy for clinicians, and even for patients, to use. Unlike the current procedure of diagnosing retinopathy, there is no need for an ophthalmologist and special cameras to employ this CDSS; clinician can refer high risk patients (based on the result of our CDSS) to ophthalmologists for more accurate examination and potential treatments. Third, our models identify factors that are most strongly related to diabetic retinopathy. By controlling these factors, doctors and patients will be able to manage the course of the disease in a more effective way. Finally, applying our CDSS can help detect this complication at early stages, and since there exists an effective laser therapy to prevent the progress of retinopathy, vision can be saved for many diabetic patients.

In conclusion, considering the prevalence of retinopathy among diabetic patients (about one third of diabetics have retinopathy [123]) and the significant proportion of individuals whose susceptibility to retinopathy remains undiagnosed (about 50% of all diabetic retinopathy patients), our CDSS provides a great value to people who suffer from diabetes all over the world. Based on the current statistics (about 30 million in the US and 415 million worldwide [2]), diabetic retinopathy remains undiagnosed in 5.5 million US citizens ($30M \times 30\% \times 50\% = 5.5M$) and in more than 62 million individuals around the globe ($415M \times 30\% \times 50\% = 62.25M$). Therefore, if our CDSS helps diagnose even a small percentage of these cases at early stages of the disease, vision, and quality of life, can be saved for a large number of diabetics. Although we cannot appraise the value of sight to an individual, we can enumerate some of the costs associated with losing it, such as medical care, assistance programs, and loss of productivity costs. Therefore, the

application of our CDSS helps save a large amount of expenses in both healthcare and welfare systems.

We admit that our study has a number of limitations. First, in the EMR data we used, we did not have information about the time a patient's disease was first diagnosed. As a results, we could not incorporate the duration of time the patients lived with diabetes into our models. Since longer duration of diabetes has been shown to be a strong predictor of retinopathy [62], our CDSS could even perform better had we had access to the patients' date of diagnosis. Fortunately, having access to several other predictors made up for the lack of this variable. Additionally, we should note that every machine learning technique has several parameters that can be adjusted. Even though we tried to systematically adjust these parameters to achieve better results, those values cannot be considered optimal.

CHAPTER V

DEVELOPING A SYNTHETIC INFORMATIVE MINORITY OVER-SAMPLING (SIMO) ALGORITHM EMBEDDED INTO SUPPORT VECTOR MACHINE TO LEARN FROM IMBALANCED DATASETS

In this chapter, we describe our proposed synthetic informative minority over-sampling (SIMO) algorithm, which is imbedded into support vector machine (SVM) for learning from imbalanced datasets. Here we discuss why we chose over-sampling versus other methods to handle the imbalanced data learning challenge, and why we chose SVM.

First, to apply a sampling method, no extra information is required other than the dataset itself [124]. However, in cost-sensitive methods, the information about the misclassification cost for each class is required, while this kind of information is unknown. The only known fact is that the misclassification cost for minority class is higher than misclassification cost for majority class ([88],[76]). Second, we apply over-sampling versus under-sampling. The major limitation in under-sampling is the possibility of losing important information by removing some parts of the data, while there is not such a problem in over-sampling.

There are three main reasons for choosing SVM as the classifier. First, this method has a very strong and at the same time simple theoretical background which makes it easy to explain intuitively [125]. Second, this method develops a hyperplane (decision boundary) that separates the data space for classifying the data points (examples). It is known that the data points near the decision boundary are more important and difficult to classify [126]. Therefore, identifying the near boundary data samples is rather easy in SVM. Finally, SVM has been shown to have a very good performance and high generalization power in many practical applications compared to other machine learning techniques ([125], [100]).

5.1. Support Vector Machine

SVM is a machine learning technique that can be applied to both regression and pattern recognition (classification) problems. For the classification, SVM develops a decision boundary that separates two classes in the data space. To build this decision boundary, SVM maximizes the separating margin between two classes in the data space while it minimizes the classification error. Figure 5.1 shows a linear SVM decision boundary. Dots and stars denote the two classes in the data. The data points that lie on the margins at both sides of the decision boundary are called support vectors. These support vectors are shown in Figure 5.1 with a circle around them. w is the normal to the decision boundary and $b/|w|$ is the perpendicular distance of the decision boundary from the origin [127]. When two classes are not completely separable, some of the examples will be misclassified. In Figure 5.1, one star data point has misclassified as a dot, the distance of this point from the decision boundary is $-\varepsilon/|w|$.

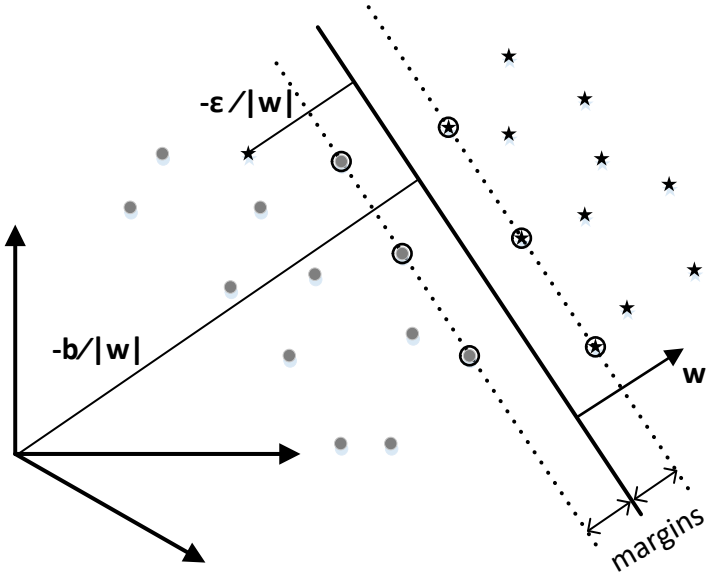


Figure 5.1-*Linear SVM hyperplane*

SVM can be applied to both linear and non-linear separable problems. When two classes are not linearly separable, kernel trick can be employed and the data is mapped to a feature space (using a mapping function $\phi(.)$), which is in a higher dimension [128]. In the feature space, two classes will be linearly separable and the problem will be handled similar to the linearly separable case.

Now we describe the SVM mathematical formulation. Let the training dataset for a two-class problem be represented as $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$. $x_1, x_2, \dots, x_N \in R^m$ are the m dimensional training data points and $y_1, y_2, \dots, y_N \in \{-1, +1\}$ are their corresponding class labels (-1 for majority class and +1 for minority class). By solving the optimization problem in Formulation 5.1, SVM develops a decision boundary that separates two classes.

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon_i \tag{5.1}$$

$s.t.$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \varepsilon_i$$

$$\varepsilon_i \geq 0$$

ε_i are positive slack variables. When a classification error occurs, these variables will be greater than 1. C is a parameter that determines the error penalty. C, which the user chooses is a tradeoff between minimizing the error and maximizing the margin.

Usually the Lagrangian formulation of SVM is solved (Formulation 5.2). The Lagrangian formulation is easier to handle because the constraints in Formulation 5.1 are replaced by Lagrangian multipliers [127].

$$L_P(w, b, \varepsilon, \alpha, \mu) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon_i - \sum_{i=1}^{N} \alpha_i (y_i (w^T \phi(x_i) + b) - 1 + \varepsilon_i) - \sum_{i=1}^{N} \mu_i \varepsilon_i \quad (5.2)$$

where α_i and μ_i are positive Lagrangian multipliers associated with first and second sets of constraints in Formulation 5.1. The Krush-Kuhn-Tucker conditions for the Lagrangian primal (Formulation 5.2) are as follows,

$$w = \sum_{i=1}^{N} \alpha_i y_i \, \phi(x_i) \quad (5.3)$$

$$0 \leq \alpha_i \leq C \quad (5.4)$$

$$\sum_{i=1}^{N} \alpha_i \, y_i = 0 \quad (5.5)$$

The Lagrangian dual form of Formulation 5.2 is obtained by replacing the Equation 5.3 in 5.2. Formulation 5.6 shows the Lagrangian dual [127],

$$\max_{\alpha} \ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \, K(x_i, x_j) \quad (5.6)$$

$$s.t.$$

$$\sum_{i=1}^{N} \alpha_i \, y_i = 0$$

$$0 \leq \alpha_i \leq C$$

where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel function that calculates the inner product of data points in the feature space.

The solution for w is given by

$$w = \sum_{j=1}^{N_s} \alpha_i y_i \, \phi(x_i) \tag{5.7}$$

Here N_s is the number of support vectors. In fact Equation 5.7 is the same as Equation 5.3, but because in the optimal solution of the Lagrangian dual only α_i corresponding to the support vectors have non-zero values, the summation in Equation 5.7 is only on support vectors [127].

To determine the class of a new sample, x, a sign function (sgn(.)) is used, it is obtained using,

$$y = sgn\{w^T \phi(x) + b\} \tag{5.8}$$

or

$$y = sgn\left\{\sum_{i=1}^{N_s} \alpha_i{}^* y_i \, K(x_i, x) + b\right\} \tag{5.9}$$

where α^* is the solution of Lagrangian formulation in 5.6.

SVM on Imbalanced Datasets

Although SVM has a very good performance on balanced datasets, when applied to imbalanced datasets, its performance deteriorates dramatically, especially on the minority class. The SVM decision boundary in an imbalanced dataset is closer toward the minority class region compared to the ideal classification decision boundary. As a result, a considerable number of minority class examples will be misclassified as the majority. Wu and Chang [129] mentioned two reasons for this decision boundary skewness. The first reason is in regard to the imbalanced training data ratio, because the negative data points outnumber the positive examples, these positive examples are further away from the "ideal" decision boundary compared to the majority examples. Second, the imbalanced supports vector ratio, because the number of the negative (majority class) support vectors is much more than the positive (minority class) support vectors, a positive test data point might have more negative support vector neighbors, and as a result will be misclassified as negative (majority) class. Akbani, et al. [97] pointed out another reason for the skewed decision boundary. The objective of the SVM model is to maximize the margin between two classes as

well as minimizing the classification errors and there is a tradeoff between these two. When the number of negative examples is much more than the positive ones, the cumulative misclassification cost of the positive points is relatively small, therefore SVM tends to maximize the margin to its highest possible degree by classifying most (sometimes all) of the examples as negative. Thus, the decision boundary will be shifted toward the minority class region. In the next section, we describe our proposed remedy to this problem.

5.2. SIMO and W-SIMO Algorithms

In this study, we developed a novel synthetic informative minority over-sampling (SIMO) algorithm embedded into SVM. As we mentioned earlier, when SVM is applied to an imbalanced dataset, the decision boundary will be closer to the minority class space in favor of the majority class examples. Therefore, a considerable portion of minority examples will be misclassified. In SIMO, we generate synthetic data points that belong to the minority class. In this way, the distribution of the dataset will be more balanced and a better performance will be expected from machine learning techniques. Research has shown the data points that are close to the boundary of classes are the important data points in forming the classifiers [76]. Therefore, in SIMO we focus on the minority data points near the boundary of two classes.

The first step in performing SIMO (Algorithm 5.1) is to partition the dataset into training and test datasets. This partitioning is conducted in a way that the imbalance ratio in training and test datasets will be the same as the imbalance ratio in the original dataset. The reason for partitioning the data is to avoid biases and to assess the SIMO performance fairly on imbalanced data with the original imbalance ratio (test dataset). Next, we calculate the imbalanced gap in the training dataset. Imbalanced gap is the difference between the number of majority examples and minority examples in the training dataset. Imbalanced gap is the upper bound for generating the synthetic data points in our algorithm. In the next stage, we develop a SVM on the original imbalanced

training dataset and evaluate this initial model by computing the G mean. As it can be seen in Figure 5.2a, the initial SVM decision boundary is close to the minority class data space in favor of majority class data space and the ideal decision boundary should be located farther away from the minority dataspace.

The next step in SIMO is to calculate the Euclidean distance of the minority data points from the SVM decision boundary. As we mentioned earlier, data points close to the boundary of classes are important and informative. In order to select the informative minority data points, we identify those that are close to the SVM decision boundary. Therefore, after calculating the Euclidean distance of the minority data points from the decision boundary, the top Δ% of them that are the closest ones to the decision boundary will be selected as informative minority data points (Figure 5.2b). Next, we generate synthetic data points in the space of the informative minority examples and append the generated data points to the training dataset. At this stage, we have a new training dataset that includes more minority examples compared to the previous training dataset (Figure 5.2c). The number of synthetically generated data points and their indices will be recorded at each iteration. Next, a new SVM will be developed on the updated training dataset. The decision boundary of this new SVM will be shifted toward the majority class data space closer to the ideal decision boundary (Figure 5.2d). The reason is that by generating synthetic minority examples, the imbalance ratio of the training dataset will be reduced and following that, the imbalance ratio of the support vectors will be alleviated. Therefore, the decision boundary will be shifted toward the majority class dataspace (As we discussed in detail in Section 5.1, the position of the SVM decision boundary only depends on the support vectors). The new SVM will be assessed by computing the G mean, and the G mean will be logged into a vector for further evaluations. Again, in the updated training dataset, the Euclidean distance of the minority data points from the new SVM decision boundary is calculated, informative ones will be selected, and new synthetic minority data points will be generated. Another SVM will be developed on the updated dataset,

the SVM will be assessed, and the results will be recorded. These steps will be repeated until the number of synthetically generated examples reaches the imbalanced gap.

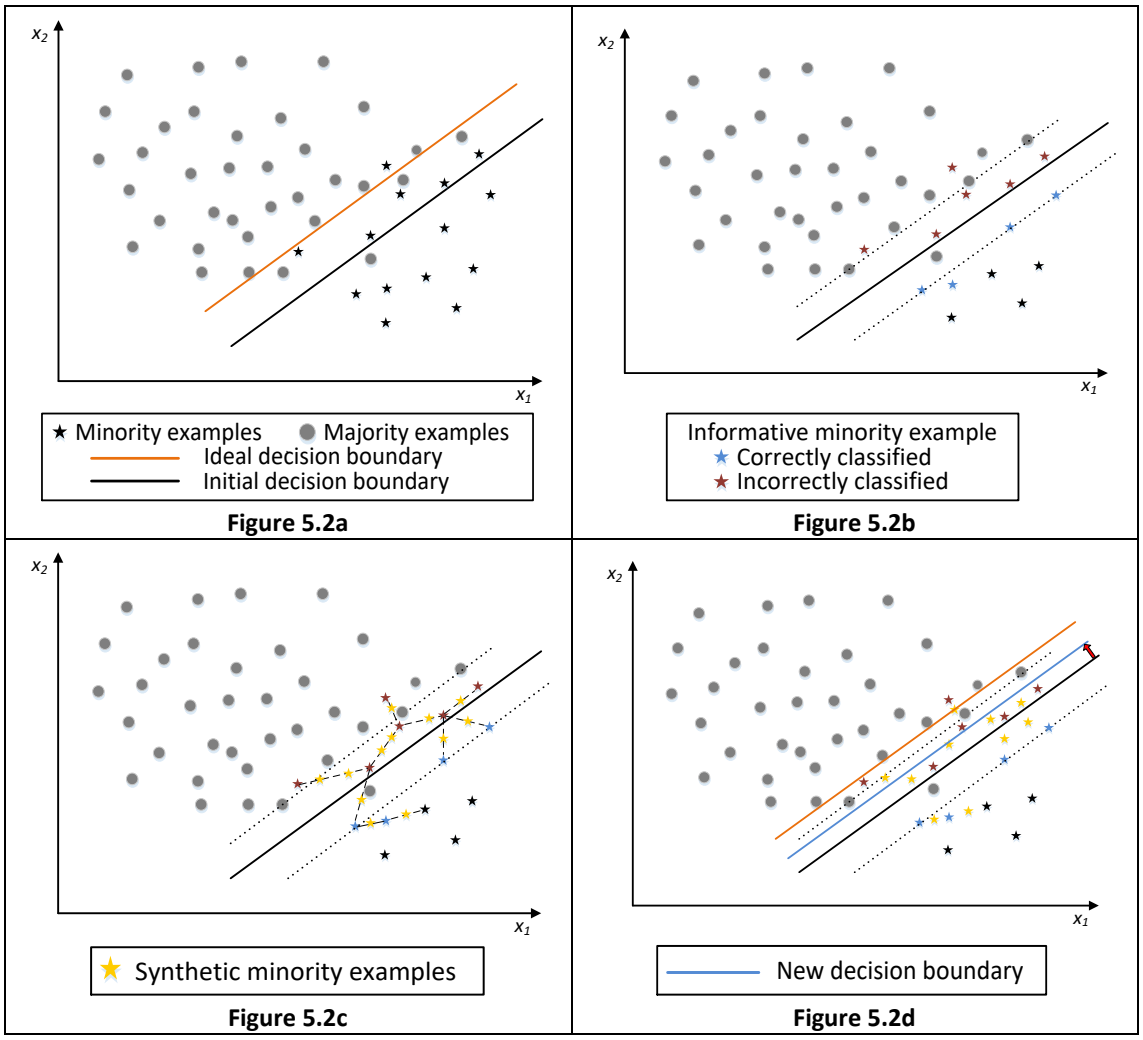


Figure 5.2- *SIMO algorithm mechanism (simplified)*

The performance of machine learning techniques highly depends on the structure and complexity of datasets. In our algorithm, in each iteration, we create a new updated dataset by generating more synthetic minority examples. Even though the performance of the SVM improves on the updated training datasets compared to the original imbalanced dataset, the improvement in the performance of the SVM in each iteration compared to the previous iteration is not guaranteed in all datasets. In the other words, the G mean might not always be increasing through the iterations.

Therefore, we keep track of the SVM performances and their corresponding training dataset in each iteration. At the end of the loop, the best performing model is identified by comparing the G mean values, and the training dataset associated with that model/iteration will be selected as the final over-sampled training dataset.

In this study, we proposed another version of SIMO that we call weighted synthetic informative minority over-sampling (W-SIMO). Steps 1 to 8 in W-SIMO (Algorithm 5.2) are the same as SIMO, i.e. an initial SVM is developed on the original imbalanced training dataset, and top Δ% minority data points close to the SVM decision boundary are identified as informative minority examples. In the next step, informative minority examples will be classified into two groups: first, those that are correctly classified ($S^{+}_{\text{inf\_c}}$) through SVM, and second, those that are incorrectly classified ($S^{+}_{\text{inf\_ic}}$). The data points in $S^{+}_{\text{inf\_ic}}$ will be over-sampled to a higher degree compared to the data points in $S^{+}_{\text{inf\_c}}$. We adopt this idea from AdaBoost, which pays more attention to the incorrectly classified examples [89]. In W-SIMO, by over-sampling the examples in $S^{+}_{\text{inf\_ic}}$ with a higher degree, we consider them even more informative compared to the examples in $S^{+}_{\text{inf\_c}}$. It means that more synthetic minority data points will be generated in the space of $S^{+}_{\text{inf\_ic}}$ examples. After over-sampling (synthetically generating) the informative minority data points in $S^{+}_{\text{inf\_ic}}$ and $S^{+}_{\text{inf\_c}}$, the reminder of the W-SIMO is similar to SIMO. Applying the SIMO and W-SIMO is not limited to the SVM. We use SVM in our algorithms to identify informative data points to over sample them, however the final over-sampled data can be used in any other machine learning technique, such as decision tree, logistic regression, and random forest. The notations of Algorithms 5.1 and 5.2 are shown in Table 5.1.

Table 5.1-*Notations for SIMO and W-SIMO algorithms*

Notations
D: Initial imbalanced dataset
$\widehat{\boldsymbol{S}}$**:** Initial imbalanced training dataset
T: Imbalanced test dataset
Δ**:** Top Δ% of minority data points close to decision boundary
$\boldsymbol{p}$**:** Oversampling degree for minority informative data points that are correctly classified at each iteration
$\boldsymbol{P}$**:** Oversampling degree for minority informative data points that are incorrectly classified at each iteration
$\boldsymbol{S\_G\_D}$**:** Synthetic generated data points count
$\boldsymbol{Max\_I}$**:** Maximum iteration number
$\boldsymbol{G\_m\_L}$**:** $\boldsymbol{G\ mean}$ variation log in each iteration

Algorithm 5.1- SIMO

Given D, Δ, p, Max_I

1. Partition D into Training $\hat{S}$, and Test T datasets
2. Calculate the $Imbalanced\_Gap$ in $\hat{S}$

$$Imbalance\_Gap = Majority\_Count - Minority\_Count$$

3. Develop the $Initial\ SVM$ model on $\hat{S}$, $Initial\ SVM$ decision boundary: $\widehat{D\_B} = \hat{w}^T x + \hat{b}$ $\{\hat{w} = \sum_{j=1}^{N_s} \hat{\alpha}_i y_i\, \phi(x_i)\}$
4. Compute $G\ mean$ for $Initial\ SVM$ on T: $Initial\_G\ Mean$
5. $S = \hat{S}$, $I = 0$ $(iteration)$, $SVM = Initial\ SVM$, $D\_B = \widehat{D\_B}$, $G\_m\_L = Initial\_G\ Mean$, $S\_G\_D = 0$

While $S\_G\_D < Imbalance\_Gap$ **AND** $I < Max\_I$

6. $I = I + 1$
7. Calculate the Euclidean distance of minority data points form D_B

$$Euc\_D(x^{k+}) = \frac{|\sum_{t=1}^{m} w_t x_t^{k+} + b|}{\sqrt{\sum_{t=1}^{m} {w_t}^2}}$$

8. Identify informative minority data points: S_{inf}^{+}
   Top Δ% of minority data points close to D_B based on the Euclidean distance
9. Over-sample data points in S_{inf}^{+} by p%, name the synthetic generated data points $\acute{S}_{inf}^{+}$
10. $S = S \cup \acute{S}_{inf}^{+}$
11. Calculate the number of synthetic generated data points

$$S\_G\_D = S\_count - \hat{S}\_count$$

12. Develop a support vector machine on S, SVM
13. Compute $G\ mean$ for SVM on T
14. Add the $G\ mean$ to the G_m_L, $(G\_m\_L = [G\_m\_L; G\ mean])$

End

15. Find the maximum $G\ mean$ and its index in G_m_L
16. Select the over-sampled training dataset associated with the maximum $G\ mean$
17. Train the model of interest on the final over-sampled training dataset
18. Evaluate the model on the test dataset by computing the $G\ mean$ and AUC

Algorithm 5.2- W-SIMO

Given $D, \Delta, p, P, (p < P)$, Max_I

1. Partition D into Training $\hat{S}$, and Test T datasets
2. Calculate the $Imbalanced\_Gap$ in $\hat{S}$

$$Imbalance\_Gap = Majority\_Count - Minority\_Count$$

3. Develop the $Initial\ SVM$ model on $\hat{S}$, $Initial\ SVM$ decision boundary: $\widehat{D\_B} = \hat{w}^T x + \hat{b}$

$$\{\hat{w} = \sum_{j=1}^{N_s} \hat{\alpha}_i y_i\, \phi(x_i)\}$$

4. Compute $G\ mean$ for $Initial\ SVM$ on T: $Initial\_G\ Mean$
5. $S = \hat{S}$, $I = 0$ $(iteration)$, $SVM = Initial\ SVM$, $D\_B = \widehat{D\_B}$, $G\_m\_L = Initial\_G\ Mean$, $S\_G\_D = 0$

While $S\_G\_D < Imbalance\_Gap$ **AND** $I < Max\_I$

6. $I = I + 1$
7. Calculate the Euclidean distance of minority data points form D_B

$$Euc\_D(x^{k+}) = \frac{|\sum_{t=1}^{m} w_t x_t^{k+} + b|}{\sqrt{\sum_{t=1}^{m} {w_t}^2}}$$

8. Identify informative minority data points: S_{inf}^{+}
   Top $\Delta\%$ of minority data points close to D_B based on the Euclidean distance
9. Classify informative minority data points using the SVM model, form:
   i. $S_{\text{inf\_c}}^{+}$, informative minority data points that are ***correctly*** classified
   ii. $S_{\text{inf\_}ic}^{+}$, informative minority data points that are ***incorrectly*** classified
10. Over-sample data points in $S_{\text{inf\_c}}^{+}$ by $p\%$, name the synthetic generated data points $\acute{S}_i^{+}$
11. Over-sample data points in $S_{\text{inf\_}ic}^{+}$ by $P\%$, name the synthetic generated data points $\acute{S}_{ii}^{+}$
12. $S = S \cup \acute{S}_i^{+} \cup \acute{S}_{ii}^{+}$
13. Calculate the number of synthetic generated data points

$$S\_G\_D = S\_count - \hat{S}\_count$$

14. Develop a support vector machine on S, SVM
15. Compute $G\ mean$ for SVM on T
16. Add the $G\ mean$ to the G_m_L, $(G\_m\_L = [G\_m\_L; G\ mean])$

End

17. Find the maximum $G\ mean$ and its index in G_m_L
18. Select the over-sampled training dataset associated with the maximum $G\ mean$
19. Train the model of interest on the final over-sampled training dataset
20. Evaluate the model on the test dataset by computing the $G\ mean$ and AUC

5.3. Numerical Experiments

In this section, we provide the results of our numerical experiments to assess the performance of SIMO and W-SIMO compared to other existing algorithms in imbalanced data learning. First, we describe the evaluation metrics that we used for the assessments. Second, we provide the characteristics of the benchmark imbalanced datasets that we used. Finally, we present the results of the numerical experiments.

Evaluation Metrics

In classification or pattern recognition problems confusion matrix plays an important role to assess the predictive models. Figure 5.3 shows a confusion matrix. As was pointed out earlier, in this study, we consider the minority class as positive, and the majority class as negative class. Accuracy of prediction (Formulation 5.10) is a common evaluation metric in the balanced datasets; however, it is misleading in assessing the predictive models when applied in imbalanced datasets. Consider an imbalanced dataset with the 10% rate of the positive examples. Because negative examples outnumber the positive ones, simply classifying all of the examples as negative will result in a 90% accuracy. Therefore, in imbalanced datasets other appropriate evaluation metrics such as sensitivity, specificity, G mean, and AUC should be applied [130].

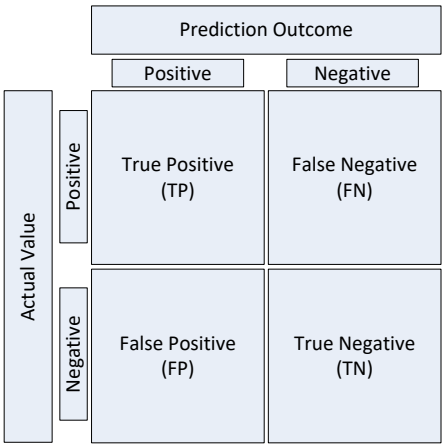


Figure 5.3-*Confusion Matrix*

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{5.10}$$

Sensitivity or true positive rate (TPR) (it is also called hit rate or recall) is a metric that evaluates the accuracy of predicting the positive examples. On the other hand, specificity or true negative rate (TNR) assesses the accuracy of detecting the negative examples. Formulations 5.11 and 5.12 show the calculation of TPR and TNR.

$$TPR = \frac{TP}{TP+FN} \tag{5.11}$$

$$TNR = \frac{TN}{FP+TN} \tag{5.12}$$

TPR and TNR assess the detection accuracy in positive and negative examples separately. Therefore, considering one of them without the other one would not be helpful, therefore, we need a metric such as G mean that incorporates these two metrics at the same time. G mean is the geometric mean of TPR and TNR (Formulation 5.13). Thus, any model with poor performance on either positive or negative examples will have a low G mean.

$$G\ mean = \sqrt{TPR \times TNR} \tag{5.13}$$

Another assessment tool that is independent of the data distribution is Receiving Operator Characteristic (ROC) chart. ROC shows the tradeoff between TPR and TNR by manipulating the decision cut-off. Decision cut-off is the threshold value for decision making based on the output of a predictive model. When the decision cut-off for a model is 0, all of the examples will be classified as positive, therefore TPR=100% but TNR=0%. On the other hand, if decision cut-off is 1, TPR=0% and TNR=100%. Thus, by changing the decision cut-off from 1 to 0, we can increase the TPR, and TNR will decrease at the same time. In ROC chart, the x-axis shows the 1-TNR and y-axis denotes the TPR, in this way the graph will be increasing. Each point on the ROC chart shows the value of TPR and 1-TNR for a specific decision cut-off value. The closer the ROC chart to the top left point, the better the performance of the classifier. Figure 5.4 shows a ROC chart, the 45-degree line is the base line model (random), the dash line corresponds to a good performing model, and dotted line is for the perfect model. An easier way to assess the models and compare different classifiers is to measure the area under the curve (AUC) in ROC chart. AUC takes values between 0 to 100%. AUC for the base line model is 50%, and therefore, classifiers with AUC below 50% are even worse than random guess. The closer the AUC of classifier to 100%, the better the performance of the classifier.

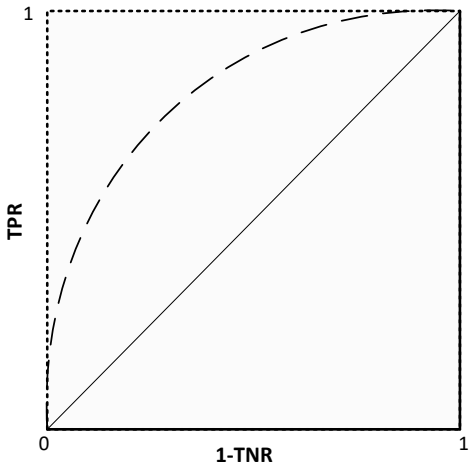


Figure 5.4-*ROC chart*

Datasets

In this study, we used 15 benchmark imbalanced datasets that are publicly available in UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). We tried to use datasets with various imbalance ratios from 1:1.38 to 1:8.9, i.e. the percentage of minority class in the benchmark datasets ranges from 42% to 10%. To test SIMO and W-SIMO on datasets with more severe imbalance ratio, we randomly removed some portions of minority class examples from Breast Cancer dataset and generated datasets with 1:3.91(BreastC20 dataset) and 1:8.9 (BreastC10) imbalance ratio. Table 5.2 shows the name and characteristics of these datasets.

Results

In this study, we compared the performance of our algorithms, SIMO and W-SIMO to six other exiting approaches in imbalanced data learning. We also provided the modeling results on the original imbalanced data for reference. For all of the algorithms, we used the parameters suggested by their developers. We assessed SIMO and W-SIMO in comparison with these algorithms: under-sampling, SMOTE, borderline SMOTE, safe-level SMOTE, cluster SMOTE, and cost sensitive SVM. In cost sensitive SVM, we assigned the error cost of the two classes

based on the imbalance ratio in the dataset. For instance, if the imbalance ratio in a data is 1:4, the error cost for the minority class is 4 times greater that the error cost for majority class.

Table 5.2 *-Benchmark datasets characteristics*

Dataset	Minority class	Majority class	# of variables	# of records	Imbalance ratio
Liver Disorders (Liver)	"1"	"2"	7	345	1 : 1.38
Ionosphere	bad	good	34	351	1 : 1.79
Pima Indians Diabetes (Pima)	"1"	"0"	8	768	1 : 1.87
Breast Cancer Wisconsin Original (BreastCO)	malignant	benign	10	699	1 : 1.91
Iris	Versicolor	All other	5	150	1 : 2
Yeast	NUC	All other	8	1484	1 : 2.6
Statlog Vehicle Silhouettes (Vehicle)	van	All other	18	846	1 : 3.25
Contraceptive Method Choice (CMC)	Long-term	All other	9	1473	1 : 3.42
Breast Cancer Wisconsin_20% (BreastC20)	malignant	benign	10	699	1 : 3.91
Connectionist Bench_Vowel Recognition (Vowel)	"0" & "1"	All other	11	990	1 : 4.5
Ecoli	pp	All other	8	336	1 : 5.46
Libras Movement_12 (Libras12)	"1" & "2"	All other	91	360	1 : 5.88
Libras Movement_34 (Libras34)	"3" & "4"	All other	91	360	1 : 6.34
Glass Identification (Glass)	"7"	All other	9	214	1 : 6.38
Breast Cancer Wisconsin_10% (BreastC10)	malignant	benign	10	699	1 : 8.9

To avoid over-fitting and fairly assess the generalizability and performance of various approaches, we applied 4-fold cross validation in our numerical experiments [74]. In a 4-fold cross validation, the original dataset is partitioned into four mutually exclusive and exhaustive subsets with equal sizes (Sub_1, Sub_2, Sub_3, and Sub_4). Then, the models are developed four times, each time the model is trained on three of the subsets, and is tested on the fourth one. The final performance will be the average of the models 1, 2, 3, and 4. Figure 5.5 shows the mechanism of 4-fold cross validation.

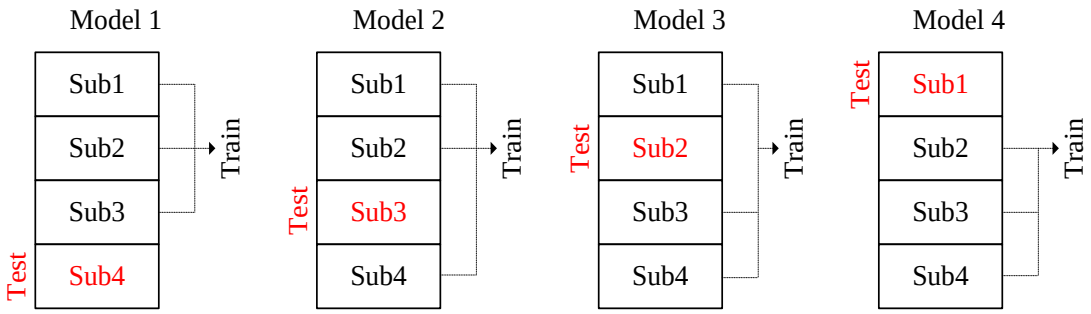


***Figure 5.5**- 4-fold cross validation mechanism*

In order to further reduce the effect of randomness, we ran each 4-fold cross validation on all approaches 10 times. Therefore, each approach has been applied to each dataset 40 times. As a result, for each evaluation metric we have both average value and 95% confidence interval. Tables 5.3 and 5.4 show the performance of all of the 8 imbalanced data learning approaches as well as the learning from original imbalanced dataset in a linear SVM classifier. The first row for each dataset in this table shows the evaluation metric average (G mean in Table 5.3 and AUC in Table 5.4), the second row shows the half of the 95% confidence interval width (HCI) for the evaluation metric, and the third row shows the performance ranking of each approach compared to other approaches.

As it can be seen in Tables 5.3 and 5.4, in all of the 15 imbalanced datasets, our proposed algorithms, SIMO and W-SIMO had the best performance compared to other approaches (approaches with ranks 1, 2, and 3 are **bolded** in Tables 5.3 and 5.4). In addition, the difference between the G mean and AUC value for SIMO and W-SIMO and other approaches is significant. To show this difference and the achieved improvement thorough applying our algorithm, we calculated the difference between the G mean and AUC of our algorithm and the G mean and AUC of the best algorithm among other approaches (the approach with rank 3 in Tables 5.3 and 5.4) in all datasets. We also calculated the difference between the G mean and AUC of the best and second best algorithms among approaches other than our algorithm (the approaches with rank 3 and 4 in Tables 5.3 and 5.4) in all 15 benchmark datasets. Table 5.5 shows the average of these

Table 5.3- *Comparing the performance of various imbalance data learning approaches (using G mean)*

		Original Data	Under Sampling	SMOTE	BorSMOTE	Safe Level SMOTE	Cluster SMOTE	Cost Sensitive	SIMO	W-SIMO
Liver	G mean	64.58%	65.25%	64.91%	64.47%	64.86%	**65.62%**	65.18%	**68.41%**	**68.81%**
	95% HCI	0.90%	1.68%	1.65%	1.24%	2.22%	1.17%	1.13%	1.08%	1.50%
	Rank	8	4	6	9	7	**3**	5	**2**	**1**
Ionosphere	G mean	82.68%	82.54%	**83.25%**	82.22%	82.43%	83.24%	83.06%	**84.69%**	**84.77%**
	95% HCI	1.96%	1.39%	1.47%	1.42%	2.10%	1.84%	1.03%	0.90%	1.76%
	Rank	6	7	**3**	9	8	4	5	**2**	**1**
Pima	G mean	70.08%	73.83%	74.25%	72.97%	**74.31%**	74.22%	74.08%	**75.40%**	**76.05%**
	95% HCI	0.90%	1.25%	0.88%	0.67%	0.75%	0.89%	1.13%	0.69%	0.65%
	Rank	9	7	4	8	**3**	5	6	**2**	**1**
BreastCO	G mean	96.56%	96.65%	96.81%	96.56%	96.87%	96.85%	**96.97%**	**97.63%**	**97.69%**
	95% HCI	0.30%	0.63%	0.35%	0.42%	0.37%	0.42%	0.32%	0.28%	0.24%
	Rank	9	7	6	8	4	5	**3**	**2**	**1**
Iris	G mean	57.68%	71.91%	74.44%	72.20%	75.14%	**75.56%**	74.19%	**77.94%**	**78.31%**
	95% HCI	7.65%	2.62%	2.58%	3.03%	1.54%	2.39%	2.75%	1.79%	1.41%
	Rank	9	8	5	7	4	**3**	6	**2**	**1**
Yeast	G mean	40.85%	70.30%	70.60%	69.12%	70.53%	**70.95%**	70.88%	**71.95%**	**71.80%**
	95% HCI	0.54%	0.77%	0.61%	0.53%	0.48%	0.54%	0.62%	0.38%	0.46%
	Rank	9	7	5	8	6	**3**	4	**1**	**2**
Vehicle	G mean	95.54%	95.78%	**95.83%**	95.60%	95.74%	95.59%	95.64%	**96.52%**	**96.68%**
	95% HCI	0.50%	0.52%	0.58%	0.58%	0.46%	0.45%	0.62%	0.30%	0.49%
	Rank	9	4	**3**	7	5	8	6	**2**	**1**
CMC	G mean	0.00%	65.20%	65.09%	64.52%	65.35%	65.10%	**65.35%**	**66.01%**	**66.36%**
	95% HCI	0.00%	1.17%	0.70%	0.47%	0.61%	0.54%	0.25%	1.00%	0.49%
	Rank	9	5	7	8	4	6	**3**	**2**	**1**
BreastC20	G mean	96.06%	95.82%	95.72%	**96.12%**	95.74%	95.85%	95.84%	**97.27%**	**97.28%**
	95% HCI	0.47%	0.81%	0.67%	0.49%	0.74%	0.42%	0.72%	0.52%	0.39%
	Rank	4	7	9	**3**	8	5	6	**2**	**1**
Vowel	G mean	86.70%	87.88%	88.90%	89.71%	88.65%	**89.71%**	88.10%	**90.94%**	**91.07%**
	95% HCI	0.55%	0.88%	0.57%	0.49%	0.98%	0.90%	0.74%	0.78%	0.97%
	Rank	9	8	5	4	6	**3**	7	**2**	**1**
Ecoli	G mean	71.17%	89.24%	89.48%	84.77%	89.79%	**89.90%**	89.75%	**91.59%**	**91.82%**
	95% HCI	1.62%	1.10%	1.12%	0.99%	0.65%	0.78%	0.79%	0.74%	0.40%
	Rank	9	7	6	8	4	**3**	5	**2**	**1**
Libras12	G mean	66.21%	33.69%	43.66%	44.41%	43.74%	48.12%	**70.42%**	**86.89%**	**85.37%**
	95% HCI	5.32%	3.51%	0.96%	2.45%	0.90%	1.31%	6.62%	1.67%	2.47%
	Rank	4	9	8	6	7	5	**3**	**1**	**2**
Libras34	G mean	83.89%	88.69%	89.30%	87.28%	**89.84%**	88.19%	89.50%	**91.39%**	**91.76%**
	95% HCI	3.24%	2.51%	1.71%	1.99%	1.56%	1.13%	1.85%	1.68%	1.31%
	Rank	9	6	5	8	**3**	7	4	**2**	**1**
Glass	G mean	91.61%	91.23%	91.35%	91.16%	90.97%	**91.79%**	91.17%	**92.57%**	**92.62%**
	95% HCI	2.29%	1.30%	2.02%	1.67%	1.57%	1.22%	1.08%	1.54%	1.14%
	Rank	4	6	5	8	9	**3**	7	**2**	**1**
BreastC10	G mean	93.84%	**95.28%**	94.49%	94.11%	94.67%	94.32%	94.56%	**95.68%**	**95.77%**
	95% HCI	1.63%	1.34%	0.84%	1.04%	1.00%	0.41%	0.85%	0.72%	0.83%
	Rank	9	**3**	6	8	4	7	5	**2**	**1**

Table 5.4- Comparing the performance of various imbalance data learning approaches (using AUC)

		Original Data	Under Sampling	SMOTE	BorSMOTE	Safe Level SMOTE	Cluster SMOTE	Cost Sensitive	SIMO	W-SIMO
Liver	AUC	**66.25%**	65.50%	65.14%	64.71%	65.11%	65.87%	65.45%	**68.76%**	**69.18%**
	95% HCI	0.70%	1.63%	1.56%	1.29%	2.06%	1.18%	1.08%	0.95%	1.38%
	Rank	**3**	5	7	9	8	4	6	**2**	**1**
Ionosphere	AUC	83.81%	83.18%	83.87%	82.66%	83.26%	**84.01%**	83.30%	**85.22%**	**85.37%**
	95% HCI	1.72%	1.18%	1.33%	1.33%	1.86%	1.65%	0.94%	0.84%	1.75%
	Rank	5	8	4	9	7	**3**	6	**2**	**1**
Pima	AUC	71.99%	74.02%	74.41%	73.20%	**74.46%**	74.37%	74.26%	**75.63%**	**76.26%**
	95% HCI	0.73%	1.21%	0.85%	0.61%	0.72%	0.88%	1.06%	0.65%	0.64%
	Rank	9	7	4	8	**3**	5	6	**2**	**1**
BreastCO	AUC	96.57%	96.66%	96.82%	96.58%	96.88%	96.86%	**96.98%**	**97.65%**	**97.72%**
	95% HCI	0.30%	0.63%	0.35%	0.42%	0.37%	0.42%	0.31%	0.28%	0.24%
	Rank	9	7	6	8	4	5	**3**	**2**	**1**
Iris	AUC	63.78%	73.37%	75.40%	73.77%	75.90%	**76.51%**	75.17%	**78.93%**	**79.25%**
	95% HCI	3.27%	2.64%	2.64%	2.94%	1.53%	2.05%	2.47%	1.44%	1.47%
	Rank	9	8	5	7	4	**3**	6	**2**	**1**
Yeast	AUC	57.33%	70.45%	70.75%	70.11%	70.62%	**71.06%**	70.97%	**72.15%**	**72.01%**
	95% HCI	0.19%	0.70%	0.49%	0.54%	0.41%	0.51%	0.54%	0.44%	0.46%
	Rank	9	7	5	8	6	**3**	4	**1**	**2**
Vehicle	AUC	95.59%	95.81%	**95.87%**	95.65%	95.78%	95.63%	95.68%	**96.54%**	**96.71%**
	95% HCI	0.49%	0.52%	0.57%	0.58%	0.45%	0.44%	0.60%	0.29%	0.48%
	Rank	9	4	**3**	7	5	8	6	**2**	**1**
CMC	AUC	50.00%	65.60%	65.41%	65.60%	65.57%	65.58%	**65.61%**	**66.30%**	**66.57%**
	95% HCI	0.00%	1.17%	0.65%	0.42%	0.56%	0.53%	0.23%	0.71%	0.61%
	Rank	9	5	8	4	7	6	**3**	**2**	**1**
BreastC20	AUC	96.09%	95.84%	95.75%	**96.13%**	95.77%	95.87%	95.86%	**97.28%**	**97.32%**
	95% HCI	0.45%	0.80%	0.66%	0.48%	0.73%	0.42%	0.72%	0.51%	0.38%
	Rank	4	7	9	**3**	8	5	6	**2**	**1**
Vowel	AUC	87.35%	87.94%	88.95%	89.82%	88.71%	**89.83%**	88.29%	**91.04%**	**91.22%**
	95% HCI	0.50%	0.86%	0.57%	0.48%	0.97%	0.88%	0.71%	0.74%	0.87%
	Rank	9	8	5	4	6	**3**	7	**2**	**1**
Ecoli	AUC	75.18%	89.42%	89.63%	85.03%	89.94%	**90.02%**	89.91%	**91.70%**	**91.91%**
	95% HCI	1.16%	1.07%	1.15%	0.96%	0.63%	0.80%	0.77%	0.72%	0.40%
	Rank	9	7	6	8	4	**3**	5	**2**	**1**
Libras12	AUC	72.67%	55.35%	55.36%	51.65%	54.76%	55.37%	**73.76%**	**87.97%**	**86.71%**
	95% HCI	3.13%	1.71%	1.41%	2.16%	1.48%	1.65%	5.36%	1.41%	1.91%
	Rank	4	7	6	9	8	5	**3**	**1**	**2**
Libras34	AUC	85.02%	88.88%	89.55%	87.59%	**90.11%**	88.52%	89.68%	**91.57%**	**91.92%**
	95% HCI	2.61%	2.50%	1.63%	1.88%	1.42%	1.07%	1.79%	1.61%	1.27%
	Rank	9	6	5	8	**3**	7	4	**2**	**1**
Glass	AUC	92.06%	91.56%	91.68%	91.50%	91.44%	**92.09%**	91.47%	**92.86%**	**92.98%**
	95% HCI	2.06%	1.21%	1.88%	1.57%	1.45%	1.14%	1.11%	1.37%	1.01%
	Rank	4	6	5	7	9	**3**	8	**2**	**1**
BreastC10	AUC	94.04%	**95.35%**	94.60%	94.23%	94.80%	94.43%	94.65%	**95.77%**	**95.86%**
	95% HCI	1.55%	1.29%	0.77%	0.92%	0.95%	0.39%	0.76%	0.69%	0.78%
	Rank	9	**3**	6	8	4	7	5	**2**	**1**

differences in all datasets. We ran a *t*-test to compare the improvement from the approach with rank 3 to our algorithm with the achieved improvement from the approach with rank 4 to the approach with rank 3 (best and second best approaches not including SIMO and W-SIMO). The *p*-values for G mean and AUC were 0.0091 and 0.0122 respectively. Therefore, the *t*-test showed that the difference between our algorithm and the best algorithm among other existing approaches was significantly greater than the difference between the approaches with rank 3 and 4 at the confidence level of 95%. Table 5.6 demonstrates the overall ranking of SIMO and W-SIMO compared to other imbalanced data learning approaches when applied to linear SVM. The overall ranking is calculated based on the average of various approaches' ranking in 15 benchmark datasets. Since W-SIMO and SIMO had the first and second places in all datasets, their overall ranking is 1.1 and 1.9 respectively.

As we mentioned earlier, the oversampled training data by SIMO and W-SIMO can be used in any other machine learning technique. Therefore, SIMO and W-SIMO can be considered pre-processing oversampling algorithms. To evaluate the performance of SIMO and W-SIMO in other data mining techniques, we applied them in SVM with RBF kernel function, logistic regression, and decision tree. Tables 5.7, 5.8, and 5.9 present the overall rankings of our algorithms as well as their counterparts when applied to SVM with RBF kernel, logistic regression, and decision tree in all benchmark datasets. As it can be seen in Tables 5.7, 5.8, and 5.9, the overall ranking of W-SIMO and SIMO is not about 1 and 2, unlike what we observed in Table 5.6. This means that our algorithms were not always the best when applied in machine learning techniques other than linear SVM. In fact, these results were expected since SIMO and W-SIMO are imbedded into linear SVM, therefore, we expected them to have a better performance in linear SVM. Even though our algorithms were not always the best ones in other machine learning techniques, their overall performance was better compared to other approaches.

As Tables 5.7, 5.8, and 5.9 show, either SIMO or W-SIMO was the best overall algorithm in SVM with RBF kernel, logistic regression, and decision tree.

Table 5.5-*Average difference between our algorithm and other approaches*

	Average difference between our algorithm and the best one among other approaches	Average difference between the best and second best ones among other approaches
G mean	2.46%	0.44%
AUC	2.28%	0.24%

Table 5.6-*Overall ranking on linear SVM*

Approach	G Mean	AUC
W-SIMO	1.1	1.1
SIMO	1.9	1.9
Cluster SMOTE	4.7	4.7
Cost Sensitive	5.0	5.2
SMOTE	5.5	5.6
Safe Level SMOTE	5.5	5.7
Under Sampling	6.3	6.3
BorSMOTE	7.3	7.1
Original Data	7.7	7.3

Table 5.7-*Overall ranking- SVM-RBF kernel*

Approach	G Mean	AUC
W-SIMO	3.3	3.0
Cluster SMOTE	3.7	3.7
SIMO	4.4	4.0
Cost Sensitive	4.7	5.1
Under Sampling	4.9	5.1
Safe Level SMOTE	5.3	5.2
SMOTE	5.3	5.4
BorSMOTE	5.9	6.3
Original Data	7.3	7.2

Table 5.8-*Overall ranking on logistic regression*

Approach	G Mean	AUC
W-SIMO	2.9	2.9
Cluster SMOTE	3.1	3.0
SMOTE	3.6	3.7
SIMO	3.9	4.1
Safe Level SMOTE	4.4	4.4
Under Sampling	5.2	5.1
BorSMOTE	5.8	5.8
Original Data	7.1	7.1

Table 5.9- *Overall ranking on decision tree*

Approach	G Mean	AUC
SIMO	3.29	3.07
Under Sampling	3.29	3.21
W-SIMO	3.36	3.29
SMOTE	4.43	4.71
Cluster SMOTE	4.50	4.71
Original Data	5.21	4.93
Safe Level SMOTE	5.21	5.36
BorSMOTE	6.86	6.71

As we noted at the beginning of this chapter, one of the reasons that we used SVM in our algorithm was its great performance and accuracy compared to other machine learning techniques. The results of the numerical experiments in logistic regression and decision tree showed that our algorithm was not always the best in all datasets in these data mining techniques. However, when we compared the best performing algorithms (imbalanced data learning algorithms, such as SIMO, SMOTE, and under-sampling) in each machine learning technique in

each dataset, it turned out that SVM always outperformed other data mining techniques. Therefore, our algorithm might not always have the best performance when applied to logistic regression and decision tree, but its performance in SVM is better and has higher G mean and AUC. Table 5.10 demonstrates these results. For each dataset, we provide the G mean and AUC of the best imbalanced data learning approach in each of the four machine learning techniques, linear SVM, SVM with RBF kernel, logistic regression, and decision tree. The **bold** numbers show the best performing machine learning technique in each dataset and the underlined numbers are the results of our algorithms. Only in three datasets, the best performing model was not incorporated with our algorithm; those cases are shown in ***italic bold***. The output of our algorithm in those cases is shown in parenthesis and they are not much lower than the best performing approaches. Overall, no one can claim that their algorithm is the best performing algorithm in all datasets, because the performance of a technique or algorithm highly depends on the distribution, size, and complexity of datasets, however, the overall performance of algorithms on multiple datasets from various domains can be a fair comparison measure.

Table 5.10-*The performance of best approach in each machine learning technique*

	SVM-Linear		***SVM-RBF***		***Logistic Regression***		***Decision Tree***	
	G mean	*AUC*	*G mean*	*AUC*	*G mean*	*AUC*	*G mean*	*AUC*
Liver	**68.81%**	**69.18%**	62.27%	64.09%	65.09%	66.68%	62.44%	63.17%
Ionosphere	84.77%	85.37%	***94.10%*** (93.96%)	***94.13%*** (94.00%)	81.54%	82.57%	87.59%	87.75%
Pima	**76.05%**	**76.26%**	70.43%	70.54%	74.51%	74.65%	69.24%	69.33%
BreastCO	**97.61%**	**97.62%**	97.04%	97.07%	96.47%	96.49%	94.75%	94.76%
Iris	78.31%	79.25%	**97.06%**	**97.12%**	75.05%	75.40%	94.28%	94.43%
Yeast	**71.95%**	**72.15%**	70.25%	70.34%	70.85%	71.08%	66.18%	66.29%
Vehicle	**96.68%**	**96.71%**	95.83%	95.95%	96.22%	96.24%	91.51%	91.55%
CMC	**66.36%**	**66.57%**	66.15%	66.28%	65.74%	65.88%	61.82%	61.98%
BreastC20	**97.28%**	**97.32%**	97.25%	97.27%	96.35%	96.38%	93.52%	93.57%
Vowel	91.07%	91.22%	**99.58%**	**99.59%**	90.31%	90.36%	95.57%	95.60%
Ecoli	91.82%	91.91%	***93.50%*** (91.83%)	***93.61%*** (92.08%)	90.65%	90.77%	86.10%	86.82%
Libras12	86.89%	87.97%	**97.64%**	**97.71%**	39.37%	42.79%	84.27%	85.76%
Libras34	91.76%	91.92%	***92.97%***	***93.07%***	82.74%	83.05%	82.74%	83.05%
Glass	**92.62%**	**92.98%**	89.33%	89.91%	91.55%	91.89%	92.37%	92.61%
BreastC10	95.77%	95.86%	**96.73%**	**96.79%**	94.68%	94.81%	92.13%	92.33%

Another advantage of our proposed algorithm is that it makes a minimal alteration to the original distribution of the dataset. While other over-sampling approaches generate enough data points to completely fill the imbalanced gap in the data, SIMO and W-SIMO only focus on the informative data points close to the decision boundary between two classes in the data, and therefore, they do not generate as many synthetic data points as other over-sampling methods. Table 5.11 demonstrates the imbalanced gap between majority and minority class in various datasets. It also shows the average number of data points generated by our algorithms as well as other over-sampling approaches. The number in parenthesis shows the amount of the synthetically generated data points as a percentage of the total imbalanced gap in the training datasets. As it can be seen, SIMO and W-SIMO usually generate less number of data points compared to other over-sampling methods. This result shows two advantages of our proposed algorithms. First, our algorithms do not dramatically change the distribution of the data from its original shape. Second, with less amount of data generated, the further computational cost in training the machine learning techniques will be lower.

Table 5.11- *Imbalanced gap and average # of synthetically generated data points (% of the imbalance gap)*

	Imbalanced Gap in Training Data	**Other Approaches**	**SIMO**	**W-SIMO**
Liver	41	41 (100%)	20 (48.8%)	22 (53.7%)
Ionosphere	75	75 (100%)	18 (24%)	22 (29.3%)
Pima	174	174 (100%)	104 (59.8%)	96 (55.1%)
BreastCO	154	154 (100%)	25 (16.2%)	30 (19.5%)
Iris	38	38 (100%)	25 (65.8%)	24 (63.1%)
Yeast	467	467 (100%)	392 (83.9%)	373 (79.9%)
Vehicle	336	336 (100%)	50 (14.9%)	46 (13.7%)
CMC	606	606 (100%)	543 (89.6%)	569 (93.9%)
BreastC20	229	229 (100%)	13 (5.7%)	18 (7.9%)
Vowel	473	473 (100%)	173 (36.6%)	152 (32.1%)
Ecoli	174	174 (100%)	60 (34.5%)	58 (33.3%)
Libras12	175	175 (100%)	29 (16.6%)	17 (9.7%)
Libras34	180	180 (100%)	20 (11.1%)	17 (9.4%)
Glass	117	117 (100%)	5 (4.2%)	6 (5.1%)
BreastC10	282	282 (100%)	35 (12.4%)	19 (6.7%)

Sensitivity Analysis

For applying SIMO and W-SIMO, their parameters, i.e. Δ, p, and P need to be specified. To evaluate the performance of SIMO in different parameters values, we performed a sensitivity analysis. In the sensitivity analysis, we considered values 10% to 50% for Δ, and 5% to 50% for p. Table 5.12 depicts the results of the sensitivity analysis for Δ=10, 20, 30, and 40% and p=10 and 40%. As shown in Table 5.12, different values of parameters do not make a considerable difference in the performance of SIMO. Therefore, SIMO is not very sensitive to the values of its parameters. Moreover, except in 4 cases, in all of the other cases, with even the worst parameters' value, SIMO had a better performance compared to the 3rd best approach. Based on this analysis, we suggest the following policy for choosing the parameters' values. When the imbalance ratio of the data is high (the minority class rate below 20%), it is better to select higher values for Δ and p, i.e. values between 30% to 40% for Δ, and values between 25% to 50% for p. The reason is that because the number of the minority data points in highly imbalanced datasets is very low, by selecting relatively higher values for Δ, we consider greater numbers of minority data points for over-sampling. Therefore, we avoid the potential overfitting. On the other hand, for datasets with lower imbalanced ratio (the minority class rate between 20-40%), choosing lower values for Δ and p will generate better results. Selecting the parameters for W-SIMO follows the same policy with one difference, and that is selecting a higher value for P compared to p. Our suggestion based on the sensitivity analysis is to choose 20% to 30% greater values for P. For example, if p=20%, values between 40% to 50% are appropriate for P.

Table 5.12- *Sensitivity analysis on SIMO parameters*

		SIMO									3rd best approach
		Δ=10%		Δ=20%		Δ=30%		Δ=40%		Best Δ & *p*	
		***p*=10%**	***p*=40%**	***p*=10%**	***p*=40%**	***p*=10%**	***p*=40%**	***p*=10%**	***p*=40%**		
Liver	G mean	**68.41%**	68.23%	68.35%	68.15%	68.15%	68.11%	67.91%	67.90%	Δ=10%	*65.62%*
	AUC	**68.76%**	68.44%	68.70%	68.36%	68.32%	68.20%	68.19%	68.08%	*p*=10%	*66.25%*
Ionosphere	G mean	84.63%	84.48%	**84.69%**	84.07%	84.33%	83.88%	83.98%	83.71%	Δ=20%	*83.25%*
	AUC	85.20%	85.09%	**85.22%**	84.22%	85.02%	84.43%	84.65%	84.40%	*p*=05%	*84.01%*
Pima	G mean	75.28%	75.19%	**75.40%**	75.32%	75.25%	75.15%	75.14%	75.09%	Δ=20%	*74.31%*
	AUC	75.48%	75.40%	**75.63%**	75.54%	75.58%	75.37%	75.36%	75.21%	*p*=10%	*74.46%*
BreastCO	G mean	97.60%	97.57%	**97.63%**	97.62%	97.55%	97.48%	97.36%	97.21%	Δ=15%	*96.97%*
	AUC	97.61%	97.59%	**97.65%**	97.63%	97.56%	97.5%	97.37%	97.22%	*p*=15%	*96.98%*
Iris	G mean	77.86%	77.97%	**78.41%**	77.67%	78.31%	77.59%	78.02%	77.56%	Δ=20%	*75.56%*
	AUC	78.56%	78.70%	**79.18%**	78.40%	79.01%	78.36%	78.78%	78.09%	*p*=25%	*76.51%*
Yeast	G mean	71.29%	**71.81%**	71.71%	71.47%	71.79%	71.26%	71.35%	71.08%	Δ=10%	*70.95%*
	AUC	71.53%	**71.99%**	71.90%	71.70%	71.94%	71.51%	71.61%	71.29%	*p*=15%	*71.06%*
Vehicle	G mean	96.30%	96.34%	96.28%	96.29%	**96.43%**	96.26%	96.34%	96.35%	Δ=30%	*95.83%*
	AUC	96.32%	96.35%	96.31%	96.31%	**96.46%**	96.29%	96.36%	96.37%	*p*=30%	*95.87%*
CMC	G mean	65.49%	65.58%	65.82%	**66.01%**	65.91%	65.87%	65.55%	65.46%	Δ=20%	*65.35%*
	AUC	65.79%	65.87%	66.05%	**66.30%**	66.19%	66.15%	65.85%	65.78%	*p*=40%	*65.61%*
BreastC20	G mean	97.26%	97.12%	**97.27%**	97.10%	96.98%	96.99%	96.93%	96.84%	Δ=12%	*96.12%*
	AUC	97.27%	97.13%	**97.28%**	97.11%	97.00%	97.00%	96.94%	96.85%	*p*=05%	*96.13%*
Vowel	G mean	90.48%	**90.94%**	90.52%	90.73%	90.69%	90.35%	90.64%	90.46%	Δ=10%	*89.71%*
	AUC	90.60%	**91.04%**	90.65%	90.86%	90.80%	90.44%	90.76%	90.58%	*p*=35%	*89.83%*
Ecoli	G mean	90.89%	90.92%	**91.59%**	91.52%	91.58%	91.29%	91.57%	91.37%	Δ=12%	*89.90%*
	AUC	91.10%	91.13%	**91.70%**	91.63%	91.68%	91.39%	91.65%	91.48%	*p*=20%	*90.02%*
Libras12	G mean	86.42%	86.59%	**86.89%**	86.27%	86.22%	85.99%	85.92%	85.98%	Δ=20%	*70.42%*
	AUC	87.60%	87.73%	**87.97%**	87.50%	87.32%	87.08%	87.17%	87.29%	*p*=10%	*73.76%*
Libras34	G mean	90.72%	90.85%	90.95%	91.06%	91.25%	**91.34%**	91.23%	91.18%	Δ=30%	*89.84%*
	AUC	90.90%	91.04%	91.24%	91.27%	91.45%	**91.52%**	91.40%	91.34%	*p*=25%	*90.11%*
Glass	G mean	92.40%	92.16%	92.11%	92.33%	92.30%	**92.57%**	92.37%	92.07%	Δ=30%	*91.79%*
	AUC	92.70%	92.47%	92.44%	92.63%	92.60%	**92.86%**	92.70%	92.42%	*p*=40%	*92.09%*
BreastC10	G mean	94.95%	95.32%	95.13%	95.35%	95.39%	95.17%	95.50%	**95.68%**	Δ=40%	*95.28%*
	AUC	95.10%	95.39%	95.27%	95.44%	95.50%	95.29%	95.61%	**95.77%**	*p*=35%	*95.35%*

5.4. Discussion and Conclusion

Imbalanced datasets are widespread in various domains such as healthcare, finance, and information system security. In an imbalanced dataset, the number of examples belonging to one class outnumbers the number of examples from the other class. Therefore, in an imbalanced dataset, there are majority and minority classes of examples. Training machine learning techniques using imbalanced datasets is a critical challenge in data analytics. The prediction accuracy of a data mining technique, especially prediction accuracy of detecting the minority class in an imbalanced dataset, is inferior to the performance of the same technique when applied to a balanced dataset. There has been an enormous effort to address the problem of imbalanced

data learning in recent years. Sampling methods along with cost sensitive approaches are among the most efficient remedies to the imbalanced data learning problem.

In this study, we proposed a synthetic informative minority oversampling (SIMO) algorithm imbedded into SVM to enhance the performance of machine learning techniques when applied to imbalanced datasets. In this algorithm, first SVM is applied to the original imbalanced dataset. In the next step, minority examples close to the SVM decision boundary are selected as the informative minority examples. Next, these examples are over-sampled to a pre-specified degree. Finally, a new SVM model is developed on the updated dataset. This process iterates until we reach a pre-specified balance level. In each iteration, we have an updated training dataset, which is formed by adding the newly generated data points to the previous dataset. Each of these training datasets is used to develop a SVM model, and the SVM model is assessed on the test dataset. At the end, the best model and its associated training dataset is selected as the final over-sampled training dataset. In this research, we also developed another version of SIMO called W-SIMO. W-SIMO is different from SIMO in the degree of over-sampling the informative minority examples. In W-SIMO, informative minority examples that are incorrectly classified are over-sampled with a higher degree compared to the informative minority examples that are correctly classified. In this way, there is more focus on incorrectly classified minority examples.

SIMO and W-SIMO have several advantages compared to other imbalanced data learning methods. First, they are embedded into SVM, which is a powerful machine learning technique in pattern recognition problems. Second, in SIMO and W-SIMO, we over-sample the minority examples rather than under-sampling the majority examples, therefore we avoid losing potentially useful information by discarding some portion of the data. Third, our focus in SIMO and W-SIMO is only on the data points (examples) near the decision boundary as the informative minority data points. This focus is even more important in W-SIMO where we over-sample the incorrectly classified examples with a higher degree. Therefore, SIMO and W-SIMO concentrate

on the informative minority examples that usually are misclassified by standard machine learning techniques. Fourth, compared to other oversampling methods, SIMO generates fewer synthetic data points. Therefore, the changes to the original distribution of the data and further computational costs will be lower compared to other oversampling approaches. Fifth, the oversampled data through SIMO can be used to train any other machine learning technique, thus its application is not limited only to SVM. Finally, SIMO and W-SIMO are not very sensitive to their parameters, even though we suggest to select higher values for Δ and p in highly imbalanced datasets and lower values in moderately imbalanced datasets.

We applied our algorithms to 15 publicly available benchmark imbalanced datasets and assessed their performance in comparison with existing approaches in the area of imbalanced data learning. These approaches were cost sensitive SVM, under sampling, SMOTE, cluster SMOTE, safe level SMOTE and borderline SMOTE as well as the original imbalanced dataset. Our algorithm had the best performance in all datasets compared to the other seven approaches in the linear SVM. In fact, the difference between our algorithm and second best algorithm was significantly greater than the difference between other algorithms (for instance, the difference between second and third best approaches). Besides linear SVM that SIMO and W-SIMO were embedded into, we also assessed SIMO and W-SIMO in other machine learning techniques such as SVM with RBF kernel, logistic regression, and decision tree. Our algorithms were not always the best in these machine learning techniques in all bench mark datasets, however their overall performances were better than all other imbalanced data learning approaches. Moreover, the results showed that the best performing machine learning technique in all datasets was either linear SVM or SVM with RBF kernel function, and except for in three datasets, our algorithms were the best ones. From the practical implication point of view, our proposed algorithm can enhance the performance of the predictive models and decision support systems in various

domains such as diagnosing diseases, detecting re-admissions, and predicting the loan defaults in financial institutions among other application domains.

Here we briefly explain the performing mechanism of our algorithms, SIMO and W-SIMO and suggest an outline for using them. SIMO and W-SIMO are over-sampling algorithms that aim to decrease the imbalance ratio in imbalanced datasets through generating synthetic data points belonging to the minority class. These algorithms first develop a SVM model on the imbalanced data, and form the SVM decision boundary. Next, minority data points near the decision boundary are identified as informative minority data points. To identify these informative minority data points, the Euclidean distance of all of the minority data points from the SVM decision boundary is computed, then the top Δ% of them, which are the closest ones to the decision boundary are selected. Next, these informative minority data points are over-sampled by generating synthetic data points in their data space. Thus, unlike other general over-sampling approaches such as SMOTE that over-sample the whole minority data points, in SIMO and W-SIMO the focus is on the data points near the boundary of the classes in the data. There are two reasons for focusing on the data points near the decision boundary. First, they are the data points that are hard to classify, and second, they will form the support vectors in SVM that are important for developing the model. After over-sampling the informative minority data-points, a new SVM model is developed on the updated data. This process iterates until we reach to a desired balanced distribution in the data.

To apply SIMO and W-SIMO, Δ, p, and P need to be specified as the parameters of the algorithms. Δ specifies the level of minority data that we want to focus on as informative minority data points. Smaller value for the Δ means that we will only focus on the data points that are very close to the decision boundary. p and P are the degree of over-sampling the informative minority data points. Greater value for p and P is indicative of generating more synthetic data points at each iteration of the algorithms. We suggest the following strategy for choosing these parameters'

value. When the imbalance ratio of the data is high (the minority class rate below 20%), it is better to select higher values for Δ and p, i.e. values between 30% to 40% for Δ, and values between 25% to 50% for p. On the other hand, for datasets with lower imbalanced ratio (the minority class rate between 20-40%), choosing lower values for Δ and p will generate better results. Finally, always select a higher value for P compared to p.

Our proposed algorithms may have a limitation that all of the over-sampling approaches face. This limitation is the computational time when the algorithms are applied to very large size datasets. Even though considering the recent advances in computational power of the computers, the computational time is not as critical as it used to be, we still need to enhance the speed of our algorithms in large size datasets. Therefore, we consider speeding up our algorithms in big data usage as one of the most important directions for future research. One way to achieve higher speed could be decreasing the size of the data thorough approaches such as variable selection before using the data in over-sampling algorithms. Another way could be improving the SVM training algorithms. We are considering another direction for future research, and that is applying our developed algorithms to develop a clinical decision support system for predicting kidney disease among diabetic patients. The dataset that we are going to use for that research contains the lab, demographic, clinical events, and comorbidity data of a large number of diabetic patients. We believe that this future research will reveal the performance and efficiency of our algorithm in a larger imbalanced dataset.

CHAPTER VI

CONCLUSION

In this dissertation we conducted three studies. In the first study, we addressed the rare items problem in association rule mining. This problem emerges when some of the items in the data are not as frequent as others. Even though these items may not occur frequently, they can be very important. Therefore, discovering their associations with other items usually is in the interest of data analysts and managers. To address this problem, we proposed a new assessment metric for evaluating association rules and called this metric adjusted_support. Applying adjusted_support, enabled us to retrieve rare rules without over-generating association rules. In this research, we used adjusted_support in order to discover the association patterns among complications of diabetes. Identifying associations among diabetes complications has several benefits. First, it could lead to better diagnoses of diabetes complications when the existence of some other diabetes-related complications is known. For example, if there exists a strong association between neuropathy and retinopathy, then a patient diagnosed with diabetic neuropathy would also be at risk for diabetic retinopathy. Second, knowledge of strong associations among frequently occurring diabetes complications can help physicians provide more effective intervention and treatment plans. Third, it may provide useful information

Begin for medical scientists to better understand the relationships among different diseases. We discovered several strong associations between various complications of diabetes that can be in the interest of physicians and clinicians. We also analyzed the comorbidity index among various demographic groups of diabetic patients. Finally, we studied the prevalence of diabetes complications in every demographic group of patients and compared them.

In the second research, we addressed the problem of low compliance rate with annual eye examination for diabetic patients that leads to a very high undiagnosed rate of diabetic retinopathy (over 50%). To address this problem, we developed a CDSS for diabetic retinopathy. Our developed CDSS for diabetic retinopathy has several advantages over the existing diagnostic systems. First and foremost, it only uses the results of a simple blood test and demographic data to predict the risk of diabetic retinopathy. Therefore, unlike the dominant approach in the extant literature that uses image processing on images of retina, it does not require eye exams, thereby addressing the low rates of compliance with annual ophthalmologic tests for diabetic patients. Equally important, our CDSS eliminates the need to have access to specialists, which is particularly critical for patients living in remote areas. Second, our decision support system is based on a large database of clinical encounters that span over several years and across several states of the US. The decisions of this system are more generalizable and valid compared to those of other systems that employ a similar approach but only use data from a few hundred patients. Finally, our CDSS uses a greater number of risk factors to predict the outcome. This not only improves the prediction results, but also sheds more light on contribution and importance of different risk factors on diabetic patients' susceptibility to retinopathy.

To develop this CDSS, we proposed a new ensemble approach to further enhance the prediction accuracy. We termed this ensemble approach confidence margin. While exiting ensemble methods assigns various weights at model level, confidence margin calculates different weights at both model and record level. To calculate the weight of each model at each record, the difference

between the model's output and the decision cutoff is calculated (confidence margin). Next, a weight is assigned to models at each record based on the calculated confidence margins. We conducted a numerical analysis to assess the performance of confidence margin ensemble in comparison to other exiting ensemble methods. The results of this analysis showed that confidence margin ensemble had the best performance in most cases (14 out of 18).

In the third research, we addressed the problem of imbalanced data learning. Data mining techniques do not have a very good performance in imbalanced datasets. In this research we developed a synthetic informative minority over-sampling algorithm imbedded into support vector machine in order to enhance the performance of predictive modeling techniques when applied to imbalanced datasets. The proposed algorithm, SIMO, generates synthetic minority data points that are located near the boundary between two classes in the data space. After applying SIMO in an imbalanced dataset, the number of minority class data points will be increased and the dataset will be more balanced. In this research, we developed another version of SIMO, which we call weighted SIMO (W-SIMO). In W-SIMO, after identifying the informative minority examples, they are grouped into two categories. First, those that are correctly classified by the SVM, and second, those that are incorrectly classified by the SVM. At the over-sampling stage, more data points are generated in the space of the minority data examples that are misclassified. The over-sampled dataset through SIMO and W-SIMO can be used by other machine learning techniques and it is not limited only to the SVM.

We performed numerical experiments to evaluate the performance of SIMO and W-SIMO in comparison to exiting imbalanced data learning approaches that are widely used in literature and practice. These approaches are cost-sensitive SVM, under sampling, SMOTE, cluster SMOTE, safe level SMOTE and borderline SMOTE. To conduct the numerical experiments, we used 15 benchmark imbalanced datasets with various imbalance ratio. The results of the numerical experiments showed that SIMO and W-SIMO had the best performance on all benchmark

datasets compared to other existing approaches in linear SVM. We replicated the same experiments in other machine learning techniques, such as decision tree, logistic regression, and SVM with RBF kernel function. Based on the results of these analyses, either SIMO or W-SIMO had the best overall performance in decision tree, logistic regression, and SVM with RBF. Another advantage of SIMO and W-SIMO compared to other over-sampling approaches is that SIMO and W-SIMO generate the least number of synthetic data points. Therefore, the alteration to the original data distribution will be minimal. In addition, because the size of the over-sampled data will be smaller, the computational cost of training predictive models will be lower.

REFERENCES

[1] A. B. Martin, M. Hartman, J. Benson, A. Catlin, and N. H. E. A. Team, "National health spending in 2014: faster growth driven by coverage expansion and prescription drug spending," *Health Affairs*, vol. 35, no. 1, pp. 150-160, 2016.

[2] "IDF Diabetes Atlas, 7th edn.," International Diabetes Federation, Brussels, Belgium International Diabetes Federation2015.

[3] "National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States," Centers for Disease Control and Prevention, Atlanta, GA: U.S. Department of Health and Human Services2014.

[4] R. Wullianallur, "Data Mining in Healthcare," in *Healthcare Informatics*: CRC Press, 2010, pp. 211-224.

[5] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, p. 1, 2014.

[6] J. R. Evans and C. H. Lindner, "Business analytics: the next frontier for decision sciences," *Decision Line*, vol. 43, no. 2, pp. 4-6, 2012.

[7] D. Delen and H. Demirkan, "Data, information and analytics as services," *Decision Support Systems*, vol. 55, no. 1, pp. 359-363, 4// 2013.

[8] D. Delen, *Real-World Data Mining: Applied Business Analytics and Decision Making*. FT Press (a Pearson Publishing Company), 2015.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, 1993, vol. 22, no. 2, pp. 207-216: ACM.

[10] D. S. Fong *et al.*, "Diabetic Retinopathy," *Diabetes Care*, vol. 26, no. suppl 1, pp. s99-s102, January 1, 2003 2003.

[11] P. K. Chan, F. Wei, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *Intelligent Systems and their Applications, IEEE*, vol. 14, no. 6, pp. 67-74, 1999.

[12] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining comorbidity: implications for understanding health and health services," *The Annals of Family Medicine*, vol. 7, no. 4, pp. 357-363, 2009.

[13] D.-C. Suh, I.-S. Choi, C. Plauschinat, J. Kwon, and M. Baron, "Impact of comorbid conditions and race/ethnicity on glycemic control among the US population with type 2 diabetes, 1988–1994 to 1999–2004," *Journal of Diabetes and its Complications*, vol. 24, no. 6, pp. 382-391, 2010.

[14] P. C. Albertsen, D. F. Moore, W. Shih, Y. Lin, H. Li, and G. L. Lu-Yao, "Impact of comorbidity on survival among men with localized prostate cancer," *Journal of Clinical Oncology*, vol. 29, no. 10, pp. 1335-1341, 2011.
[15] A. D. Hanchate, K. M. Clough-Gorr, A. S. Ash, S. S. Thwin, and R. A. Silliman, "Longitudinal patterns in survival, comorbidity, healthcare utilization and quality of care among older women following breast cancer diagnosis," *Journal of general internal medicine*, vol. 25, no. 10, pp. 1045-1050, 2010.
[16] H. M. Zolbanin, D. Delen, and A. Hassan Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decision Support Systems*, vol. 74, pp. 150-161, 6// 2015.
[17] S. Hill *et al.*, "Survival disparities in Indigenous and non-Indigenous New Zealanders with colon cancer: the role of patient comorbidity, treatment and health service factors," *Journal of epidemiology and community health*, vol. 64, no. 2, pp. 117-123, 2010.
[18] H. Teppo and O.-P. Alho, "Comorbidity and diagnostic delay in cancer of the larynx, tongue and pharynx," *Oral oncology*, vol. 45, no. 8, pp. 692-695, 2009.
[19] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000, pp. 1-15: Springer.
[20] E. Dimitriadou, A. Weingessel, and K. Hornik, "A cluster ensembles framework, Design and application of hybrid intelligent systems," ed: IOS Press, Amsterdam, The Netherlands, 2003.
[21] G. Pandey, G. Atluri, F. Gang, R. Gupta, M. Steinbach, and V. Kumar, "Association analysis techniques for analyzing complex biological data sets," in *2009 IEEE International Workshop on Genomic Signal Processing and Statistics*, 2009, pp. 1-4.
[22] C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, and O. Gandrillon, "Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data," *Genome Biology*, vol. 3, no. 12, p. 1, 2002.
[23] S. Naulaerts *et al.*, "A primer to frequent itemset mining for bioinformatics," *Briefings in bioinformatics*, vol. 16, no. 2, pp. 216-231, 2015.
[24] D. Gnatyshak, D. I. Ignatov, A. Semenov, and J. Poelmans, "Gaining insight in social networks with biclustering and triclustering," in *International Conference on Business Informatics Research*, 2012, pp. 162-171: Springer.
[25] R. Rathipriya, K. Thangavel, and J. Bagyamani, "Binary particle swarm optimization based biclustering of web usage data," *arXiv preprint arXiv:1108.0748*, 2011.
[26] P.-N. Tan and V. Kumar, "Mining association patterns in web usage data," 2002.
[27] C. Potter *et al.*, "Global teleconnections of climate to terrestrial carbon flux," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D17, 2003.
[28] C. Potter *et al.*, "Understanding global teleconnections of climate to regional model estimates of Amazon ecosystem carbon fluxes," *Global Change Biology*, vol. 10, no. 5, pp. 693-703, 2004.
[29] H. Xiong, P.-N. Tan, and V. Kumar, "Mining strong affinity association patterns in data sets with skewed support distribution," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 387-394: IEEE.
[30] J. Liu, X. Fan, and Z. Qu, "A new interestingness measure of association rules," in *Genetic and Evolutionary Computing, 2008. WGEC'08. Second International Conference on*, 2008, pp. 393-397: IEEE.
[31] X. Yan, C. Zhang, and S. Zhang, "Confidence Metrics for Association Rule Mining," *Applied Artificial Intelligence*, vol. 23, no. 8, pp. 713-737, 2009.
[32] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," presented at the Proceedings of the eighth ACM SIGKDD

international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, 2002.
[33] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," *GESTS International Transactions on Computer Science and Engineering,* vol. 32, no. 1, pp. 71-82, 2006.
[34] B. Liu, W. Hsu, and Y. Ma, "Mining association rules with multiple minimum supports," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, pp. 337-341: ACM.
[35] H. Yun, D. Ha, B. Hwang, and K. H. Ryu, "Mining association rules on significant rare data using relative support," *Journal of Systems and Software,* vol. 67, no. 3, pp. 181-191, 2003.
[36] K. Wang, Y. He, and J. Han, "Pushing support constraints into association rules mining," *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, no. 3, pp. 642-658, 2003.
[37] M. Seno and L. Karypis, "An algorithm for finding frequent itemsets using length-decreasing support constraint. 2001 IEEE Intl," in *Conf on Data Mining, San Jose*, 2001, pp. 505-512.
[38] E. R. Omiecinski, "Alternative interest measures for mining associations in databases," *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, no. 1, pp. 57-69, 2003.
[39] R. U. Kiran and P. K. Re, "An improved multiple minimum support based approach to mine rare association rules," in *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, 2009, pp. 340-347: IEEE.
[40] Y. S. Koh and S. D. Ravana, "Unsupervised Rare Pattern Mining: A Survey," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 10, no. 4, p. 45, 2016.
[41] G. J. Simon, J. Schrom, M. R. Castro, P. W. Li, and P. J. Caraballo, "Survival association rule mining towards type 2 diabetes risk assessment," in *AMIA Annual Symposium Proceedings*, 2013, vol. 2013, p. 1293: American Medical Informatics Association.
[42] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, "Extending association rule summarization techniques to assess risk of diabetes mellitus," *IEEE Transactions on Knowledge and Data Engineering,* vol. 27, no. 1, pp. 130-141, 2015.
[43] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, and F. Hadaegh, "An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database," *International journal of endocrinology and metabolism,* vol. 13, no. 2, 2015.
[44] M. D. Kamalesh, K. H. Prasanna, B. Bharathi, R. Dhanalakshmi, and R. A. Canessane, "Predicting the Risk of Diabetes Mellitus to Subpopulations Using Association Rule Mining," in *Proceedings of the International Conference on Soft Computing Systems*, 2016, pp. 59-65: Springer.
[45] A. M. Shin *et al.*, "Diagnostic analysis of patients with essential hypertension using association rule mining," *Healthcare informatics research,* vol. 16, no. 2, pp. 77-81, 2010.
[46] F. Valent, S. Tillati, and L. Zanier, "Prevalence and comorbidities of known diabetes in northeastern Italy," *Journal of diabetes investigation,* vol. 4, no. 4, pp. 355-360, 2013.
[47] H. S. Kim, A. M. Shin, M. K. Kim, and Y. N. Kim, "Comorbidity study on type 2 diabetes mellitus using data mining," *The Korean journal of internal medicine,* vol. 27, no. 2, pp. 197-202, 2012.
[48] I. Torre-Díez, B. Martínez-Pérez, M. López-Coronado, J. R. Díaz, and M. M. López, "Decision Support Systems and Applications in Ophthalmology: Literature and

Commercial Review Focused on Mobile Apps," *Journal of Medical Systems,* journal article vol. 39, no. 1, pp. 1-10, 2014.
[49] A. Karma, S. Gummerus, E. Kujansuu, and T. Pitkäjärvi, "Predicting Diabetic Retinopathy," *Acta Ophthalmologica,* vol. 65, no. S182, pp. 136-139, 1987.
[50] R. Klein, B. K. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets, "GLycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy," *JAMA,* vol. 260, no. 19, pp. 2864-2871, 1988.
[51] P. Kahai, K. R. Namuduri, and H. Thompson, "A Decision Support Framework for Automated Screening of Diabetic Retinopathy," *International Journal of Biomedical Imaging,* vol. 2006, p. 8, 2006, Art. no. 45806.
[52] A. Paunksnis, V. Barzdziukas, D. Jegelevicius, S. Kurapkiene, and G. Dzemyda, "The use of information technologies for diagnosis in ophthalmology," *Journal of Telemedicine and Telecare,* vol. 12, no. suppl 1, pp. 37-40, July 1, 2006 2006.
[53] K. Marsolo, M. Twa, M. A. Bullimore, and S. Parthasarathy, "Spatial Modeling and Classification of Corneal Shape," *IEEE Transactions on Information Technology in Biomedicine,* vol. 11, no. 2, pp. 203-212, 2007.
[54] C. L. Tsai *et al.*, "Automated Retinal Image Analysis Over the Internet," *IEEE Transactions on Information Technology in Biomedicine,* vol. 12, no. 4, pp. 480-487, 2008.
[55] K. Noronha, U. Acharya, K. Nayak, S. Kamath, and S. Bhandary, "Decision support system for diabetes retinopathy using discrete wavelet transform," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine,* p. 0954411912470240, 2012.
[56] S.-E. Bursell, L. Brazionis, and A. Jenkins, "Telemedicine and ocular health in diabetes mellitus," *Clinical and Experimental Optometry,* vol. 95, no. 3, pp. 311-327, 2012.
[57] S. J. J. Kumar and M. Madheswaran, "An Improved Medical Decision Support System to Identify the Diabetic Retinopathy Using Fundus Images," *Journal of Medical Systems,* journal article vol. 36, no. 6, pp. 3573-3581, 2012.
[58] D. Xiao, J. Vignarajan, J. Lock, S. Frost, M.-L. Tay-Kearney, and Y. Kanagasingam, "Retinal image registration and comparison for clinical decision support," *The Australasian medical journal,* vol. 5, no. 9, p. 507, 2012.
[59] M. R. K. Mookiah, U. R. Acharya, C. K. Chua, C. M. Lim, E. Ng, and A. Laude, "Computer-aided diagnosis of diabetic retinopathy: A review," *Computers in biology and medicine,* vol. 43, no. 12, pp. 2136-2155, 2013.
[60] P. Prasanna, S. Jain, N. Bhagat, and A. Madabhushi, "Decision support system for detection of diabetic retinopathy using smartphones," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, 2013, pp. 176-179: IEEE.
[61] A. Bourouis, M. Feham, M. A. Hossain, and L. Zhang, "An intelligent mobile based decision support system for retinal disease diagnosis," *Decision Support Systems,* vol. 59, pp. 341-350, 2014.
[62] R. Klein, B. E. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets, "The Wisconsin Epidemiologic Study of Diabetic Retinopathy: III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years," *Archives of ophthalmology,* vol. 102, no. 4, pp. 527-532, 1984.
[63] R. Klein, B. E. Klein, S. E. Moss, M. D. Davis, and D. L. DeMets, "The Wisconsin Epidemiologic Study of Diabetic Retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years," *Archives of ophthalmology,* vol. 102, no. 4, pp. 520-526, 1984.

[64] R. J. Tapp *et al.*, "The prevalence of and factors associated with diabetic retinopathy in the Australian population," *Diabetes care,* vol. 26, no. 6, pp. 1731-1737, 2003.
[65] B. E. Klein, S. E. Moss, R. Klein, and T. S. Surawicz, "The Wisconsin Epidemiologic Study of Diabetic Retinopathy: XIII. Relationship of serum cholesterol to retinopathy and hard exudate," *Ophthalmology,* vol. 98, no. 8, pp. 1261-1265, 1991.
[66] E. Y. Chew *et al.*, "Association of elevated serum lipid levels with retinal hard exudate in diabetic retinopathy: Early Treatment Diabetic Retinopathy Study (ETDRS) Report 22," *Archives of ophthalmology,* vol. 114, no. 9, pp. 1079-1084, 1996.
[67] D. S. Fong *et al.*, "Retinopathy in Diabetes," *Diabetes Care,* vol. 27, no. suppl 1, pp. s84-s87, 2004.
[68] R. Klein, B. E. Klein, S. E. Moss, and K. J. Cruickshanks, "The Wisconsin epidemiologic study of diabetic retinopathy: XVII: The 14-year incidence and progression of diabetic retinopathy and associated risk factors in type 1 diabetes11Proprietary interest: none," *Ophthalmology,* vol. 105, no. 10, pp. 1801-1815, 1998.
[69] R. Klein, M. D. Knudtson, K. E. Lee, R. Gangnon, and B. E. Klein, "The Wisconsin Epidemiologic Study of Diabetic Retinopathy XXIII: the twenty-five-year incidence of macular edema in persons with type 1 diabetes," *Ophthalmology,* vol. 116, no. 3, pp. 497-503, 2009.
[70] M. Skevofilakas, K. Zarkogianni, B. G. Karamanos, and K. S. Nikita, "A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010, pp. 6713-6716.
[71] V. Balakrishnan, M. R. Shakouri, and H. Hoodeh, "Developing a hybrid predictive system for retinopathy," *Journal of Intelligent & Fuzzy Systems,* Article vol. 25, no. 1, pp. 191-199, 2013.
[72] M. S. Roy, R. Klein, B. J. O'Colmain, B. E. Klein, S. E. Moss, and J. H. Kempen, "The Prevalence of Diabetic Retinopathy Among Adult Type1 Diabetic Persons in the United States," *Archives of Ophthalmology,* vol. 122, no. 4, pp. 546-551, 2004.
[73] J. W. Tukey, "Exploratory data analysis," 1977.
[74] G. Seni and J. F. Elder, "Ensemble methods in data mining: improving accuracy through combining predictions," *Synthesis Lectures on Data Mining and Knowledge Discovery,* vol. 2, no. 1, pp. 1-126, 2010.
[75] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review,* journal article vol. 33, no. 1, pp. 1-39, 2009.
[76] H. He and E. A. Garcia, "Learning from imbalanced data," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 21, no. 9, pp. 1263-1284, 2009.
[77] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 2, pp. 539-550, 2009.
[78] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* pp. 321-357, 2002.
[79] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Advances in intelligent computing*: Springer, 2005, pp. 878-887.
[80] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," in *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings*, T.

Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 475-482.
[81] D. A. Cieslak, N. V. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets," in *GrC*, 2006, pp. 732-737.
[82] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405-425, 2014.
[83] H. He, Y. Bai, E. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2008, pp. 1322-1328: IEEE.
[84] A. Pourhabib, B. K. Mallick, and Y. Ding, "Absent data generating classifier for imbalanced class sizes," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2695-2724, 2015.
[85] B. Wang and N. Japkowicz, "Imbalanced data set learning with synthetic samples," in *Proc. IRIS Machine Learning Workshop*, 2004, p. 19.
[86] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, 2001, vol. 17, no. 1, pp. 973-978: LAWRENCE ERLBAUM ASSOCIATES LTD.
[87] R. Longadge and S. Dongre, "Class Imbalance Problem in Data Mining Review," *arXiv preprint arXiv:1305.1707*, 2013.
[88] M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in *ICML-2003 workshop on learning from imbalanced data sets II*, 2003, vol. 2, pp. 2-1.
[89] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*, 1995, pp. 23-37: Springer Berlin Heidelberg.
[90] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2007.
[91] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: misclassification cost-sensitive boosting," in *Icml*, 1999, pp. 97-105.
[92] W. Lee, C.-H. Jun, and J.-S. Lee, "Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification," *Information Sciences*, vol. 381, pp. 92-103, 3// 2017.
[93] C. Drummond and R. C. Holte, "Exploiting the cost (in) sensitivity of decision tree splitting criteria," in *ICML*, 2000, vol. 1, no. 1.
[94] M. Kukar and I. Kononenko, "Cost-Sensitive Learning with Neural Networks," in *ECAI*, 1998, pp. 445-449.
[95] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the international joint conference on AI*, 1999, pp. 55-60.
[96] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC*, 2003, pp. 49-56.
[97] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *European conference on machine learning*, 2004, pp. 39-50: Springer.
[98] B. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowledge and information systems*, vol. 25, no. 1, pp. 1-20, 2010.

[99] J. Mathew, M. Luo, C. K. Pang, and H. L. Chan, "Kernel-based SMOTE for SVM classification of imbalanced datasets," in *Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE*, 2015, pp. 001127-001132: IEEE.

[100] H. Yu, C. Mu, C. Sun, W. Yang, X. Yang, and X. Zuo, "Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data," *Knowledge-Based Systems*, vol. 76, pp. 67-78, 2015.

[101] Y. Tang and Y.-Q. Zhang, "Granular SVM with repetitive undersampling for highly imbalanced protein homology prediction," in *2006 IEEE International Conference on Granular Computing*, 2006, pp. 457-460: IEEE.

[102] R. Batuwita and V. Palade, "Efficient resampling methods for training support vector machines with imbalanced datasets," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1-8: IEEE.

[103] M. A. H. Farquad and I. Bose, "Preprocessing unbalanced data using support vector machine," *Decision Support Systems*, vol. 53, no. 1, pp. 226-233, 4// 2012.

[104] P.-N. Tan, S. Michael, and V. Kumar, "Chapter 6. association analysis: Basic concepts and algorithms," *Introduction to Data Mining. Addison-Wesley. ISBN*, vol. 321321367, 2005.

[105] J. D. Piette and E. A. Kerr, "The Impact of Comorbid Chronic Conditions on Diabetes Care," *Diabetes Care*, vol. 29, no. 3, pp. 725-731, 2006.

[106] J. N. Struijs, C. A. Baan, F. G. Schellevis, G. P. Westert, and G. A. van den Bos, "Comorbidity in patients with diabetes mellitus: impact on medical health care utilization," *BMC Health Services Research*, journal article vol. 6, no. 1, p. 84, 2006.

[107] D. Delen, A. Oztekin, and L. Tomak, "An analytic approach to better understanding and management of coronary surgeries," *Decision Support Systems*, vol. 52, no. 3, pp. 698-705, 2// 2012.

[108] H. V. Jagadish *et al.*, "Big data and its technical challenges," *Communications of the ACM*, vol. 57, no. 7, pp. 86-94, 2014.

[109] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine learning*, vol. 30, no. 2-3, pp. 195-215, 1998.

[110] R. B. Rao, S. Krishnan, and R. S. Niculescu, "Data mining for improved cardiac care," *ACM SIGKDD Explorations Newsletter*, vol. 8, no. 1, pp. 3-10, 2006.

[111] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *Intelligent Systems and their Applications, IEEE*, vol. 14, no. 6, pp. 67-74, 1999.

[112] H. Haibo and E. A. Garcia, "Learning from Imbalanced Data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263-1284, 2009.

[113] D. W. Hosmer Jr and S. Lemeshow, *Applied logistic regression*. John Wiley & Sons, 2004.

[114] P. Hájek, "Municipal credit rating modelling by neural networks," *Decision Support Systems*, vol. 51, no. 1, pp. 108-118, 4// 2011.

[115] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *International Journal of Medical Informatics*, vol. 77, no. 2, pp. 81-97, 2// 2008.

[116] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. De Jesús, *Neural network design*. PWS publishing company Boston, 1996.

[117] S. Lee, "Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls," *Decision Support Systems*, vol. 49, no. 4, pp. 486-497, 11// 2010.

[118] L. Breiman, "Random Forests," (in English), *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001/10/01 2001.

[119] P. Liu, Y. Wang, L. Cai, and L. Zhang, "Classifying skewed data streams based on reusing data," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010, vol. 4, pp. V4-90-V4-93: IEEE.
[120] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42-47, 2012.
[121] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 10, pp. 993-1001, 1990.
[122] C. Sabanayagam, W. Yip, D. S. Ting, G. Tan, and T. Y. Wong, "Ten Emerging Trends in the Epidemiology of Diabetic Retinopathy," *Ophthalmic Epidemiology*, pp. 1-14, 2016.
[123] R. Lee, T. Y. Wong, and C. Sabanayagam, "Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss," *Eye and Vision*, vol. 2, p. 17, 2015.
[124] P. Liu, Y. Wang, L. Cai, and L. Zhang, "Classifying skewed data streams based on reusing data," in *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, 2010, vol. 4, pp. V4-90-V4-93: IEEE.
[125] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, 1998.
[126] A. Anand, G. Pugalenthi, G. B. Fogel, and P. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino acids*, vol. 39, no. 5, pp. 1385-1391, 2010.
[127] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121-167, 1998.
[128] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144-152: ACM.
[129] G. Wu and E. Y. Chang, "Adaptive feature-space conformal transformation for imbalanced-data learning," in *ICML*, 2003, pp. 816-823.
[130] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data mining and knowledge discovery handbook*: Springer, 2005, pp. 853-867.

VITA

Saeed Piri

Candidate for the Degree of

Doctor of Philosophy

Thesis: DEVELOPING AND DEPLOYING DATA MINING TECHNIQUES IN HEALTHCARE

Major Field: Industrial Engineering and Management

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in July, 2017.

Completed the requirements for the Master of Science in Industrial Engineering at Sharif University of Technology, Tehran, Iran in 2011.

Completed the requirements for the Bachelor of Science in Industrial and System Engineering at Amirkabir University of Technology, Tehran, Iran in 2008.

Experience:

Research Associate, Center for Health Systems Innovation, Spears School of Business, Oklahoma State University, 2014-2017.

Research Associate, School of Industrial Engineering and Management, Oklahoma State University, 2013- 2014.

Professional Memberships:

Institute for Operations Research and the Management Sciences (INFROMS)

Institute of Industrial and System Engineering (IISE),

Decision Science Institute (DSI)

Alpha Pi Mu