Computation of an Enriched Set of Predictors for Type 2 Diabetes Prediction

Noushin Hajarolasvadi

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Eastern Mediterranean University
June 2016
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Cem Tanova
Acting Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Computer Engineering.

Prof. Dr. Işık Aybay
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Computer Engineering.

Asst. Prof. Dr. Ahmet Ünveren
Co-Supervisor

Prof. Dr. Hakan Altınçay
Supervisor

Examining Committee

1. Prof. Dr. Hakan Altınçay
2. Prof. Dr. Doğu Arifler
3. Prof. Dr. Hasan Kömürcügil
4. Assoc. Prof. Dr. Ekrem Varoğlu
5. Asst. Prof. Dr. Ahmet Ünveren

ABSTRACT

According to World Health Organization, about 422 million people worldwide have diabetes, vast majority of whom belong to Type 2. In addition to this population, a noticeable percentage of people has either undiagnosed Type 2 diabetes or prediabetes. Since this disease causes death mainly through physiological complications such as cardiovascular disease, it is highly crucial to diagnose it in an early stage. The medical diagnosis is done by three invasive blood tests which make it almost impossible to periodically screen the whole population. As an alternative approach, development of automated systems that can identify patients having Type 2 diabetes using non-invasive predictors such as age, waist circumference, family history and body mass index is extensively studied. In this thesis, the use of an enriched set of predictors including symptoms, diagnoses, lifestyle habits and medications is considered for improving the detection performance. The main motivation for this study is that the complications due to the onset of the disease might occur before medical diagnosis. The performance of various classifiers including logistic regression and support vector machines, and feature selection schemes such as mRMR and Relief are investigated. The experiments conducted have shown that additionally defined features provide better area under the receiver operating characteristic curve scores.

Keywords: Type 2 Diabetes Classification, Feature Extraction, Feature Selection, Filters, Wrappers, Embedded Feature Selection

ÖZ

Dünya sağlık örgütüne göre dünya çapında, büyük çoğunluğu tip 2 olmak üzere yaklaşık 422 milyon insan diyabet hastasıdır. Bu gruba ek olarak, önemli sayıda tespit edilmemiş tip 2 diyabet veya öndiyabet hastası mevcuttur. Bu hastalık kardiyovasküler hastalıklar gibi fizyolojik komplikasyonlar yüzünden ölüme sebebiyet verdiğinden, erken teşhis son derece önemlidir. Tıbbi teşhis üç farklı kan testi ile yapıldığından tüm nüfusu periyodik olarak taramak mümkün değildir. Alternatif bir yaklaşım olarak, yaş, bel çevresi, aile tarihçesi ve vücut kitle indisi gibi prediktörler kullanarak tip 2 diyabet hastalarını bulabilen otomatik sistemlerin geliştirilmesi konusunda yoğun olarak çalışılmaktadır. Bu tezde, tanıma başarımını artırmak için semıtomlar, teşhisler, yaşam tarzı ve kullanılan ilaçlar gibi bilgiler içeren zenginleştirilmiş bir prediktör kümesi kulanımı üzerinde çalışılmıştır. Bu çalışmanın esas motivasyonu, hastalığın başlangıcından dolayı oluşan komplikasyonların tıbbi teşhis yapılmadan önce başlamasının mümkün olmasıdır. Lojistik regresyon ve destek vektör makinaları gibi sınıflandırıcıları da içeren birçok sınıflandırıcının ve mRMR ile Relief gibi birçok öznitelik seçme yönteminin başarımları incelenmiştir. Yapılan deneysel çalışmalar, ek olarak tanımlanmış prediktörlerin karar vericinin etkinliği eğri altı alanını iyileştirdiğini göstermiştir.

Anahtar Kelimeler: Tip 2 Diyabet Sınıflandırma, Öznitelik Çıkarımı, Öznitelik Seçimi, Fitreler, Sarmalılar, Gömülü Öznitelik Seçme

To My Parents

ACKNOWLEDGMENT

First of all, I would like to represent my deepest allegiant thanks to my knowledgeable supervisor Prof. Dr. Hakan Altınçay who guided me through different steps of this survey patiently. Besides, I would like to thank Prof. Dr. Ahmet Ünveren for sharing his time and experience in favor of this thesis.

Special thanks to my parents and my siblings who always supported me and encouraged me to continue my education toward master and doctoral degree. I also would like to express my appreciation to my friend Saeed Mohammad Zadeh who enhanced my motivation by his presence.

Last but not least, I wish to thank all the faculty members at the department of Computer Engineering, especially the chairman, Prof. Dr. Işık Aybay.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ADA	American Diabetes Association
AUC	Area Under the Curve
CFS	Correlation-based Feature Selection
CMIM	Conditional Mutual Information Method
EM	Expectation Maximization
FPG	Fasting Plasma Glucose
GA	Genetic Algorithms
HbA1c	Glycosylated Hemoglobin
kNN	k Nearest Neighbor Classifier
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Logistic Regression
ML	Maximum Likelihood
mRMR	Maximum Relevance Minimum Redundancy
NHANES	National Health and Nutrition Examination Survey
OGTT	Oral Glucose Tolerance Test
SFS	Sequential Forward Selection
SVM	Support Vector Machines
T2DM	Type 2 Diabetes Mellitus

Chapter 1

INTRODUCTION

1.1 Introduction

According to World Health Organization, the worldwide population having diabetes was about 422 million people in 2014, vast majority of whom belong to Type 2 [1]. This chronic disease has two major types. Type I diabetes, also known as juvenile diabetes, occurs due to malfunctioning of pancreas in producing insulin. Insulin is a hormone which moves glucose (sugar) to cells so that they produce energy. This type of diabetes has no cure because pancreas does not produce insulin [2]. However, it is possible to control it. Type 2 Diabetes Mellitus (T2DM) is more common. In this type, pancreas is producing insulin but it is the body cells that cannot process or absorb the produced insulin. As a result, the body suffers from insulin deficiency. Unfortunately, T2DM has an asymptomatic phase which leads to progressive complications due to untreated hyperglycemia [3]. Minimum duration of this phase is estimated to be 4 to 7 years. As a result, 30-50% of the population of T2DM remain undiagnosed [4]. This means an individual may not realize he/she has high blood glucose for considerable amount of time which leads to developing short-term and long-term complications. Also, late diagnosing causes financial burden for both the patient and the health care system. Short-term complications caused by T2DM are very low blood glucose (hypoglycemia) and very high blood glucose (Hyperosmolar Hyperglycemic Nonketotic Syndrome) [2]. Long-term complications are diabetic retinopathy, nephropathy (kidney problems), diabetic neuropathy (foot ulcer, etc.)

and cardiovascular problems. In addition, there is an intermediary phase between being normal and having T2DM. It is defined as the period in which the level of blood sugar of an individual is higher than normal but not high enough to be considered as having T2DM. Fortunately, studies show that T2DM/prediabetes can be controlled, prevented or delayed [5] by losing weight, changing life style, increasing physical activity, etc. [6]. Therefore, like many other diseases, screening and early detection of T2DM/prediabetes is important. Timely detection of this disease can be achieved by invasive blood tests.

According to the most updated American Diabetes Association (ADA) guidelines published in 2016, diagnosis of prediabetes and T2DM is based on three different plasma glucose measurements: The fasting plasma glucose (FPG), the oral glucose tolerance test (OGTT) and the level of Glycosylated Hemoglobin (HbA1c). Using these invasive blood tests, an individual can be categorized in one of these three categories as follow:

Table 1.1: Diagnosis of prediabetes and T2DM using three invasive blood tests

Normal	**Prediabetes**	**Diabetes**
FPG < 100 and OGTT < 140 and HbA1c < 5.7%	140 ≤ OGTT ≤ 199 mg/dl or 100 ≤ FPG ≤ 125 mg/dl or 5.7% ≤ HbA1c < 6.5%	FPG ≥ 126 mg/dl or OGTT ≥ 200 mg/dl or HbA1c ≥ 6.5%

Undiagnosed diabetes and undiagnosed prediabetes mean any of an individual's lab test results is within the aforementioned ranges but he/she is not aware of it. A good solution for early detection of diabetes is periodic screening. In this solution, FPG, OGTT at 2h and HbA1c level of the patient are tested. In case of testing FPG or

OGTT, a fasting hour criteria must be abided (8 ≥ and < 24 hours). Thus, HbA1c has this advantage over FPG and OGTT that it does not need fasting. Unfortunately, periodic screening is not applicable to everybody since many people may not be willing to be regularly examined by invasive and expensive blood tests [5]. As an alternative solution, development of an automated system that can predict T2DM in the absence of invasive lab tests may be considered.

In recent years, plenty of novel algorithms have been suggested to detect people having T2DM or prediabetes with acceptable accuracy. As a pattern classification task, detection of people having prediabetes or T2DM at earlier stages is very important so as to avoid consequent complications. In order to design an automated system to detect prediabetes/T2DM, reliable predictors should be identified. The conventionally used predictors are risk factors of this disease. World Health Organization defines risk factors of a disease as any attribute of an individual that increases the probability of developing that disease [1]. Therefore, when one knows more risk factors related to a specific disease such as T2DM, early diagnosis becomes more successful.

The risk factors of T2DM can be split into two categories:

- **Non-modifiable risk factors** which include mostly physiological characteristics like age, gender, genetic-predisposition, etc.
- **Modifiable risk factors** that one can control like unhealthy diet, tobacco use and physical inactivity.

The automated systems developed so far employ predictors from both of these groups. Taking into account the fact that, the patients may suffer from the

complications caused by prediabetes/T2DM before being diagnosed, it is aimed in this thesis to compute an enriched set of predictors including symptoms, diagnosis, lifestyle habits and medications used so as to improve the performance of prediabetes/T2DM detection. A questionnaire-based dataset which includes a wide range of questions about the participants is considered for this purpose. Hundreds of novel features are evaluated using a wide set of feature selection schemes to compute an enriched set of predictors. Experimental results have shown that better performance scores can be achieved with the use of an enriched set of predictors.

The thesis is organized as follows. Next in Chapter 2, details about the machine learning algorithms used for imputation, feature selection and classification are provided. Chapter 3 presents the procedure applied in the definition of an enriched set of predictors. This is followed by Chapter 4 which provides a comprehensive evaluation of the selected set of additional predictors. Finally, conclusions and directions for future work are presented in Chapter 5.

Chapter 2

PATTERN CLASSIFICATION PROBLEM

2.1 Introduction

Pattern Classification is the task of labeling input samples as one of the predefined groups known as classes [7], [8]. The first step in solving a classification problem is to prepare a dataset by measuring physical and non-physical descriptors of the samples (patterns) known as features. This way, each sample in the dataset is represented by a vector of features or variables. In general, there are two types of features, numerical and categorical.

In general, features need to be preprocessed. For instance, numerical features may need to be discretized or normalized. Discretization is the process of transforming a numerical value to a categorical value [9]. In case of categorical features, it is often necessary to use the dummy representation for transforming each categorical feature into a set of binary features. More specifically, a categorical feature with m different values will be represented by $(m-1)$ dummy features after one of the categories is selected as the reference. In case when m is equal to two, the feature is called binary and it can be represented using 0 and 1. In addition, dealing with noise, redundancy and outlying samples can be done in this step. Outliers are samples that are significantly inconsistent with the remaining samples of the data set. That is, they do not comply the general pattern of the data.

After preprocessing, discriminative features with respect to the domain of the problem must be selected/extracted. The classification performance heavily depends on the features employed. Using a small number of features may lead to poor models that underfit the given data whereas utilizing larger number of features may cause unnecessary model complexity and hence lead to overfitting.

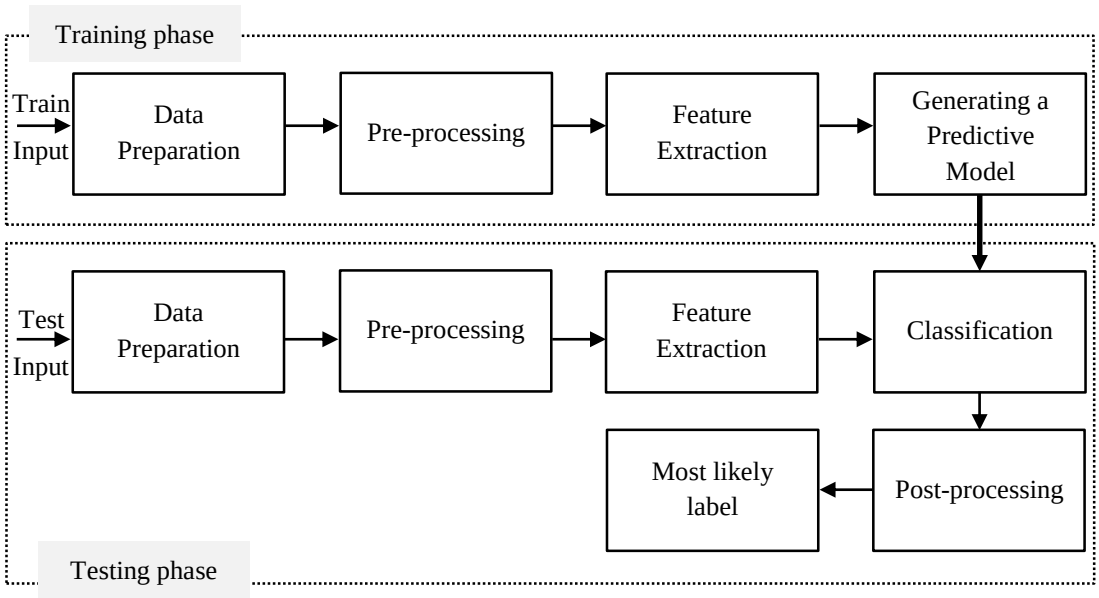

Figure 2.1: Components of a pattern classification system

Figure 2.1 shows the block diagram of a pattern classification system. As we see in the figure, the next step after feature extraction is to design a predictive model that can properly define the patterns of the data so that it can later be used to classify unseen or test data. In other words, it is aimed to learn discriminative information about different classes using the training data. Adjusting the complexity of the decision models is highly crucial in achieving satisfactory level of test performance. In particular, overfitting may occur if the complexity of the selected model is higher than the data under concern. As a matter of fact, one should select a model that is not so simple which cannot discriminate the classes and not so complex that memorizes

the data instead of learning and generalization. In the testing phase, the performance of the models generated will be tested using another set of data which is hidden from the training phase.

In generation of a predictive model, both parametric and non-parametric models are used. In parametric approaches, a functional form is selected which corresponds to making assumptions about underlying distribution of the data. In such cases, model generation corresponds to estimating the model parameters using the training data. Alternatively, in non-parametric approaches, classification models are generated using the proximities among the samples within and between different classes. In both approaches, the model is finalized by minimizing a performance metric such as error rate or misclassification cost [10].

After the training phase is completed, it is necessary to evaluate the performance of the designed system using a test data set. There are various methods that may be considered to generate train/test splits. One of these approaches is k fold cross validation. In this approach, the given set of samples is divided to k folds of similar size. The first fold is held out as the test set and the other $(k-1)$ folds are used as training data to generate the model. Then, the model is evaluated by testing with the samples in the held-out set. This procedure is repeated k times so that k different performance scores are obtained for the metric under concern. These scores are then averaged and used as the overall performance score for the model considered.

2.2 Missing Value Imputation

In real world data, missing values may happen due to various reasons. For example, test results of a patient may not be available because hospital lacks the required

medical equipment. In some cases, the record keeping may not be well-established. In order to generate an effective scheme for imputation, the source of missing value should be known. In some cases, the data is missed completely at random (MCAR). This means that there is no systematic cause for the missingness. In such cases, the probability of missingness is independent of the value of the variable [7]. For example, the blood test tube of a patient breaks accidentally. Second type of missingness is when the data is missed at random (MAR). In this case, the probability of missingness is independent of the value of the variable but it happens based on a pattern and it can be predicted using other variables [7]. In the third type, the data is not missed at random (NMAR). In this case, the pattern of missing data depends on the variable itself and it cannot be predicted using other variables [7]. In case of MCAR and MAR, the missing value is imputable using simple methods because the reason of missingness is ignorable [7]. However, imputation is an important task because wrong imputation of missing values may lead to misleading models.

Most of the previously used techniques for missing value imputation rely on statistical analysis and machine learning methods [7]. Some of the most important imputation methods in these two groups are discussed below.

The notation used in the following context can be summarized as follows. Assume that a labeled dataset of N samples and d features is given. Then, $\boldsymbol{x}_i$ denotes the vector of features corresponding to the *i*th sample and it can be shown as

$$\boldsymbol{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix} \in c_i, \qquad i = 1, 2, ..., C, \tag{2.1}$$

where c_i is the class it belongs. If there exists only two classes, the classification problem is binary.

2.2.1 Statistical Methods

Mean imputation is one of the most widely used statistical method due to its simplicity and efficiency. The main idea of this method is to impute the missing values of a feature with the mean of all observed (available) values of that feature. For instance, the missing values of the *j*th feature is imputed using

$$\frac{1}{N_{o,\mathrm{j}}} \sum_{i=1}^{N} (1 - m_{ij}) x_{ij}, \tag{2.2}$$

where $N_{o,\mathrm{j}}$ is the number of samples with an existing value for the *j*th feature. m_{ij} is 1 if the value of the *j*th feature in the *i*th sample is missing and zero otherwise. This method is useful when feature type is numerical. Also, if the data has outliers mean imputation is compromised [11].

Median imputation is more robust to outliers [11]. In this approach, the median of observed data for the *j*th feature is used to impute the missing values of the *j*th feature. The feature value to impute the missing values is computed as

$$\underset{\substack{i=1,\ldots,\mathrm{N} \\ x_{ij} \neq NA}}{median} \{x_{ij}\}, \tag{2.3}$$

where NA represents a missing value. When the feature is binary or categorical, mean and median are not applicable. Thus another statistical method known as mode

imputation is generally used. In this approach, the most frequently observed value of the feature is used to replace the missing values of that feature.

Hot and cold deck imputation are two other statistical methods for imputation of missing values. In the hot deck method, the complete sample which is the most similar to the sample with missing values is found. Then, missing values are imputed with the matching components of the complete sample. The drawback of this method is that the imputation of all missing values of a sample is done using a single complete sample [7], [11].

Cold deck imputation approach is similar to hot deck in terms of methodology. However, a data source other than current is employed. More specifically, the missing values are imputed using the most similar sample from an external data source. One disadvantage of this method is that the external data source may differ from the main data source in some sense such as the methodology of data collection. This may cause more inconsistency and bias in the performance of the classifier [11].

2.2.2 Machine Learning Methods

Machine learning methods are more complex than the statistical approaches because they estimate the missing values by creating a predictive model. k Nearest Neighbor (kNN) is one of these methods. In fact, kNN is a hot deck imputation method. In this method, k nearest neighbors of the sample with missing values are selected from complete samples by using a distance metric. After selecting the nearest k neighbors, the missing values are imputed using mean or mode of the neighbors. A better approach is to assign a weight to each neighbor based on its distance from sample $\boldsymbol{x}_i$ so that a closer neighbor contributes more to the imputation task than the

others. Another important parameter of this method is the selection of the distance metric. In general, both categorical and numerical features may be available. In such cases, the heterogeneous Euclidean overlap metric can be employed [7]. Let $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ represent a pair of samples, then the distance between $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ can be computed as

$$D(\boldsymbol{x}_a, \boldsymbol{x}_b) = \sqrt{\sum_{j=1}^{d} D_j(x_{aj}, x_{bj})}, \tag{2.4}$$

where $D_j(x_{aj}, x_{bj})$ is the distance function which calculates the distance between two samples for the feature and it can be expressed as follow:

$$D_j(x_{aj}, x_{bj}) = \begin{cases} 1, & (1-m_{aj})(1-m_{bj}) = 0 \\ D_{cat}(x_{aj}, x_{bj}), & x_j \; is \; a \; categorical \; feature \\ D_{num}(x_{aj}, x_{bj}), & x_j \; is \; a \; numerical \; feature \end{cases} \tag{2.5}$$

If any of the input values $\boldsymbol{x}_{aj}$ or $\boldsymbol{x}_{bj}$ is unknown, the distance value is 1. If the value of the categorical inputs is the same, the distance function $D_{cat}(x_{aj}, x_{bj})$ returns a value of 0, otherwise it returns 1. $D_{num}(x_{aj}, x_{bj})$ is a normalized distance function used for numerical features. It uses the maximum and minimum values of observed samples in the training data for the feature under concern as

$$D_{num}(x_{aj}, x_{bj}) = \frac{\left| x_{aj} - x_{bj} \right|}{\max\limits_{i=1,\ldots,N}(x_{ij}) - \min\limits_{i=1,\ldots,N}(x_{ij})} \tag{2.6}$$

Many studies report that kNN outperforms other methods such as mean imputation or other machine learning based algorithms such as decision-trees (C4.5) [7], [11], [12]. When compared to the other methods, the main advantage of kNN is that only the most similar samples affect the value of imputation. However, the computational

cost of kNN is high because it searches the whole set of the training data to find the most similar samples.

2.3 Modeling Techniques

Many algorithms are developed to design automated classification systems and most of them use a statistical method to find the decision boundaries which divide the data set into two or more classes. The relative performance of a particular scheme depends on the domain of the problem since each classification task has its distinguishing characteristics such as the amount of training data and underlying distribution of data. Two of the most well-known classifiers namely, Logistic Regression and Support Vector Machines are employed in this thesis. These methods are presented in Sections 2.3.1 and 0, respectively.

2.3.1 Logistic Regression Classifier

The logistic regression (LR) classifier computes a linear decision boundary between two classes of data. In LR, the main aim is to represent the probability that the given sample belongs to a predefined category. More specifically, let x denote a predictor and c denote a binary response variable whose value is either *positive* or *negative*. LR models the probability that a given sample belongs to a specific category as

$$p(c = positive \mid x) = p(x) = \beta_0 + \beta_1 x \tag{2.7}$$

where β_0 and β_1 are the intercept and slope of the linear model. The values of these design parameters should be estimated using the training data. It is important to note that the probability must be between 0 and 1. Thus, logistic function is used in LR to satisfy this constraint. Logistic function is defined as

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{2.8}$$

The values β_0 of β_1 and can be estimated using the maximum likelihood method [10]. Eq. (2.8) can be re-written as

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}. \tag{2.9}$$

The left side of the Eq. (2.9) is called odds. By taking the logarithm of both sides of Eq. (2.9), we obtain

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x \tag{2.10}$$

The left side is log-odds which is positive when $p(x) > 0.5$. This corresponds to selection of the positive class as the most likely when $\beta_0 + \beta_1 x$.

In general, there is more than one predictor or feature. For example, multiple factors such as age, ethnicity and waist circumference are contributing in determining whether an individual has T2DM or not. Assuming that there are *d* predictors, the multivariate logistic regression is defined as

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_d x_d \tag{2.11}$$

where x_j and β_j are the *j*th feature and the coefficient of the *j*th feature, respectively. As in the case of univariate modeling, the parameters can be computed using the maximum likelihood method.

It is obvious that a linear decision boundary is obtained when LR is used. When the decision boundary is more complex, enlarging the feature space using quadratic or cubic terms solves the nonlinearity problem.

2.3.2 Support Vector Machines

Support Vector Machine (SVM) computes both linear and nonlinear decision boundaries to separate different classes. SVM is a supervised learning method and generally, it is used for binary classification.

When two predictors are utilized, a linear boundary corresponds to a line in two dimensional feature space. It corresponds to a hyperplane when 3 or more predictors are considered. A hyperplane is a subspace having one dimension less than that of the feature space employed [10]. Thus, the mathematical definition of a hyperplane in a d dimensional space is

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d = 0 \tag{2.12}$$

A point in the space can be either on the hyperplane or not. Thus, it is clear that any $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ for which Eq. (2.12) holds true is a point on the hyperplane. If the point is not on the hyperplane, then it satisfies either Eq. (2.13) or Eq. (2.14) based on the value of $\boldsymbol{x}$. In this case, the point lies to either side of the hyperplane.

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d > 0 \tag{2.13}$$

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d < 0 \tag{2.14}$$

In other words, a hyperplane divides the feature space into two subspaces and each point which is not on the hyperplane belongs to one of these subspaces.

When the classes are linearly separable, the optimal decision boundary is defined by SVM as the hyperplane which has the maximal margin. More specifically, the margin is defined as minimum of the perpendicular distances of all training samples

to the hyperplane. The separating hyperplane with the largest margin is named as the maximal margin hyperplane [10]. Figure 2.2 shows the maximal margin hyperplane on a hypothetical data set for two features.

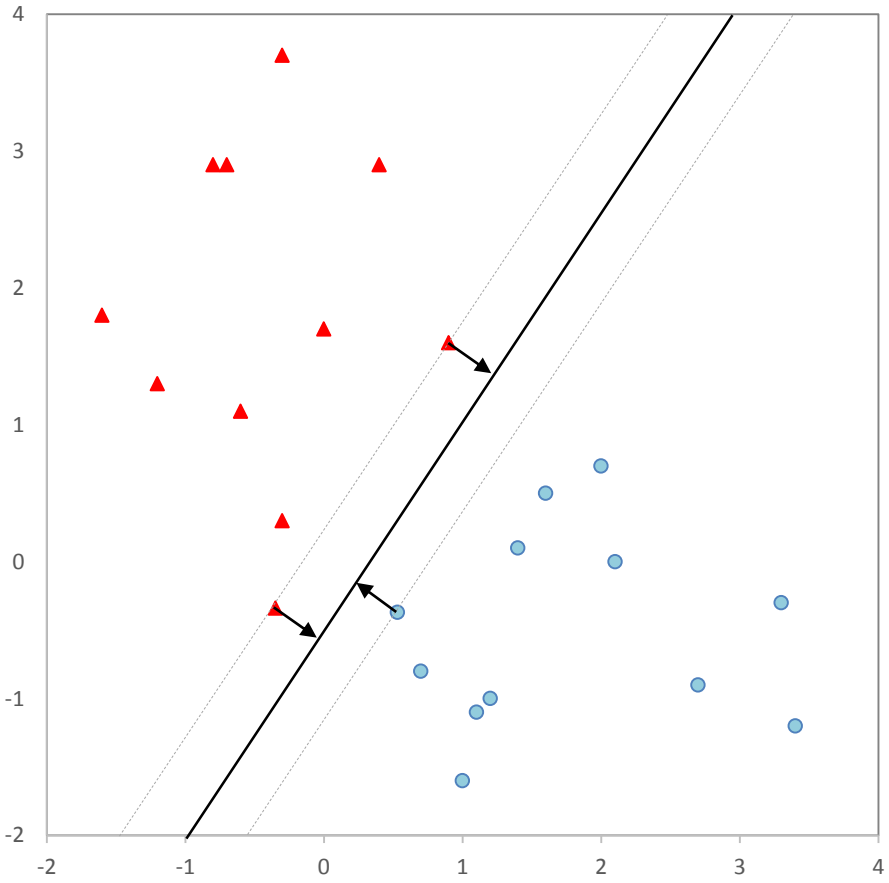

Figure 2.2: The solid line: maximal margin hyperplane, points on dashed lines: support vectors

In this figure, a scatter plot for 2 classes denoted by▲ and ●. The three samples with equal distance from the decision boundary (the bold line) show the width of margin and they are called support vectors because they are vectors in a d dimensional space and they support the hyperplane in the sense that if they move slightly the maximal margin hyperplane will move as well. In fact, the maximal margin hyperplane depends only on these support vectors and not on the whole training samples. In

order to categorize the samples into two classes, i.e. *positive* and *negative*, SVM supports to find a solution that maximizes the margin.

Let the *positive* and *negative* classes be represented using +1 and -1, respectively. In case of linearly separable classes, the decision boundary or the separating hyperplane is defined as

$$\begin{cases} \boldsymbol{\beta}^T \boldsymbol{x}_i + \beta_0 \geq +1 \ \ if\ c_i = -1 \\ \boldsymbol{\beta}^T \boldsymbol{x}_i + \beta_0 \leq -1 \ \ if\ c_i = +1 \end{cases} \tag{2.15}$$

β_0 is the offset from the origin and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_d]$ is the weight vector for the hyperplane. Combining the two equations into one, SVM supports to find $\boldsymbol{\beta}$ and β_0 such that

$$c_i(\boldsymbol{\beta}^T \boldsymbol{x}_i + \beta_0) \geq +1 \tag{2.16}$$

In case of linearly separable classes, the separating hyperplane found by SVM must have the maximum distance from the closest training samples. The maximum distance can be calculated as $\frac{2}{\|\boldsymbol{\beta}\|}$. If we calculate $\frac{\|\boldsymbol{\beta}\|^2}{2}$ instead of $\frac{2}{\|\boldsymbol{\beta}\|}$ we can convert the maximization problem to minimization problem. This helps us to formulate the problem as follow

$$\begin{aligned} &minimize\ \frac{\|\boldsymbol{\beta}\|^2}{2} \\ &subject\ to \begin{cases} \boldsymbol{\beta}^T \boldsymbol{x}_i + \beta_0 \geq +1 \ \ if\ c_i = -1 \\ \boldsymbol{\beta}^T \boldsymbol{x}_i + \beta_0 \leq -1 \ \ if\ c_i = +1 \end{cases} \end{aligned} \tag{2.17}$$

We can solve Eq. (2.17) using Lagrange multipliers and the dual problem. After some manipulations, Eq. (2.17) can be written as

$$L_p = \left(\frac{\|\boldsymbol{\beta}\|^2}{2} \right) - \sum_{i=1}^{N} \alpha_i \left(c_i \left(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 \right) - 1 \right)$$
$$= \left(\frac{\|\boldsymbol{\beta}\|^2}{2} \right) - \sum_{i=1}^{N} \alpha_i c_i \left(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0 \right) + \sum_{i=1}^{N} \alpha_i \quad (2.18)$$

α_i are the Lagrange multipliers. That is for each sample there exists a Lagrange multiplier and the constraint $\alpha_t \geq 0$ should set to restrict its weight to be non-negative. It is important to note that the Lagrange multiplier is zero for those $\mathbf{x}_i$ that are not located on the hyperplane. Thus, support vectors can be defined as those $\mathbf{x}_i$ with non-zero Lagrange multiplier, *i.e.* $\alpha_t \geq 0$. Other samples can be removed without causing any change in the location of the optimal hyperplane. Solution for β_0 can be found as

$$\beta_0 = c_i - \boldsymbol{\beta}^T \mathbf{x}_i \quad (2.19)$$

For a given test sample, the class label is identified by

$$c_t = \sum_{i=1}^{N} \alpha_i c_i \mathbf{x}_i^T \mathbf{x}_t + \beta_0 \quad (2.20)$$

where $\mathbf{x}_t$ and c_t are the test sample and its label, respectively and $\sum_{i=1}^{N} \alpha_i c_i \mathbf{x}_i$ is the solution for $\boldsymbol{\beta}$.

In practice, the classes may not always be linearly separable. In such cases, linear SVM does not provide the best-fitting boundary. In such cases, the problem is converted to a linearly separable one by using non-linear mapping to convert the sample space of d dimensions to an l dimensional space where $l > d$. SVM can then search for a linear decision boundary within the l dimensional space.

In order to implement this, a kernel function must be used. For example, the polynomial function can be used as the kernel which is defined as

$$k(\boldsymbol{x}_i, \boldsymbol{x}_k) = \left(\boldsymbol{x}_i^T \boldsymbol{x}_k + 1\right)^p \tag{2.21}$$

2.3.3 Performance Evaluation Metrics

In order to be able to compare the performance of different classifiers, various metrics are developed. These metrics are based on the correct and incorrect classification of the tested samples. This information can be summarized by a contingency table, as shown in Table 2.1 which is also called confusion matrix [13]. In this table, the number of true positives and true negatives are shown by TP and TN. Similarly, FP and FN denote the number of false positives and false negatives, respectively.

Table 2.1: Confusion matrix

		Predicted Labels	
		Positive	Negative
True Labels	Positive	TP	FN
	Negative	FP	TN

TP represents the number of positive samples which are correctly classified whereas FN gives the number of misclassified positives. Similarly, FP and TN represent the number of misclassified and correctly classified negative samples. It should be noted that $TN + TP + FP + FN = N$ where N is the total number of samples. Although the confusion matrix is enough to understand the classifier performance, some other measures are defined so as to represent the performance on different

classes in a clear way. Examples of such measurements are accuracy, sensitivity and specificity that are defined as follows:

$$
\begin{aligned}
Accuracy &= \frac{(TP + TN)}{TN + TP + FP + FN} \\
Sensitivity &= \frac{TP}{TP + FN} \\
Specificity &= \frac{TN}{TN + FP}
\end{aligned}
\tag{2.22}
$$

Accuracy shows the percentage of correctly classified samples, ignoring their class labels. For a binary classification problem with *positive* and *negative* classes, sensitivity or True Positive Rate (TPR) is the proportion of correctly classified positive samples. Specificity denotes the number of correctly classified negative samples. Also, it is important to note that False Positive Rate (FPR) can be defined as $(1 - specificity)$.

The scores obtained using these metrics correspond to only one particular operating point. In general, the operating point value is selected so that the probability of error is minimized. However, instead of minimizing rate of misclassification, we may need to minimize the misclassification of a particular class. Alternatively, we may need to compute the performance on different classes as a function of different decision thresholds. For such cases, TPR and FPR are computed and resultant scores are plotted for each different threshold. The resultant set of scores are plotted to construct the Receiver Operating Characteristic (ROC) curve. Area Under the ROC Curve (AUC) is generally used as an alternative evaluation metric since it takes into account the performance on both classes for different decision thresholds.

2.4 Feature Selection Schemes

The main goal in feature selection is to compute a set of features that are relevant with the target and have minimum dependency with other features. In general, feature subsets perform better due to several reasons such as avoiding overfitting and model simplification [14]. There are three types of feature selection methods, namely filter, wrapper and embedded methods.

2.4.1 Filter Methods

These methods use a statistical evaluation metric such as mean and standard deviation to assign a score or weight to each feature based on their relevance to the target response. Some of the filter methods (univariate) consider only the correlation of features with the target class whereas others (multivariate) take into account both the correlation among different features and the correlation with the target class. The correlation among features represents the pair-wise similarity of different features [15]. The univariate filter methods are known to be fast, scalable and independent of the classifier. However, they ignore the existence of dependency among features and their interactions with the classifier. The multivariate feature selection schemes are slower than univariate ones but they take into account the dependency among features [14]. As an example, t-test, chi-square and information gain are univariate whereas correlation-based feature selection (CFS) is multivariate.

2.4.1.1 t-test

The t-test examines the difference of two populations in two different classes using the mean and standard deviation of each population. The t-test score of a given feature can be expressed as

$$\frac{|\mu_1 - \mu_2|}{\sqrt{\frac{(\sigma_1)^2}{N_1} - \frac{(\sigma_2)^2}{N_2}}}, \tag{2.23}$$

where μ_i and σ_i are mean and standard deviation of the *i*th class and N_i denotes the number of samples in that class, for $i = 1,2$. The results from t-test are more reliable when the sample space is large enough and the variances are small. The main disadvantage of this method is that the correlation among different features is not considered.

2.4.1.2 Chi-square

Chi-square goodness of fit test is a common hypothesis testing based scheme that compares a sample of a feature against a population with known parameters [16]. For a binary classification problem, if f_i and e_i denote the actual count of the observed samples and the expected number of samples in a given class, then chi-square test score is computed as

$$\chi^2 = \sum_{i=1,2} \frac{(f_i - e_i)}{e_i}. \tag{2.24}$$

This test is applicable to categorical features. In case of applying chi-square on numerical features, discretization must be applied as a pre-processing step.

2.4.1.3 Information Gain

Information gain measures the importance of a feature by using entropy which is a measure of uncertainty of a random variable X, defined as

$$H(X) = -\sum_i p(x_i) \log p(x_i). \tag{2.25}$$

where $p(x_i)$ is the probability that $X = x_i$. In the current context, each feature is considered as a random variable and the set of discrete values that the feature may take forms the sample space of the random variable.

The conditional entropy of the random variable Y denoted by $H(Y \mid X = x_i)$ shows the entropy of $\boldsymbol{Y}$ among those samples in which X has value x_i. Information gain can be defined as the amount of reduction in entropy caused by dividing the samples to different groups based on a specific feature.

Let c denote the target class. For a given feature x_j, the information gain is defined as

$$IG(c \mid x_j) = H(c) - H(c \mid x_j) \tag{2.26}$$

2.4.1.4 Minimum Redundancy Maximum Relevance (mRMR)

mRMR is a multivariate method proposed by Peng *et al*. to select a subset of features which have maximum relevance with the target response and minimum mutual information with each other [17]. Relevance or dependency is often measured in terms of mutual information. The mutual information between two random variables X and Y is defined as

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \tag{2.27}$$

where $p(x)$ and $p(y)$ are probability density functions and $p(x,y)$ is their joint probability density function.

mRMR first searches for the most relevant feature with the target class by considering $I(x_j;c)$. Then, other features are added to the previously selected subset in an iterative manner. In order to find the next feature to be added, Eq. (2.28) is used

$$\max_{x_j \in S_{m-1}} [I(x_j;c) - \frac{1}{m-1} \sum_{x_k \in S_{m-1}} I(x_j;x_k)]. \quad (2.28)$$

$I(x_j;c)$ denote the mutual information between feature x_j and the target response whereas $I(x_j;x_k)$ is the mutual information between x_j and x_k. Also, S_{m-1} denotes the previously selected subset of $m-1$ features. The algorithm stops when the size of the selected subset satisfies the predefined stopping criteria. Similar to chi-square, numerical features need to be discretized.

2.4.1.5 Relief

Relief is an iterative scheme [18]. In each iteration, it picks a random sample from the training set. The nearest sample from the same class (hit) and the closest sample from the other class (miss) are then identified using the Euclidean distance function. The algorithm assigns a weight to each feature using the distances to the nearest hit and nearest miss samples. As a final step, relief filters features with weight scores less than the selected threshold α. The pseudo code for relief algorithm is given in Algorithm 2-1 [18].

Algorithm 2-1: Relief algorithm

Relief (S, N, α)
Begin:
Separate *S into S^+ (positive samples) and S^- (negative samples)*
Let *w = (0, 0, . . . , 0)*
For i = 1 to N
Pick *a random instance $X \in S$*
Compute *Z^+ : the nearest sample in S^+*
Compute *Z^- : the nearest sample in S^-*
if *(X belongs to positive class)* ***then***

nearest-hit = Z^+; *nearest-miss* = Z^-

else

nearest-hit = Z^-; *nearest-miss* = Z^+

Update-weight *(W , X ,nearest-hit, nearest-miss)*

Set *relevance* = $\left(\frac{1}{N}\right)W$

For j = 1 to d

if($relevance_j \geq \alpha$) ***then***

x_j *is a relevant feature*

else

x_j *is an irrelevant feature*

End;

Update-weight (w, x, nearest-hit, nearest-miss)

For j = 1 to d

$$W_j = W_j - diff(x_j, nearest-hit)^2 + diff(x_j, nearest-miss)^2$$

In this algorithm, N , d and α are the total number of samples, the number of features and a selected threshold, respectively and x_j denotes the *j*th feature. This algorithm is noise-tolerant.

2.4.1.6 Correlation-based Feature Selection (CFS)

Most of the traditional filter methods rank features only by taking into account the predictive power of each individual feature. This means the correlation among features is not considered which may cause poor performance of the classifier due to selecting redundant features. Correlation-based Features Selection (CFS) proposed by Hall in 1999 evaluates the effectiveness of a subset of features by taking into account both the importance of each feature within the selected subset and the correlation among features in the subset [19]. In general, the algorithm tries to ignore irrelevant features because they do not bring in useful information while they cause a larger computational cost. An important advantage of this algorithm is that it does not incur the computational cost associated with iterative algorithms because it

computes the correlation matrix of feature to class and feature to feature in the first iteration and then uses the best first search algorithm to search within the feature subset space [9], [19].

Originally, CFS is designed to measure the correlation between nominal features but not numerical ones. Thus, it is necessary to discretize numeric features for CFS algorithm. Given a subset of m features, the merit of the set is computed as

$$\frac{\sum_{j=1}^{m} I(x_j;c)}{\sqrt{m+\sum_{j=1}^{m}\sum_{k=1}^{m} I(x_j;x_k)}}. \tag{2.29}$$

2.4.1.7 Conditional Mutual Information Maximization (CMIM)

Conditional Mutual Information Maximization (CMIM) ensures to select a small subset of features by maximizing the conditional mutual information between features and the target response. Conditional mutual information shows the amount of shared information between two random variables.

Let S_{m-1} denote previously determined subset of features. The next feature x_j to be added must be selected from the set of not previously selected feature $(S \setminus S_{m-1})$ as

$$arg\max_{x_j \in \Delta S}[\min_{x_k \in S_{m-1}} I(c;x_j \mid x_k)]. \tag{2.30}$$

2.4.2 Wrapper Methods

Wrapper algorithms select, evaluate and compare the performance of different subsets of features by taking into account a particular classification scheme. In other words, a predictive model evaluates different subsets of features using the training data and assigns a score to each subset based on a performance metric. Theses algorithms are either deterministic or randomized. Both of the types have the advantage of interacting with the classifier and taking into account feature

dependencies [14]. Although the deterministic algorithms are simpler than randomized algorithms, the risk of converging to a local optima is higher among them. On the other hand, randomized algorithms have a higher risk of overfitting. Genetic Algorithm (GA), Stepwise Forward Selection (SFS) and Stepwise Backward Selection (SBS) are three widely used wrapper methods.

2.4.2.1 Genetic Algorithm

Genetic Algorithm (GA) is an optimization method inspired from genetic selection which computes the best feature subset using heuristic search [20]. In each iteration of the algorithm, more powerful individuals are selected because they often survive and dominate the weaker ones in natural selection. GA benefits from two rules dominating in natural selection, namely crossover and mutation.

In order to solve the problem of feature selection, GA converts the problem of searching for an optimal solution to looking for an extrema (a maximum or a minimum) in the search space where each subset of features is represented by a point. It starts by generating a population of randomly selected individuals known as chromosomes. Then, a fitness function is selected based on which each individual of the population is evaluated and ranked. Higher-ranked individuals are selected to mate and generate a new population. Generation of new individuals is performed using crossover and mutation. Selection, fitness evaluation and generating new population steps are then repeated until the optimization objective is satisfied.

Crossover is the process of producing a new individual from two high-ranked individuals. The next operator to be applied is mutation. Mutation aims to make simple but random modification on the offspring. For instance, it can be defined as flipping 0 to 1 or vice versa for the case of binary chromosomes.

Generally, the fitness function is defined as a metric to quantify the performance of the classifier. For example, AUC can be employed as the fitness function. Also, to ensure that GA converges, a convergence criteria is needed. Usually, this stopping criteria is defined as the number of times GA executes without any improvement in the best value obtained from fitness function. Algorithm 2-2 shows the pseudo code of GA.

Algorithm 2-2: Simple Genetic Algorithm

Begin:
Let: *{ $l, M, R_e, P_m, P_c, Max_{iter}$ } be the design parameters*
Initialize *the population : current_pop*
For 1 to Max_{iter}
Evaluate *the population using a fitness function*
Select *pairs of individuals from current_pop: parents*
Elitism(R_e ***)***
Mutation(P_m ***)***
Crossover(P_c ***)***
Generate *a new population: new_pop*
current_pop = new_pop
End For;
End;

In this algorithm, $l, M, R_e, P_m, P_c, Max_{iter}$ denote the chromosome length, population size, elitism rate, mutation probability, crossover probability and maximum number of iterations (convergence criteria) respectively. GA controls the rate and type of selection, crossover and mutation using these tuning parameters.

2.4.2.2 Stepwise Forward Selection

Stepwise Forward Selection (SFS) is an iterative algorithm that starts with an empty set of features. Then, SFS adds predictors one by one to this initial model. It evaluates each candidate feature in terms of the classification performance using a metric such as AUC. The variable with maximum additive improvement to the

previously selected set is selected in each iteration [10]. The algorithm stops if adding a new feature does not improve the performance of the classifier. Let R_d denote the whole set of features then algorithm 2-3 shows the pseudo code of SFS.

Algorithm 2-3: Stepwise Forward Selection

Begin

Let $R_s = \varnothing$, $AUC_{best} = 0$, $AUC_{cand} = 0$

Do

Let *found = false;*

For $x_j \in R_d - R_s$:

if($Evaluate(R_s \cup \{x_j\}) > AUC_{cand}$ ***) then***

$AUC_{cand} = Evaluate(R_s \cup \{x_j\})$

$x_{cand} = x_j$

End For;

if($AUC_{cand} > AUC_{best}$ ***) then***

$AUC_{best} = AUC_{cand}$

$R_s = R_s \cup \{x_{cand}\}$

found = true;

While*(found);*

Return R_s *as the best subset;*

End*;*

In this algorithm, R_s denoted the selected subset of features. x_j and AUC_{cand} denote the next candidate feature and the *AUC* obtained from adding it to R_s, respectively. AUC_{best} is the best *AUC* obtained using the selected features and difference operator (‘ \ ’) is used to show that the selected feature x_j is dropped from R_d so that it would not be used in next iterations. The algorithm stops if adding any of the remaining features does not improve AUC_{best}.

2.4.2.3 Stepwise Backward Selection

Stepwise Backward Selection (SBS) works similar to SFS except that it starts with the whole set of predictors and drops one predictor at time. The selection of the next

feature to be eliminated is based on the improvements achieved in the AUC score with respect to the previously selected feature set [10]. Let R_d denote the whole set of features then algorithm 2-4 shows the SBS.

Algorithm 2-4: Stepwise backward selection

Begin

Let $R_s = R_d$, $AUC_{best} = 0$, $AUC_{cand} = 0$

Do

Let *found = false;*

For $x_j \in R_s$:

if($Evaluate(R_s \setminus \{x_j\}) > AUC_{cand}$ ***) then***

$AUC_{cand} = Evaluate(R_s \setminus \{x_j\})$

$x_{cand} = x_j$

End For;

if($AUC_{cand} > AUC_{best}$ ***) then***

$AUC_{best} = AUC_{cand}$

$R_s = R_s \setminus \{x_{cand}\}$

found = true;

While*(found);*

Return R_s *as the best subset;*

End*;*

2.4.3 Embedded Methods

In this group of schemes, the task of feature selection is embedded into the training of the classifier. Embedded methods are similar to wrapper methods in the sense that they interact with classifier. However, embedded methods do not need intensive computation [14], [21].

2.4.3.1 Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) generates a linear model such that the coefficients of correlated features are set to zero so that they do not contribute the classification task [22]. As a result, the model decides to use only a

subset of features. The coefficients are estimated by minimizing the following objective function:

$$\begin{aligned} &\textit{Minimize} \sum_{i=1}^{N} (c_i - \beta_0 - \boldsymbol{\beta} \boldsymbol{x}_i^T)^2, \\ &\textit{subject to} \sum_{j=1}^{d} \left|\beta_j\right| \le t, \quad \boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_d] \end{aligned} \tag{2.31}$$

where $\boldsymbol{x}_i$ and c_i are the sample and the class label, respectively and β_j denote the coefficient of the *j*th feature. $t \ge 0$ is a tuning parameter by which the amount of coefficient shrinkage is controlled.

Chapter 3

EXTRACTION OF ADDITIONAL PREDICTORS

3.1 Introduction

As it is already mentioned, the aim of this study is to compute an enriched feature set to develop an improved automated system for detection of people having undiagnosed T2DM or prediabetes. In order to achieve this, a data set that is rich in terms of risk factors, symptoms, laboratory tests, diagnoses, life style habits and medication is needed. The Pima Indian Diabetes data set (collected and published by National Institute of Diabetes and Digestive and Kidney Diseases) is commonly used for diabetes classification. Table 3.1 shows a brief list of publication on this dataset. For example, Temurtas et al. used multilayer neural network to classify Pima Indian data [23]. Lekkas et al. applied a fuzzy approach on the same data and accuracy of 79.37% was achieved [24].

Table 3.1: Comparison of different studies on diabetes classification

Study	Method	Metric	Score	Data set
Lekkas et al. [24]	eClass	Accuracy	79.37%	Pima Indian
Temurtas et al. [23]	MLNN with LM	Accuracy	82.37%	Pima Indian
Polat et al. [25]	GDA–LS-SVM (10-CV)	Accuracy	79.16%	Pima Indian
Meng et al. [26]	AIRS	Accuracy	67.40%	Pima Indian
Kayaer et al. [27]	GRNN	Accuracy	80.21%	Pima Indian

AIRS: Artificial Immune Recognition System
GDA: Generalized Discriminant Analysis
GRNN: General Regression Neural Network
LM: Levenberg–Marquardt algorithm
LS-SVM: Least Square Support Vector Machine
MLNN: Multilayer Neural Networks

Although the Pima Indian Diabetes dataset is commonly used for diabetes classification, characteristics of this data set make it inappropriate for our work from two different aspects. Firstly, this data set does not include a wide range of features. Each patient is represented using only eight predictors, all of which are well-known risk factors of diabetes. Secondly, the size of the data set is very small (only 768 samples).

3.2 The Dataset Employed

After conducting extensive survey on previously published work, a dataset collected as a part of National Health and Nutrition Examination Survey (NHANES) program is selected. This program is conducted by National Center for Health Statistics to represent demographic and biologic characteristics of U.S. non-institutionalized population. Although this dataset is not collected specifically for diabetes classification, it is rich in terms of relevant and potentially useful features. The collection of data in NHANES is questionnaire based. The collection of the NHANES data is done every two years. Each NHANES wave includes detailed information about health status and characteristics of participants categorized in five different groups namely, demographic data, dietary data, examination data, laboratory data and questionnaire data. Each group includes tens to hundreds of questions. Each question is represented by a question code (QCode). This multilayer categorization makes the job of finding a specific information easier.

Different subsets of the NHANES based datasets are previously utilized to study T2DM from different perspectives. Heikes et al. used logistic regression and classification tree models to develop a screening tool which can be used by public so that they determine whether they need to visit a health professional for further

examination. Their resulting screening tool includes 8 features namely, age, waist circumference, gestational diabetes, height, race, hypertension, family history and exercises. They obtained sensitivity and specificity of 77.65% and 51.36% [5]. Yu et al. in another research used 14 features to evaluate SVM performance on two different classification schemes which are different in terms of distribution of diabetic people. Best performance of SVM on first scheme was obtained using 8 features namely family history, age, race, weight, height, waist circumference, BMI, hypertension. In case of second scheme two more features namely, sex and physical activity are used. The AUC scores obtained from this study was 0.835 and 0.732 for the first and second scheme, respectively [28].

3.3 Feature Extraction from NHANES

Each participant is represented by an ID which makes it possible to find values of different variables for each patient. For example, in order to find age, ethnicity and education level of the participants, demographic data group which includes only one data file named as Demographic Variables & Sample Weights or in brief DEMO_F.XPT is utilized. In this data file there are 42 questions, each being represented by a distinct QCode. The Qcodes RIDAGEYR, RIDRETH1 and DMDEDUC2 provide the data we are looking for. The individuals who participated in NHANES had completed a household interview questionnaire. These individuals are defined as “interviewed”. Then, all interviewed participants completed one or more examination components in the Mobile Examination Center (MEC). These individuals are called “MEC examined”.

NHANES 2009-2010 is selected for this study. This wave includes 13,272 participants. However, only 10,253 of them satisfy the condition of being both

“interviewed” and “MEC examined”. Therefore, the target population of our research is extracted from this list of participants. As we have mentioned above, this data set is not collected for T2DM detection system development. As a matter of fact, some preprocessing should be performed to make it more compatible with our main objective. The preprocessing steps we followed are in parallel with previous efforts of data extraction for diabetes classification [29], [30], [31]. Initially, pregnant women are discarded due to probable gestational diabetes using the variable RIDEXPRG. Also, participants aged less than 20 years are excluded using the variable RIDAGEYR. A data set of size 5991 participants is obtained by applying these general rules. This group of people also includes those having diagnosed prediabetes or diagnosed T2DM that are to be discarded from further studies. Identification of diagnosed people is done using QCodes DIQ010 and DIQ160 from DIQ_F data file in the questionnaire category. These questions are used to ask if the participant is already diagnosed with diabetes or prediabetes by a doctor or other health professionals. Positive respondents of these questions were excluded from the population (n=880). Negative respondents are examined using laboratory tests to be classified as normal (no diabetes), undiagnosed T2DM or undiagnosed prediabetes.

The samples are labeled as negative or positive using the respective laboratory tests namely, FPG, OGTT and HbA1c as presented in Chapter 1. Similar to many other surveys, NHANES has missing values. Participants who do not have any of the aforementioned laboratory test results are also discarded (n=789).

The laboratory tests of samples who answered negatively to questions DIQ010 and DIQ160 are evaluated in the following order:

- The participant is evaluated for having undiagnosed T2DM using his/her HbA1c level.

- If HbA1c level is missing, FPG or OGTT with respect to the fasting hour criteria is examined for having undiagnosed T2DM.
- If none of the previous steps led to labeling the patient, the two previous steps are repeated for having undiagnosed prediabetes.
- If neither of the above cases occurred, the participant is categorized as normal provided that all three laboratory test satisfy the ranges of being normal.

In this study, three risk assessment tools, namely, FINDRISC [32], CANRISK [33] and ADArisk [34] are used to compute a list of 10 predictors. Systolic blood pressure and diastolic blood pressure are also added to this list [6]. These features are represented in Table 3.2 with their matching QCodes in NHANES. After dummy representation of the categorical features, this set of features corresponds to an 18-dimensional feature vector and it is used as the benchmark feature set. In the rest of the thesis, it is called the reference feature vector.

Table 3.2: List of the predictors used in this study

Feature	**Type**	**DATAFILE.QCODE**
Age	numerical	DEMO_F.RIDAGEYR
Gender	binary	DEMO_F.RIAGENDR
Ethnicity	categorical	DEMO_F.RIDRETH1
Education	categorical	DEMO_F.DMDEDUC2
Body Mass Index *(kg/m²)*	numerical	BMX_F.BMXBMI
Waist circumference *(cm)*	numerical	BMX_F.BMXWAIST
Family history	binary	MCQ.MCQ300C
High blood sugar	binary	DIQ_F.DIQ160
Systolic Blood Pressure	numerical	BPX_F.BPXSY1,BPX_F.BPXSY2,BPX_F.BPXSY3,BPX_F.BPXSY4
Diastolic Blood Pressure	numerical	BPX_F.BPXDI1,BPX_F.BPXDI2,BPX_F.BPXDI3,BPX_F.BPXDI4
Hypertension	binary	BPQ_F.BPQ020,BPQ_F.BPQ050A, Mean of Systolic BP, Mean of Diastolic BP
Physical activity	binary	PAQ_F.PAQ610,PAQ_F.PAQ615,PAQ_F.PAQ625,PAQ_F.PAQ630,PAQ_F.PAQ640,PAQ_F.PAQ645,PAQ_F.PAQ655,PAQ_F.PAQ660,PAQ_F.PAQ

		670,PAQ_F.PAQ675

After discarding diagnosed people and participants with missing information for any of the predictors considered and excluding diagnosed prediabetes and T2DM patients, 4322 samples remained. The distribution of samples in NHANES 2009-2010 is shown in Table 3.3.

Table 3.3: Number of samples within each population

Categories	#Samples	Class definition
Normal	2059	Negative class
Undiagnosed prediabetes	2003	Positive class
Diagnosed prediabetes	130	Excluded
Undiagnosed diabetes	260	Positive Class
Diagnosed diabetes	715	Excluded

3.4 Computation of an Enriched Set of Features

In addition to standard features, a wide range of features are extracted from the questionnaire category to identify additional predictors of diabetes using feature selection methods. These features are selected to cover various aspects of life style and biological characteristics such as alcohol use, consumption of organic food, kidney condition, mental health, drug use and many other various disease. iptions are given in Appendix B.

Table 3.4 shows the numbers and types of features selected from each category for evaluation. These features are called additional features in the following context. Appendix A presents the complete list of the categories and the questions employed

to define novel features. The list of all questions used in our study and their corresponding descriptions are given in Appendix B.

Table 3.4: Number of questions and extracted features from NHANES

Category	**#Questions**	**#Extracted Features**
Alcohol Use	8	1 binary 3 numerical
Arthritis	128	27 binary
Audiometry	16	3 binary 1 categorical
Blood Pressure and Cholesterol	20	7 binary
Bowel Health	11	4 categorical
Cardiovascular Disease	16	4 binary 1 categorical
Consumer Behavior	17	2 numerical
Consumer Behavior Phone Follow-up	88	3 binary 23 categorical
Current Health Status	13	1 categorical 3 numerical
Dermatology	8	2 numerical
Diabetes	19	1 binary
Diet Behavior and Nutrition	54	2 binary 4 categorical
Drug Use	41	2 binary
Health Insurance	15	1 binary
Hepatitis C Follow Up	37	1 categorical
Hospital Utilization and Access to Care	9	1 binary 1 categorical
Income	14	1 numerical
Kidney Conditions – Urology	15	5 binary 3 categorical
Medical Conditions	80	24 binary
Mental Health - Depression Screener	10	10 categorical
Occupation	23	4 binary
Oral Health	8	3 binary
Osteoporosis	91	4 binary
Physical Functioning	34	4 binary 19 categorical
Prescription Medications	6	1 binary 1 numerical
Reproductive Health	57	6 binary

Respiratory Health & Disease	14	6 binary 3 categorical 2 numerical
Sexual Behavior	48	5 binary
Sleep Disorders	3	1 binary 1 numerical
Smoking – Cigarette Use	36	1 binary 1 categorical
Smoking – Household Smokers	4	1 binary
Smoking – Recent Tobacco Use	23	1 numerical
Weight History	62	1 numerical

Chapter 4

EXPERIMENTAL RESULTS

4.1 Experimental Results

Four different classifiers are used to build models using the reference feature vector. These classifiers are evaluated using 10-fold cross-validation. The average AUC scores, accuracy, sensitivity and specificity values provided by these classifiers are reported in Table 4.1. It can be seen that the performance scores of these four classifiers are comparable. These results are used as our reference performances in the rest of the experiments.

Table 4.1: The performance scores of different classifiers using the reference feature vector

Classifier	AUC	Accuracy	Sensitivity	Specificity
Logistic Regression	0.7645	70.52%	72.74%	68.09%
LASSO	0.7645	70.55%	72.56%	68.33%
Linear SVM	0.7642	70.45%	72.87%	67.80%
Polynomial SVM	0.7645	70.80%	73.66%	67.65%

In order to evaluate the association between the predictors and the positive class, chi-square is used for categorical features and t-test is applied on numerical variables. The corresponding p-values achieved for the reference predictors are shown in Table 4.2. Features with p-value less than 0.0001 can be considered as being significantly related to diabetes. Ethnicity is a categorical feature with 5 values namely, Mexican American, non-Hispanic white, non-Hispanic black, other Hispanic and other races. Non-Hispanic white value is selected as a reference for dummy

representation. Similarly, in education, high school level is selected as the reference for the dummy representation. p-values are not available for the reference categories.

Table 4.2: p-values of the reference predictors

Feature	**p-value (X^2/t-test)**	**Feature**	**p-value (X^2/t-test)**
Age	**< 0.0001**	Systolic BP	**< 0.0001**
BMI (Kg/m^2)	**< 0.0001**	Diastolic BP	0.7327
Waist circumference (cm)	**< 0.0001**	Family history	**< 0.0001**
Gender	0.0734	Hypertension	**< 0.0001**
High Blood Sugar	**< 0.0001**	Physical Activity	**< 0.0001**
Ethnicity		Education	
Mexican American	0.0750	< 9th Grade	**< 0.0001**
Other Hispanic	0.6313	9-11th Grade	0.0010
NH White		High School	0.1598
NH Black	**< 0.0001**	Some College	0.0028
Other Race	0.7366	College Graduate	

The aim of this study is to improve the classifier performance by finding an enriched set of features which are either directly or inversely related to diabetes. Various feature selection algorithms from filter, wrapper and embedded categories are applied on additional features represented in Table 3.4 to select a useful subset of them. Since additional features suffer from missing value problem, mean and mode imputation are performed using the available samples. Also, all numerical features are normalized using zero mean unit variance normalization.

In the first set of experiments, five filter approaches having a similar experimental setting are studied. These are namely chi-square, IG, relief, t-test and CFS. In all experiments, logistic regression is used as our predictive model. For all five schemes, the 200 top-ranked additional features are considered together with the reference

features. In the first step, the classifier is trained using only the reference feature vector. Then, during next steps, the 200 top-ranked additional features are appended in groups of 1, 5 and 50 features. To be more specific, appending the additional features is as follow: the 10 top-ranked additional features are appended one by one. Then, additional features ranked between 10 and 50 are appended in groups of 5 and finally those features ranked between 50 and 200 are appended by groups of 50. Figure 4.1 presents the average AUC scores achieved using the 15 top-ranked features. Since the performance deteriorates, the scores corresponding to larger number of features are not presented in this figure.

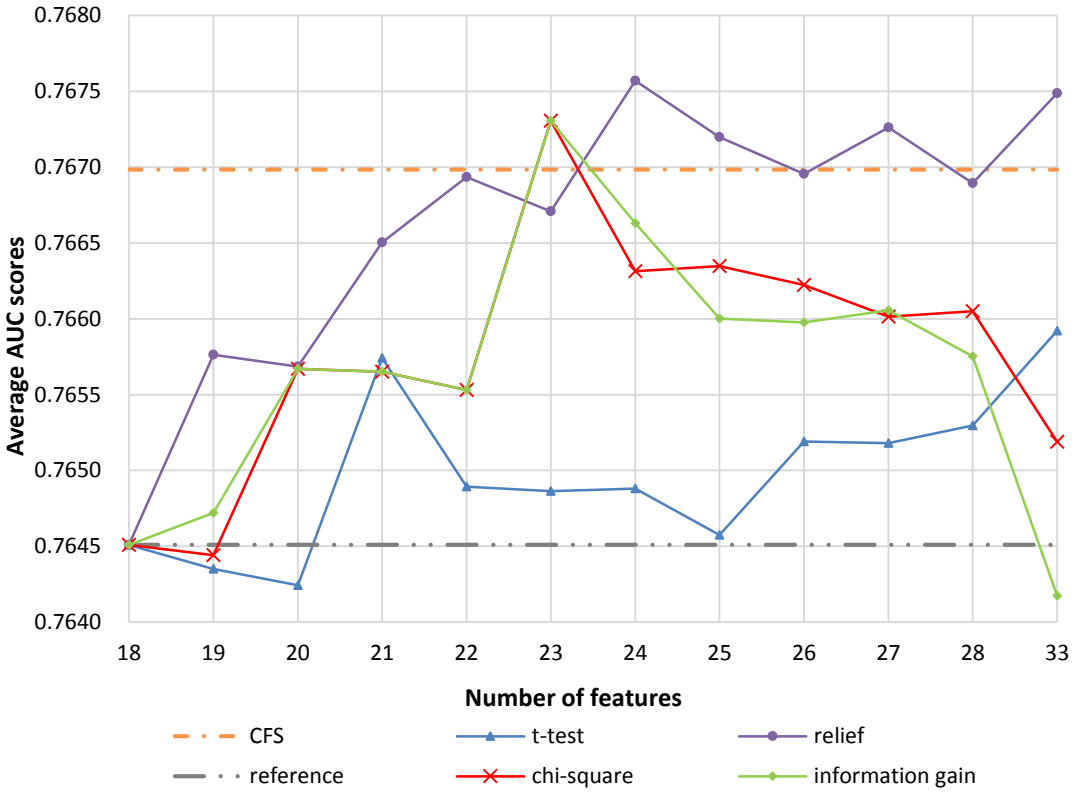

Figure 4.1: Average AUC scores achieved by fitting LR method on additional features

As it can be seen in the figure, all feature selection methods improved the reference model performance. Relief achieved the best performance improvement with AUC of

0.7676 using 6 additional features. Both chi-square and information gain used 5 additional features to achieve maximum AUC of 0.7673. CFS also improved the reference model performance by using 6 additional features. In the simulation studies, top-ranked 10, 50, 100 and the whole set of additional features are used for evaluating the effectiveness of CFS on NHANES dataset. It is observed that the best AUC is obtained using 10 additional features.

It is important to note that CFS removes useless features. In addition, according to Table 4.2, we observed that some of the reference features like gender and diastolic blood pressure are not significantly important. In order to prevent CFS from removing non-significant reference features, only additional features are used as the input for this algorithm. That is, the selected subset of additional features by CFS is added to the complete vector of reference features and they are used together as the input of the classifier. The performances are observed to degrade in general as more features are employed.

Table 4.3 shows the 10 top-ranked features of the univariate methods (CFS is multivariate) that also provide the ranks of the features. The rank of each predictor is also presented. It can be seen in the table that the intersection set of the features computed by these four univariate methods includes 6 features. These features cover a wide range of characteristics of the individuals from lifestyle habits and medications to symptoms. Except for t-test, all other three methods detected question DED120 either as the first or the second important feature. It corresponds to the following question: “How much time did you usually spend outdoors between 9 in the morning and 5 in the afternoon on the days that you worked or went to school?” The range of valid answers is in terms of number of minutes from 14 to 480 and the

average answer is 117.30 and 87.78 for negative and positive classes, respectively. This means people with diabetes spend less hours outside.

Table 4.3: Top 10 additional features computed by each method

Qcode	chi-square		IG		Relief		t-test	
	Score	rank	score	rank	score	rank	Score	Rank
DED120	328.4	1	0.045	1	0.023	2		
DED125	328.0	2	0.043	2	0.030	1	9.1	9
BPQ040A	208.3	3	0.028	3	0.013	7	15.1	1
BPQ050A	188.8	4	0.025	4	0.014	6	14.3	2
BPQ100D	147.2	5	0.020	5	0.015	3	12.6	3
BPQ080	117.3	6	0.015	6			11.1	4
MCQ160A	114.2	8	0.015	9	0.007	10	11.0	6
CDQ010	113.7	9	0.015	8			11.0	5
DUQ200	113.2	7	0.014	10	0.014	5	10.7	7
RXDCOUNT	97.2	10	0.015	7				
BPQ100A	82.6	11					9.3	8
CDQ001	73.6	13					8.8	10
OHQ850	36.1	31			0.014	4		
HSQ480	25.5	50			0.010	9		
RHQ420	13.4	115			0.012	8		

Question DED125 asks about the same life style behavior but this time it is about spending time outdoor between 9 in the morning and 5 in the afternoon on the days that the participant did not go to work/school. Again, normal people spend more time outdoor which means this group are more active.

Questions BPQ040, BPQ050 and BPQ100D belong to Blood Pressure and Cholesterol data file of questionnaire category in NHANES 2009-2010. BPQ040A asks whether the participant have ever been told to take prescribed medicine for high blood pressure/hypertension. BPQ050A and BPQ100D asks if the participant followed the advice and currently uses the prescribed medicine for high blood

pressure/hypertension. Based on different risk assessment tools hypertension can be considered as a symptom of T2DM [32], [33], [34].

Another question from the intersection subset of feature selection methods represented in Table 4.3 is MCQ160A. This question asks “Doctor ever said you had arthritis?” Arthritis is a way of referring to any kind of joint disease and it can leads to disability of the patient. Untreated diabetes causes pain and stiffness in joints which is known as diabetic arthropathy (arthritis). This proves that arthritis is an informative complication of T2DM.

Next question from the aforementioned intersection is DUQ200 which asks about using drugs such as Marijuana or Hashish. Alshaarawy et al. reported that T2DM is inversely associated with recently active cannabis smoking behavior of patients [35].

RHQ420 is found to be important only by one of the algorithms (relief). This question asks the participant to answer whether she has ever taken birth control pills. Berenson et al. conducted a survey to study effect of using two different birth control methods on glucose and insulin levels of postmenopausal women with various ethnicities. Their experimental results show slight elevation of insulin and glucose levels either the postmenopausal women used injectable or pill medications as a mean of pregnancy prevention [36].

The maximum AUC scores achieved and the corresponding numbers of features are presented in Table 4.4. In all experiments, logistic regression classifier is utilized. Slight improvements are generally achieved using the additional features. This is most probably due to highly sparse and noisy dataset employed in this study. For

example, some of the features such as ARQ034C and DBQ925F have 99 percent missing values which makes them highly unreliable.

Table 4.4: Maximum AUC results obtained by fitting LR in this study

Feature selection method	Maximum AUC	#Features on maximum AUC	#Additional features used
Reference	0.7645	18	0
T-test	0.7659	33	200
Chi-square	0.7673	23	200
IG	0.7673	23	200
Relief	**0.7676**	24	200
CFS	0.7670	6	10
mRMR (1st Exp.)	0.7668	53	50
CMIM	0.7668	63	50
GA	0.7663	24	50
SFS	0.7644	**5**	50
LASSO	0.7617	31	960

It can be seen in the table that the best AUC score is provided by relief and the least number of additional features is used by SFS. AUC scores obtained from chi-square and information gain are very close to that of relief.

The best scores achieved using mRMR and CMIM that take into account both relevance and redundancy of the predictors are also presented in Table 4.4. Since mRMR is a multivariate method, two different experiments are conducted. In the first experiment, using the chi-square test, 50 top-ranked additional features are computed. Then, this subset is appended to the reference feature vector and the composite vector is considered as the input of mRMR. After the ranked list is obtained, reference features are discarded from the ranked list so that a list of additional features ranked by mRMR is obtained. Finally, taking into account their

ranks, the additional features are added to that model, incrementally. The scores presented in Table 4.4 correspond to this experimental setup.

In the second experiment, 50 top-ranked additional features obtained from applying chi-square test is appended to reference feature vector. This subset is the input of mRMR as in the previous experiment. In this case, 18 top-ranked features obtained from mRMR algorithm are used to generate a baseline model and the remaining features which may include reference features are added to them in an incremental manner. The experimental results are presented in Figure 4.2. It can be seen in the figure that the results are generally inferior to those computed using the univariate methods such as chi-square.

The intersection of features found by mRMR with the other four univariate methods includes only one question that is DUQ200. Other features obtained from mRMR include questions related with medication, oral health status, eating styles, depression and physical functioning characteristics. For instance, taking Atenolol which is a medication for treating angina, hypertension and heart attack is found to be highly relevant with diabetes. Another relevant medication found by this method is using the birth control pills at the time of answering the questionnaire. Having hysterectomy, treatment for gum disease, buying organic fruits and feeling bad about yourself are also among features selected by mRMR which are very different from those computed using the univariate filter methods. Hysterectomy is a surgical operation to remove uterus. Appiah et al. studied the association between hysterectomy and incidence of diabetes among postmenopausal women using NHANES data set. The results from their study show that the incidence of diabetes

among women with no hysterectomy is less than women who had hysterectomy operation [37].

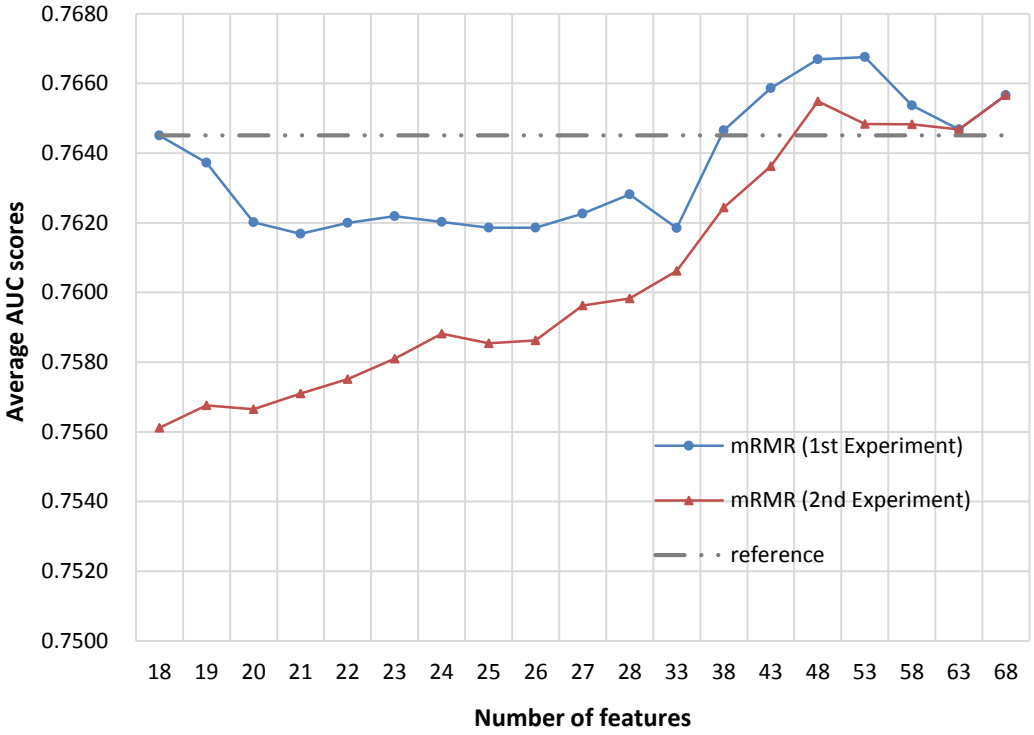

Figure 4.2 The average AUC scores obtained using mRMR in two different experimental setups.

In our simulations, various values are evaluated for the tuning parameters of GA. The best scores are achieved for the following settings:

- length of chromosome = 50
- size of population = 100
- maximum number of iterations = 20
- rate of elitism = 0.5
- probability of mutation = 0.1
- probability of crossover = 0.6

GA is run on the top-ranked (using chi-square) 50 features. Figure 4.3 presents the average and best AUC values obtained in the first fold. It can be seen in the figure

that 20 iterations were enough for the GA to converge. Using more number of iterations did not yield in improvement therefore, in order to benefit from computational time we selected the tuning parameter with 20 iterations. This is true for almost all folds.

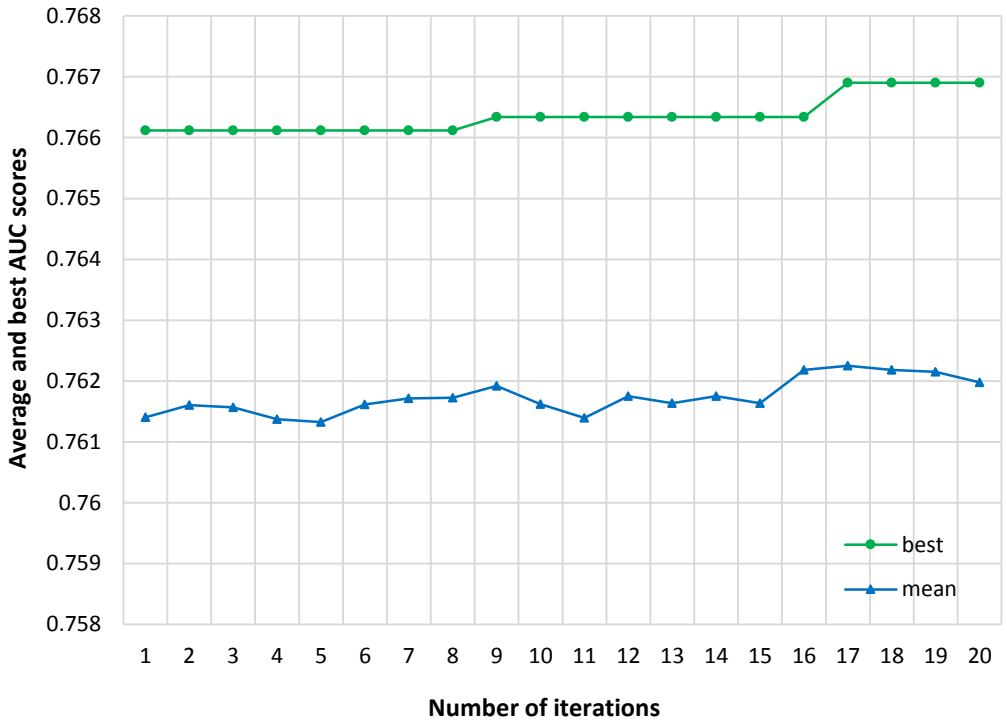

Figure 4.3: The average and best AUC scores achieved by GA in the first fold

The whole set of features including the reference and additional ones are used as the input for LASSO classifier. The optimal feature subset selected by LASSO include both the reference and additional features. That is, the coefficient of some of the reference features were set to zero during generation of the model and hence they did not contribute in the classification task. For example, physical activity, systolic blood pressure and having some college level of education are not considered to be important by LASSO in most of the folds.

More experiments are conducted using the intersection and union of the 30 top-ranked features obtained from t-test, chi-square, relief and mRMR to study T2DM prediction. Since the AUC of information gain is similar to that of chi-square and their top-ranked features entirely overlap, information gain is not considered. We used the first 30 features because, among all methods, the maximum number of features which resulted in maximum AUC is 30. Two experiments are conducted. In the first, the intersection set is appended to the reference features and, in the second, the union set is appended to reference features. Results of these experiments are presented in Table 4.5.

Table 4.5: AUC scores achieved using intersection and union sets

	Reference	**Intersection set**	**Union set**
LR	0.7645	0.7649	0.7691
LASSO	0.7645	0.7649	0.7687
Radial SVM	0.7645	0.7605	0.7648

It can be seen from the table that using union subset with 65 features improves the AUC results more than using the intersection subset with only four features namely, BPQ040A, BPQ100D, DUQ200 and MCQ160A.

4.2 Generating Models Using Diagnosed T2DM

As we mentioned before, diagnosed T2DM patients are generally discarded from the population under study because they might bias the classification results [29]. In this thesis, further experiments are performed to study the possibility of employing data from diagnosed T2DM patients as an additional data source during model generation for classification of undiagnosed T2DM patients.

Two binary classification problems are defined for this purpose. In the first problem (Problem I), samples having either undiagnosed T2DM or undiagnosed prediabetes are the positive class whereas people having normal glucose levels are the negative class. The positive class of the second problem (Problem II) includes individuals having undiagnosed T2DM and the negative class consists of people who are either normal or have prediabetes. For both problems, 10-fold cross validation is applied on the data and in each iteration the performance of the classifier is studied with and without enriching the training data using the diagnosed patients. The experimental setup for the first fold is illustrated in Figure 4.4.

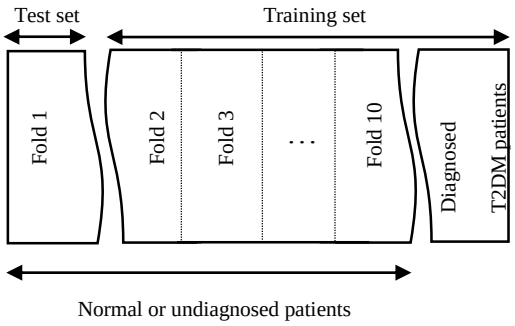

Figure 4.4: Employing data from diagnosed T2DM patients during model generation

The distribution of participants for two different problems are presented in Table 4.6.

Table 4.6: Number of samples with respect to problem definition

Categories	#Samples	Problem I	Problem II
Normal	2059	Negative	Negative
Undiagnosed prediabetes	2003	Positive	Negative
Diagnosed prediabetes	130	Excluded	Excluded
Undiagnosed diabetes	260	Positive	Positive
Diagnosed diabetes	117 or 845	Additional positive	Additional positive

The experiments are conducted on the reference predictors. For the diagnosed participants, there are many with missing values. Thus, using 117 samples without

any missing values and the whole population of 845 diagnosed T2DM patients are studied separately.

Logistic regression is used for model generation as before. The results are presented in Table 4.7. It can be seen that the performance of the reference system is not improved by using data from diagnosed patient.

Table 4.7: Average measurement result obtained by fitting LR model

	External Data	**AUC**
Problem I	Reference	0.7645
	Complete	0.7606
	Incomplete	0.7616
Problem II	Reference	0.8161
	Complete	0.8144
	Incomplete	0.8116

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusions

In this thesis, we studied detection of people having undiagnosed prediabetes or T2DM using different machine learning techniques. The use of data from diagnosed T2DM population as an additional source of data for model generation is also addressed. Data extracted from NHANES 2009-2010 is used for conducting all the experiments.

In this thesis, the main goal was to enrich the reference list of predictors with other life style habits, medications, diagnoses and symptoms for improving the detection performance. The reference features are known to be highly related with this disease. By applying feature selection methods, we tried to identify additional predictors which may contribute to the detection and the use of drugs such as Marijuana and Hashish are found to contribute to the detection of diabetes. Using organic foods and taking medications like Atenolol and birth control pills are some of the other features found to be discriminative between normal people and patients having either prediabetes or T2DM. Number of medications used by the patient is another risk factor found to be related with diabetes mellitus.

Using data from diagnosed patients of this chronic disease as an external source of data during classifier training is not found to be useful. This is mainly due to the

differences in the characteristics of this already diagnosed and undiagnosed patients. In particular, diagnosed patients already suffer from many complications and symptoms that are caused by diabetes whereas the undiagnosed patients either does not have these characteristics or the disease is not progressed enough to be identified.

It is also observed that the definition of the positive and negative classes results in different performance measures. In general, when the negative class is defined to cover undiagnosed T2DM and undiagnosed prediabetes, the performance scores are inferior to the case when it includes only the undiagnosed T2DM. This is mainly due to the fact that the characteristics of patients having prediabetes are more similar to that of normal people than those having T2DM.

Since the dataset employed is collected using questionnaires, it is noisy and sparse. In addition, a considerable percentage of some of the features are missing. This can be considered as one of the key factors limiting the improvements achieved.

5.2 Future Work

We believe that the 10 top-ranked features found by different feature selection methods in this study need to be studied more meticulously. For example, it is important to understand the exact relationship between these features and diabetes mellitus. We suggest that using an Electronic Health Record based dataset may help to identify novel predictors that improve the results significantly as the data in health records are more reliable and less sparse. The use of more complex imputation methods such as kNN may also contribute to the detection performance.

REFERENCES

[1] "Global report on diabetes", *WHO Press*, Geneva, 2016.

[2] "Complications", [Online]. Available: http://www.diabetes.org

[3] A. M. Spijkerman, J. M. Dekker, G. Nijpels, M. C. Adriaanse, P. J. Kostense, D. Ruwaard, C. D. Stehouwer, L. M. Bouter and R. J. Heine, "Microvascular complications at time of diagnosis of type 2 diabetes are similar among diabetic patients detected by targeted screening and patients newly diagnosed in general practice: the hoorn screening study", *Diabetes care*, vol. 26, no. 9, pp. 2604-2608, 2003.

[4] M. I. Harris, R. Klein, T. A. Welborn and M. W. Knuiman, "Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis", *Diabetes Care*, vol. 15, no. 7, pp. 815-819, 1992.

[5] K. E. Heikes, D. M. Eddy, B. Arondekar and L. Schlessinger, "Diabetes risk calculator", *Diabetes Care*, vol. 31, no. 5, pp. 1040-1045, 2008.

[6] L. Geiss, C. James, E. Gregg, A. Albright, D. Williamson and C. Cowie, "Diabetes risk reduction behaviors among U.S. adults with prediabetes", *American Journal of Preventive Medicine*, vol. 38, no. 4, pp. 403–409, 2010.

[7] P. J. Garcia-Laencina, J.-L. Sancho-Gomez and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review", *Neural Computing and Applications,* vol. 19, no. 2, pp. 263-282, 2010.

[8] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification", 2nd ed., *Wiley Interscience*, pp. 680, 2000.

[9] M. A. Hall, "Correlation-based feature selection for machine learning", *Hamilton: The University of Waikato*, pp.198, 1999.

[10] G. James, D. Witten, T. Hastie and R. Tibshirani, "An introduction to statistical learning with applications in R", *Springer Text in Statistics*, p. 426, 2013.

[11] D. Banks, L. House, F. McMorris, P. Arabie and W. Gaul, "Classification, clustering, and data mining applications", *Springer Berlin Heidelberg*, pp. 639-647, 2004.

[12] T. D. Pigott, "A review of methods for missing data", *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353-383, 2001.

[13] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.

[14] Y. Saeys, I. Inza and P. Larranaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[15] X.-Q. Zeng and G.-Z. Licorresponding, "Supervised redundant feature detection for tumor classification", *BMC Med Genomics*, vol. 7, no. 2, pp. S2:S5, 2013.

[16] T. M. Franke, T. Ho and C. A. Christie, "The chi-square test: often used and more often misinterpreted", *American Journal of Evaluation*, vol. 33, no. 3, pp. 448-458, 2012.

[17] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[18] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm", *AAAI-92*, pp. 129–134, 1992.

[19] M. A. Hall, "Correlation-based feature Selection for discrete and numeric class machine learning", *17th International Conference on Machine Learning*, pp. 359-366, 2000.

[20] M. Melanie, "An introduction to genetic algorithms", vol. 24, *The MIT Press*, pp. 293-315, 1999.

[21] Y. Huanga, P. McCullaghb, N. Blackb and R. Harperc, "Feature selection and classification model construction on type 2 diabetic patients data", *Artificial Intelligence in Medicine,* vol. 41, no. 3, pp. 251-262, 2007.

[22] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society,* vol. 58, no. 1, pp. 267-288, 1996.

[23] H. Temurtas, N. Yumusak and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", *Expert Systems with Applications,* vol. 36, no. 4, pp. 8610-8615, 2009.

[24] S. Lekkas and L. Mikhailov, "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases", *Artificial Intelligence in Medicine,* vol. 50, no. 2, pp. 117-126, 2010.

[25] K. Polat and S. Güneş, "Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm", *Expert Systems with Applications,* vol. 34, no. 1, pp. 773–779, 2008.

[26] L. Meng, P. v. d. Putten and H. Wang, "A comprehensive benchmark of the artificial immune recognition system (AIRS)", *First International Conference on Advanced Data Mining and Applications,* pp. 575-582, 2005.

[27] "Medical diagnosis on Pima Indian diabetes using general regression neural networks", *International Conference on Artificial Neural Networks and Neural Information Processing,* pp. 181-184, 2003.

[28] W. Yu, T. Liu, R. Valdez, M. Gwinn and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes", *BMC Medical Informatics and Decision Making,* vol. 10, no. 16, pp. 1, 2010.

[29] T. M. Dall, K. M. V. Narayan, K. B. Gillespie, P. D. Gallo, T. D. Blanchard, M. Solcan, M. O. Grady and W. W. Quick, "Detecting type 2 diabetes and prediabetes among asymptomatic adults in the united states: modeling american diabetes association versus US preventive services task force diabetes screening guidelines", *Popul Health Metrics,* vol. 12, no. 1, pp. 1, 2014.

[30] J. Aponte, "Prevalence of normoglycemic, prediabetic and diabetic A1c levels", *World Journal of Diabetes,* vol. 4, no. 6, pp. 349-357, 2013.

[31] L. Zhang, Z. Zhang, Y. Zhang, G. Hu and L. Chen, "Evaluation of Finnish diabetes risk score in screening undiagnosed diabetes and prediabetes among U.S. adults by gender and race: NHANES 1999-2010", *PLoS ONE,* vol. 9, no. 5, pp. e97865, 2014.

[32] J. Lindstrom and J. Tuomilehto, "The diabetes risk score: a practical tool to

predict type 2 diabetes risk.", *Diabetes Care,* vol. 26, no. 3, pp. 725-731, 2003.

[33] C. Robinson, G. Agarwal and K. Nerenberg, "Validating the CANRISK prognostic model for assessing", *CHRONIC DIS CAN,* vol. 32, pp. 19-31, 2011.

[34] "American Diabetes Association. standards of medical care in diabetes-2016", *Diabetes Care,* vol. 39, no. 1, pp. S1-S106, 2016.

[35] O. Alshaarawy and J. Anthony, "Cannabis smoking and diabetes mellitus: results from meta-analysis with eight independent replication samples", *PubMed,* vol. 26, no. 4, pp. 597-600, 2015.

[36] A. B. Berenson, P. V. D. Berg, K. J. Williams and M. Rahman, "Effect of injectable and oral contraceptives on glucose and insulin levels", *PubMed Central,* vol. 117, no. 1, pp. 41-47, 2011.

[37] D. Appiah, S. J. Winters and C. A. Hornung, "Bilateral oophorectomy and the risk of incident diabetes in postmenopausal women", *Diabetes Care,* vol. 37, no. 3, pp. 725-733, 2014.

APPENDICES

Appendix A: Question Codes of Additional features

Category	QCODE
Alcohol Use	ALQ120Q, ALQ120U, ALQ130, ALQ140Q, ALQ150
Arthritis	ARQ010, ARQ020A, ARQ020B, ARQ020C, ARQ020D, ARQ020E, ARQ020F, ARQ020G, ARQ021BA, ARQ022AA, ARQ022AB, ARQ022AC, ARQ022AD, ARQ022AE, ARQ022AF, ARQ022AG, ARQ024A, ARQ024B, ARQ024C, ARQ024D, ARQ024E, ARQ024F, ARQ024G, ARQ030A, ARQ030B, ARQ030C, ARQ030D, ARQ030E, ARQ034A, ARQ034B, ARQ034C, ARQ034D, ARQ034E, ARQ070, ARQ080, ARQ110, ARQ112A, ARQ112B, ARQ125C, ARD125A, ARQ125D, ARQ125E
Audiometry	AUQ131, AUQ191, AUQ260, AUQ270
Blood Pressure and Cholesterol	BPQ020, BPQ040A, BPQ050A, BPQ057, BPQ080, BPQ100A, BPQ100C, BPQ100D
Bowel Health	BHQ010, BHQ020, BHQ030, BHQ040
Cardiovascular Disease	CDQ001, CDQ002, CDQ003, CDQ008, CDQ010
Consumer Behavior	CBD160, CBQ190
Consumer Behavior Phone Follow-up Module - Adult	CBQ505, CBQ550, CBQ660, CBQ665, CBQ670, CBQ675, CBQ680, CBQ700, CBQ685, CBQ790, CBQ795, CBQ800, CBQ805, CBQ815, CBQ820, CBQ825, CBQ785, DBQ780, DBQ750, DBQ760, DBQ770, CBD710, CBD715, CBD720, CBD725, CBD730, CBD735
Current Health Status	HSD010, HSQ480, HSQ490, HSQ496
Depression Screener	DPQ010, DPQ020, DPQ030, DPQ040, DPQ050, DPQ060, DPQ070, DPQ080, DPQ090, DPQ100
Dermatology	DED120, DED125
Diabetes	DIQ010, DIQ160, DIQ170
Diet Behavior and Nutrition	DBQ235A, DBQ235B, DBQ235C, DBQ915, DBQ925a, DBQ925b, DBQ925c, DBQ925d, DBQ925e, DBQ925f, DBQ925g, DBQ925h,

	DBQ925i, DBQ925j
Drug Use	DUQ200, DUQ240, DUQ250
Health Insurance	HIQ011
Hepatitis C Follow Up	HCQ100
Hospital Utilization and Access to Care	HUQ040, HUQ050, HUQ090
Income	IND235
Kidney Conditions – Urology	KIQ005, KIQ010, KIQ022, KIQ025, KIQ026, KIQ042, KIQ046, KIQ480
Medical Conditions	MCQ010, MCQ035, MCQ051, MCQ070, MCQ082, MCQ086, MCQ140, MCQ160A, MCQ160B, MCQ160C, MCQ160D, MCQ160E, MCQ160F, MCQ160G, MCQ160K, MCQ160L, MCQ160M, MCQ160N, MCQ170K, MCQ170L, MCQ170M, MCQ220, MCQ300A, MCQ300B
Occupation	OCQ510, OCQ530, OCQ550, OCQ570
Oral Health	OHQ850, OHQ855, OHQ860
Osteoporosis	OSD110a, OSD110b, OSD110c, OSQ060, OSQ130, OSQ150
Physical Functioning	PFQ020, PFQ030, PFQ049, PFQ051, PFQ054, PFQ057, PFQ059, PFQ061B, PFQ061C, PFQ061D, PFQ061E, PFQ061F, PFQ061G, PFQ061H, PFQ061I, PFQ061J, PFQ061K, PFQ061L, PFQ061M, PFQ061N, PFQ061O, PFQ061P, PFQ061Q, PFQ061R, PFQ061S, PFQ061T
Prescription Medications	RXDDRGID (See Appendix 2), RXDCOUNT
Reproductive Health	RHQ131, RHQ172, RHQ420, RHQ510, RHD280, RHD442
Respiratory Health & Disease	RDQ031, RDQ070, RDQ080, RDQ090, RDQ100, RDQ134, RDQ135, RDQ137, RDD120, AGQ030
Sexual Behavior	SXQ260, SXQ265, SXQ270, SXQ272, SXQ753
Sleep Disorders	SLD010H, SLQ060
Smoking – Cigarette Use	SMQ020, SMQ040, SMQ050Q, SMD650
Smoking – Household Smokers	SMD410
Smoking – Recent Tobacco Use	SMQ720
Weight History	WHD140

Appendix B: List of All Extracted Questions

QCode	Question Description
ALQ120Q	In the past 12 months, how often did {you/SP} drink any type of alcoholic beverage? PROBE: How many days per week, per month, or per year did {you/SP} drink?
ALQ120U	UNIT OF MEASURE.
ALQ130	In the past 12 months, on those days that {you/SP} drank alcoholic beverages, on the average, how many drinks did {you/he/she} have?
ALQ140Q	In the past 12 months, on how many days did {you/SP} have 5 or more drinks of any alcoholic beverage? PROBE: How many days per week, per month, or per year did {you/SP} have 5 or more drinks in a single day?
ALQ150	Was there ever a time or times in {your/SP's} life when {you/he/she} drank 5 or more drinks of any kind of alcoholic beverage almost every day?
ARQ010	These next questions are about pain in the back, neck or hip area that {you/SP} may have had. Please look at this hand card. HAND CARD ARQ1 {Have you/Has SP} ever had pain, aching or stiffness in any of these locations almost every day for at least 6 weeks in a row? Include pain even if it was mild.
ARQ020A	Pain Diagram Area 1: Neck
ARQ020B	Pain Diagram Area 2: Upper Back
ARQ020C	Pain Diagram Area 3: Mid Back
ARQ020D	Pain Diagram Area 4: Low Back
ARQ020E	Pain Diagram Area 5: Buttocks
ARQ020F	Pain Diagram Area 6: Anterior Hips
ARQ020G	Pain Diagram Area 7: Sternum and Anterior Rib Cage
ARQ021BA	Next we are going to ask you a series of questions about the location{s} you just mentioned. Which specifically did {you/SP} have in {your/his/her} NECK? Was it...
ARQ022AA	Do {you/SP} still have NECK pain, aching or stiffness?
ARQ022AB	Do {you/SP} still have UPPER BACK pain, aching or stiffness?
ARQ022AC	Do {you/SP} still have MID BACK pain, aching or stiffness?
ARQ022AD	Do {you/SP} still have LOW BACK pain, aching or stiffness?
ARQ022AE	Do {you/SP} still have BUTTOCKS pain, aching or stiffness?
ARQ022AF	Do {you/SP} still have HIP pain, aching or stiffness?
ARQ022AG	Do {you/SP} still have RIB CAGE pain, aching or stiffness?
ARQ024A	Was there one time when {you/SP} had pain, aching or stiffness in

	{your/his/her} NECK on almost every day for 3 or more months in a row?
ARQ024B	Was there one time when {you/SP} had pain, aching or stiffness in {your/his/her} UPPER BACK on almost every day for 3 or more months in a row?
ARQ024C	Was there one time when {you/SP} had pain, aching or stiffness in {your/his/her} MID BACK on almost every day for 3 or more months in a row?
ARQ024D	Was there one time when {you/SP} had pain, aching or stiffness in {your/his/her} LOW BACK on almost every day for 3 or more months in a row?
ARQ024E	Was there one time when {you/SP} had pain, aching or stiffness in {your/his/her} BUTTOCKS on almost every day for 3 or more months in a row?
ARQ024F	Was there one time when {you/SP} had pain, aching or stiffness in {your/his/her} HIP on almost every day for 3 or more months in a row?
ARQ024G	Was there one time when {you/SP} had pain, aching or stiffness in {your/his/her} RIB CAGE on almost every day for 3 or more months in a row?
ARQ030A	For {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness, {have you/has s/he} ever taken any of the following medicines? Ibuprofen (eye-byu-proh-fen), Motrin, or Advil
ARQ030B	For {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness, {have you/has s/he} ever taken any of the following medicines? Aleve, Naprosyn (na-proh-sen), Anaprox (an-a-proks), Naproxen (na-prox-sen)
ARQ030C	For {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness, {have you/has s/he} ever taken any of the following medicines? Indocin (in-doh-sen), Indomethacin (in-doh-meth-a-sen)
ARQ030D	For {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness, {have you/has s/he} ever taken any of the following medicines? Celebrex, Vioxx
ARQ030E	For {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness, {have you/has s/he} ever taken any of the following medicines? Aspirin, Bufferin, Ecotrin, or Vanquish (Please do not count Tylenol .)
ARQ034A	How much did this medicine help to relieve {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness? Would you say it relieved...
ARQ034B	How much did this medicine help to relieve {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness? Would you say it relieved...

ARQ034C	How much did this medicine help to relieve {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness? Would you say it relieved...
ARQ034D	How much did this medicine help to relieve {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness? Would you say it relieved...
ARQ034E	How much did this medicine help to relieve {your/SP's} {back/{or} neck/{or} buttocks} pain, aching or stiffness? Would you say it relieved...
ARQ070	If {you/SP} {don't/doesn't/didn't} take any medicine, {do/does/did} {you/he/she} often wake up from sleep because of pain, aching or stiffness?
ARQ080	Does/Did} {your/SP's} pain, aching or stiffness usually get better when {you/he/she} {do/does/did} either walking or stretching for a half hour?
ARQ110	Please look at this hand card. HAND CARD ARQ2. Besides injuries or fractures, {have you/has SP} ever had pain that is just in one of these two areas every day for at least two weeks?
ARQ112A	Was the pain at Location A on the diagram (the plantar aspect of the heel)?
ARQ112B	Was the pain at Location B on the diagram (the posterior Achilles tendon area)?
ARQ125C	Next are some questions about conditions that affect the eyes, the intestines, or bones and joints. Has a doctor or other health professional ever told {you/SP} that {you/s/he} had ulcerative colitis (ulcer-a-tive co-light-us)?
ARD125A	Next are some questions about conditions that affect the eyes, the intestines, or bones and joints. Has a doctor or other health professional ever told {you/SP} that {you/s/he} had iritis (eye-right-us)or uveitis (you-vee-eye-t-us)?
ARQ125D	Next are some questions about conditions that affect the eyes, the intestines, or bones and joints. Has a doctor or other health professional ever told {you/SP} that {you/s/he} had Crohn's (crow-n-z) disease?
ARQ125E	Next are some questions about conditions that affect the eyes, the intestines, or bones and joints. Has a doctor or other health professional ever told {you/SP} that {you/s/he} had ankylosing spondylitis (ank-eh-low-s-ing spawn-d-light-us)?
AUQ131	These next questions are about {your/SP's} hearing. Which statement best describes {your/SP's} hearing (without a hearing aid)? Would you say {your/his/her} hearing is excellent, good, that {you have/s/he has} a little trouble, moderate trouble, a lot of trouble, or {are you/is s/he} deaf?

AUQ191	In the past 12 months, {have you/has SP} been bothered by ringing, roaring, or buzzing in {your/his/her} ears or head that lasts for 5 minutes or more?
AUQ260	{Are you/Is SP} bothered by ringing, roaring, or buzzing in {your/his/her} ears or head only after listening to loud sounds or loud music?
AUQ270	{Are you/Is SP} bothered by ringing, roaring, or buzzing in {your/his/her} ears or head when going to sleep?
BPQ020	{Have you/Has SP} ever been told by a doctor or other health professional that {you/s/he} had hypertension, also called high blood pressure?
BPQ040A	Because of {your/SP's} (high blood pressure/hypertension), {have you/has s/he} ever been told to . . . take prescribed medicine?
BPQ050A	HELP AVAILABLE (Are you/Is SP) now taking prescribed medicine
BPQ057	{Have you/Has SP} ever been told by a doctor or other health professional that {you have/s/he has} high normal blood pressure or borderline hypertension?
BPQ080	{Have you/Has SP} ever been told by a doctor or other health professional that {your/his/her} blood cholesterol level was high?
BPQ100A	(Are you/Is SP) now following this advice to eat fewer high fat or high cholesterol foods?
BPQ100C	(Are you/Is SP) now following this advice to increase (your/his/her) physical activity or exercise?
BPQ100D	(Are you/Is SP) now following this advice to take prescribed medicine?
BHQ010	Next, we'd like to talk to you about bowel health. We'll start with accidental bowel leakage. There are four types of bowel leakage that can happen: leakage (passing) of gas, leakage of mucus, leakage of liquid stool, and leakage of solid stool. We will ask you about leakage of each of these one at a time. How often during the past 30 days have you had any amount of accidental bowel leakage that consisted of gas? Would you say . . .
BHQ020	How often during the past 30 days have you had any amount of accidental bowel leakage that consisted of mucus?
BHQ030	How often during the past 30 days have you had any amount of accidental bowel leakage that consisted of liquid stool?
BHQ040	How often during the past 30 days have you had any amount of accidental bowel leakage that consisted of solid stool?
CDQ001	{Have you/Has SP} ever had any pain or discomfort in

	{your/her/his} chest?
CDQ002	{Do you/Does she/Does he} get it when {you/she/he} walk uphill or hurry?
CDQ003	{Do you/Does she/Does he} get it when {you/she/he} walk at an ordinary pace on level ground?
CDQ008	Have {you/she/he} ever had a severe pain across the front of {your/her/his} chest lasting for half an hour or more?
CDQ010	{Have you/Has SP} had shortness of breath either when hurrying on the level or walking up a slight hill?
CBD160	During the past 7 days, how many times did {you or someone else in your family/you} cook food for dinner or supper at home?
CBQ190	How many of these meals were cooked at home?
CBQ505	{Great. I'll tell you when you will need it.} For the first few questions, please answer yes or no. In the past 12 months, did you buy food from fast food or pizza places?
CBQ550	[For the following questions, please answer yes or no.] In the past 12 months, did you eat at a restaurant with waiter or waitress service?
CBQ660	{For the next set of questions, please use hand card 4.} When you buy food from a grocery store or supermarket, how important is "price"? Would you say very important, somewhat important, not too important, or not at all important?
CBQ665	How about "nutrition"? When you buy food from a grocery store or supermarket, how important is "nutrition"? [Would you say very important, somewhat important, not too important, or not at all important?]
CBQ670	How about "taste"? [When you buy food from a grocery store or supermarket, how important is "taste"?] [Would you say very important, somewhat important, not too important, or not at all important?]
CBQ675	How about "how easy the food is to prepare"? [When you buy food from a grocery store or supermarket, how important is "how easy the food is to prepare"?] [Would you say very important, somewhat important, not too important, or not at all important?]
CBQ680	How about "how well the food keeps after it's bought"? [When you buy food from a grocery store or supermarket, how important is "how well the food keeps after it's bought [in other words, how soon it spoils]"?] [Would you say very important, somewhat important, not too important, or not at all important?]
CBQ700	{Now turn the page to use hand card 5.} Many food packages contain an expiration date such as "use by" or "sell by". How often do you use the expiration date when deciding to buy a product? Would you say always, most of the time, sometimes, rarely, or never?
CBQ685	How about the information on the percent daily value? [HAND

	CARD #5][How often do you use information on the percent daily value on a food label, {such as the part colored in blue on hand card 5,} when deciding to buy a food product?] [Would you say always, most of the time, sometimes, rarely, or never?]
CBQ790	In the past 30 days, when you bought fruits, how often did you buy organic fruits? {Using hand card 13} Would you say always, most of the time, sometimes, rarely, or never?
CBQ795	How about organic vegetables? [In the past 30 days,] when you bought vegetables, how often did you buy organic vegetables? Would you say always, most of the time, sometimes, rarely, or never?
CBQ800	How about organic milk and other dairy products? [In the past 30 days,] [when you bought milk and other dairy products, how often did you buy organic milk and other dairy products? Would you say always, most of the time, sometimes, rarely, or never?]
CBQ805	How about organic eggs? [In the past 30 days,] [when you bought eggs, how often did you buy organic eggs? Would you say always, most of the time, sometimes, rarely, or never?]
CBQ815	How about organic poultry, such as chicken or turkey? [In the past 30 days,] [when you bought poultry, such as chicken or turkey, how often did you buy organic poultry? Would you say always, most of the time, sometimes, rarely, or never?]
CBQ820	How about organic meats? [In the past 30 days,] [when you bought meats, how often did you buy organic meats? Would you say always, most of the time, sometimes, rarely, or never?]
CBQ825	{Now, please look at hand card 14. This is a picture of the USDA Organic seal. Have you ever seen this seal on a food product?}
CBQ785	The interview was completed in:
DBQ780	Some food packages contain health claims about the benefits of nutrients or foods {like the examples on hand card 8}. How often do you use this kind of health claim when deciding to buy a product? Using hand card 9, would you say always, most of the time, sometimes, rarely, or never?
DBQ750	{For the next few questions you'll use hand card 6 to respond, but first please look at hand card 5 which shows an example of the food label. The "Nutrition Facts" panel of a food label is everything on this page except the list of ingredients in pink. How often do you use the Nutrition Facts panel when deciding to buy a food product?} Would you say always, most of the time, sometimes, rarely, or never?
DBQ760	How about the list of ingredients? [HAND CARD #5] How often do you use the list of ingredients on a food label, {such as the part colored in pink on hand card 5,} when deciding to buy a food product? Would you say always, most of the time, sometimes, rarely, or never?

DBQ770	How about the information on the serving size? [HAND CARD #5] [How often do you use information on the serving size on a food label, {such as the part colored in green on hand card 5,} when deciding to buy a food product?] Would you say always, most of the time, sometimes, rarely, or never?]
CBD710	Now think about the types of food products you buy using food labels. How often do you look for nutrition information on the food label when you buy snack items like chips, popcorn, or pretzels? Would you say always, most of the time, sometimes, rarely, or never?
CBD715	How about "breakfast cereals"? [How often do you look for nutrition information on the food label when you buy breakfast cereals?] [Would you say always, most of the time, sometimes, rarely, or never?
CBD720	How about "salad dressings"? [How often do you look for nutrition information on the food label when you buy salad dressings?] [Would you say always, most of the time, sometimes, rarely, or never?]
CBD725	How about "raw meat, poultry, or fish"? [How often do you look for nutrition information on the food label when you buy raw meat, poultry, or fish?] [Would you say always, most of the time, sometimes, rarely, or never?]
CBD730	How about "processed meat products like hot dogs or bologna"? [How often do you look for nutrition information on the food label when you buy processed meat products like hot dogs or bologna?] [Would you say always, most of the time, sometimes, rarely, or never?]
CBD735	How about "bread"? [How often do you look for nutrition information on the food label when you buy bread?] [Would you say always, most of the time, sometimes, rarely, or never?]
HSD010	{First/Next} I have some general questions about {your/SP's} health. Would you say {your/SP's} health in general is . . .
HSQ480	Now thinking about {your/SP's} mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was {your/his/her} mental health not good?
HSQ490	During the past 30 days, for about how many days did poor physical or mental health keep {you/SP} from doing {your/his/her} usual activities, such as self-care, work, school or recreation?
HSQ496	During the past 30 days, for about how many days {have you/has SP} felt worried, tense, or anxious?
DBQ235A	Now, I'm going to ask you how often {you/SP} drank milk at different times in {your/his/her} life. How often did {you/SP} drink any type of milk, including milk added to cereal when {you were/s/he was} a child between the ages of 5 and 12 years old? Would you say...?

DBQ235B	Now, I'm going to ask you how often {you/SP} drank milk at different times in {your/his/her} life. How often did {you/SP} drink any type of milk, including milk added to cereal when {you were/s/he was} a teenager between the ages of 13 and 17 years old? Would you say...?
DBQ235C	Now, I'm going to ask you how often {you/SP} drank milk at different times in {your/his/her} life. How often did {you/SP} drink any type of milk, including milk added to cereal when {you were/s/he was} a young adult between the ages of 18 and 35 years old? Would you say...?
DBQ915	{Do you/Does SP} consider {yourself/himself/herself} to be a vegetarian?
DBQ925a	What foods {are you/is SP} allergic to?
DBQ925b	What foods {are you/is SP} allergic to?
DBQ925c	What foods {are you/is SP} allergic to?
DBQ925d	What foods {are you/is SP} allergic to?
DBQ925e	What foods {are you/is SP} allergic to?
DBQ925f	What foods {are you/is SP} allergic to?
DBQ925g	What foods {are you/is SP} allergic to?
DBQ925h	What foods {are you/is SP} allergic to?
DBQ925i	What foods {are you/is SP} allergic to?
DBQ925j	What foods {are you/is SP} allergic to?
DUQ200	The following questions ask about use of drugs not prescribed by a doctor. Please remember that your answers to these questions are strictly confidential. The first questions are about marijuana and hashish. Marijuana is also called pot or grass. Marijuana is usually smoked, either in cigarettes, called joints, or in a pipe. It is sometimes cooked in food. Hashish is a form of marijuana that is also called 'hash.' It is usually smoked in a pipe. Another form of hashish is hash oil. Have you ever, even once, used marijuana or hashish?
DUQ240	Have you ever used cocaine, crack cocaine, heroin, or methamphetamine?
DUQ250	The following questions are about cocaine, including all the different forms of cocaine such as powder, 'crack', 'free base', and coca paste. Have you ever, even once, used cocaine, in any form?
HIQ011	The (first/next) questions are about health insurance. {Are you/Is SP} covered by health insurance or some other kind of health care plan? [Include health insurance obtained through employment or purchased directly as well as government programs like Medicare and Medicaid that provide medical care or help pay medical bills.]
HCQ100	Which of the following statements describes most closely what (your/SP's) doctor told you about (your/his/her) hepatitis C test result?

HUQ040	What kind of place {do you/does SP} go to most often: is it a clinic, doctor's office, emergency room, or some other place?
HUQ050	{During the past 12 months, how/How} many times {have you/has SP} seen a doctor or other health care professional about {your/his/her} health at a doctor's office, a clinic, hospital emergency room, at home or some other place? Do not include times {you were/s/he was} hospitalized overnight.
HUQ090	During the past 12 months, that is since {DISPLAY CURRENT MONTH} of {DISPLAY LAST YEAR}, {have you/has SP} seen or talked to a mental health professional such as a psychologist, psychiatrist, psychiatric nurse or clinical social worker about {your/his/her} health?
IND235	Monthly family income (reported as a range value in dollars).
KIQ005	Many people have leakage of urine. The next few questions ask about urine leakage. How often {do you/does SP} have urinary leakage? Would {you/s/he} say . . .
KIQ010	How much urine {do you/does SP} lose each time? Would {you/s/he} say . . .
KIQ022	{Have you/Has SP} ever been told by a doctor or other health professional that {you/s/he} had weak or failing kidneys? Do not include kidney stones, bladder infections, or incontinence.
KIQ025	In the past 12 months, {have you/has SP} received dialysis (either hemodialysis or peritoneal dialysis)?
KIQ026	{Have you/Has SP} ever had kidney stones?
KIQ042	During the past 12 months, {have you/has SP} leaked or lost control of even a small amount of urine with an activity like coughing, lifting or exercise?
KIQ046	During the past 12 months, {have you/has SP} leaked or lost control of even a small amount of urine without an activity like coughing, lifting, or exercise, or an urge to urinate?
KIQ480	During the past 30 days, how many times per night did {you/SP} most typically get up to urinate, from the time {you/s/he} went to bed at night until the time {you/he/she} got up in the morning. Would {you/s/he} say
MCQ010	Has a doctor or other health professional ever told {you/SP} that {you have/s/he/SP has} asthma?
MCQ035	{Do you/Does SP} still have asthma?
MCQ051	During the past 3 months, {have you/has SP} taken medication prescribed by a doctor or other health professionals for asthma?
MCQ070	{Have you/Has SP} ever been told by a doctor or other health care professional that {you/s/he} had psoriasis (sore-eye-asis)?
MCQ082	Has a doctor or other health professional ever told {you/SP} that {you have/s/he/SP has} celiac (sele-ak) disease, also called or sprue

	(sproo)?
MCQ086	{Are you/is SP} on a gluten-free diet?
MCQ140	{Do you/Does SP} have trouble seeing, even when wearing glasses or contact lenses, if {you/he/she} wear{s} them?
MCQ160A	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had arthritis?
MCQ160B	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had congestive heart failure?
MCQ160C	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had coronary heart disease?
MCQ160D	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had angina, also called angina pectoris?
MCQ160E	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had a heart attack (also called myocardial infarction)?
MCQ160F	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had a stroke?
MCQ160G	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had emphysema?
MCQ160K	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had chronic bronchitis?
MCQ160L	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had any kind of liver condition?
MCQ160M	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had a thyroid problem?
MCQ160N	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had gout?
MCQ170K	{Do you/Does SP} still . . . have chronic bronchitis?
MCQ170L	{Do you/Does SP} still . . . have any kind of liver condition?
MCQ170M	{Do you/Does SP} still . . . have a thyroid problem?
MCQ220	{Have you/Has SP} ever been told by a doctor or other health professional that {you/s/he} had cancer or a malignancy of any kind?
MCQ300A	Including living and deceased, were any of {SP's/your} close biological that is, blood relatives including father, mother, sisters or brothers, ever told by a health professional that they had a heart attack or angina (an-gi-na) before the age of 50?
MCQ300B	Including living and deceased, were any of {SP's/your} close biological that is, blood relatives including father, mother, sisters or brothers, ever told by a health professional that they had asthma (az-ma)?
OCQ510	The next questions ask about being exposed to dust in (your/SPs) work. Being exposed to dust means that {you/SP} breathed in the dust or had dust on {your/his/her} clothes, skin or hair. In any job, {have you/has SP} ever been exposed to dust from rock, sand,

	concrete, coal, asbestos, silica or soil?
OCQ530	In any job, {have you/has SP} ever been exposed to dust from baking flours, grains, wood, cotton, plants or animals?
OCQ550	The next questions ask about being exposed to fumes in {your/SP's} work. Being exposed to fumes means that {you/SP} breathed in fumes or had a lasting smell on {your/his/her} clothes, skin or hair. In any job, {have you/has SP} ever been exposed to exhaust fumes from trucks, buses, heavy machinery or diesel engines?
OCQ570	In any job, {have you/has SP} ever been exposed to any other gases, vapors or fumes? Examples are vapors from paints, cleaning products, glues, solvents, and acids; or welding/soldering fumes.
OHQ850	{Have you/Has SP} ever had treatment for gum disease such as scaling and rootplaning, sometimes called "deep cleaning"?
OHQ855	{Have you/Has SP} ever had any teeth become loose on their own, without an injury?
OHQ860	{Have you/Has SP} ever been told by a dental professional that {you/s/he} lost bone around [your/his/her} teeth?
OSD110a	How old {were you/was SP} when {you/SP} fractured {your/his/her} (fracture site selected in OSQ100a) for the first time after age 20?
OSD110b	How old {were you/was SP} when {you/SP} fractured {your/his/her} (fracture site selected in OSQ100b) for the first time after age 20?
OSD110c	How old {were you/was SP} when {you/SP} fractured {your/his/her} (fracture site selected in OSQ100c) for the first time after age 20?
OSQ060	Has a doctor ever told {you/SP} that {you/s/he} had osteoporosis, sometimes called thin or brittle bones?
OSQ130	{Have you/has SP} ever taken any prednisone or cortisone pills nearly every day for a month or longer? [Prednisone and cortisone are types of steroids.]
OSQ150	Including living and deceased, were either of {your/SP's} biological parents ever told by a health professional that they had osteoporosis or brittle bones?
PFQ020	{Do you/Does SP} have an impairment or health problem that limits {your/his/her} ability to {walk, run or play} {walk or run}?
PFQ030	Is this an impairment or health problem that has lasted, or is expected to last 12 months or longer?
PFQ049	The next set of questions is about limitations caused by any long-term physical, mental or emotional problem or illness. Please do not include temporary conditions, such as a cold [or pregnancy]. Does a physical, mental or emotional problem now keep {you/SP} from working at a job or business?
PFQ051	{Are you/Is SP} limited in the kind or amount of work {you/s/he} can do because of a physical, mental or emotional problem?
PFQ054	Because of a health problem, {do you/does SP} have difficulty

	walking without using any special equipment?
PFQ057	{Are you/Is SP} limited in any way because of difficulty remembering or because {you/s/he} experience{s} periods of confusion?
PFQ059	{Are you/Is SP} limited in any way in any activity because of a physical, mental or emotional problem?
PFQ061B	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . walking for a quarter of a mile [that is about 2 or 3 blocks]?
PFQ061C	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . walking up 10 steps without resting?
PFQ061E	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . lifting or carrying something as heavy as 10 pounds [like a sack of potatoes or rice]?
PFQ061F	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . doing chores around the house [like vacuuming, sweeping, dusting, or straightening up]?
PFQ061G	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . preparing {your/his/her} own meals?
PFQ061H	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . walking from one room to another on the same level?
PFQ061I	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . standing up from an armless straight chair?
PFQ061J	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . getting in or out of bed?
PFQ061K	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . eating, like holding a fork, cutting food or drinking from a glass?
PFQ061L	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . dressing {yourself/himself/herself}, including tying shoes, working zippers, and doing buttons?
PFQ061M	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . standing or being on {your/his/her} feet for about 2 hours?
PFQ061N	By {yourself/himself/herself} and without using any special

	equipment, how much difficulty {do you/does SP} have . . . sitting for about 2 hours?
PFQ061O	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . reaching up over {your/his/her} head?
PFQ061P	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . using {your/his/her} fingers to grasp or handle small objects?
PFQ061Q	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . going out to things like shopping, movies, or sporting events?
PFQ061R	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . participating in social activities [visiting friends, attending clubs or meetings or going to parties]?
PFQ061S	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . doing things to relax at home or for leisure [reading, watching TV, sewing, listening to music]?
PFQ061T	By {yourself/himself/herself} and without using any special equipment, how much difficulty {do you/does SP} have . . . pushing or pulling large objects like a living room chair?
RHQ131	The next questions are about {your/SP's} pregnancy history. {Have you/Has SP ever been pregnant? Please include (current pregnancy,) live births, miscarriages, stillbirths, tubal pregnancies and abortions.
RHQ172	{Did {your/SP's} delivery/Did any of {your/SP's} deliveries} result in a baby that weighed 9 pounds (4082 g) or more at birth? (Please count stillbirths as well as live births.)
RHQ420	Now I am going to ask you about {your/SP's} birth control history. {Have you/Has SP} ever taken birth control pills for any reason?
RHQ510	{Have you/Has SP} ever used Depo-Provera or injectable to prevent pregnancy?
RHD280	{Have you/Has SP} had a hysterectomy, including a partial hysterectomy, that is, surgery to remove {your/her} uterus or womb?
RHD442	{Are you/Is SP} taking birth control pills now?
RDQ031	{Do you/Does SP} usually cough on most days for 3 consecutive months or more during the year?
RDQ070	In the past 12 months {have you/has SP} had wheezing or whistling in {your/his/her} chest?
RDQ080	[In the past 12 months], how many attacks of wheezing or whistling {have you/has SP} had?
RDQ090	[In the past 12 months], how often, on average, has {your/SP's} sleep been disturbed because of wheezing? Would you say this happens . . .

RDQ100	[In the past 12 months], has {your/SP's} chest sounded wheezy during or after exercise or physical activity?
RDQ134	(In the past 12 months), (have you/has SP) taken medication, prescribed by a doctor, for wheezing or whistling?
RDQ135	During the past 12 months, how much did {you/SP} limit {your/his/her} usual activities due to wheezing or whistling? Would you say...
RDQ137	During the past 12 months, how many days of work or school did {you/SP} miss due to wheezing or whistling?
RDD120	[In the past 12 months], how many times {have you/has SP} gone to the doctor's office or the hospital emergency room for one or more of these attacks of wheezing or whistling?
AGQ030	During the past 12 months, {have you/has SP} had an episode of hay fever?
SXQ260	Has a doctor or other health care professional ever told you that you had genital herpes?
SXQ265	Has a doctor or other health care professional ever told you that you had genital warts?
SXQ270	In the past 12 months, has a doctor or other health care professional told you that you had gonorrhea, sometimes called GC or clap?
SXQ272	In the past 12 months, has a doctor or other health care professional told you that you had chlamydia?
SXQ753	Has a doctor or other health care professional ever told you that you had human papillomavirus or HPV?
SLD010H	The next set of questions is about your sleeping habits. How much sleep {do you/does SP} usually get at night on weekdays or workdays?
SLQ060	{Have you/Has SP} ever been told by a doctor or other health professional that {you have/s/he has} a sleep disorder?
SMQ020	These next questions are about cigarette smoking and other tobacco use. {Have you/Has SP} smoked at least 100 cigarettes in {your/his/her} entire life?
SMQ040	{Do you/Does SP} now smoke cigarettes. .
SMQ050Q	How long has it been since {you/SP} quit smoking cigarettes?
SMD650	During the past 30 days, on the days that {you/SP} smoked, about how many cigarettes did {you/s/he} smoke per day?
SMD410	I would now like to ask you a few questions about smoking. Does anyone who lives here smoke cigarettes, cigars, or pipes anywhere inside this home?
SMQ720	During the past 5 days, on the days {you/he/she} smoked, how many cigarettes did {you/he/she} smoke each day?
WHD140	Up to the present time, what is the most {SP has} ever weighed?