PREDICTING THE DISEASE OF ALZHEIMER (AD) WITH SNP BIOMARKERS AND CLINICAL DATA BASED DECISION SUPPORT SYSTEM USING DATA MINING CLASSIFICATION APPROACHES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ONUR ERDOĞAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF HEALTH INFORMATICS

SEPTEMBER 2012

PREDICTING THE DISEASE OF ALZHEIMER (AD) WITH SNP BIOMARKERS AND CLINICAL DATA BASED DECISION SUPPORT SYSTEM USING DATA MINING CLASSIFICATION APPROACHES

Submitted by **ONUR ERDOĞAN** in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Informatics Institute**

Assist.Prof. Dr. Yeşim Aydın Son _____
Head of Department, **Health Informatics**

Assist.Prof. Dr. Yeşim Aydın Son _____
Advisor, **Health Informatics**

Examining Committee Members:

Prof. Dr. Nazife Baykal _____
Information Systems, Middle East Technical University

Assist.Prof. Dr. Yeşim Aydın Son _____
Health Informatics, Middle East Technical University

Assist.Prof. Dr.Aybar Can Acar _____
Health Informatics, Middle East Technical University

Dr.Levent Çarkacıoğlu _____
Central Bank of the Republic of Turkey

Assist.Prof. Dr.Tülin Yanık _____
Department of Biological Sciences, Middle East Technical University

Date: 07.09.2012

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name and Surname : Onur Erdoğan

Signature :

ABSTRACT

PREDICTING THE DISEASE OF ALZHEIMER'S (AD) WITH SNP BIOMARKERS AND CLINICAL DATA BASED DECISION SUPPORT SYSTEM USING DATA MINING CLASSIFICATION APPROACHES

Erdoğan, Onur
M.Sc., Department of Health Informatics
Supervisor: Assist. Prof. Dr. Yeşim Aydın Son

September 2012, 152 pages

Single Nucleotide Polymorphisms (SNPs) are the most common DNA sequence variations where only a single nucleotide (A, T, C, G) in the human genome differs between individuals. Besides being the main genetic reason behind individual phenotypic differences, SNP variations have the potential to exploit the molecular basis of many complex diseases. Association of SNPs subset with diseases and analysis of the genotyping data with clinical findings will provide practical and affordable methodologies for the prediction of diseases in clinical settings. So, there is a need to determine the SNP subsets and patients' clinical data which is informative for the prediction or the diagnosis of the particular diseases. So far, there is no established approach for selecting the representative SNP subset and patients'

clinical data, and data mining methodology that is based on finding hidden and key patterns over huge databases. This approach have the highest potential for extracting the knowledge from genomic datasets and to select the number of SNPs and most effective clinical features for diseases that are informative and relevant for clinical diagnosis. In this study we have applied one of the widely used data mining classification methodology: “decision tree” for associating the SNP Biomarkers and clinical data with the Alzheimer’s disease (AD), which is the most common form of “dementia”. Different tree construction parameters have been compared for the optimization, and the most efficient and accurate tree for predicting the AD is presented.

Keywords: Data Mining, Single Nucleotide Polymorphism, Integrating Genotype and Phenotype Data, Decision Tree, Alzheimer’s Disease

ÖZ

ALZHEIMER (AD) HASTALIĞININ VERİ MADENCİLİĞİ SINIFLANDIRMA YAKLAŞIMLARI KULLANARAK SNP BİYOLOJİK GÖSTERGELERİ VE KLİNİK VERİLERLE KARAR DESTEK SİSTEMLERİNE DAYALI TAHMİN EDİLMESİ

Erdoğan, Onur
Yüksek Lisans, Sağlık Bilişimi Bölümü
Tez Yöneticisi: Yard. Doç. Dr. Yeşim Aydın Son

Eylül 2012, 152 sayfa

Tek Nükleotit Polimorfizmi (SNP), insan genomundaki tek nükleotitin (A, T, C, G) bireyler arasında değişiklik gösterdiği en yaygın DNA dizisi çeşitliliğidir. SNPler, bireysel fenotipik farklılıkların arkasındaki temel genetik neden olmak dışında birçok kompleks hastalıklarında altında yatan sebep olabilir. Tek nükleotit değişimlerinin hastalıkla ilişkilendirilmesi ve klinik bulgularla birlikte bireylerin genotip verilerinin analizi, klinik açıdan hastalığın tahmin edilmesi için ekonomik ve pratik bir metodoloji sağlayacaktır. Bu yüzden, belirli bir hastalığın tespiti veya tahmin edilebilmesi için bilgi verici bir SNP kümesinin ve klinik verilerin belirlenmesi gerekir. Şimdiye kadar, klinik verilerle temsilci bir SNP kümesinin seçilmesi ve çok

büyük veri tabanlarından gizli ve anahtar örüntülerin bulunması temeline dayanan veri madenciliği metodolojisi için yerleşik bir yaklaşım bulunmamaktadır. Bu yaklaşım genom veri setlerinde bilgi keşfi için ve ayrıca klinik teşhislerde hastalıkla alakalı bilgi verici SNP sayısını ve klinik özellikleri seçmek için en yüksek potansiyele sahiptir. Bu çalışmada, bunamanın en yaygın hali olan Alzheimer (AD) hastalığı ile SNP biyolojik göstergeleri ve klinik verileri ilişkilendirmek için, yaygınca kullanılan veri madenciliği sınıflandırma yöntemlerinden "karar ağacı" metodolojisi uygulanmıştır. Farklı karar ağacı oluşturma parametreleri, ağacı en optimal duruma getirmek üzere karşılaştırılmıştır ve Alzheimer (AD) hastalığını doğru tahmin eden karar ağacı sunulmuştur.

Anahtar Kelimeler: Veri Madenciliği, Tek Nükleotid Polimorfizmi, Genotip ve Fenotip Verilerin Birleştirilmesi, Karar Ağacı, Alzheimer Hastalığı

To My Family

ACKNOWLEDGMENTS

I appreciate Prof. Dr. Nazife BAYKAL who has given me great ideas and made me meet with my advisor Assist. Prof. Dr. Yeşim AYDIN SON when I was about to lose my hope. She has offered me interesting notions on a new point of view. Endless thanks to my advisor Assist. Prof. Yeşim AYDIN SON who trusted and supported me from the beginning to the end. Without her guidance I would be lost in this study because of my basis about subject.

I want to add my gratitude to my examining committee members Assist. Prof. Dr. Aybar C. ACAR, Dr. Levent ÇARKACIOĞLU and Assist. Prof. Dr. Tülin YANIK for their valuable insight and feedback on this work.

I am also grateful to my managers Kadriye ÖZBAŞ ÇAĞLAYAN and Ahmet DİKİCİ at TÜBİTAK, who supported me for getting this degree. Their tolerance and positive approach made me encourage. I am also deeply grateful to colleagues who always motivate and cheer me up.

Thanks to my dear friends Hülya YÖRÜKOĞLU, Pınar ALTUNAY, Remzi ÇELEBİ, İnci BARUT and Nazik ÇELİK for sharing their ideas and time for me. They motivated and supported me during my thesis work.

Endless thanks to my parents, my sister and my grandmother. Whenever I needed them they were with me. Especially endless thanks to my dear mother who wakes up with me even in the early of every morning and brings me up to these days. Her being is beyond words.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A, T, C, G	Adenine, Thymine, Cytosine, Guanine
AD	Alzheimer's Disease
AHP	The Analytic Hierarchy Process
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
APOE	Apolipoprotein E
APP	Amyloid Precursor Protein
BMI	Body Mass Index
CART	Classification and Regression Tree
CHAID	Chi-squared Automatic Interaction Detector
CHGR	Center for Human Genetic Research
CHOL	Cholesterol
DNA	Deoxyribonucleic Acid
E	Entropy
GWAS	Genome Wide Association Study
HDL	High-density Lipoprotein
ID3	Iterative Dichotomiser 3
IG	Information Gain
LDL	Low-density Lipoprotein
mRNA	Messenger Ribonucleic Acid
PS1	Presenilin1
PS2	Presenilin2
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
WBC	White Blood Cell

PREFACE

SNPs are DNA sequence variations that are distributed throughout the whole genome. Many SNPs are associated with susceptibility to complex diseases such as diabetes, heart diseases, joint illnesses, schizophrenia, or Alzheimer's disease (AD). Single altered genes are the molecular basis of only a small portion of diseases. Most chronic diseases are multifactorial, and might be explained by combined effects of SNPs on different genomic locations. Hence, identification of both statistically and biologically important SNPs associated with different conditions can provide decision making opportunity based on genotypic feature of an individual and aids the prevention, prediction and diagnosis of the condition. The main purpose of this study is to construct a decision tree based on the most informative and representative selected SNPs and clinical data records of individuals with Alzheimer's disease for supporting the clinical diagnosis.

CHAPTER 1

INTRODUCTION

Data mining is an analytic process designed to explore patterns and relationships between variables, to extract hidden and unknown knowledge over large amount of data. The findings may then be validated by applying the discovered patterns or relations to the subsets of related data. In the last decades, developments of data mining methods have become a promising approach in bioinformatics in order to solve biological problems [1]. Advancements in technology and acceleration in the number of research in the field of genomics have resulted in accumulation of great amount of data. In order to analyze and obtain results from these data, artificial intelligent and computational analysis is an essential.

Analyzing the biological data sets requires understanding the data by deducing structure of data fields. Statistical modeling for the prediction of a particular disease based on microarray data or case associated SNPs set selection can be given as examples of analyzing the genomic data. Such applications present the great potential and the necessity of the interplay between data mining and bioinformatics [2].

Data mining methodology is applied to data sets for two purposes. The main aim is the classification of the data, and clustering of the data based on the similarities or differences. Before using any of the data mining modeling algorithms, preprocessing is needed for the preparation of the data for further analysis. Preprocessing involves data cleaning, data integration, data transformation and data subset selection which is also called dimension reduction. In order to handle preprocessing steps and then

extracting novel, interesting and useful information by using preprocessed data, advanced computational and statistical methods are used. The final step is the visualization and representation of findings. It has been recently shown that data mining tools are extremely useful for the analysis of high dimensional data such as whole human genome data, which comprises around 3.4 billion base pairs [1].

In human DNA, where 99,9% of base pairs are the same, a small percentage less than 0,1% varies between individuals. So, once every 100 to 300 nucleotides may differ from one to another in human genome. These signatures are defined as "Single Nucleotide Polymorphisms (SNPs)". In other words, SNPs are single nucleotide alterations in genomic DNA and cause personal differences in phenotypes such as psychological and physical characteristics. SNPs can change the structure of a protein, its regulation or expression, which alters normal biological processes. SNPs can be used as genomic markers revealing individuals susceptibility to certain disease to produce new approaches for treatment applications and to take prohibitive precaution. SNP association studies are widely done to determine possible relations between genetic variations and diseases. Such studies aim to reveal individual SNPs that have interaction with particular diseases.

Outputs of mining methods in genetic studies have revealed interesting findings inheritable tendency to contract specific diseases [3]. One of these heritable diseases under analysis due to genetic markers or clinical data is Alzheimer's disease (AD), which is a complex and genetic disorder. It has been discussed that AD is appeared at early ages if it is mainly based on genetic factors, on the other hand it has not known yet whether disease is occurred due to genetic factors or not at elderly people over 60 years of age [4]. Diagnosis of AD in this late-onset group especially presents a challenge as early clinical findings are often undistinguishable from dementia.

In this study in order to produce rules for the diagnosis of late-onset AD patients based on genotype and clinical information, we have applied decision tree algorithms after the prioritization of the genome-wide association study results, expecting to increase the accuracy rate of classification.

CHAPTER 2

BIOLOGICAL and COMPUTATIONAL BACKGROUND INFORMATION

BIOLOGICAL BACKGROUND

2.1 Human Genome: Individual Identity

Human genome can be denominated as one's significant and individual identities. It is the complete set of DNA (deoxyribo nucleic acid) of a living being. It carries the information needed for biological functionalities. It is made of chemical molecules which is called nucleotides, structured in pairs. A nucleotide is composed of a nucleobase {purine bases: A, G or pyrimidine bases: C, T}, a five-carbon sugar and phosphate groups. Figure 2.1 shows the structure of nucleobases.

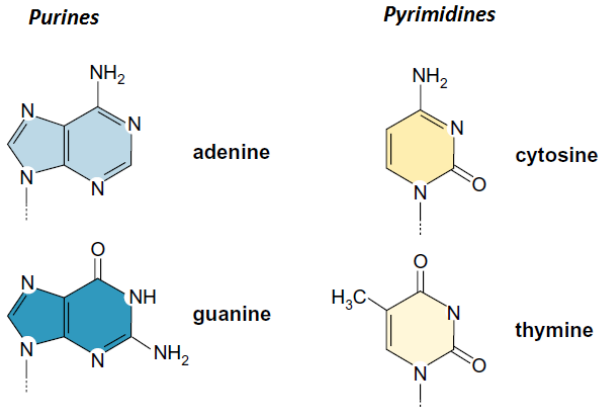

Figure 2.1 Structures[5] of DNA Bases: A, C, G, T.

DNA is comprised of two long polymers of lined up nucleotides with backbone made of sugar and phosphate groups. Nucleotides are told as two types of nucleo-bases: purines (Adenine (A), Guanine (G)) and pyrimidines (Cytosine(C), Thymine (T)). RNA (ribonucleic acid) uses uracil instead of thymine.

Since there are hydrogen bonds between each base, each base is linked to a companion base on the other chain. This pairing is specific; adenine pairs with thymine, and guanine with cytosine, which gives the “the double helix” structure to the DNA. Figure 2.2 visualizes the formation of double helix structure of DNA.

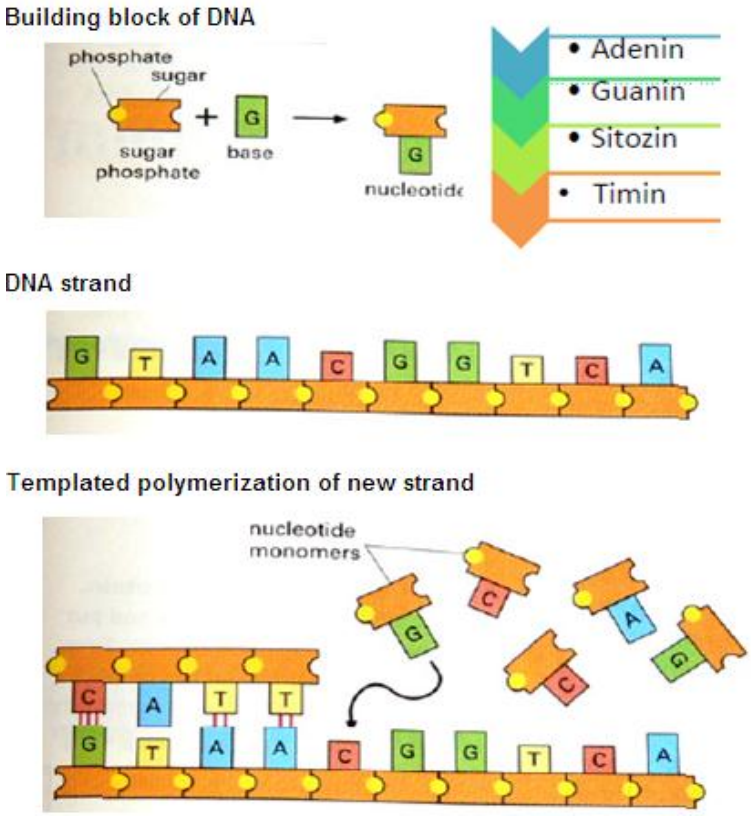

Figure 2.2 Construction of Double Helix Structure [6] of DNA

Approximately 3.2 billon base pairs in human genome scattered into 23 pairs of chromosomes. The order of the nucleotides in the DNA sequence is important as the biological information is carried in this manner. The information carried by DNA is kept in the sequence of blocks which are called genes. These genes encode specific proteins for maintaining the human life and conducting cellular activities. Protein

coding region in a human genome is only around 2% of a human genome, and rest are called non-coding region and has no function assigned so far. Recent studies collected under the ENCODE Project have assigned biological, structural or regulatory function to the 80% of the whole genome, and other studies are still in process to better understand the functional elements in the genomic sequences and the interactions between these elements regulating the genomic functions [7], [8].

2.2 Transcription and Translation

The transcription of DNA into RNA is first described by US biologist Phillip Sharp and British biologist Richard Roberts in 1977. Both discovered that before the translation to protein, a mid-product RNA is formed, through which cells make a copy of both exonic and intronic sequences of genes, and then the non-coding intronic sections that are not translated into protein are removed. Only the exons make up the mRNA [9]. By splicing together, different combinations of the exons builds altenative transcripts of genes as shown in Figure 2.3. Exons make their DNA sequence in a group of words to make protein. By spicing out the introns (nonsense words) and combining different groups of exons (words) they may end up with alternative protein products.

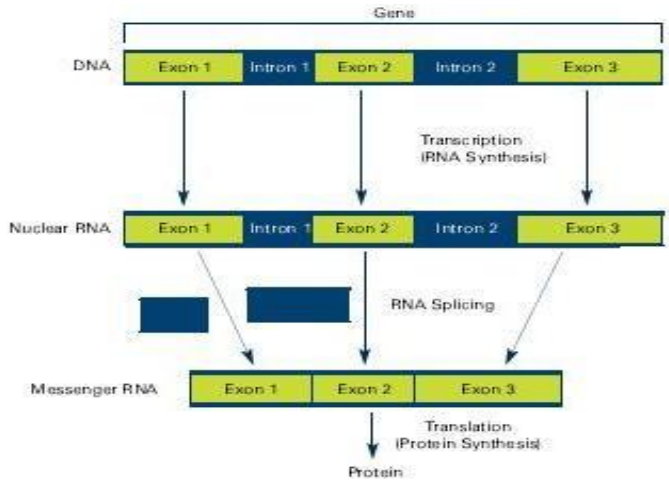

Figure 2.3 RNA Synthesis and Processing for Growing the Protein Chain[1]

[1] http://www.purpleopurple.com/inventions-and-inventors/gene-exons-and-introns.html

In this aspect, gene expression starts with transcription. Transcription is the progress of replication of nucleotide sequence of DNA by RNA polymerase enzyme into RNA sequence. The codons of a gene are copied into messenger RNA. This is the information transfer from DNA to RNA. Transcription initializes the transformation from genetic information to protein sequences. Then the mRNA is used as a template to synthetize proteins during the translation stage of the gene expression process. Based on the genetic code carried by messenger RNA, polypeptides and chain of amino acids are produced in ribosomes. Infrastructure for the protein synthesis is shown in Figure 2.4.

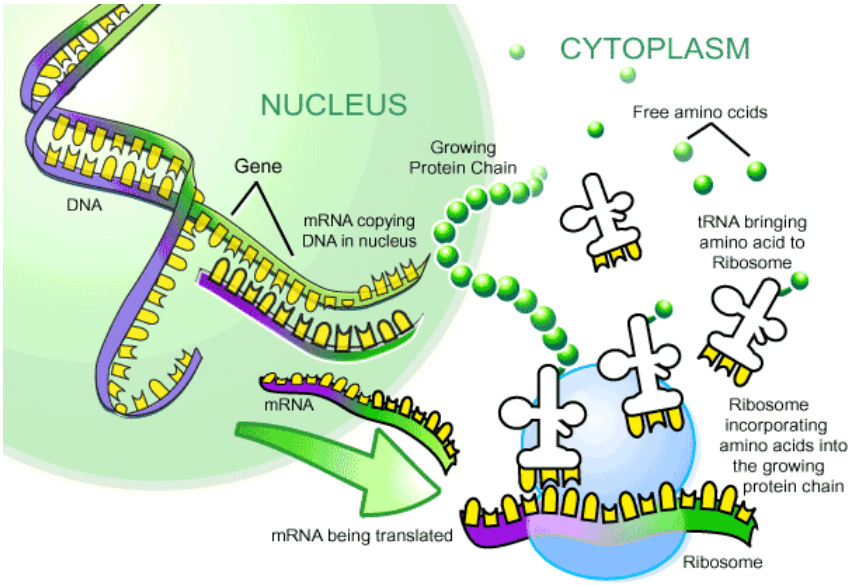

Figure 2.4 Transcription and Translation Diagram: Growing the Protein Chain[2]

Proteins are the informational macromolecules of a cell. Every protein is build up from multiple polypeptides chains formed during the translation process where each three nucleotides (e.g. TTT, CAG), also known as a “codon” is translated into one amino acid based on the genetic code.

Since each codon consists of 3 letters (e.g. ACC), there are 64 possible codon combinations. These combinations encode 20 standard amino acids in human cells. There are 64 possible codons but there are only 20 amino acids.

[2] http://nnhsbiology.pbworks.com/f/1280666569/transcription%20translation%20diagram.png

The AUG codes Methionine which is one of the amino acid. AUG, also known as the start codon, initiates the progress of protein building. There are also three stop codons that end progress; these are UAG, UGA, and UAA codons (Table 2.1). They act as a signal to terminate the transcription of DNA into RNA.

Table 2.1 mRNA Codon / Amino Acid Chart [10]

mRNA Codon/Amino Acid Chart

First Base	Second Base				Third Base
	U	C	A	G	
U	UUU, UUC – Phenylalanine (Phe) UUA, UUG – Leucine (Leu)	UCU, UCC, UCA, UCG – Serine (Ser)	UAU, UAC – Tyrosine (Tyr) UAA, UAG – Stop	UGU, UGC – Cysteine (Cys) UGA – Stop UGG – Tryptophan (Trp)	U C A G
C	CUU, CUC, CUA, CUG – Leucine (Leu)	CCU, CCC, CCA, CCG – Proline (Pro)	CAU, CAC – Histidine (His) CAA, CAG – Glutamine (Glu)	CGU, CGC, CGA, CGG – Arginine (Arg)	U C A G
A	AUU, AUC, AUA – Isoleucine (Ile) AUG – Start Methionine (Met)	ACU, ACC, ACA, ACG – Threonine (Thr)	AAU, AAC – Asparagine (Asn) AAA, AAG – Lysine (Lys)	AGU, AGC – Serine (Ser) AGA, AGG – Arginine (Arg)	U C A G
G	GUU, GUC, GUA, GUG – Valine (Val)	GCU, GCC, GCA, GCG – Alanine (Ala)	GAU, GAC – Aspartic Acid (Asp) GAA, GAG – Glutamic Acid (Glu)	GGU, GGC, GGA, GGG – Glycine (Gly)	U C A G

Based on the DNA sequence changes, biological functions may change. When the order of sequence alters, consequently different kind of amino acids come together and variation occurs. As a result, serious malfunctioning of a protein may occur. So, mutations on DNA sequence or single nucleotide polymorphisms in coding regions (critical location) has the potential to make significant changes in the shape or functionality of produced proteins.

2.3 Genetic Variation

Human genome approximately consists of 3,2 billion nucleotides where 99,9% of genome is similar for every individual [11]. Only 0,1% of the genome sequence is responsible for the differences between people that can be observed at every 300 to 1000 bases [12]. These variations in genome may be the results of repetitive elements, mutations or single nucleotide polymorphisms, which are covered in the following sections.

Variations in the sequence of a gene affect the protein production and directly the trait for the biological process. Physical traits are characteristics and physical makeup of someone such as hair color, eye color, skin color, height, weight as well as the common chronic conditions such as heart disease, diabetes, and cancer. Behavioral traits are characteristics of how ones personality and psychological status [13]. The genotype of an individual, which is defined by the variations it carries, can identify these phenotypic characteristics (physical or behavioral) of that individual, such as appearance or susceptibility to diseases.

Genotyping analysis, in order to obtain genetic variation data of individuals, can be studied by microarray or advanced sequencing technologies. These new emerging technologies with lower cost and high throughput results allows clinical researcher to design and analyze large case-control data sets where genetic variations thousands of individuals can be studied. These studies aims to define genetic basis of individual's risk of developing multifactorial complex hereditary disorders, such as cancer, heart disease, diabetes as well as determining some genetic disorders such as Alzheimer's Disease, Rheumatoid Arthritis [12].

2.4 Mutations

A mutation is any hereditary change in the sequence of DNA. This alteration in the order of nucleotides may affect the phenotype or traits of the individual. This may occur inherently or by outside influences with a DNA damaging agent such as X-

rays, ultraviolet light or toxic chemicals [14].

Mutations can occur at the level of chromosome due to insertion or deletion or they can be detected as as point mutations.

2.4.1 Point Mutations

There are many types of point mutations such as transition, transvertion, silent, neutral, missense and nonsense. The results of mutation of base alteration in the region where a gene encodes a protein distinguish based on the place of gene or the new base that makes changes [15].

If the new base which causes the mutation does not bring a new amino acid to the protein sequence, it is called "silent mutation". For example, think two codons contain letters of GCA and GCG. These two codons encode "arginine" in messenger RNA. So, alteration in the third base, G instead of A, does not affect the protein synthesis. But in some cases it may still have a phenotypic effect by speeding up or slowing down protein synthesis, or by affecting splicing. Silent mutation can be shown as follows.

Silent Mutation:

	5	T	C	T	C	A	A	A	A	A	T	T	T	A	C	G	3
Individual 1	3	A	G	A	G	T	T	T	T	T	A	A	A	T	G	C	5
		Ser			Gln			**Lys**			Phe			Thr		
	5	T	C	T	C	A	A	A	A	G	T	T	T	A	C	G	3
Individual 2	3	A	G	A	G	T	T	T	T	C	A	A	A	T	G	C	5
		Ser			Gln			**Lys**			Phe			Thr		

In some cases, single base alteration in the sequence may finalize with the production of new amino acid to the protein sequence. If the new amino acid has the similar chemical features with the previous one, this is identified as "neutral mutation".

Neutral Mutation:

	5	T	C	T	C	A	A	A	A	A	T	T	T	A	G	G	3
Individual 1	3	A	G	A	G	T	T	T	T	T	A	A	A	T	C	C	5
		Ser			Gln			**Lys**			Phe			Thr			
	5	T	C	T	C	A	A	A	G	A	T	T	T	A	G	G	3
Individual 2	3	A	G	A	G	T	T	T	C	T	A	A	A	T	C	C	5
		Ser			Gln			**Arg**			Phe			Thr			

When a base alteration occurs and a different type of amino acid is linked up to the protein sequence which must not be linked up in fact, this brings out different protein with different functionality depending on whether the change is “conservative” or “nonconservative”.

This type of mutation is called “missense mutation”. For example, CTC code in DNA, which refers to GAG in RNA, expresses the “glutamate” remnant in the structure of protein. If a change occurs in DNA such as CAC which corresponds to GUG in RNA, this expresses the “valine” remnant in betaglobulin protein. Finally, this type of mutation causes sickle cell anemia.

Base alteration in the region where a gene encodes a protein sometimes makes amino acid codon turn into a STOP codon, resulting in premature termination of translation.

In this case, with the stop codon which comes too early, a short sequence of amino acids forms the protein. This type of mutation is called “nonsense mutation “as shown in figure below. The effects of nonsense mutations change according to how much a protein is curtailed and how much protein is needed for the functionality.

Nonsense Mutation:

	5	T	C	T	C	A	A	A	A	A	T	T	T	A	C	G	3
Individual 1	3	A	G	A	G	T	T	T	T	T	A	A	A	T	G	C	5
		Ser			Gln			**Lys**			Phe			Thr			
	5	T	C	T	C	A	A	T	A	A	T	T	T	A	C	G	3
Individual 2	3	A	G	A	G	T	T	A	T	T	A	A	A	T	G	C	5
		Ser			Gln			**Stop**									

2.4.2 Frameshift Mutations (Insertion or Deletion Mutations)

This type of mutation occurs when a new base is inserted to or removed from a gene that encodes a protein. This alteration influences the reading of triplet messenger RNA during protein synthesis.

2.4.2.1 Insertion

A new nucleotide is inserted in the gene and the order of the sequence may change. This might introduce premature STOP codons or amino acid changes, as in the an example sequence below, visualizing the effect of inserting a new base [15].

Insertion:

	3	C	T	G	C	T	G	C	T	G	C	T	G			3
Individual 1	5	G	A	C	G	A	C	G	A	C	G	A	C			5
			Leu			Leu			Leu			Leu				
	3	C	T	G	C	A	T	G	C	T	G	C	T	G		3
Individual 2	5	G	A	C	G	T	A	C	G	A	C	G	A	C		5
			Leu			**His**			Ala			Ala				

This mutation downstreams the sequence of amino acid and produces very different type of protein which is also nonfunctional. On the other hand, with the premature stop codon reading, a curtailed protein may occur.

2.4.2.2 Deletion

The order of the sequence can change when a single nucleotide is deleted from the genomic sequence of a gene. In this case, a new combination mRNA is translated and the protein being produced may be useless or have premature STOP.

Deletion:

The figure above shows an example of a frameshift deletion mutation: In the second codon the deletion of 'c' causes a shift in reading frame and multiple amino acid substitutions in the subsequent protein.

2.5 Single Nucleotide Polymorphism (SNP)

Genomic variations are called polymorphism if observed in a population at a rate more than 1%. The most common type of polymorphisms is single nucleotide polymorphism (SNP, as pronounced “snip”), which can be described as the substitution of a single nucleotide with another one at a homologous site in a population [16]. SNPs are abundant and highly distributed within the individual’s genome [11]. The nucleotide position where a SNP positions is called an allele. The allele whose occurrence in the population is less frequent, so which is not dominant is called the minor allele. The proportion of minor allele to whole is called minor allele frequency [17]. SNP with the minor allele frequency greater than 1% can be identified at every 300 to 1000 base pair in human genome. As a result, total amount of SNP in genome can be estimated as 30 million SNPs [18].

Traditional analysis methods for determining disease-related genes and loci are not implementable for these multifactorial complex diseases. These common diseases may be caused by multiple genes and multiple nongenetic factors (environmental factors) at the same time [11]. In this aspect defining SNPs and mapping them is extremely important in terms of associating genotypes with presence of complex diseases and tendency to a diseases such as high blood pressure, diabetes or heart disease. Large amount of SNPs identified in human genome provides an opportunity to link genetic variations to phenotypic variations by association studies. In a such study, minor allele frequencies are calculated for both case and control groups, so results are compared [16]. By the help of this technique, genetic markers differs significantly among groups can be identified. So far, many SNPs are associated with both individual phenotypes, susceptibility to particular complex diseases and individual's response to certain medicine as a result of genome wide association studies [11]. However, although some studies have revealed genetic associations between one or more SNPs and a complex disease, some of them have been found hard to replicate [19].

2.6 Alzheimer's Disease (AD)

Alzheimer's disease (AD) is a slow progressing complex mental disorder and it is one of the heritable fatal diseases. AD causes loss of intellectual abilities such as memory and the mental break down, especially in elderly. It is firstly recognized by a German physician Dr. Alois Alzheimer. Dr. Alzheimer identified a mass of brain cell abnormalities as a disease during the autopsy soon after one of his patient died. Dense twisted bands of fibers (tangles) were observed surrounding nerve cells inside the brain [20].

As AD progresses the two abnormal protein fragments called plaques and tangles accumulates, killing the brain cells which directly affects the daily life of humans. AD starts at the hippocampus region where memories are first formed. Over many years, the plaques and tangles destroy the hippocampus slowly and making forming new memories harder. As disease progresses the plaques and tangles accumulates at

the different regions of brain compromising other functions. The level of AD as visualized in Figure 2.5 depends on how widely tangles and plaques are spread. From the hippocampus, tangles and plaques spread to the region where language functions are managed. This causes in humans to find right words while talking. When the frontal lope of the brain is affected where logical thinking is controlled, the ability to solve problems is compromised. Next, patients may lose their emotions, causing to lose control and feelings at the same time. Later, AD patients' lose their senses, hearing, sight and smelling function gets weak as the plaques spread to the regions controlling these areas. Finally, tangles and plaques move to the back side of brain, which is the hardest level of AD for patients and caregivers. Consequently, AD compromises the person's balance and coordination and in the very last stage, it destroys the part of brain that regulate breathe and heart functions.

AD affects 10% of the people over 65 years old and approximately 50% of those over the age 85 [21]. AD seen at early ages mainly has genetic basis. However it is still debated how much inherited genetic factors play roles for the elderly people (age>60) with AD [4], as distinguishing AD from dementia becomes very challenging as both shows similar symptoms like memory loss and there is no laboratory or imaging analysis that can diagnose AD at its early stages. Only after further mental impairments develop in later stages of the AD patients can be identified.

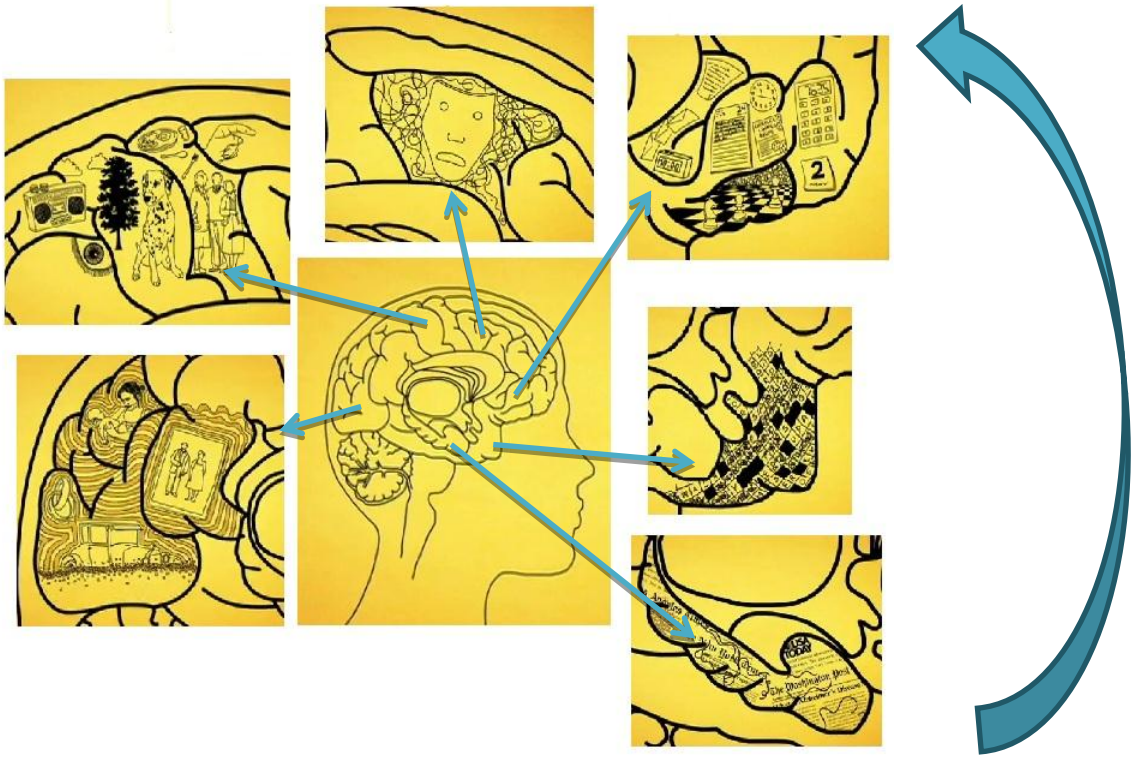

Figure 2.5 Levels of Alzheimer's Disease

Although there are many ongoing studies with promising results, which provides important information about the disease, what fully causes onset of the AD, the precursors and underlying etiology of the disease is still not known. Few genes and loci are identified that play a significant role in revelation and the development of AD [22]. Currently there is neither a definite method for the diagnosis of the disease nor treatment or cure for AD once it is developed. Association of SNP genotypes with AD can enlighten us about the molecular and genetic etiology of the disease and might offer a genomic based diagnosis technique for the differential diagnosis.

2.7 The Genetics of Alzheimer's Disease

So far, there are 4 locuses found which are related to the etiology of AD. These are amyloid precursor protein (APP) in chromosome 21., presenilin1 (PS1) gene in chromosome 14., presenilin2 (PS2) gene in chromosome 1. and APOE locus in chromosome 19. [4].

Also, scientists have studied with single-nucleotide polymorphisms (SNPs) to identify other genomic regions of DNA where changes have existed. Most SNPs don't actually have direct influences on Alzheimer's disease. But some may cause significant situation in terms of AD. In this case, person's tendency of developing Alzheimer's disease depends on the nucleotide order variation within coding genes.

Some studies have shown the associations of SNPs with Alzheimer Disease. The lists of SNPs related to AD and their chromosomal distribution related to AD are given in the Appendix A.

2.8 Genome Wide Association Studies (GWAS)

SNPs are common genetic variations and currently they offer a high potential to associate genomic information of individuals with their risks and susceptibilities to multifactorial chronic diseases[17]. These genetic variations may change the protein functionality or regulation of genes; as a result complex diseases develop. Determining the polymorphisms provides both prediction and diagnosing for the complex diseases and can reveal how individual patients will react to different drug therapies by comparing the case and control groups [17].

In literature, there are two types of approaches exist for identifying the disease responsible genetic variants.

1. Candidate gene based approach
2. Non-candidate gene based approach (GWAS)

In genetic epidemiology, analyzing the DNA sequences is an investigation of many common genetic variations in different individuals for seeing whether any of these nucleotide variations is associated with a disease. This examination is called genome wide association study (GWAS). GWAS finds correlation between SNPs and a disease by comparing the DNA of two groups of individuals: people with disease (cases) and similar people (region, nationality etc.) without disease (controls) for the entire genome as visualized in Figure 2.6. Millions of SNP variations can be

analyzed in one study with today's high throughput genotyping array or whole genome sequencing technologies and regions which are altered more frequently in case group in contrast to the controls can be associated with the condition through GWAS.

In a GWAS, studies are implemented in 5 steps.

1. Large number of case/control groups' genotype data is read from SNPs chips.
2. Data is controlled for quality of analyze by terms of observed missing values.
3. Statistical methods are applied to data comparing cases and controls.
4. Associations based on statistical tests are identified.
5. Mapping and biological interpretation is required.

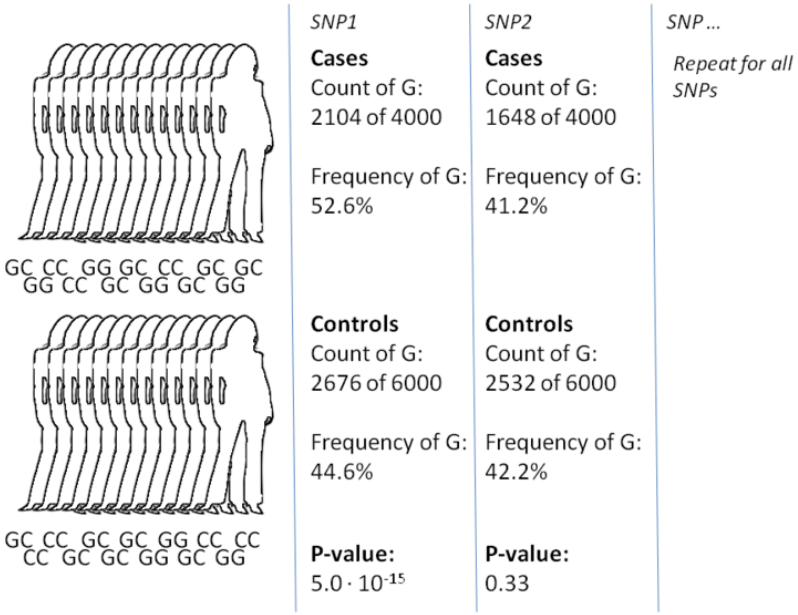

Figure 2.6 Implementation of Genome Wide Association Studies

2.9 SNP Prioritization

In genetic studies such as determining the SNPs causing to a particular disease, the main goal is the identification of significant nucleotide variants within the hundreds of thousands SNPs across control and case groups. GWAS, in this point, tests SNPs statistically but there can be enormous number of correlated SNPs available, where

some of these correlations are false associations. This means that not all the correlations detected at the common threshold of significant level (e.g. p=0,001) are biologically significant. Therefore, focusing on the statistical significance alone is not a valid approach as at the end of correlation analyses tens of thousands of SNP markers can be identified as significant.

In order to pick the right subset of SNPs for validation or apply further analysis on the association data and to develop diagnostics based on the SNP genotyping, the number of associated SNPs should be reduced to a manageable number. Hence, soon after GWAS a SNP prioritization step is included for selecting a subset based on both statistical and biological importance of SNPs responsible for the disease [17].

Prioritization of SNPs is ranking of the SNPs which have the highest potential to effect functions of genes biologically. In addition to statistical correlation findings, use of biological information such as functional effects of SNPs is used to prioritize SNPs and to form a base for selecting a subset of SNPs. There are many software tools than can prioritize SNPs after GWAS using and combining statistical information with biological data. Best known systems are SPOT and SNPLogic. The Analytic Hierarchy Process (AHP) Scoring system is based on statistical significance of associations (p-value) and biological importance of SNPs [17] for prioritization. This approach is recently proposed by our group [17] and implemented in the METU-SNP software. In this study, we have used AHP Scoring for the prioritization of SNPs of the GENADA data.

METU-SNP allows researchers to calculate AHP score, which is based on the structured prioritization of statistically significant SNPs after GWAS, following the hierarchy tree to reveal "functionally and biologically important SNPs associated with the condition". Detailed information about tree structure involving integrated pathways, functional effects, disease annotation data and statistical information can be reached in [17]. AHP scored prioritization scheme also provides SNP to gene, gene to disease and gene to biological pathway integration as it integrates the needed information from primary public databases [17].

COMPUTATIONAL BACKGROUND INFORMATION

2.10 Data Mining

In the last two decades, biomedical researches and biotechnology have been started to study widely because of the explosive increase of biological data [23]. On the other hand, fast and efficient progress in using data mining methodology has become a new approach to analyze high dimensional of genetic or biologic data for mining the novel and interesting patterns in large datasets by applying advanced classification or clustering methods [23].

Human genetics has been studied for many years by using biochemistry, biostatistics, epidemiology, molecular biology, physiology and other disciplines to identify the relationship between DNA sequence information and measures of human health. In comparison to the past, it is now possible to find DNA sequence variations and analyze them with the help of emerging technology and techniques [24].

Today, the focus of statistical analysis of high throughput data have shifted towards computer sciences that provides intelligent solutions or machine learning for mining patterns of genetic variations that are associated with susceptibility to common human diseases. Significant predictors of a disease sometimes identified by the combination of SNPs or environmental factors that causes changes in the order of nucleotides. Beyond these, clinical factors are also able to be considered as significant predictors in terms of genetic diseases. Under these conditions, the learning algorithm is searching a genetic needle in a genomic and clinical haystack [24].

Models implemented in data mining can be studied under two main headings, predictive models and descriptive models [25]. Predictive models are used to predict the class of a data whose class label is unknown using a model formed by the historical data that have already class labels. On the other hand, clustering models define patterns that can be guided in decision making by exploring similarities and dissimilarities in dataset.

A classification is data analysis method which can predict the categorical class labels or future data tendency. There are kinds of techniques which are used in classification. These can be counted as Decision Trees, Artificial Neural Networks, Genetic Algorithms and Naive Bayesian Classification methods.

Using classification methods, a new data whose class label is unknown can be predicted. In this scope, alternative examples are as follows;

- Whether a patient has the risk of cancer or not.
- Whether a person will stay alive or not after he gets out of intensive care.
- Whether a person or organization can be given credit or not.

Classification has a wide use in applications in marketing, medical domain such as medical diagnosis, and fraud detection. Moreover, this methodology provides us to predict the performance of the model which is built up.

In this chapter, one of the most popular data classification methods, decision tree methodology will be explained. The construction of tree, significant attribute selection measure, performance measure, overfitting condition and finally the interpretation of tree structure will be taken in hand**.**

2.11 Decision Tree

The decision tree is a commonly used and one of the strongest classification methods of data mining frequently used in order to generate rules from data. Rules are quite

straightforward and clear. Decision tree has branches and it is like as tree structure. Tree structure is generally used for prediction by historical data in operations research, especially in decision analysis to identify a strategy for attaining an aim. Since decision trees describe rules, they are more popular and charming among other supervised learning methods such as Neural Networks, Bayesian Networks etc. At the same time, because decision trees are cheaper to construct for solving a specific problem, easier to interpret by generating rules, easier to be integrated with data base systems and have better reliability, they are the widely used techniques among the classification models. There are several of algorithms for constructing decision trees such as C4.5, ID3 and CHAID.

A decision tree starts with a root node and consists of leaf nodes, branches and decision nodes. At first all the data samples (real data[3]) are in root node. Decision nodes determine the test to be carried out on a single attribute. At the end of the execution of the test tree is departed into branches without losing any data. Process of branching in each node is executed consecutively and this operation is dependent to upper-level branches. Each executed branch is candidate to complete classification. If classification can not be made at the bottom of a branch, a decision node exists there. But if any class occurs, there is a leaf bottom of this branch. A leaf node indicates the value of the target attributes (class[4]) of examples.

Decision tree classifies a new instances starting from the root node and moving it until a leaf node is reached, which is the label of a class.

Classifying the data by using decision tree technique is a two stage operation. First stage is the learning stage. At this stage, a training set whose class labels are known before is analyzed by the classification algorithms in order to construct a model. Trained model is indicated as classification rules or tree structure. The second stage is classification. In classification step, testing data is used to define the accuracy rate

[3] Hundreds or thousands of training cases.
[4] The categories that examples are assigned to.

of model. If the accuracy rate is acceptable, rules can be used to classify new data whose class label is unknown.

The accuracy rate of a model which is applied to test data is the proportion of true classified data's to all classes in test data. Observed classes in each sample in test data are compared to expected classes which are predicted by the model. If the accuracy rate is enough admissible, the model can predict the unknown classes of new data using inputs variables (Figure 2.7).

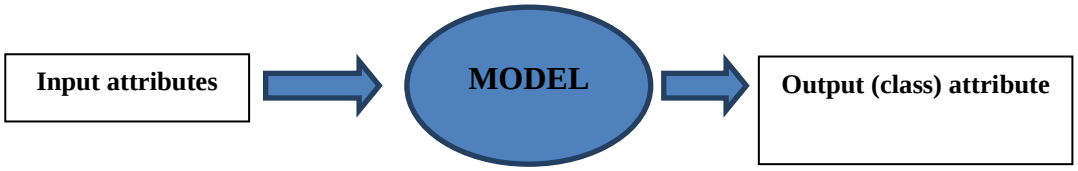

Figure 2.7 Data Mining Methodology General Components

For instance, a model can be constructed by investigating a training data in order to predict the class label of a patient whether he is at risk or not with respect to a specific disease. A classification rule emerged in this model is;

IF age> "65" AND rs7161889= "A_G" AND rs7166325= "C_G" THEN condition_of_patient= "CASE"

In accordance with this rule, people under the research whose age is greater than 65 and rs7161889 equals to "A_G" and rs7166325 equals to "C_G" have risk in terms of Alzheimer's disease.

Decision tree induction is extensively used in applied fields as diverse as medicine (diagnosis), computer science (data structures), botany (classification) categorizing various states into high, medium, low risk groups, generating rules for predicting future cases and identifying the relationships or associations unique to particular sub classes [26].

Decision trees maybe the most effective way to make decisions for the state of the patients in medical sector by utilizing medical observation data in addition to demographic features.

2.11.1 ID3

ID3 is a decision tree learning algorithm developed by Ross Quinlan. Using the ID3 algorithm decision tree is constructed in top-down manner with greedy search by testing each attribute at every tree node. A metric called information gain is used for selecting the most informative attribute in a given node. Entropy and information gain are the important concepts for ID3 algorithm since they are widely used. The algorithm constructs the decision tree based on the calculated information gain ratios considering each attributes. The highest information gain is chosen as the best split criterion [27]. Splitting process is begun to be implemented by choosing the attribute whose information gain is the highest.

The ID3 algorithm is used to build a decision tree, given a set of non-categorical attributes C_1, C_2, ..., C_n, the categorical attribute C, and a training set T of records. The bottleneck in using ID3 algorithm is that numerical data can not be handled with ID3 algorithm. Moreover, missing values must be filled in or cleaned before ID3 algorithm runs. Considering these deficits, Quinlan extended the domain of ID3 to real valued output, such as numerical data, missing valued data.

2.11.2 C4.5

Because ID3 is restricted in dealing with discrete sets of values and missing values, C4.5 algorithm is developed by Quinlan. All of the features of ID3 is inherited to C4.5 [26], [28]. C4.5 algorithm constructs the prediction model with divide and conquer strategy as similar to ID3.

In real world data, data type can be numerical, nominal or ordinal. As mentioned above, ID3 is adequate for nominal data. But numerical data brings a new approach

to analyze of data. Calculating the information gain can be thought as difficult at first sight. However, what is needed to be done is to find a threshold that can separate data into groups: according to threshold values. C4.5 algorithm ranks numeric data. Let the ranked data indicate as $\{v_1, v_2, \ldots, v_m\}$. Assume that the threshold is chosen between v_i and v_{i+1}. In this case $\{v_1, v_2, \ldots, v_i\}$ and $\{v_{i+1}, v_{i+2}, \ldots, v_m\}$ are obtained as two groups, respectively. Considering this example, someone can identify m-1 threshold value. For the splitting criterion, threshold can be calculated using the given formula:

$$threshold = \frac{v_i + v_{i+1}}{2}$$

With this method, the problem is turned out as if data is split according to a particular criterion (smaller than threshold, greater than threshold). Thus, information gain that is applied to nominal data structure can also be applied here. Algorithm considers all the thresholds and chooses the threshold whose information gain is the highest than others. Let e is a threshold with the highest information gain, in this aspect data points provide the condition of $v_i<e$ and $v_i>e$ are divided into two group within the decision tree if the attribute's information gain is the highest.

The other newness in C4.5 is related to handling of missing values. In real-world, data can contain missing values due to some several factors. Collection of data may be difficult, or while transferring the data into electronic environment mistake may be made. Moreover, if the study is related to medical domain, some information may not be obtained such as lab results or demographic data. In this case, instead of trying to process all the data manually (especially if the amount of missing value rate is too high), some new methods is better to add to algorithm. The algorithm firstly detects missing values. Secondly, the median or mean value is put instead of missing value. However, missing value handling approach is generally gives best results if the missing value rate is low in data attribute. Otherwise results and accuracy of the model can be affected negatively [26].

2.11.3 CHAID

CHAID is a classification algorithm used to study the relationship between a dependent variable and a series of predictor variables. It gets independent variables as inputs and determines how variables best combine to address the outcome in a dependent variable [29].

CHAID analysis especially deals with categorized values instead of continuous value. However, algorithm can run even with numerical data. For the categorized datasets, CHAID analysis is a perfect tool to discover the relationship between variables.

For qualitative independent variables, a series of chi-square analyses are implemented between the dependent and independent variables. For quantitative independent variables, analysis of variance (ANOVA) methods are used. If there are differences between the categories of dependent variable splitting conditions are determined optimally for the independent variables so as to maximize the ability to explain a dependent measure in terms of variance components[5].

CHAID technique essentially involves automatically constructing many cross-tabs, and decides statistical significance of the proportions. The most significant relationships are used to construct tree diagram [30].

2.12 Constructing Decision Tree

Instances in data set are inquired based on relevant features and the rules are developed. The object in here is solving the relationships between attributes which are assumed independent from each other. In each data set there are hidden and useful rules and these rules are revealed for decision making while constructing a decision tree.

[5] http://www.themeasurementgroup.com/definitions/chaid.htm

The most considerable step in setting-up a decision tree is which question will be asked respectively. Decision tree is constructed by being interrogated with the question that has the most powerful feature recursively. For splitting data, information gain or gain ratio of each feature is clarified by entropy calculation. Calculating information gain for categorical data type is partially easier than calculating for numeric data type. In decision trees, decisions are made in leaf nodes and some appropriate conditions are expected to occur attaining the leaves. When there are no questions to be asked, this means sub groups are almost pure. In other words, all the instances after splitting the data belong to the same class. Consequently, particular conditions are obtained and leaf is appeared there. The related label information is given to the leaf node.

2.13 Divide and Conquer

There are different types of decision tree construction methods. The most important criteria while building a tree is to be ensured about having enough and reliable data. The crucial point is the implementation of "divide and conquer" step. Divide and conquer is the method that Hunt used [26], [31]. Figure 2.8 expresses the basic tree structure construction algorithm by Hunt. Despite that Hunt used this algorithm for decision making purposes, some improvements have been done.

The most important ones are ID3 (Iterative Dichotomiser) which is the enhancement of Quinlan during the late 1970s and early 1980s, C4.5 (a successor of ID3) which is the second development of Quinlan subsequently. Another algorithm is called CART which is the generation of binary decision trees (classification and regression tree) developed by a group of statisticians Breiman, Freidman, R. Olshen and C. Stone in 1984 [26], [32].

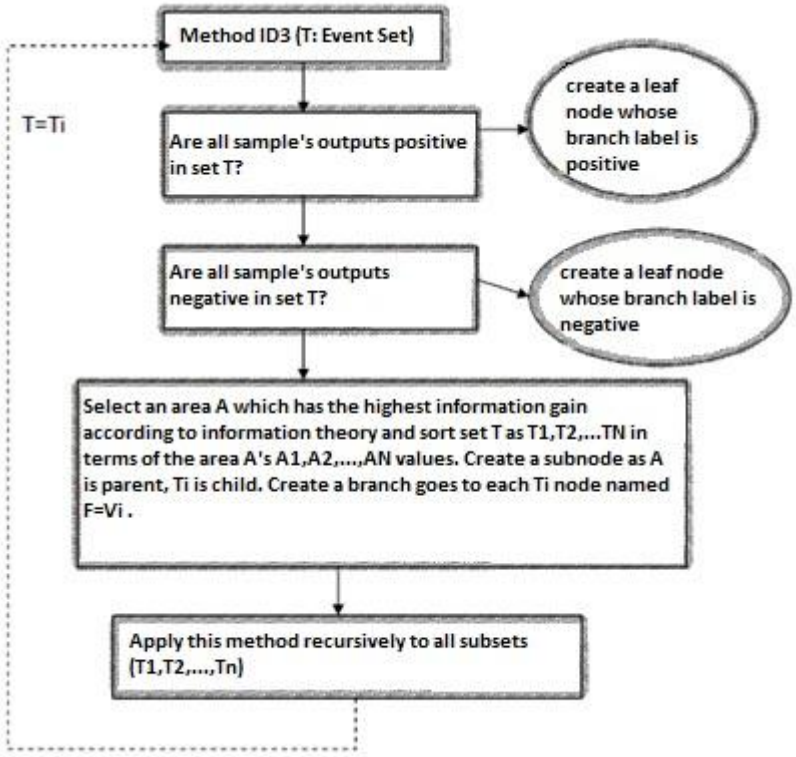

Figure 2.8 Decision Tree Construction Method of Hunt

Input:

D[6] means data partition which is a set of training rows with their class labels.

Attribute list[7], the set of independent attributes, also called candidate attributes. Attribute selection method[8] is a procedure to identify best splitting criterion data into individual classes.

Output:

A decision tree with final decision nodes.

[6] D is data partiton. It refers to complete training set tuples and their class labels.

[7] Attribute list is a whole list that contains the the name of all independent attributes. Attributes are the variables that change from tuple to tuple.

[8] Attribute selection method defines heuristic splitting criteria. The choosen procedure discriminates the tuples and justifies that it builds a tree structure more accurate and better than other ones. The procedure employes a measure of attribute selection such as information gain, gini index or gain ratio.

The given chart above shows the Hunt's method while constructing the tree in a recursive approach. But the most significant case here is selecting of attributes. If the attributes are not randomly selected, the tree will be straightforward and clear to interpret as well as its accuracy will be more crucial and certain. In this case, some enhancements and modifications provided a better tree construction.

2.14 Measures for Attribute Selection

In data mining applications, generally attribute selection measures appear as a heuristic approach for determining the splitting criterion which "best" separates a data whose class labels are known into individual homogenous classes [32]. Measures result in choosing the criterion among the list of independent attributes that are considered the most relevant to dependent variable [33]. As mentioned, all the classes, which are occurred after the splitting criterion are ideally supposed to be pure where tuples are distributes less randomly. This means, all the data samples which fall into a given partition belong to the same class.

Since attribute selection measures establish how the data are to be split into pure sub classes, they are also called as splitting rules. These measures are calculated for each attribute in a test node and ranked. The attribute whose calculated score is the best is chosen as a splitting attribute [32].

The most popular measures for attribute selection in constructing decision tree are information gain, gain ratio and gini index.

2.14.1 Entropy

The most important step while constructing decision tree using ID3 or C4.5 is to choose the significant splitting attribute. The splitting attribute is determined calculating the entropy and information gain based on entropy.

Entropy makes the construction of tree easy with less effort depending on computational limitations. What this means is that choosing the most appropriate tree construction within all the possible constructed trees using all the attributes in learning set is not a good work. Instead of, at the beginning entropy measure and information gain is calculated, so the attribute whose information gain is the highest can be used as splitting attribute.

Entropy is the measuring homogeneity of a variable in the training set due to the presence of more than one possible classification [34]. If the impurity or randomness is high, this means entropy is high. But if there is no randomness with respect to the target classifier, this means entropy is zero.

Assume that there are K categories of dependent variable. It is possible to determine proportion of instances with classification i by p_i for i=1 to k. p_i is the proportion of number of occurrences of class i to the number of instances. p_i changes between 0 and 1.

Let E be the denotation of entropy calculation. The formula for the calculation of entropy is given below. The unit of the calculation is in "bits" of information [34].

$$E = -\sum_{i=1}^{K} p_i \log_2 p_i$$

Consider that S is a sample of training examples. Dependent variable has two categories: positive and negative. Let p+ be the proportion of positive examples in S and p- be the proportion of negative examples in S. Entropy measures the purity of S using the formula given above. For the every probability of p_i, the entropy curve is obtained as shown in Figure 2.9. The peak of the curve implies the highest entropy when the class probabilities are 0,5 for p+ and 0,5 for p- of the dependent variable. Entropy takes the minimum value if and only if all the samples in data set have the same class label [34].

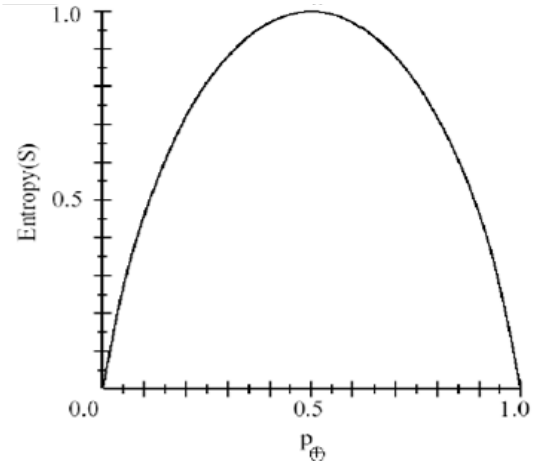

Figure 2.9 Entropy Ratio Based On Proportion of Examples In Dataset

If samples belong to the same class, in this case Entropy = 0.

If samples distributed within classes equally, in this case Entropy = 1.

If samples distributed randomly within classes, in this case 0 < Entropy < 1.

In order to find the best splitting attribute, attribute selection criterion is required which is a metric for how well one attribute classifies the training data.

2.14.2 Information Gain

ID3 and C4.5 uses the information gain which is based on pioneering work by Claude Shannon on information theory as attribute selection measure [32].

As the value of independent variable X_i is known, the information gain for a given attribute X_i with respect to the dependent attribute Y is the reduction in uncertainty [32]. Due to the information gain, dependent variable Y is got pured at the following nodes in every splitting iteration. The uncertainty of Y is measured by its entropy.

Assume that X_i and Y are discrete variables getting values $\{y_1, y_2, \ldots, y_k\}$ and $\{x_1, x_2, \ldots, x_n\}$. We explained how to calculate entropy before. In this case, entropy of Y is calculated with the formula below.

$$E(Y) = -\sum_{i=1}^{k} P(Y = y_i) log_2 (P(Y = y_i))$$

When the value of X_i is known, the uncertainty of Y is the conditional entropy based on given x_i, E(Y/X). The conditional entropy of Y given X_i is calculated with the formula below.

$$E(Y / X) = -\sum_{i=1}^{l} P(X = x_j)\ E(Y / X = x_j)$$

For each X_i in training set, information gain is calculated using the formula below until the entropy of each of these subsets is zero (i.e. each one has instances drawn from only a single class).

$$I(Y ; X) = E(Y) - E(Y / X)$$

The log function to the base 2 is used, because the information is encoded in bits [32]. The information gain ($I(Y ; X)$) determines the test variable (X_i) in each node. Based on the chosen attribute, tree is constructed in top-down manner. Hence, the process of the tree construction by consecutively division based on independent attributes equals to the partitioning the initial training data into smaller homogenous training sets repeatedly [34]. In other words, I(Y; X) tells how much would be gained by branching on X_i [32]. I(Y; X) can be also explained as the expected reduction in the information requirement by knowing the values of one of the independent variable X_i. So, highest information gain is chosen as the splitting attribute at node N. X_i is called "the best classification".

2.14.3 Gain Ratio

The information gain is biased when an independent attribute have many outcomes. This means that information gain selects attributes having a large number of distinct values. As an example, it is possible to make understandable this situation considering an attribute that acts as an identifier such as person's identification number. Since each identification number is one tuple, each partition is exactly pure in terms of class labels. In other words, decision tree may learn the training set too well and biased tree is constructed as a result with overfitting problems.

C4.5 decision tree algorithm uses gain ratio to determine split information and to find the most important attributes [35]. This overcomes the bias. Because it applies normalization to information gain using the split information value given below. This calculation tell about potential information found by splitting training dataset D into p partitions [32].

$$SplitInfo_{Attribute_i}(Dataset) = -\sum_{j=1}^{p} \frac{|D_j|}{|D|} \; log\left(\frac{|D_j|}{|D|}\right)$$

The gain ratio calculation is shown as below.

$$Gain\ ratio\ (Attribute_i\ ) = \frac{Information\ Gain(Attribute_i) = I(Y\ ;X)}{SplitInfo_{Attribute_i}(Dataset)}$$

The attribute with maximum gain ratio is chosen as splitting attribute.

2.15 Overfitting

The top-down construction of a decision tree is a widely used method in classification problems. When a tree is constructed using training data, incorrect predictions may revealed in test set depending on the noise or outliers in training data which affects the accuracy of the model.

During the process of rule extraction, basic decision tree algorithm may grow each branch of the tree just deeply enough to sufficiently classify the training examples. The classifier classifies training data by partitioning the dataset recursively until all the data in a subset belongs to only one class or no further splitting test is available. It results often a complex tree with excessive rules set that overfits the training data [36] and very low predictive power for previously unseen data. In other words, decision tree is constructed depending on the irrelevant attributes of the training samples with the results that it performs very well on the training data but poorly on test data which is not used when tree learns patterns.

Overfitting is realistically occurs since the training set does not contain all the possible samples. This means all the features of unseen data samples in test set may not be included in training data during learning of decision tree. In this condition, decision tree learns rules from training set and fits training samples effectively but test data is unable to be classified correctly because no patterns are obtained from training data that may provide matching test data attributes. Overfitting only becomes a problem when the classification accuracy on test data is significantly downgraded [37].

While dealing with decision trees for making prediction, ways to reduce overfitting must be sought as well as the possibility of significant overfitting must be considered. The leading objective is that there is always a tradeoff between constructing a model that fits the training data as well as possible, and a model that generalizes well to new sample that are not seen in learning process [38].

The figure below demonstrates the effect of overfitting tracking the performance of tree. X-axis holds the information of the number of nodes in the tree (a measure of its complexity) and y-axis is the percentage of correct classifications on the training data and also on test set of data that is not used for classifier learning. The figure says that the performance of training set goes on improving while the test set performance is expected to peak before the complete tree is grown. Afterwards, test set accuracy begins to decrease as the depth of the tree increases [38].

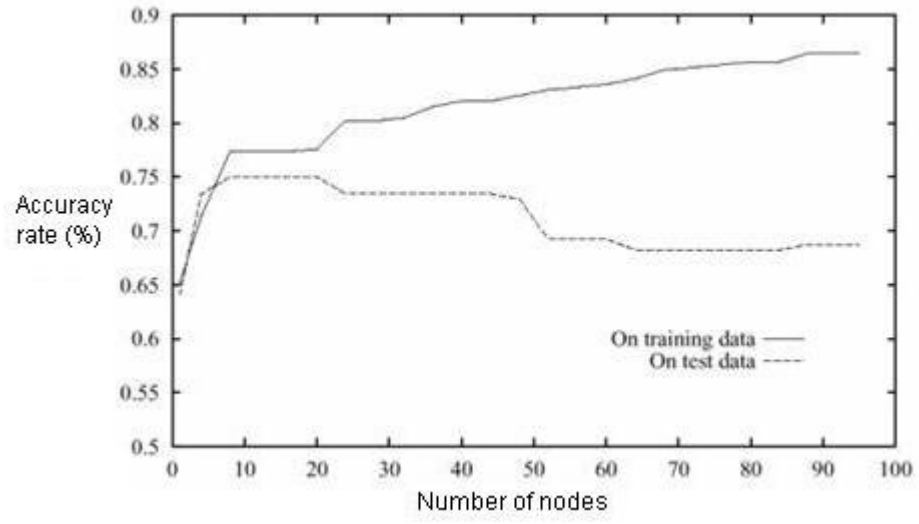

Figure 2.10 Effects of number of nodes on accuracy rate

Accuracy as a function of the number of tree nodes: on the training data it may grow up to 100%, but the final results may be worse than for the majority classifier.

Pruning of a decision tree aims to prevent the overfitting. This stops splitting earlier, or dismiss the branches that have no positive contribution on the accuracy of the decision tree.

- Pre-pruning (or forward pruning): Prevent the generation of non-significant branches.
- Post-pruning (or backward pruning): Generate the decision tree and then remove non-significant branches.

2.15.1 Pre-Pruning

Prepruning halts splitting by deciding the goodness of a split. While implementing a pre-pruning strategy, the subset is first tested to determine whether a termination condition is applied or not. If partitioning the tuples at a node would result in a split that falls below a prespecified threshold (minimal information gain), then further partitioning is stopped.

Prepruning uses the results of attribute selection. The algorithm selects the attributes in each node that are relevant to predict the class for splitting within given a set of attributes whose gain ratio are greater than minimal gain that is prespecified. Thus, the aim of pre-pruning is to determine whether an attribute is significantly correlated to the class. Statistical significance tests such as the chi-squared test are applied by algorithm [39]. The prepruning problem is now simplified to depict the variable with the optimum value of the splitting criterion to split on among all variables for which the statistical test such as chi-square shows a significant association with the class.

In addition, algorithm checks the sample size at a node. If the size of a node is smaller than the minimum split size, partitioning stops as well. Upon stopping, a node is created as a leaf node. The class label of the node is given most frequent class among the subset tuples. If it does not, a further term is generated as usual. If it does, the rule is pruned.

The results obtained clearly show that pre-pruning can reduce the number of terms generated and in some cases can also increase the predictive accuracy.

2.15.2 Post-Pruning

In order to remove the least important branches, postpruning (backward pruning) can also be used. This approach clears subtrees from a "fully grown" tree. Branches at a given node are removed and a leaf node is created here by labeling with the most frequent class among the subtree being replaced [32].

The general steps [decision tree learning] of post pruning:

1. Construct the decision using training set samples, growing the tree until the training data is fit as well as possible. Overfitting may be allowed to occur.
2. Convert the learned tree into set of rules starting from root to the end (leaf) node.
3. Prune (generalize) each rule by removing any preconditions that result in improving its estimated performance (accuracy).

For making decision of whether to replace a node or not, is done by calculating the estimated error using the pessimistic error estimation measure of a particular node and comparing it with its potential replacement leaf. C4.5 uses pessimistic error rate calculation to evaluate performance based on training set.

Error estimate for a sub-tree is weighted sum of error estimates for all its leaves. The error estimate (e) for a node is:

$$e = \frac{\left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right)}{\left(1 + \frac{z^2}{N}\right)}$$

where;

f is the error on the training data,

N is the number of instances covered by the leaf,

z from normal distribution.

Pessimistic error rate is calculated before a rule is deleted, and after a rule is deleted. If the removal of a node (subtree replacement) contributes to the performance of the model, that rule is removed from the tree and all the parent node is labeled with the most frequent class value.

2.16 Accuracy and Error Measures

After a model is constructed over training samples, the most important step is the determination of the model accuracy. As an example, one may use historical data of humans to train a classifier in order to predict the condition of a person in terms of specific disease susceptibility. In this case, estimating of how accurately the classifier predicts the disease susceptibility of a patient in the near future is considerable important. This means, patient data on which the classifier has not been trained is given a label by the classifier. Thus, the accuracy of this prediction must be measured.

For accuracy estimation, there are some measures and these measures are calculated by splitting the dataset into training and test sets. In order to evaluate the accuracy, some partition techniques such as holdout, random subsampling, k-fold cross validation methods are widely used [32].

2.16.1 Classifier Accuracy Measures

Classifier is always derived from learning data. By using the learning data, both training of a classifier and measuring the accuracy of the classifier result misleading estimates. Instead, test set consisting of class-labeled samples which are not used in learning samples is used in order to estimate the accuracy of the model. The accuracy of a classifier on a given test set is defined as the percentage of test samples which are correctly classified [32]. Accuracy is the indicator of how well classifier detects the samples of various classes.

If the accuracy rate of classifier M is Acc(M), the error rate or misclassification rate can be defined as 1-Acc(M). In order to calculate the accuracy rate of the model, confusion matrix is a useful tool. Confusion matrix tells how well the model classifies data samples flawless. Each cells in the matrix ($CM_{i,j}$) indicates the number of samples in class i which are classified by the classifier as the label j. A classifier ideally has a small error rate if most of the samples in test set are represented along the diagonal of the confusion matrix starting from $CM_{1,1}$ to $CM_{m,m}$ with the rest of the cells converge to zero.

Assume that dependent variable contains two classes. In this case, diagonal of the confusion matrix is given at Table 2.2.

Table 2.2 Confusion Matrix

		Predicted Class	
		Class1	*Class2*
Actual Class	*Class1*	true positives($CM_{1,1}$)[9]	false negatives($CM_{1,2}$)[10]
	Class2	false positives($CM_{2,1}$)[11]	true negatives($CM_{2,2}$)[12]

However, the accuracy rate of the model is not always acceptable in the circumstances of the classifier classifies only positives correctly or only negatives correctly. Assume that the accuracy of the model is calculated as 90%. This rate seems quite high but only one of the class labels may have caused this. Assume that 3-4% of the training samples are labeled as C_1 and rest of the training samples are labeled as C_2. In this case, true classified C_2 samples may increase the accuracy of the classifier even if all C_1 samples are classified incorrectly. As a result, C_1 samples are not classified efficiently. So, highest accuracy rate can be ignorable. Instead of accuracy, it is possible to calculate of how well the classifier can recognize the C_1 and C_2 samples. For this purposes, sensitivity and specificity measures can be used, respectively [32]. Sensitivity is also called as the true recognition (positive) rate. Specificity is the true negative, as well. In addition, precision can be used to obtain the rate of labeled as C_1 when the actual value is also C_1.

These measures can be calculated by the given formulas below;

$$Sensitivity = \frac{t\_positive}{positive}$$

$$Specificity = \frac{t\_negative}{negative}$$

[9] True positives are the positive samples that are correctly classified by the model.
[10] False negatives are the positive samples that are classified incorrectly by the classifier.
[11] False positives are the negative samples that are classified incorrectly by the classifier.
[12] True negatives are the negative samples that are correctly classified by the model.

$$Precision = \frac{t\_positive}{(t\_positive + f\_positive)}$$

Where t_positive is the number of true positives, positive is the number of positive samples in training set, t_negative is the number of true negatives, negative is the number of negative samples and f_positive is the number of false positives.
The formula of accuracy as the functions of sensitivity and specificity;

$$accuracy = sensitivity \frac{positive}{(positive + negative)} + specificity \frac{negative}{(negative + positive)}$$

2.16.2 Evaluating the Accuracy of a Classifier

In order to estimate the classifier accuracy some common techniques are commonly used in classification problems. Holdout, simple random sampling, cross-validation and leave-one-out methods are used for evaluating the accuracy based on randomly sampled partitions of the dataset [32]. Considering the overall computation time and produced accuracy, one of these methods is useful for model selection.

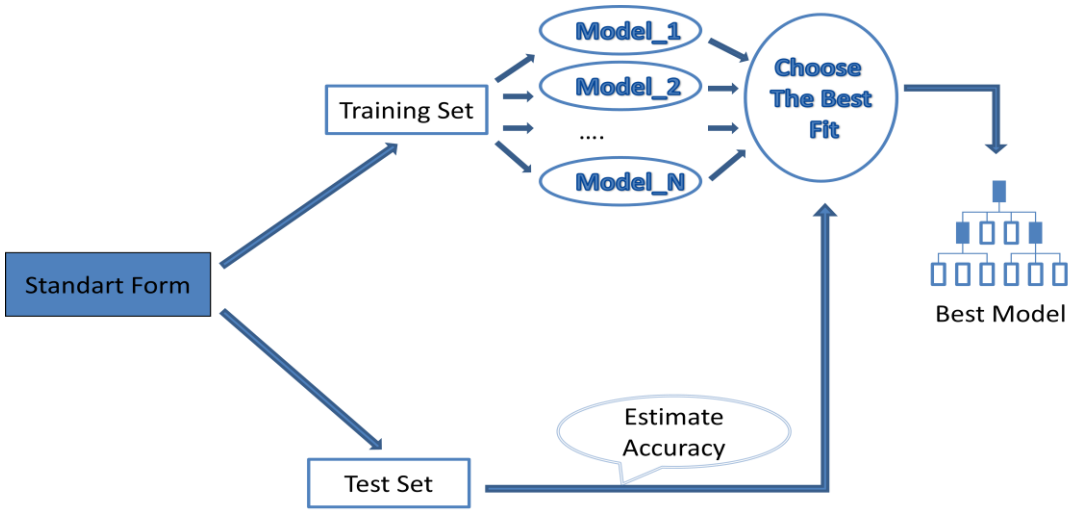

Figure 2.11 Best Model Selection

2.16.2.1 Holdout Method

The simplest method is to partition original dataset into two randomly selecting instances for training that is usually 2/3 of dataset and test that contains the rest of the original dataset. Model is constructed by using training data and accuracy of the model is estimated using test set. However, the estimation is generally poor since the estimation is done using only one portion of the original data.

2.16.2.2 Random Sampling (Repeated Holdout Method)

For increasing the reliability of holdout method, it can be applied to data set k times, iteratively. Original dataset is partitioned randomly for selecting the training and test sets each time. The accuracy rates are obtained from each iteration using selected test data. Finally, overall accuracy is determined averaging the accuracy rates calculated in each iteration.

2.16.2.3 Cross Validation

Cross validation is a statistical model evaluation method which is frequently preferred in predicting the model accuracy. In k fold cross validation, the initial dataset is partitioned into k exclusive subsets which are also called as "folds" randomly. The folds are D_1, D_2, …, D_n and these folds are equal to each other in numbers. In this method, model is tested k times. In the i^{th} iteration, D_i is kept as test set and the rest of the data partitions (k-1) are reserved as training set in order to train the classifier. This procedure is illustrated in the Figure 2.12.

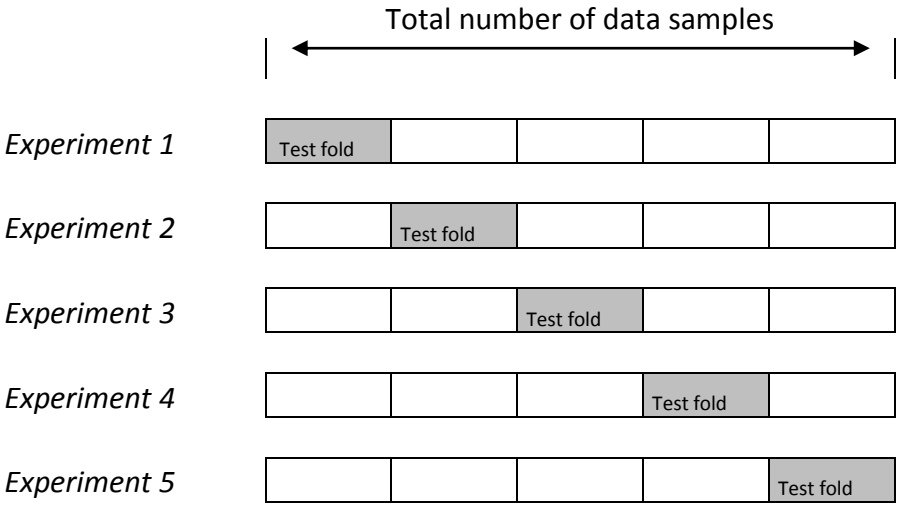

Figure 2.12 Cross Validation

The overall accuracy is calculated by averaging the accuracy rates of folds. The variance of the overall accuracy estimation is reduced as fold number (k) is increased.

$$Overall\ Accuracy = \frac{\sum Accuracy\ of\ Experiment\ i}{k}$$

The advantage of using k fold cross validation is that all the samples in dataset are used for both training and testing. On the other hand, the disadvantage is related to the computational time. Since the method runs k times, evaluating the accuracy takes much time. Unlike the holdout and repeated hold out methods, each sample in dataset is used equally for both training and testing [32].

2.16.2.4 Leave One Out Method

This approach is the special form of cross validation. One data sample is left out during each iteration for the test set. Rest of the data is used for training the classifier. That is, let X_1, X_2, …, X_n data samples in dataset. In the i^{th} iteration, X_i is

the only test instance. Data points excluding X_i train the model. In this case, method runs n times as shown in Figure 2.13.

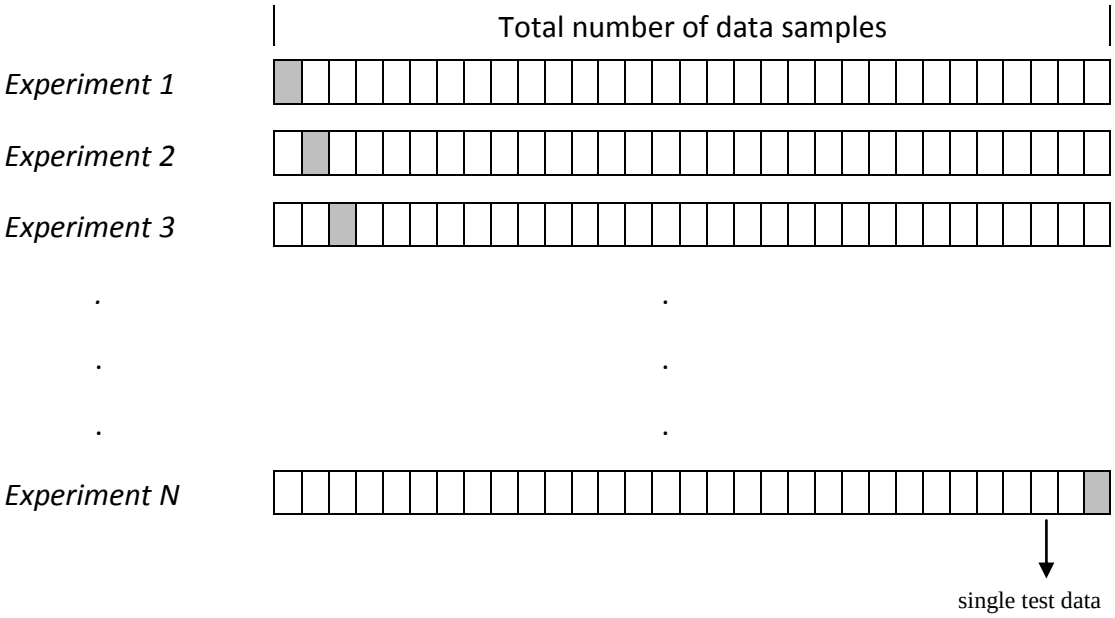

Figure 2.13 Leave One Out

The overall accuracy is calculated by averaging the accuracy rate of every iterations obtained from test example.

The evaluation of accuracy rate of the model by leave-one-out cross validation is good, but it seems very expensive to compute since it calculates accuracy rate n times.

CHAPTER 3

LITERATURE REVIEWS

To the best of our knowledge, currently there is not any study that combines genotyping and clinical information for building a decisive model for the diagnosis of AD.

However there are some cases where limited number of SNP information is analyzed together with clinical and phenotypical features like in the example of pre-eclampsia [3], breast cancer [40], prostate cancer [41], autism spectrum disorder [42] and chronic fatigue syndrome (CFS) [43]. On the other hand, S. Shah and A. Kusiak [44] used these data in order to reduce the dimension by data mining methods and Barkur S. Shastry [11] and Moore and Ritchie used classification methods in order to find novel and hidden interactions between genetic variations to predict the complex disease occurrence.

Linda Fiaschi et al. studied [3] a database related to pre-eclampsia of babies. They emphasized that there are many alternative data mining techniques that may be used for genotype-phenotype association studies. They stated that in genetic studies these mining techniques have discovered interesting findings, especially in the genetic predisposition related to a specific disease. The results were found from the original one by deleting the uninformative attributes for the population of babies. The CBC (corrected birth-weight centile) attribute was chosen as the predictive class which also means dependent to genotypic and phenotypic attributes. The features of each individual comprise both genetic and clinical data. Study was conducted by using 53 SNPs attributes and 6 clinical attributes such as “fetal disease status”, “gestation at

birth(weeks)", "gestation at birth(days)", "weight of the infant", "live at birth" and "CBC" of 372 babies. ID3, ADTree and C4.5 decision tree classification algorithms were applied to database by Linda Fiaschi et al. The goal of this research was both to propose a valid method for SNP analysis by benchmarking the results obtained from a variety of decision tree algorithms in order to identify commonality between trees and to discover any possible association, either genetic or phenotypic, with the specific disease for clinical use. Their results showed the validity of this methodology to find a subset of attributes associated with the predictive attribute, achieving a reduction in the dimension of dataset. From the clinical perspective, study emerged at least two important findings. The first is the significance of the threshold CBC of 10. CBC of 10 gave the highest proportion of agreement (Kappa) that classifies each data in an efficient way using genetic and phenotypic features. The second finding is the dependency of the CBC on the "week of delivery" parameter. From the results of this analysis on PE disease, an association between these two parameters has been found: women with pre-eclampsia who deliver before 35 weeks of pregnancy are more likely to give birth to babies with a CBC under the value of 10. They highlighted that the generic framework explained and applied in this study will lighten the researchers analyzing such data by using classification methodology.

Jennifer Listgarten et al. created [40] predictive models for breast cancer susceptibility by using Single Nucleotide Polymorphisms. They obtained 98 relevant SNPs distributed over 45 genes in order to identify breast cancer etiology and compared these with control group whose members do not have breast cancer. Their aim was to identify a subset of SNPs in order to distinguish between breast cancer and controls by using support vector machines, decision trees and naive Bayesian. SVMs as predictive models achieved 69% to classify data correctly. On the other hand, decision tree approach also performed as much similar as SVMs. Jennifer Listgarten et al. found 3 SNPs sites most important in the model (Figure 3.1). The first one is the 4536 T/C site of the aldosterone synthase gene CYP11B2 at amino acid residue 386 Val/Ala (T/C) (rs4541), the second is the 4328 C/G site of the aryl hydrocarbon hydroxylase CYP1B1 at amino acid residue 293 Leu/Val (C/G)

(rs5292) and the last one is the 4449 C/T site of the transcription factor BCL6 at amino acid 387 Asp/Asp (rs1056932).

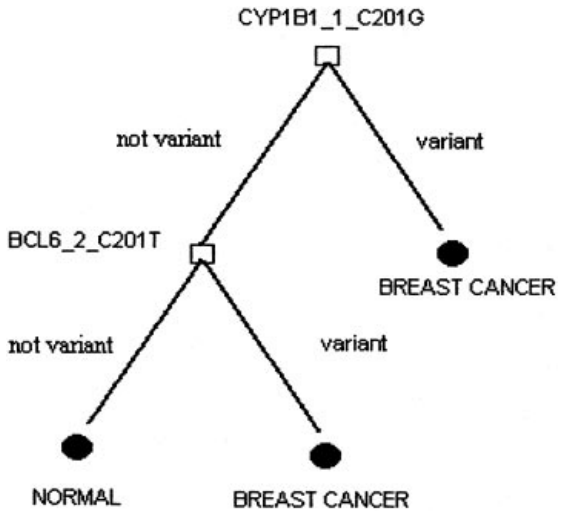

Figure 3.1 Predictive Model For Breast Cancer Susceptibility

According to Jennifer Listgarten et al., since technology is improved fast in terms of identifying more SNPs, it is possible to predict with higher accuracy and a useful clinical tool can be developed to predict cancer cases. They expressed in their study that the decision tree had more balanced errors than the other in the prediction of both cancer and noncancerous persons.

Jill S. Barnholtz-Sloan et al. has leaded [41] a study in order to associate the risk of prostate cancer with inherited variability in genes. The main goal of their study was to reveal interactions that cause prostate cancer by using classification and regression tree (CART) modeling. They used not only genetic information but also interactive effects for the prediction of risk. They maintained the study over 1084 prostate cancer cases and 94 controls. They stratified all the samples in terms of race, and analyses were made (Figure 3.2 and Figure 3.3). Finally, they compared the unconditional logistic regression results with the race-stratified CART results calculating the area under ROC curve. In the study, however the models constructed for each races differed from each other, there were found some common features such as age, family history in respect to prediction of the disease. Another finding in the study was that the specific genotypes and/or haplotypes differed between races.

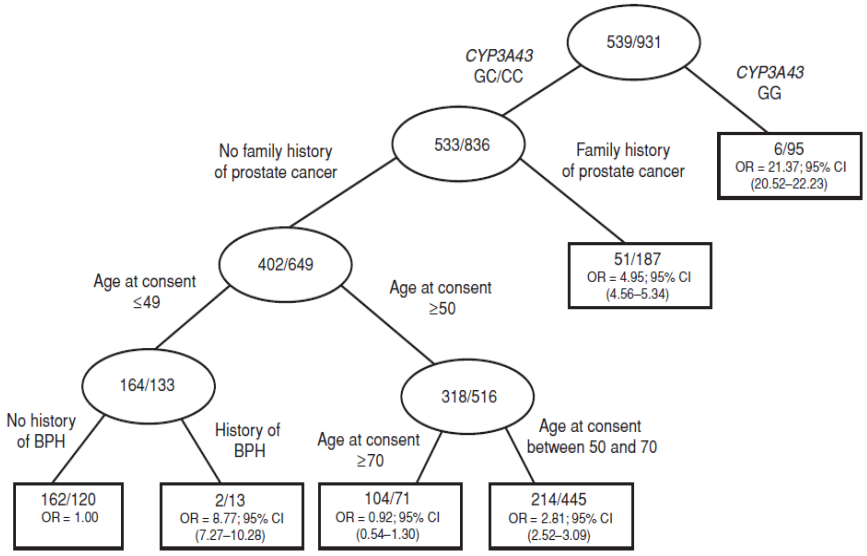

Figure 3.2 Pruned Classification Tree For Androgen Pathway in European Americans

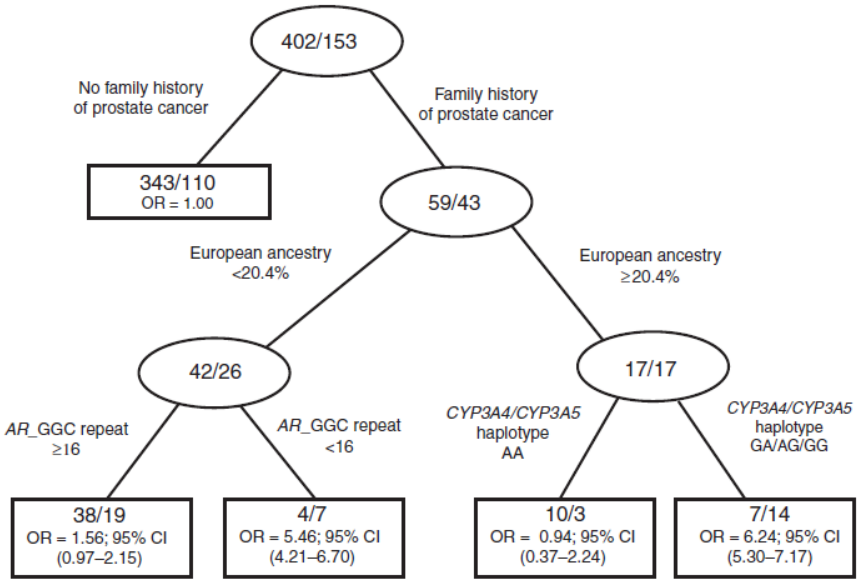

Figure 3.3 Pruned Classification Tree For Androgen Pathway in African Americans

As Figure 3.2 illustrates, for European Americans, people with particular CYP3A43 genotypes, family history of prostate cancer, age factor and history of BPH could have prostate risk. On the other hand, according to the study carried on by Jill S. Barnholtz-Sloan, if African Americans were considered (Figure 3.3), interactions

associated with risk would exist for those with family history of prostate cancer, individual European ancestry proportion, number of GGC AR repeats, and CYP3A4/CYP3A5 haplotypes based on constructed decision tree. They suggested that their findings could be used to select patients for PSA screening in the case of the detection of heritable risk. So, associated features might be appropriately guided to decision maker in terms of clinical decision support.

Yun Jiao et al. goaled [42] to determine whether SNP-based predictive models can predict the severity of a specific disease which is called Autism Spectrum Disorder[13] (ASD). High heritability of phenotypes of ASD was showed by Freitag [45] and Geschwind [46] in family based genetic studies. Moreover, Belmonte et al. (2004) showed in their study significant associations between ASD and genetic markers. Freitag (Freitag 2007; Freitag et al. 2010) maintained SNP studies to determine ASD-related altering on DNA and reported that SNPs in chromosomes 2, 3, 4, 6, 7, 10, 15, 17, X and Y were associated with ASD. Other findings from Belmonte et al. (2004) and Kim et al. (2008) were that many changes in genes such as GABRA4, GABRA2, GABRB1, GABRB2, GABRB3, TDO2, SLC25A12 were associated with ASD. From the literature review, Yun Jiao et al. selected nine ASD-related genes and 29 SNPs depended on these genes (GABRA4, GABRB1, TDO2, GABRB2, GABRA2, GABRB3, GABRA5, SLC25A12, and BDNF) for their study. Tree-based modeling have been commonly used in SNP-based classification (Bureau et al. 2005; Huang et al. 2004; Nunkesser et al. 2007; Park and Hastie 2008), so Yun Jiao et al. also applied three supervised learning methods such as decision stumps (DSs), alternating decision trees (ADTrees), and FlexTrees to generate diagnostic models. They applied learning algorithms to generate diagnostic models. Yun Jiao et al. used 118 instances of 29 SNP variables for SNP-based diagnostic model and they defined ASD symptom severity as class variable (Figure 3.4). Yun Jiao et al. finally identified important markers for overall symptom severity based on SNP-based diagnostic models. They reported that the SNP rs878960 in GABRB3 was selected by all models; this means it is the most significant SNP that has a role in prediction of the disease. As a latest step, they measured the performances of models and they

[13] Autism-spectrum disorder (ASD) is a pervasive neurodevelopmental disorder characterized by abnormal social behavior, impaired communication, and repetitive/stereotyped behavior.

obtained an accuracy of 67%, sensitivity of 88% and specificity of 42% when used DS and FlexTree. From the study of Yun Jiao et al., supervised learning methods such as decision trees have been suggested as an effective way of predicting a complex disease using genetic markers of individuals.

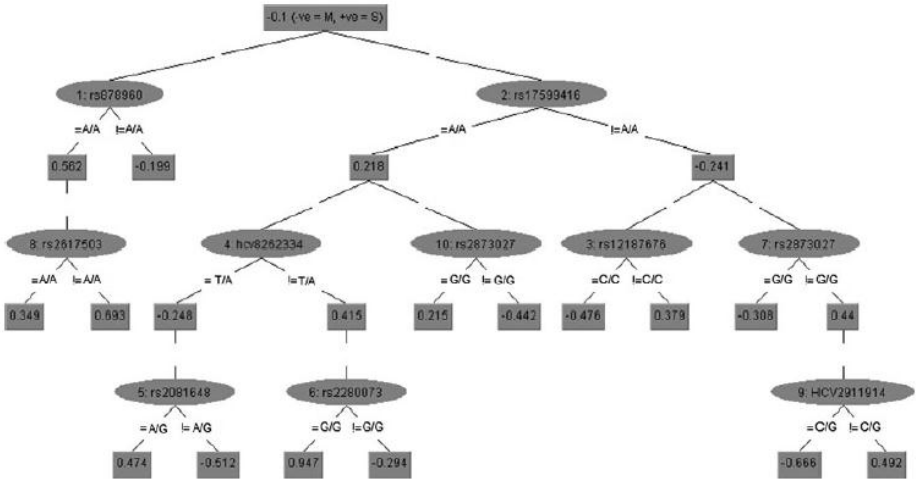

Figure 3.4 Tree Model Generated by ADTree.

Lung-Cheng Huang et al. studied [43] to predict chronic fatigue syndrome (CFS) using SNPs data and benchmarked computational tools with and without feature selection considering 42 SNPs by the Chronic Fatigue Syndrome Research Group. In order to obtain results between CFS and SNPs, they applied three different data mining classification algorithms such as naive Bayesian, support vector machines and C4.5 decision tree. In addition, they utilized feature selection methods to find the subset of representative SNPs. They used hybrid feature selection, information gain and wrapper-based approaches in dataset. Considering the correctly classified rates, naive Bayesian classification with wrapper-based feature selection performed best within predictive models. Naive Bayesian with feature selection out performed by 0,70 in terms of area under curve(AUC). For the naive Bayes model with the wrapper-based approach, only 8 SNPs out of 42 was identified, including rs4646312 (COMT), rs5993882 (COMT), rs2284217 (CRHR2), rs2918419 (NR3C1), rs1866388 (NR3C1), rs6188 (NR3C1), rs12473543 (POMC), and rs1386486 (TPH2).

S. Shah and A. Kusiak explained [44] that one of the important area in bioinformatics was the identification of gene/SNP patterns not only for impacting cure/drug development but also for associating genotyping and phenotyping information for various diseases. Since genomic studies give extensive amount of data with the number of single nucleotide polymorphisms (SNPs) ranging from thousands to hundred thousands, S. Shah and A. Kusiak used data mining methods in order to reduce the dimension of this data for supporting clinical diagnosis and identifying relationships between genotypic and phenotypic information as well as the determination of SNPs related to a specific disease analyzing of genomic data. S. Shah and A. Kusiak employed a global search mechanism, weighted decision tree, decision tree based wrapper, a correlation-based heuristic, and the identification of intersecting feature sets for selecting significant genes. Their reduction aimed methods resulted in 85% reduction of the number of SNPs related to the disease. Hence, the relative increase in cross-validation accuracy and specificity for the significant gene/SNP set was 10% and 3.2%, respectively according to their study. This showed while the number of SNPs has decreased significantly, the quality of knowledge obtained was increased due to the decision tree modeling. They emphasized in their study that feature selection could be achieved by various supervised and unsupervised methods such as k nearest neighborhood, decision tree, multi, layer perceptron, self-organizing maps.

Barkur S. Shastry made a research [11] to identify the patterns of SNPs in conditions such as diabetes, schizophrenia, and blood pressure homeostasis. By the experience of Barkur S. Shastry in the study, he explained that common disorders are caused by the combined effects of multigenes and nongenetic environmental factors (multifactorial). Therefore, it is likely that sequence variation alone is not sufficient to predict the risk of disease susceptibility. According to Barkur S. Shastry, determining how SNPs affect an individual’s health and transforming this knowledge into the development of new medicine to run decision support system, which requires the correlation of SNPs with specific diseases, will help the treatment revolution of most common diseases. Finally, this knowledge captured from SNPs and phenotype associations may give clinicians more insight into the disease and change the

definition of some disorders in the future.

Moore and Ritchie [24] summarized three important points that must be considered for a successful genetic prediction of a disease using genome wide approach. First, a powerful and appropriate data mining method must be developed in order to model the relationship between DNA order variation and disease existing, statistically. The second challenge is the selection of subset among SNPs that should be included for analyze. The final challenge is the interpretation of results. On the other hand, Moore and Ritchie (2004) emphasized that making etiological inferences from intelligent models may be the most difficult step of all because many called this progress as a needle in a haystack.

For all the disease studied in the literature, success rate is not above 70%. However as there is not any study with AD data yet we can not present any classification as benchmark.

CHAPTER 4

MATERIAL METHODS

The main goal of this study is to identify DNA sequence alterations (genotype) which are associated with the clinical findings to predict whether someone is at risk in contracting Alzheimer’s disease (AD) or not.

In addition to clinical features, some demographic features of individuals are considered to contribute in phenotype-genotype associations, as well. Such a study requires a well described and characterized clinical database including people's full blood count with LDL, HDL, WBC, cholesterol, body mass index etc. and genotype data obtained through blood sample, too.

In order to clarify the nucleotide alterations that can trigger AD and clinical findings relevant to the emergence of AD, a large sample size of case-control pairs are used in this study.

4.1 Dataset

The AD genotyping and phenotyping data is obtained from GENADA study through the dbGAP database, which is a multi-site collaborative study involving GlaxoSmithKline Inc. and medical centers in Canada. The data was planned to collect from approximately 1000 Alzheimer's disease patients and 1000 ethnically-matched controls in order to associate DNA sequence variations in genes with Alzheimer's disease phenotypes. However, with authorized access, 1718 study participant's individual level data is now available through this study. GENADA is

reported that this study is planned to begin first quarter of 2002 and end late of September 2003.

From the five of nine medical centers in Canada, eligible individuals who have Alzheimer's disease at the level of mild to moderate, a group of controls who are not yet considered AD and siblings who may be both affected or not affected were initially examined and data were obtained.

4.1.1 Selection Criteria for AD Patients

In the study AD patients must have satisfied the following conditions. Such patients were included within the study.

- ADRDA/NINCDS14 criteria for diagnosis of possible Alzheimer's disease.
- Positive diagnosis of Dementia of the Alzheimer's Type (DAT) on the DSM-IV check list, with a diagnostic code based on age of onset (to be determined by first symptoms noted by friends and family) and predominant clinical features: early onset (age 65 or less) uncomplicated (290.10), with delirium as mode of presentation (290.11), with delusions as mode of presentation (290.12), with depressed mood as mode of presentation (290.13), versus late onset (after age 65) uncomplicated (290.0), with delirium as mode of presentation (290.3), with delusions as mode of presentation (290.20), with depressed mood as mode of presentation (290.21).
- Global Deterioration Scale of Reisberg et al., score of 3 to 7.
- Have clinical data available confirming Alzheimer's diagnosis.

[14] The NINCDS-ADRDA Alzheimer's Criteria were proposed in 1984 by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and Related Disorders Association (now known as the Alzheimer's Association) and are among the most used in the diagnosis of Alzheimer's disease (AD). These criteria require that the presence of cognitive impairment and a suspected dementia syndrome be confirmed by neuropsychological testing for a clinical diagnosis of possible or probable AD; while they need histopathologic confirmation (microscopic examination of brain tissue) for the definitive diagnosis.

4.1.2 Selection Criteria for AD Controls

Control individuals were matched ethnically and they were eligible to the cases by the gender and age in order to satisfy homogeneity. In this study controls were included if the following conditions applied.

- Not possess the history of memory problems.
- Mini Mental State Examination[15] higher than the appropriate threshold dementia score taking into account ages of individuals.
- DRS -2 AEMSS[16] (Age and education adjusted MOANS scale score) of 9 or higher (after adjustment for age and education).
- Clock test (11:10) with a score equal or greater than 14.

The Figure 4.1 shows the number of participants with the availability of their genotype and phenotype data.

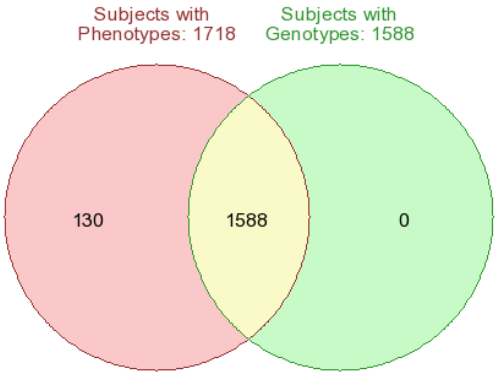

Figure 4.1 Availability of the Genotype/Phenotype Data of Participants

[15] The mini–mental state examination (MMSE) or Folstein test is a brief 30-point questionnaire test that is used to screen for cognitive impairment. It is commonly used in medicine to screen for dementia. It is also used to estimate the severity of cognitive impairment at a specific time and to follow the course of cognitive changes in an individual over time.

[16] The DRS-2 Total Score is a composite score comprising the five DRS-2™ subscales: Attention (ATT), Initiation/Perseveration (I/P), Construction (CONST), Conceptualization (CONCEPT), and Memory (MEM). Client obtained a DRS-2™ Total Score of 117 out of a possible 144 points, which corresponds to an Age-Corrected MOANS Scaled Score of 3 (1 percentile range) and indicates a severely impaired level of performance. This Total Score also corresponds to an Age- and Education-corrected MOANS Scaled Score of 3 (1 percentile range) and indicates a severely impaired level of performance.

All participants provided their genetic and clinical data voluntarily and they provided their consent of the use of data by self and legal representative.

Using the clinical and genotype data, prediction of the disease for the future provides an important point of view in terms of clinical decision making. For this reason, both clinical data and single nucleotide polymorphisms are associated with the Alzheimer's disease.

In our study, clinical data obtained from the blood samples of controls and cases contain cholesterol (mmol/l), hemoglobin (g/l), HBA1C_PCT, HDL Cholesterol (mmol/l), LDL cholesterol (mmol/l), Triglycerides (mmol/l) and amount of White Blood Cells. We used all these variables in our study as shown at Table 4.1 since any of them may be relevant to Alzheimer's disease.

In addition to clinical and genotype parameters that are obtained from the individuals, medical records are also saved such as age of onset of first symptoms noted of AD cases, body mass index at the moment of first diagnose.

Since the initial goal of the study is to reveal if Alzheimer's disease susceptibility is changed by polymorphic variation in specific genes, hundreds of thousands of probes are arrayed on a chip. This chip determines many SNPs interrogating simultaneously [47]. In this study 410907 SNPs are captured for each individual and by interrogating the DNA sequences of people SNP alleles are gained where the allele frequency is greater than 1% in the population.

- **Data access provided by:** dbGaP Authorized Access
- **Data Access Committee (DAC):** JAAMHDAC@mail.nih.gov
- **Release Date:** January 21, 2010
- **Embargo Release Date:** July 21, 2010

Table 4.1 Clinical Data Attributes of Individuals

Variable Name	***Description***
age	Diagnosis age
age_on	Onset age of AD
gender	Sex of the individuals.
case/control	Information about affection status
Body Mass Index (BMI)[17]	Body fatness for individuals.
CHOL (mmol/l)[18]	Amount of fat lipid carried in the blood by molecules called lipoproteins.
HB (g/l)[19]	Amount of proteins that are found in red blood cells.
HBA1C_PCT[20]	Percentage of hemoglobin in red blood cells (erythrocytes) that are tied up to glucose.
HDLCH (mmol/l)[21]	Amount of high density lipoprotein.
LDLCH (mmol/l)[22]	Amount of low density lipoprotein.
TRIG (mmol/l)[23]	Amount of triglycerides in blood plasma.
WBC (giga/l)[24]	The number of leukocytes in the blood.

[17] **Body mass index** is defined as the individual's body mass divided by the square of his or her height. It is used to screen for weight categories that may lead to health problems.

[18] **Cholesterol** is a fat (lipid) which is produced by the liver and is crucial for normal body functioning. It is essential for determining which molecules can pass into the cell and which cannot. It is essential for the production of hormones released by the adrenal glands (cortisol, corticosterone, aldosterone, and others)

[19] **Hemoglobin** is responsible for carrying oxygen from the lungs to all other tissues of the body. Hemoglobin also refers to a blood test that indicates the amount of hemoglobin in the blood. In this case, how well the red blood cells are able to carry oxygen can be inferred.

[20] **HLA1C** is an important analyze for the follow-up of diabetes. It tells the average glucose amount of last 2-3 months of an individual. Normal values are between 4-6%.

[21] **HDLCH** is also referred as bad cholesterol. LDL carries cholesterol from the liver to cells. If too much is carried, too much for the cells to use, there can be a harmful buildup of LDL. This lipoprotein can increase the risk of arterial disease if levels rise too high. Most human blood contains approximately 70% LDL.

[22] **LDLCH** is also referred as good cholesterol. HDL prevents arterial disease. HDL does the opposite of LDL - HDL takes the cholesterol away from the cells and back to the liver. It is either broken down or expelled from the body as waste in the liver.

[23] **Triglycerides** are the chemical constructions in which most fat exists in the body, as well as in food. Too much of this type of fat can cause hardening and narrowing of arteries. High triglycerides often occur along with high levels of Cholesterol.

[24] **WBC** gives information about a wide range of disease and conditions. WBC helps diagnose an infection or inflammatory process since some diseases trigger a response by the immune system and cause an increase in the number of WBCs.

4.2 Preprocessing of Dataset

Data preprocessing is the most important step in knowledge extraction methods. Cabena et al. [48] say that preparation of data takes 60% of efforts of whole knowledge extraction process.

In order to get good results, high quality of data is needed. Thus, incomplete, noisy, inconsistent data must be cleaned from the database. In this aspect, data cleaning, data integration, data transformation and data reduction can be counted as the components of data preprocessing as shown in Figure 4.2.

The data preparation step in knowledge discovery covers all the activities for constructing the final data from the raw data. Data preparation tasks may be applied repeatedly but there is no need to apply all the tasks in any designated order [49].

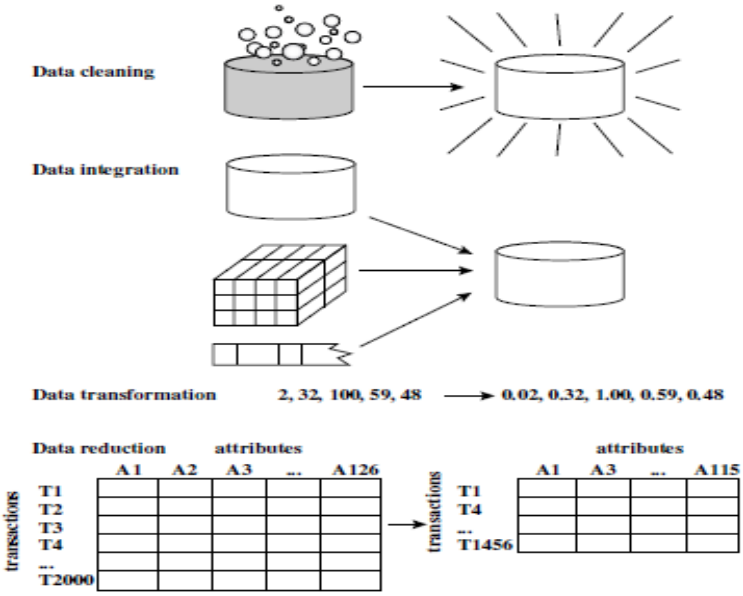

Figure 4.2 Data Mining Preprocessing Phases

4.2.1 Data Cleaning

This step is related to handling missing values, identify outliers and smooth out noisy data as well as correcting inconsistent data. Incomplete, inconsistent or noisy data are commonly seen in large, real world databases.

In our study, some lab findings for individuals are missing in the dataset. This may be because test is unable for person or data is not entered by laboratory. We eliminated the tuples whose attribute values are unknown because the size of dataset is quite enough. Even if we remove tuples with missing values, the accuracy of the model will not be affected much. Furthermore, inconsistent data is corrected. For example, if age onset is entered for a person who does not have AD symptom, this is not correct and inconsistency occurs here. Such data problems were handled as one of the step of preprocessing phase.

The Table 4.2 shows the missing value counts of clinical attributes in the dataset.

Table 4.2 Missing Value Counts of Clinical Data

	CHOL (MMOL_L)	HB (G_L)	HBA1C_ PCT	HDLCH (MMOL_L)	LDLCH (MMOL_L)	TRIG (MMOL_L)	Age_on for "Case" group
Missing	22	36	35	27	61	22	14

All the genotype data is fully filled in, so SNP alleles for each individual is both defined and no need to imputation. When we clear the missing value data, there are 1480 data tuples left for the construction of the prediction model. In addition, onset ages of 14 people in the case group are not kept in the dataset. In this case, we fill the missing values of age-onset attribute with the mean of age of cases.

4.2.2 Data Integration

In today's relational databases or object oriented programming approaches, data is segmented since keeping the all types of data in one table is theoretically

unacceptable. Correspondingly, GENADA contains different repositories in terms of genotype data or laboratory results. Combining the research results from different sources becomes significant.

In this step, data from different sources are combined. But there may be entity identification problems and data value conflicts may be occurred.

In GENADA, genotype data is stored in a table with individual IDs (subject ID) which may also be considered as primary key. In this way, it is possible to match people's SNPs data with other features specified in other data tables through using structured query language (SQL).

Laboratory results corresponding to each subject ID are also kept in another data table. There may be more than one tuple for a specific subject ID. In this condition the first visit values are taken into account since the diagnosis is made and determined in the first visit. Other visits are just for the monitoring and tracking the condition of people.

Database management systems help combining high dimensional data with structured query language. We used MS Access to transfer data tables and to create relationships between keys.

4.2.3 Data Transformation

Normalization of data may be required during analyze if parametric statistical methods are used such as nearest neighbors or neural networks to obtain better and more accurate results. There are many data normalization methods such as max-min normalization, standardization etc. We do not deal with normalization techniques during data analyze.

4.2.4 Data Reduction

Data analysis problems are more prevalent in bioinformatics. Since biological data contains so many features, in such large datasets, dimension reduction is generally beneficial for not only computational efficiency but also improvement of accuracy of the analysis. In such a case, reducing the dimension of data is an effective way to prepare data for the further analyze. This means to describe current information in less attributes. There are two major techniques of dimension reduction. One of the methods is called feature selection and the other is called feature extraction. Feature selection is a process of choosing the optimal subset of features based on an objective function or method. This directly affects the mining performance, speed of learning, reliability of the analyze and simplicity and comprehensibility of results found. On the other hand, feature reduction refers to the mapping of the original data to a lower-dimensional data by using some statistical methods such as principal component analysis, regression attribute selection criteria such as forward selection, backward elimination, stepwise or defining the correlations between each attributes and dependent variable are the most widely known and valid reduction methods.

The other face of data reduction is about the deletion of data tuples. In this aspect, data with missing values that can not be predicted may be deleted if there is enough data to analyze. In addition, statistical sampling methods such as stratified sampling or simple random sampling can be used to reduce data, but the rest of samples must represent the data well.

In our study, from the high dimension of data attributes, the subset which contains the highest significant data attributes is chosen. This gives an opportunity that with a small set of data attributes faster and robust inductions can be done instead of using all the attributes. On the other hand, the accuracy rate of classification does not significantly decrease.

4.2.4.1 GWAS

The first implementation is determination of statistically significant SNPs associated with the Alzheimer's disease. In order to reveal most important SNPs, we first used PLINK software. PLINK is an open source and free software, developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT. The aim of this free tool is to analyze genotype-phenotype data associations of case/control groups in a computationally efficient manner.

PLINK runs in the command prompt environment and it is sustained by 3 parameters [50]. We used default parameters in order to obtain results. For more interests in use of PLINK, official documents [50] are available.

As a result, a range of features such as data management, summary statistics, population stratification and basic association analysis are given. Furthermore, we obtained statistical significant SNPs by their p-values. For each SNP within 410969 SNP biomarkers, asymptotic p value is calculated by PLINK and an output file whose extension is ".assoc" is created at the end of this transaction. P value identifies the statistically most significant disease associated SNPs. So, PLINK associates SNPs statistically; this means PLINK does not consider the biological significance in association studies. So each calculated P-values are unadjusted p values. Not only statistically but also biological significance must be taken into account during defining the most relevant SNP attributes in dimension reduction since the significant SNPs will determine Alzheimer disease in a more accurate and effective way.

4.2.4.2 SNP Prioritization

Reducing the number of SNPs by taking biological and statistical importance into consideration for selecting the representative SNPs from hundreds of thousands SNPs is one of the current challenges. In our study, after we have applied GWAS, we have obtained p-values of SNPs that are statistically significant. Then, the major

problem was reducing the dimention by selecting the most important SNPs related to disease. AHP based scoring algorithm was used. METU-SNP [17] ranked the SNP scores according to importance and SNPs with a score above 0,40 are selected. There were 958 biologically and statistically significant SNPs defined with an AHP Score ranking between 0.409511-0.717559. RsID's of selected SNPs are provided in the Appendix B.

4.3 Data Analyze Using Rapid Miner Tool

Rapid Miner is an open source tool for data mining. Rapid Miner is generally preferred in data mining applications since it has a powerful graphical interface for both design of analysis process and interpretation of the results. It presents efficient data handling such as data loading, data transformation, data modeling using operators and data visualization. Java infrastructure allows integrating other data mining tools such as Weka, R, as well. Moreover, Rapid Miner supports error recognition and quick fixes, hence the error can be found easily during analyze [51].

In this study, we used Rapid Miner tool to construct a probabilistic decision support model by C4.5 decision tree algorithm using clinical and genetic features of 744 case and 746 control subjects. Decision support model will assist deciding about the condition of the patient. As mentioned before, C4.5 has some advantages such as handling missing values or using both numerical and categorical data.

4.3.1 Constructing the Model

In the present study, we have preprocessed the raw data and constructed two different decision trees in order to see how much genotyping and clinical information of patients contributes to the prediction of Alzheimer's disease and if increases the accuracy rate of the model.

The first decision tree inputs only representative SNPs and the second construction contains both SNPs and clinical information as inputs.

The C4.5 algorithm constructs decision trees in top-down manner by using the best single feature test as mentioned before. The criterion of the best single feature test is gain ratio for C4.5 to split data into subsets. The test selects the highest value of gain. This process is such a ranking of features implementation considering each features is independent and there are no interactions between them.

Prediction of the susceptibility to AD applying data mining classification algorithm is the goal of this study. Hence, any attribute which included in the model must contribute to the accuracy of the prediction. Considering this, we chose the representative SNPs subset within hundreds of thousands SNPs. In addition to genetic markers, laboratory test results were used in the analysis.

To investigate the accuracy of C4.5 prediction model k-fold cross validation was used. For k-fold cross validation method, data set randomly divided into distinct parts. k-1 folds were used for training the data and 1 fold was used to estimate the predictive performance of the model. This procedure was repeated k times. Finally, the average estimate of overall iterative runs was calculated as the accuracy of the model. The process of data classification that models our implementation is given in Figure 4.3.

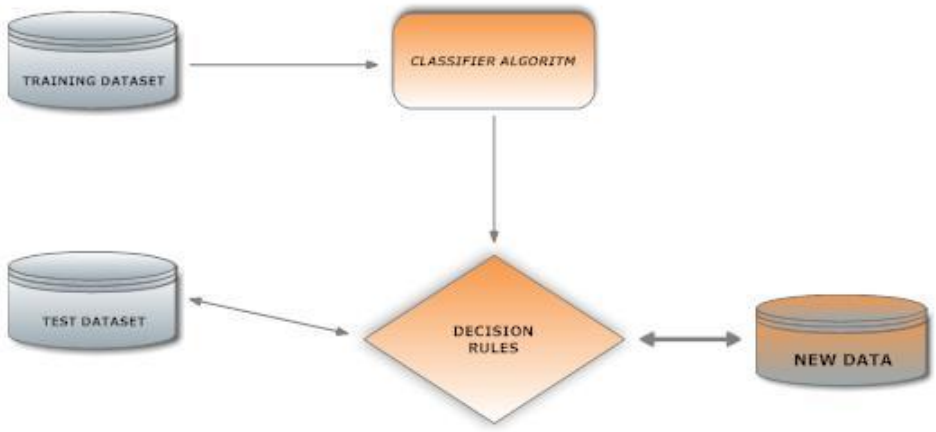

Figure 4.3 Data Classification Process

Since the construction of a decision tree depends on entropy reduction, the splitting approach may overfit the training set and may lead to poor accuracy in the test set. In response to the overfitting problem, pruning strategy is used during the construction of decision tree. Rapid Miner allows us using both prepruning and postpruning. The prepruning checks the goodness measure (gain ratio, information gain etc.) if it falls below a threshold or not. In addition, the minimal size in a node for splitting can be identified. If the node has samples with the number smaller than the minimal split size, construction stops there. On the other hand, at each node, beholding the number of instances that are misclassified on a test set is possible by propagating errors upwards from leaf nodes. This can be benchmarked to the error rate if the node was replaced by the most common class. If the difference is a reduction in error, then the subtree at the node can be considered for pruning. This calculation is performed for all nodes in the tree. This strategy is called postpruning and this may result more accurate rule extraction.

4.3.1.1 C4.5 with Genetic Markers (Representative SNPs)

Only statistically and biologically significant SNPs subset was chosen as input data. We obtained the subset using AHP scoring as mentioned before. We applied C4.5 algorithm to dataset with its 958 independent variables considering the given selected genotype data in order to predict whether a person is case or control.

We used gain ratio to select best single feature which divide dataset into partitions. The minimal gain ratio is identified as 0,01 since gain ratio greater than 0,01 does not have a role to divide data into subsets. The 11-fold cross validation has been choosenas it should satisfy the requirements for the volume of our dataset for making unbiased error prediction over test set.

Extracting Rules from Decision Tree

Based on Figure 4.3, we firstly identified a validation method to be used in modeling phase. The next step is to apply learning algorithm to training data and to generate

rules. Decision trees are preferred since they are easy to understand. Decision trees have visual impressions over humans. But as the depth of tree increases, they become large and difficult to interpret. In this condition, IF_THEN rules are extracted from decision trees. IF_THEN rules make interpretation of decision tree easier [32].

To extract rules, every rule is created for each path from the root node to leaf node using logical operator "AND" within IF parts. THEN part holds the information of class prediction.

We used Rapid Miner to generate rules. The summary of the decision tree model is given in Table 4.3.

Table 4.3 Decision Tree Construction Summary for Representative SNPs

Learning Algorithm: C4.5
Attribute Selection Criterion: Specifies the used it for selecting attributes. We chose the gain ratio for the criterion term.
Inputs: 958 SNPs chosen based on AHP Scoring.
Output: If a person is AD or not AD.
Minimal Gain: Which must be achieved in order to produce a partition = 0,01.
Maximal Depth: The maximum tree depth = 12.
Validation Method: 11-fold Cross Validation.
Minimal Size for Split: The minimal size of a node in order to allow a partition = 5.
Minimal Leaf Size: The minimal size of all leaves = 2.
Confidence: Used for the pessimistic error calculation of pruning = 0,25.
Number of Pre-pruning Alternatives: The number of alternative nodes tried when pre pruning would prevent a split = 3.

The tree structure for the prediction of Alzheimer's disease is given in the electronic format in Appendix C in a CD for the further interests since the depth of tree is 12. Constructed tree structure using only SNPs is attached to Appendix D in top-down tree manner.

The predictive decision nodes for “Cases” are given in the table in Appendix E in IF-THEN format with their probabilities.

4.3.1.2 C4.5 with Genetic Markers (Representative SNPs) and Clinical Data

Not only statistically and biologically significant SNPs subset but also clinical and demographic data of individuals were used as input data. The SNPs subset was identified as mentioned before using AHP scoring. In addition to SNPs which constructed the tree in the previous analyze, we targeted to add some clinical features identified in Table 4.1. Under this condition, we can easily benchmark whether clinical data contribute to the prediction of Alzheimer’s disease or not. We applied C4.5 algorithm to dataset with its 958 independent variables considering the selected genotype data, 8 clinical and 2 demographic features in order to predict whether a person is case or control.

Gain ratio was used to select best single feature that divide dataset into partitions. The minimal gain ratio is identified as 0,01 since gain ratio greater than 0,01 does not have a role to divide data into subsets. For the efficient and confidential benchmarking, we tried 11 fold cross validation to dataset. Next, we compared results whether addition of clinical attributes contribute to the prediction of the complex disease or not.

Extracting Rules from Decision Tree

We firstly determined validation method which we use in modeling phase. The next step is to apply learning algorithm to training data for generating rules. As the depth of tree increases, they become large and difficult to interpret. In this condition, we gave IF_THEN rules from decision tree. IF_THEN rules make interpretation of decision tree easier [32].

To extract rules, every rule is created for each path from the root node to leaf node using logical operator “AND” within ‘IF’ parts. ‘THEN’ part holds the information of class prediction.

We used Rapid Miner for generating rules. The summary of the decision tree model is given Table 4.4.

Table 4.4 Decision Tree Construction Summary for Representative SNPs and Clinical Data

Learning Algorithm: C4.5
Attribute Selection Criterion: Specifies the used it for selecting attributes. We chose the gain ratio for the criterion term.
Inputs: 958 SNPs chosen based on AHP Scoring + 8 clinical information + 2 demographic information of individuals in the study.
Output: If a person is AD or not AD.
Minimal Gain: Which must be achieved in order to produce a partition = 0,01.
Maximal Depth: The maximum tree depth = 12.
Validation Method:11-fold Cross Validation.
Minimal Size for Split: The minimal size of a node in order to allow a partition = 5.
Minimal Leaf Size: The minimal size of all leaves = 2.
Confidence: Used for the pessimistic error calculation of pruning = 0,25.
Number of Pre-pruning Alternatives: The number of alternative nodes tried when pre pruning would prevent a split = 3.

The tree structure for the prediction of Alzheimer’s disease is given in the electronic format in Appendix C in a CD for the further interests since the depth of tree is 12. Constructed tree structure using SNPs and clinical data attributes is attached to Appendix F in top-down tree manner.

The predictive decision nodes for “Cases” are given in the table in Appendix G in IF-THEN format with their probabilities.

CHAPTER 5

RESULTS

Even though combinations of a variety of genetic and other factors are suspected to complex diseases, the exact cause of AD is still not clear. Hence the ongoing studies to unveil the genetic and molecular basis of AD is very important to understand the risk factors, develop prevention methods, diagnosis tools and new therapies.

In this study we have attempted to construct a decision tree for the diagnosis of late-onset AD based on patients genotyping and clinical data. We have carried out analysis using two different input dataset: the first data set contains only SNPs identified as associated to Alzheimer's disease (AD) after GWAS of GENADA data and the second data set that contains both SNPs and clinical and demographic information of individuals as inputs for the predictive model.

Accuracy rates of the C4.5 models based on the prepruning and postpruning parameters with the use of 11 fold cross validation method are as follows. Table 5.1 and Table 5.2 denote accuracy rates of pruned trees. In order to avoid overfitting, parameters that give the best result in terms of performance of the model is chosen considering the test set accuracy. As mentioned before, training set accuracy increases as the depth of the tree is expanded. In parallel with this expansion, the accuracy of test set also increases until a particular depth; then it starts to decrease due to overfitting of training set. In this point construction should stop.

The Table 5.1 and Figure 5.1 below show the accuracy rates of each tree with different pruning parameters using only representative SNPs data. In each decision

tree, performance evaluation was made over test set that is not seen by learning algorithm for rule (knowledge) extraction. Referring the Table 5.1, we decided to construct a tree with the depth 12 and we tried 3 alternative nodes when prepruning would prevent a split for increasing the performance.

Table 5.1 Accuracy Rates For Pruning Parameters by 11 Fold Cross Validation: Representative SNPs

		Prepruning Alternatives						
		3	**5**	**10**	**15**	**20**	**25**	**30**
Depth of Decision Tree	**2**	**52.64% ± 3.33%**	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%
	3	**51.83% ± 3.14%**	51.83% ± 3.14%	51.76% ± 3.22%	51.76% ± 3.22%	51.76% ± 3.22%	51.76% ± 3.22%	51.76% ± 3.22%
	4	**51.96% ± 2.86%**	51.96% ± 2.86%	51.90% ± 2.92%	51.90% ± 2.92%	51.90% ± 2.92%	51.90% ± 2.92%	51.90% ± 2.92%
	5	53.25% ± 3.14%	**53.32% ± 3.16%**	53.25% ± 3.24%	53.25% ± 3.24%	53.25% ± 3.24%	53.25% ± 3.24%	53.25% ± 3.24%
	6	52.84% ± 4.01%	52.91% ± 3.98%	**52.98% ± 4.10%**	52.98% ± 4.10%	52.98% ± 4.10%	52.98% ± 4.10%	52.98% ± 4.10%
	7	**53.59% ± 4.11%**	53.59% ± 4.08%	53.38% ± 4.22%	53.38% ± 4.22%	53.38% ± 4.22%	53.38% ± 4.22%	53.38% ± 4.22%
	8	54.26% ± 2.37%	**54.39% ± 2.39%**	54.19% ± 2.78%	54.19% ± 2.78%	54.19% ± 2.78%	54.19% ± 2.78%	54.19% ± 2.78%
	9	54.73% ± 2.28%	**54.86% ± 2.30%**	54.80% ± 2.64%	54.73% ± 2.71%	54.73% ± 2.71%	54.73% ± 2.71%	54.73% ± 2.71%
	10	55.74% ± 1.97%	**55.94% ± 2.02%**	55.88% ± 2.18%	55.88% ± 2.18%	55.88% ± 2.18%	55.88% ± 2.18%	55.88% ± 2.18%
	11	**55.54% ± 1.72%**	54.87% ± 2.12%	54.87% ± 2.39%	55.07% ± 2.25%	55.00% ± 2.26%	55.00% ± 2.26%	55.00% ± 2.26%
	12	**56.08% ± 1.96%**	55.88% ± 1.91%	55.74% ± 2.23%	55.88% ± 2.05%	55.81% ± 2.02%	55.81% ± 2.02%	55.81% ± 2.02%
	13	**55.81% ± 2.18%**	55.54% ± 2.11%	55.41% ± 2.26%	55.54% ± 2.31%	55.47% ± 2.28%	55.47% ± 2.28%	55.47% ± 2.28%

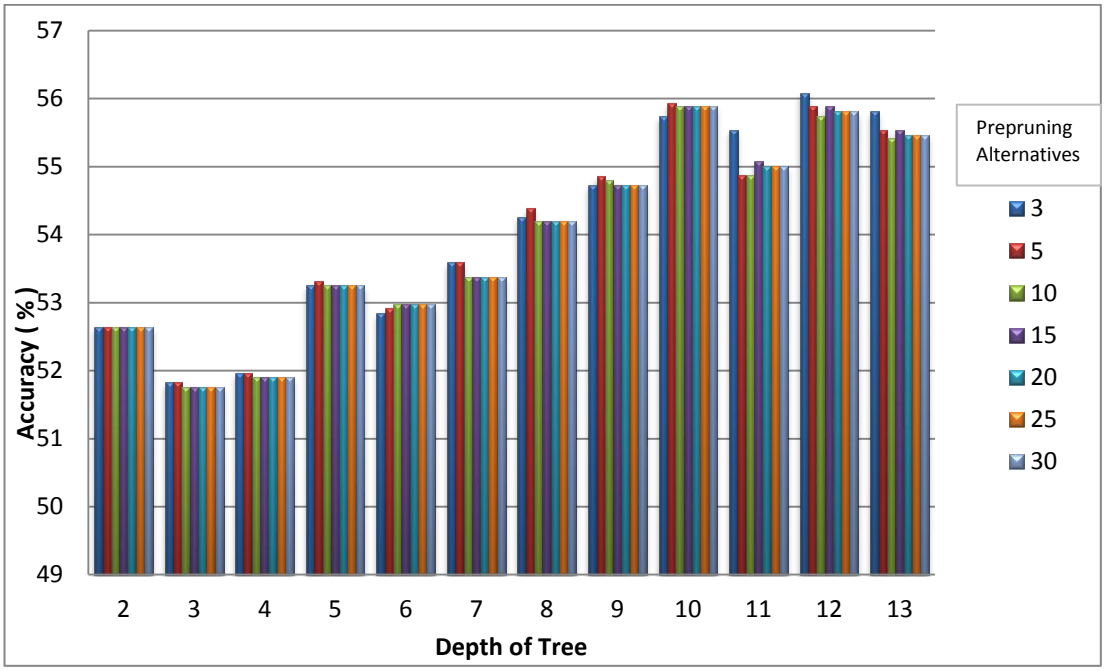

Figure 5.1 Graphical Presentation of Accuracy Rates obtained from Representative SNPs

The Table 5.2 and Figure 5.2 give information about the accuracy rates of each tree with different pruning parameters using representative SNPs data and clinical information of individuals. Referring the Table 5.2 we decided to construct a tree with the depth 12 and we tried 3 alternative nodes when prepruning would prevent a split for increasing the performance.

Table 5.2 Accuracy Rates For Pruning Parameters by 11 Fold Cross Validation: Representative SNPs and Clinical Data

		Prepruning Alternatives						
		3	**5**	**10**	**15**	**20**	**25**	**30**
Depth of Decision Tree	**2**	**52.64% ± 3.33%**	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%	52.64% ± 3.33%
	3	**52.24% ± 3.06%**	52.24% ± 3.06%	52.24% ± 3.06%	52.17% ± 3.16%	52.17% ± 3.16%	52.17% ± 3.16%	52.17% ± 3.16%
	4	**52.30% ± 3.05%**	52.30% ± 3.05%	52.30% ± 3.05%	52.24% ± 3.11%	52.24% ± 3.11%	52.24% ± 3.11%	52.17% ± 3.10
	5	**53.65% ± 3.08%**	53.59% ± 3.11%	53.65% ± 3.12%	53.59% ± 3.21%	53.59% ± 3.21%	53.59% ± 3.21%	53.52% ± 3.25%
	6	52.98% ± 3.63%	52.91% ± 3.64%	**53.04% ± 3.55%**	53.04% ± 3.70%	53.04% ± 3.70%	53.04% ± 3.70%	52.98% ± 3.72%
	7	**53.25% ± 3.93%**	53.18% ± 3.94%	53.18% ± 3.90%	53.11% ± 4.06%	53.11% ± 4.06%	53.11% ± 4.06%	53.04% ± 4.09%
	8	**54.06% ± 2.98%**	53.85% ± 3.00%	53.78% ± 3.13%	53.72% ± 3.34%	53.72% ± 3.34%	53.72% ± 3.34%	53.65% ± 3.39%
	9	**54.39% ± 3.08%**	54.19% ± 3.06%	54.06% ± 3.27%	53.85% ± 3.55%	53.85% ± 3.55	53.85% ± 3.55	53.78% ± 3.58%
	10	**54.87% ± 2.51%**	54.59% ± 2.48%	54.46% ± 2.78%	54.32% ± 3.07%	54.32% ± 3.07%	54.32% ± 3.07%	54.25% ± 3.12%
	11	**54.93% ± 2.52%**	53.92% ± 2.45%	54.06% ± 2.85%	53.99% ± 3.09%	53.99% ± 3.09%	53.99% ± 3.09	53.92% ± 3.14%
	12	**55.07% ± 2.49%**	54.26% ± 2.22%	54.32% ± 2.85%	54.12% ± 3.04%	54.12% ± 3.04%	54.12% ± 3.04%	54.05% ± 3.11%
	13	**54.53% ± 2.25%**	53.99% ± 1.66%	54.12% ± 2.35%	53.85% ± 2.75%	53.85% ± 2.75%	53.85% ± 2.75%	53.78% ± 2.80%

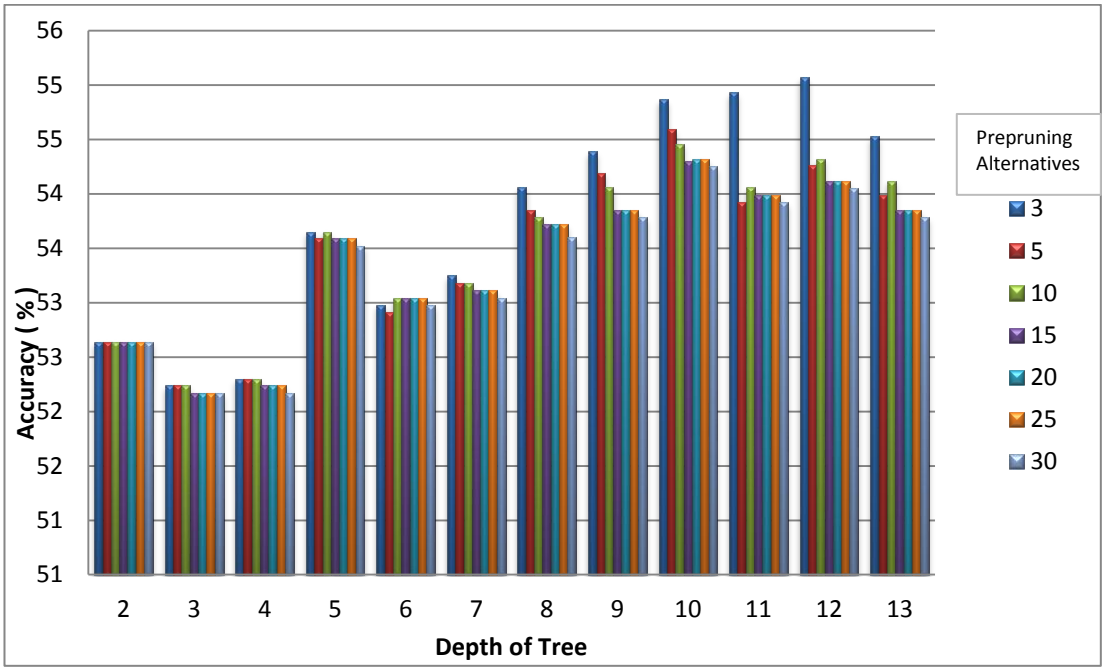

Figure 5.2 Graphical Presentation of Accuracy Rates obtained from Representative SNPs and Clinical Data

As far as tables represents, clinical information does not have a serious effect on the prediction of Alzheimer's disease when we compare test set accuracy results. Concisely, same pruning parameters and same validation strategy are used for both tree constructions. While the highest accuracy rate for decision tree constructed using only representative SNPs is 56,08%, the highest accuracy for the tree constructed using SNPs and clinical data is calculated as 55,07% from test set.

According to decision tree using only SNPs data, the Figure 5.3 shows the distribution of SNPs among chromosomes chosen based on the attribute selection criteria within biologically related top 958 SNPs. We can say that genetic variations that cause Alzheimer's disease are accumulated on chromosomes 1., 2., 6. and 9. in general.

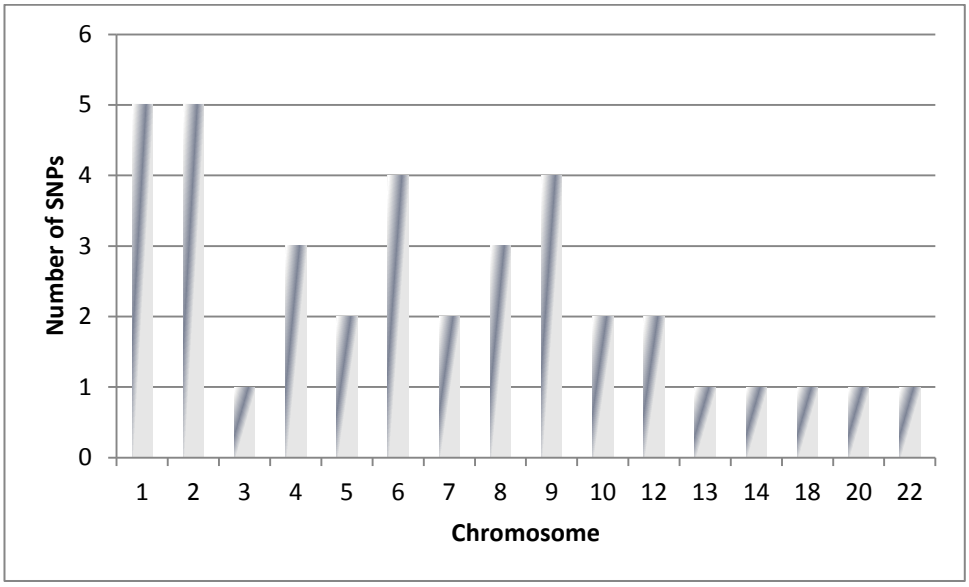

Figure 5.3 Chromosomal Distribution of Decision Tree SNPs

In addition to SNPs data, use of clinical information while constructing the tree, the Figure 5.4 shows the distribution of SNPs among chromosomes chosen based on the attribute selection criteria within biologically and statistically significant top 958 SNPs. We can say that genetic variations that are associated withAlzheimer's disease are accumulated in chromosomes 1., 6. and 9. in general.

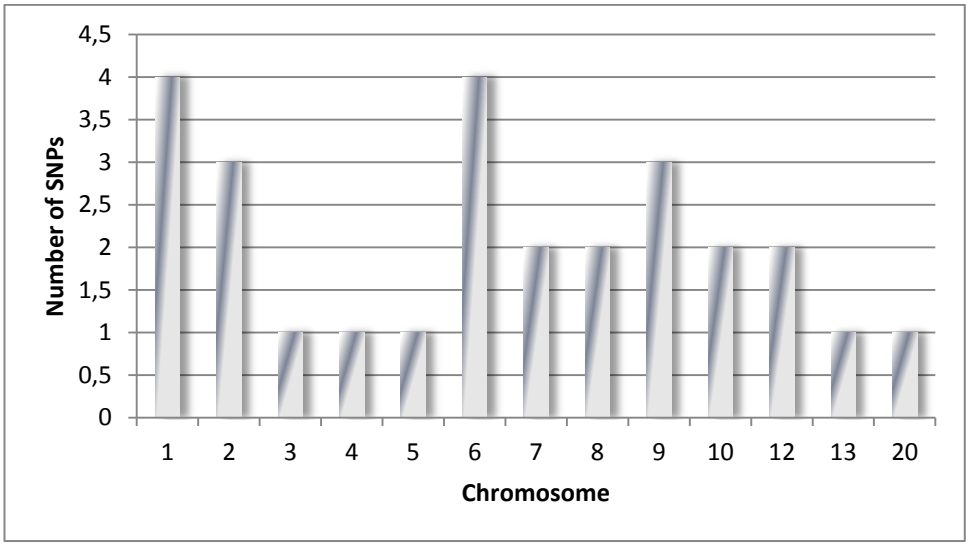

Figure 5.4 Chromosomal Distribution of Decision Tree SNPs

In the first decision tree (Figure 5.5), we used top 958 SNPs related to AD selected based on AHP scoring. Based on attribute selection criterion, 38 SNPs were chosen for the prediction of AD. The information such as reference SNP alleles, minor allele frequency, minor allele count, chromosomal position related to SNPs chosen as splitting attributes in decision tree is listed in Appendix H. Based on these 38 SNPs, decision tree generated 26 rules to predict the disease in humans. The tree paths show SNP combinations and interactions which give information in disease prediction. The location of each SNP was given in the Appendix E. According to findings, genetic variations located on genes such as *ABCC4, ANGPT2, ANGPT2, ARHGAP26, ATG5, C9orf3, DBT, DDO, DISC1, ENPP6, FGD4, FMNL2, FOXO3, GABBR2, GSN, KCNN3, KIF26B, LIPH, MAML3, NBN, PDZD8, PLCB1, PTPRM, SEMA3C, SEMA3C, SEMA5A, SLC35A3, SNW1, SYN3, TLL2, TPO, TRHDE, C9orf3, CAMKK2, DOK1, HMGA1, PIKFYVE, STK39*. Furthermore, rules generated from decision tree algorithm points out which of SNPs subsets is more predominant in the etiology of the disease. In other words, variations in some candidate genes together can help to determine patients' risk to have AD.

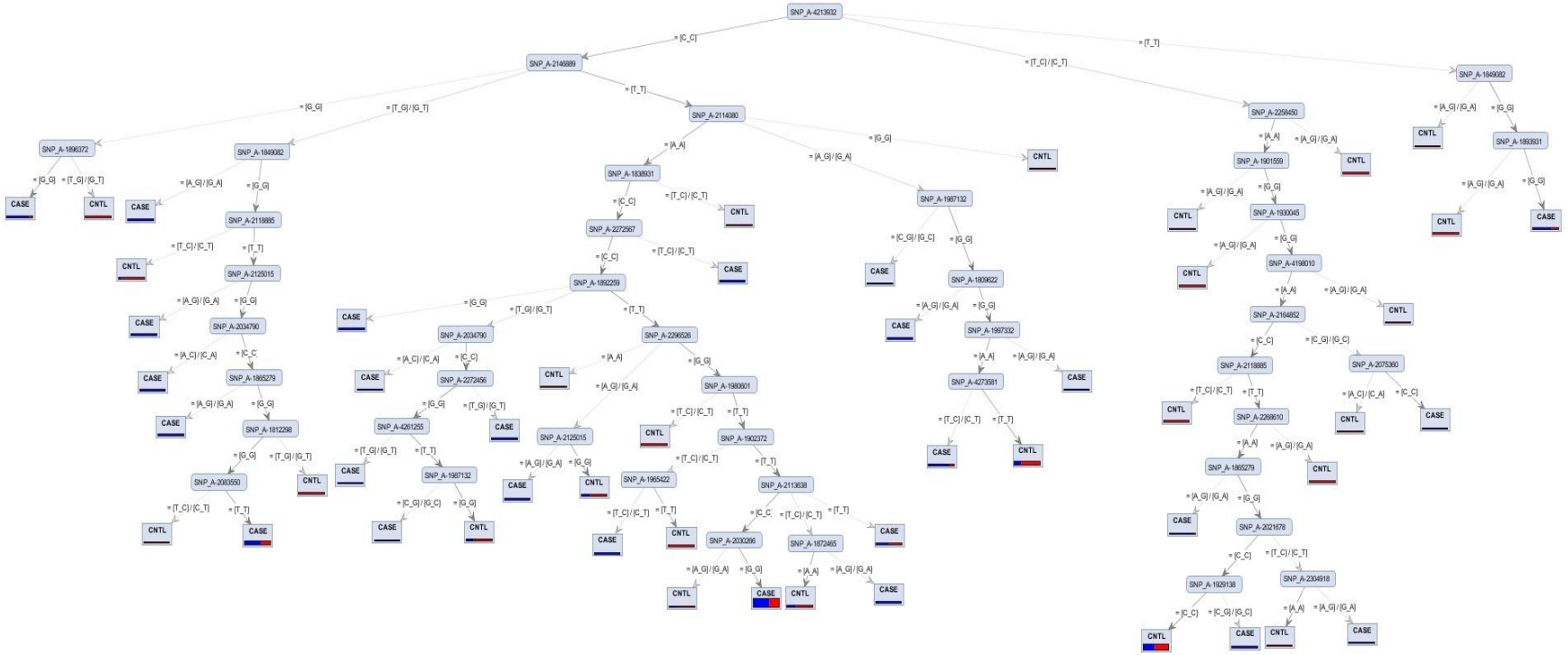

Figure 5.5 Visualization of Decision tree for Representative SNPs

Decision tree obtained using only representative SNPs is presented in Figure 5.5. According to the results, the most significant genetic variations are SNP_A4213932, SNP_A2146889, SNP_A2258450, SNP_A1849082 when we consider the top levels of the tree. For more detailed information, the electronic format of decision tree can be viewed through any web browsers (e.g. Internet Explorer) or flash player.

When we include the relevant clinical data in addition to SNP data, attribute selection criterion chose 27 SNPs out of 958 SNPs and clinical features such as HBA1C_PCT, Body Mass İndex, Hemoglobin, WBC, Trig, Chol and HDL. These clinical features decreased the number of predictive SNPs included in the decision tree (Figure 5.6). Hence, in the presence of the clinical features of individuals, less SNPs were identified in the decision tree for the prediction of disease. 27 SNPs in candidate genes were determined and based on splitting criterion, there are 29 rules generated from the decision tree for the prediction of disease. The information such as reference SNP alleles, minor allele frequency, minor allele count, chromosomal position related to SNPs chosen as splitting attributes in decision tree is given in Appendix J. Genetic variations in genes which are also given in Appendix G are including *ABCC4, ANGPT2, ATG5, C9orf3, DBT, DDO, DISC1, ENPP6, FMNL2, FOXO3, GABBR2, GSN, HMGA1, KCNN3, KIF26B, LIPH, LUM, NBN, PDZD8, PLCB1, SEMA3C, SEMA3C, SEMA5A, STK39, TLL2, TPO, TRHDE.*

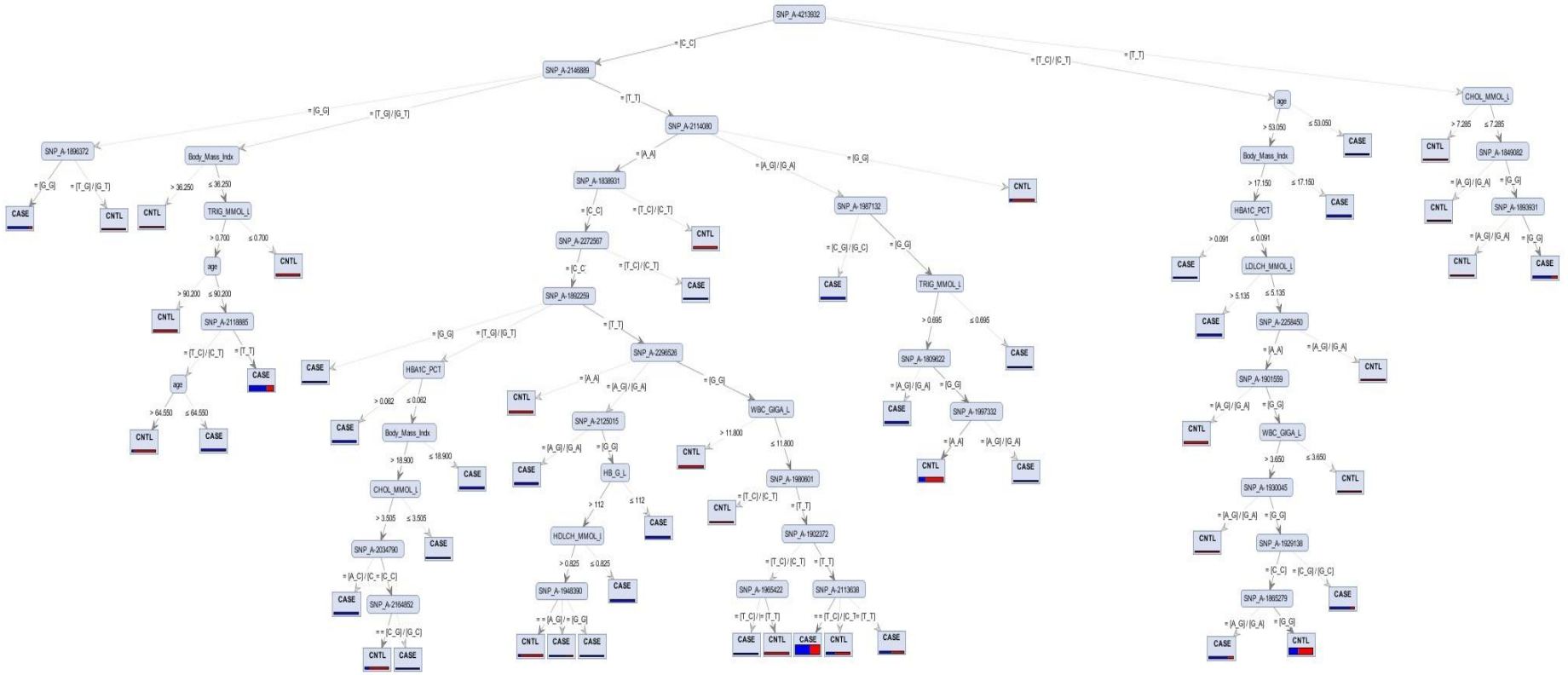

Figure 5.6 Visualization of Decision tree for Representative SNPs and Clinical Information

Decision tree obtained using representative SNPs and clinical information of individuals is given in Figure 5.6. According to tree, the most significant genetic and clinical variations are SNP_A4213932, SNP_A2146889, age and cholesterol (mmol) when we have a look at the first two level of the tree. For more detailed information, the electronic format of decision tree can be viewed through any web browsers (eg Internet Explorer) or flash player.

The performance of classifiers for both decision trees is calculated from the confusion matrix. The first decision tree that inputs only representative SNPs classifies test samples accurately at 56,08% using 11 fold cross validation strategy. According to Table 5.3, classifier classifies 428 cases and 402 controls correctly out of 1480 people.

Table 5.3 Confusion Matrix For Only Representative SNPs

	True CASE	**True CNTL**	**class precision**
Prediction. CASE	428	334	**56,17%**
Prediction. CNTL	316	402	**55,99%**
class recall	***57,53%***	***54,62%***	

On the other hand, decision tree that inputs representative SNPs and clinical information classifies test samples at 55,07% using 11 fold cross validation strategy. According to Table 5.4, classifier classifies 461 cases and 354 controls correctly out of 1480 people.

Table 5.4 Confusion Matrix for Representative SNPs and Clinical Data

	True CASE	**True CNTL**	**class precision**
Prediction. CASE	461	382	**55,69%**
Prediction. CNTL	283	354	**55,57%**
class recall	***61,96%***	***48,10%***	

As a result, integration of clinical information didn't improve the accuracy rate of the model. Hence, only using genotype data, Alzheimer's disease can be predicted for the new data samples.

CHAPTER 6

CONCLUSIONS and FUTURE WORK

6.1 Discussion

Here, we have revealed the AD associated SNPs with decision tree approach and also investigated whether use of clinical information increases the prediction accuracy or not for the late-onset AD for GENADA Study.

The link between AD and genes on four chromosomes, 1, 14, 19, and 21 have been described in previous studies so far. Best known association among these genes is the APOE gene on chromosome 19 that has a role in the molecular etiology of late-onset AD. The SNP variations rs10524523, rs429358, rs7412, rs4420638 located on APOE that are show to be associated with late onset AD patients in three different studies [52], [53]. As the Affymetrix 500K Set comprises Mapping250K_Nsp and Mapping250K_Sty Arrays platform, which used for the genotyping of subjects in the GENADA study does not include these specific SNPs we cannot validate these results in our study.

Even though APOE's relation to AD is well established, it is also known that it can only account for a very minor percent of the genetic of AD. Supporting this observation when we have reviewed the SNPs mapping to the APOE gene considering their overall p-value association the APOE gene was ranked at 2291 with a combined p=0.2 among 9731 that we can collect information in this study. As the overall statistical analysis of APOE gene and SNPs located in that region didn't show strong association it was not included in the further analysis. There are many other

factors for developing AD, and the additional genes that may play a role in the etiology of AD is still under investigation.

On the other hand, some cases of early-onset Alzheimer disease are caused by genetic variation recently identified on chromosomes 1 and 14. Researchers have emphasized that this form of the disorder can result from genetic variations in one of three genes: APP (Amyloid Precursor Protein) on chromosome 21, PSEN1 (Presenilin-1) on chromosome 14, or PSEN2 (Presenilin-2) on chromosome 1. This is because when any of these gene structures is changed, extensive amounts of a toxic protein fragment called amyloid beta peptide are produced in the brain. This peptide can build up in the brain to create clumps that are called amyloid plaques, which are the indicators of Alzheimer disease. A buildup of toxic amyloid beta peptide and amyloid plaques conduce to the death of nerve cells in the brain and lead to the progressive symptoms of Alzheimer's disease [54]. Just a few researchers consider that the investigation for Alzheimer's genes must be ended. Almost all researchers claims that there are more genes involved in Alzheimer's disease. Moreover they are convinced that other conditions must also be taken into account for the disease to develop. The linked genes and SNPs that found so far related to Alzheimer's disease are given in the Appendix G.

From the previous studies, some risky genes and SNPs located on these genes were identified. While expanding the list of associated SNPs our study furthermore revealed linkages between SNPs and their allele variants linked to AD risk. Consequently, based on decision tree rule inductions, we can show the SNPs pathway that lead to AD diagnosis.

Up to date, many studies on the genetics of AD have been represented in Alzgene website[25]. The site summarizes the each result from meta-analysis studies. Apart from APOE that is highly correlated and accepted, the previously published GWA studies do not provide consistency in the candidate genes or location of the genetic variations being associated with the disease [54].

[25] http://www.alzgene.org/largescale.asp

In clinical settings, current diagnosis methods for Alzheimer Disease have a high cost with low accuracy. On the other hand, the tecniques available for the diagnosis of the AD is not easy for the patient or the specialists. Considering the difficulties of diagnostic methods, implementation of decision tree using genotypic information of individuals can support distinguishing the AD patients from dementia in clinic since the accuracy rate of classification is quite acceptable with 56,08%.

This study is the debut of a fresh approach in classification of AD patients. With the guidance of the study, other data mining methods can be implemented to higher dimensional genome databases in order to extract novel and important patterns for predicting any complex diseases.

In this study, we have constructed two prediction models. The first one inputs only significant SNPs and the second one inputs clinical information in addition to significant SNPs information. The aim of the use of clinical information is to show whether clinical information contributes to the prediction of Alzheimer's disease. In contrast to the expectations, clinical information was not a good predictor; hence the accuracy rate of model did not improve. Still with the guidance of the results, we can still infer that some significant clinical parameters, that can be selected by gain ratio calculation, regulates clinical indicators such as cholesterol, and HLA. In other words, some SNPs that were found to be associated with AD, which are already used to build the genotyping based decision tree, have a role in the lipid metabolism and effects levels of cholesterol, LDL or HDL. So, the information based on SNP genotyping already reflects the relation of the clinical parameters with the AD, thus clinical information can not add any further gain on the decision.

6.2 Conclusion

Medical diagnosis does not begin until the patient has visited doctor with various symptoms for many complex diseases. By this time it may be late in the natural history of the diseases. The most common diagnosis of Alzheimer's disease is a through examination that includes complete medical and psychiatric history, a

neurological exam, laboratory tests to rule out anemia, vitamin deficiencies, and other conditions, a mental status exam to evaluate the person's thinking and memory and talking with family members or caregivers [55]. The alternative disease recognition method is the use of genetic variations, which is becoming easier and cheaper way to predict the diseases.

Information obtained from the human genome alterations (SNPs) changes the progress of clinical practice. It provides us to understand of disease mechanisms. Use of genetic information will allow early and more accurate prediction of disease progression.

The disease prediction model will contribute to enhanced clinical practice for many complex diseases. Converting this knowledge into daily applied diagnostics will be a challenge in health domain. Thus, constructing a model creates a new approach to clinical practice with many benefits or patients.

This thesis presents a disease computational study of Alzheimer’s disease based on DNA molecule information provided by SNPs. We applied decision tree methodology to predict AD in a specific medical domain. In order to obtain more informative data, the disease associated genetic variations were used as input data. Firstly, we found statistically significant SNPs and calculated correlations of each SNPs. Secondly, using p values and biological information such as functional effects of SNPs were used to prioritize SNPs and form a base for selecting a subset of SNPs. For this purpose, scoring and prioritizing of SNPs Analytic Hierarchy Process (AHP) was used for the dimension reduction as a preprocessing step. In modeling phase, attribute selection criterion identified the predictors. We constructed two decision trees. The first one was constructed using 958 SNPs and the second one was constructed using not only 958 SNPs but also clinical information of individuals. In the first tree, 38 SNPs was chosen by attribute selection criterion. On the other hand, clinical information of individuals are important in decision making in the second tree, so apart from clinical features 27 SNPs was chosen by the selection criterion.

The results show that clinical information does not play much role in determining the disease when we benchmark the accuracy rates of each tree. What we mean is that, decision tree build with only SNPs information has an accuracy rate of 56,08%. On the other hand, decision tree constructed with clinical data and SNPs data has an accuracy rate of 55,07%. In this study we have shown that diagnosis of AD can be done only based on genotyping information. This also supports a strong genetic basis for the development of late onset AD.

6.3 Future Work

Decision tree algorithm is one of the widely used classification method in data mining. We have applied the C4.5 algorithm for 744 cases and 746 controls. However, this study has some limitations: First, our study needs more data containing different populations to obtain more reliable results. Hence, more financial support is necessary to obtain SNP data of individuals. As the information of more genetic variation becomes accessible, the method probably improves prediction performance. Second, Naïve Bayesian Classification and Artificial Neural Network approaches can be applied for the same purpose and the performance indicators can be benchmarked in order to select the most accurate prediction model.

In the context of determining the complex diseases, we are planning to enhance the classification algorithm to accommodate candidate factors other than SNPs, such as race or other environmental variables that may affect susceptibility to disease, as such required data become available.

Additionally after a decision support algorithm is developed based on the genotyping and clinical information hopefully with much higher prediction power, the system should be tested in a pilot study on patients whom need differential diagnosis between AD and dementia. Doctors and patients will benefit from the outcomes of such research only if the prediction models suggested are validated and implemented as decision support systems for clinical use.

REFERENCES

[1] A. Bar-or, A. Schuster, R. Wolff, and D. Keren, "Decision Tree Induction in High Dimensional, Hierarchically Distributed Databases," in *Proceedings of SDM'05 Newport Beach*, 2005.

[2] K. Raza, "Application of Data Mining in Bioinformatics," *Indian Journal of Computer Science and Engineering*, vol. 1, no. 2, pp. 114–118, 2009.

[3] L. Fiaschi, "A Framework for the Application of Decision Trees to the Analysis of SNPs Data," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2009, pp. 106–113.

[4] U. D. B. Nazlıel, "Alzheimer Hastalığı ve Genetik," 1999, pp. 45–51.

[5] "The Structures of DNA and RNA." [Online]. Available: http://biology.kenyon.edu/courses/biol63/watson_06.pdf. [Accessed: 11-Sep-2012].

[6] E. Balcan, "DNA Yapısı." [Online]. Available: http://www2.bayar.edu.tr/erdal.balcan/Ders2 ve 3 Nukleik Asitler nukleozom transpozonlar2012.pdf. [Accessed: 11-Sep-2012].

[7] K. Y. Yip, C. Cheng, N. Bhardwaj, J. B. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder, and M. Gerstein, "Classification of Human Genomic Regions Based on Experimentally Determined Binding Sites of More Than 100 Transcription-related Factors.," *Genome Biology*, vol. 13, no. 9, p. R48, Sep. 2012.

[8] I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. a. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, and R. Kaul, "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[9] P. Sharp and R. Roberts, "The Complexity of Eukaryotic Genomes." 1977.

[10] I. B. S. Resource, "mRNA Codon / Amino Acid Chart," *Cirriculum Framework/How do mutations affect living organisms?,How does protein synthesis occur?*, 2003. [Online]. Available: http://www.leydenscience.org/mfumagalli/Honors Biology/mRNA codon chart.pdf. [Accessed: 11-Sep-2012].

[11] B. Shastry, “SNP Alleles in Human Disease and Evolution,” *Journal of Human Genetics*, vol. 47, no. 11, pp. 561–6, Jan. 2002.

[12] *Human Genetic Variation*. National Human Genome Research Institute, 2007, p. 126.

[13] “Genetic Variation.” [Online]. Available: https://www.migeneticsconnection.org/genomics/Genetic Variation/GeneticVariation.htm. [Accessed: 11-Sep-2012].

[14] M. Marinus, “Mutation,” in *Molecular Genetics of Bacteria*, 2. ed., 2003, pp. 1–7.

[15] B. Debele, K. Ege, and B. A. Dal, “Mutasyon , DNA Hasarı ,Onarım Mekanizmaları ve Kanserle İlişkisi,” *J. Fac. Pharm, Ankara*, vol. 35, no. 2, pp. 149–170, 2006.

[16] S. Nakken, I. Alseth, and T. Rognes, “Computational Prediction of the Effects of non-Synonymous Single Nucleotide Polymorphisms in Human DNA Repair Genes.,” *Neuroscience*, vol. 145, no. 4, pp. 1273–9, Apr. 2007.

[17] G. Üskünkar, “An Integrative Approach To Structured SNP Prioritization and Representative SNP Selection For Genome-wide Association Studies:Algorithms and Systems,” Middle East Technical University, 2008.

[18] W. Africa, “A Map of Human Genome Variation From Population-Scale Sequencing.,” *Nature*, vol. 467, no. 7319, pp. 1061–73, Oct. 2010.

[19] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn, “Meta-Analysis of Genetic Association Studies Supports a Contribution of Common Variants to Susceptibility to Common Disease.,” *Nature Genetics*, vol. 33, no. 2, pp. 177–82, Feb. 2003.

[20] “A History of Alzheimer Disease.” [Online]. Available: www.ahaf.org/alzheimers/about/understanding/history.html. [Accessed: 10-Jun-2012].

[21] E. E. Tripoliti, D. I. Fotiadis, M. Argyropoulou, and G. Manis, “A Six Stage Approach for the Diagnosis of the Alzheimer’s Disease Based on fMRI Data.,” *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 307–20, Apr. 2010.

[22] NIH, “Alzheimer’s Disease Genetics,” 2008.

[23] J. Shasha Wang, M. Zaki, T. Hannu, and D. Shasha, *Data Mining in Bioinformatics*. 2005.

[24] J. Y. Chen and S. Lonardi, Eds., “Mining Patterns of Epistasis in Human Genome,” in *Biological Data Mining*, 1st ed., pp. 187–200.

[25] S. Özekes, “Veri Madenciliği Modelleri ve Uygulama Alanları,” *İstanbul Ticaret Üniversitesi Dergisi*, pp. 65–82.

[26] S. Yıldırım, “Tümevarım Öğrenme Tekniklerinden C4.5’in İncelenmesi,” İstanbul Teknik Üniversitesi, 2003.

[27] W. Peng, J. Chen, and H. Zhou, “An Implementation of ID3 - Decision Tree Learning Algorithm,” *From web. arch. usyd. edu. au/wpeng/ …*. Sydney, Australia, 2009.

[28] J. R. Quinlan, “Learning Decision Tree Classifiers,” vol. 28, no. 1, pp. 71–72, 1996.

[29] “The CHAID Analysis,” 2007. [Online]. Available: http://www.smres.com/CHAIDAnalysis.pdf. [Accessed: 06-Feb-2012].

[30] R. Hoare, “Using CHAID for Classification Problems,” *New Zealand Statistical Association 2004 conference*, 2004.

[31] R. Kohavi and R. Quinlan, “Decision Tree Discovery,” *Citeseer*, vol. 3, 1999.

[32] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second Edi. 2006, p. 743.

[33] I. Jenhani, N. B. Amor, and Z. Elouedi, “Decision Trees as Possibilistic Classifiers,” *International Journal of Approximate Reasoning*, vol. 48, no. 3, pp. 784–807, Aug. 2008.

[34] J. D. Hand, H. Mannila, and Smyth Padhraic, “Decision Tree Induction: Using Entropy for Attribute Selection 4.1,” in *Principles of Data Mining*, 2001, pp. 51–64.

[35] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection,” *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271–277, 2010.

[36] A. S. Al-Hegami, “Classical and Incremental Classification in Data Mining Process,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 7, no. 12, pp. 179–187, 2007.

[37] T. Mitchell and M. Hill, “Decision Tree Learning,” in *Machine Learning*, 1997.

[38] A. Pfeffer and D. Parkes, “CS181 Lecture 3 — Overfitting , Description Length and More on the ID3 Algorithm,” 2010. [Online]. Available: http://www.seas.harvard.edu/courses/cs181/docs/lecture3-notes.pdf. [Accessed: 05-Sep-2012].

[39] H. Shi, “Best-first Decision Tree Learning,” The University of Waikato, 2007.

[40] J. Listgarten, “Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms,” *Clinical Cancer Research*, vol. 10, no. 8, pp. 2725–2737, Apr. 2004.

[41] J. S. Barnholtz-Sloan, X. Guan, C. Zeigler-Johnson, N. J. Meropol, and T. R. Rebbeck, “Decision Tree-based Modeling of Androgen Pathway Genes and Prostate Cancer Risk,” *American Association for Cancer Research*, vol. 20, no. 6, pp. 1146–55, Jun. 2011.

[42] Y. Jiao, R. Chen, and X. Ke, “Single Nucleotide Polymorphisms Predict Symptom Severity of Autism Spectrum Disorder.,” *Journal of Autism and Developmental Disorders*, Jul. 2011.

[43] L. Huang, S. Hsu, and E. Lin, “Fatigue Syndrome Based on Genetic Data,” *Journal of Translational Medicine*, vol. 7:81, pp. 1–8, 2009.

[44] S. C. Shah and A. Kusiak, “Data Mining and Genetic Algorithm Based Gene/SNP Selection.,” *Artificial Intelligence in Medicine*, vol. 31, no. 3, pp. 183–96, Jul. 2004.

[45] C. M. Freitag, “The Genetics of Autistic Disorders and Its Clinical Relevance: A Review of the Literature,” *Molecular Psychiatry*, vol. 12, no. 1, pp. 2–22, Jan. 2007.

[46] D. H. Geschwind, “Advances in Autism,” *Annual Review of Medicine*, vol. 60, pp. 367–80, Jan. 2009.

[47] R. Rapley and S. Harbron, Eds., “Overview of Microarrays in Genomic Analysis,” in *Molecular Analysis and Genome Discovery*, vol. 5, West Sussex, England: John Wiles & Sons, LTD, 2004.

[48] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, “Data Mining: From Concept to Implementation,” New Jersey, 1998.

[49] O. Marban, G. Mariscal, and J. Segovia, “A Data Mining and Knowledge Discovery Process Model,” in *Data Mining and Knowledge Discovery in Real Life Applications*, no. February, J. Ponce and A. Karahoca, Eds. 2009, pp. 1–17.

[50] S. Purcell, "PLINK (1.07) Documentation," 2010. [Online]. Available: http://pngu.mgh.harvard.edu/~purcell/plink/dist/plink-doc-1.07.pdf. [Accessed: 12-Sep-2012].

[51] "RapidMiner Brochure," *http://rapid-i.com/component/option,com_docman/task,doc_download/gid,46/Itemid,165*, 2008. .

[52] a D. Roses, M. W. Lutz, H. Amrine-Madsen, a M. Saunders, D. G. Crenshaw, S. S. Sundseth, M. J. Huentelman, K. a Welsh-Bohmer, and E. M. Reiman, "A TOMM40 Variable-length Polymorphism Predicts the Age of Late-onset Alzheimer's disease," *The Pharmacogenomics Journal*, vol. 10, no. 5, pp. 375–84, Oct. 2010.

[53] L. Bertram, C. Lange, K. Mullin, M. Parkinson, M. Hsiao, M. F. Hogan, B. M. M. Schjeide, B. Hooli, J. Divito, I. Ionita, H. Jiang, N. Laird, and T. Moscarillo, "Genome-wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in Addition to APOE.," *American Journal of Human Genetics*, vol. 83, no. 5, pp. 623–32, Nov. 2008.

[54] P. Momeni and R. Ferrari, "Genetic and Blood Biomarkers of Alzheimer' s Disease," *The Open Nuclear Medicine Journal*, vol. 2, pp. 12–24, 2010.

[55] "AAN Guideline Summary for Patients and Their Families Alzheimer's Disease," *American Academy of Neurology*, Montreal, p. 2.

GLOSSARY

Allele—A form of a gene. Each person receives two alleles of a gene, one from each biological parent. This combination is one factor among many that influence a variety of processes in the body. On chromosome 19, the apolipoprotein E (APOE) gene has three common forms or alleles: e2, e3 and e4.

Apolipoprotein E (APOE) gene—A gene on chromosome 19 involved in making a protein that helps carry cholesterol and other types of fat in the bloodstream. The APOE e4 allele is considered a risk-factor gene for Alzheimer's disease and appears to influence the age at which the disease begins.

Chromosome—A compact structure containing DNA and proteins present in nearly all cells of the body. Chromosomes carry genes, which direct the cell to make proteins and direct a cell's construction, operation, and repair. Normally, each cell has 46 chromosomes in 23 pairs. Each biological parent contributes one of each pair of chromosomes.

DNA (deoxyribonucleic acid)—The hereditary material in humans and almost all other organisms. Almost all cells in a person's body have the same DNA. Most DNA is located in the cell nucleus.

Gene—A basic unit of heredity. Genes direct cells to make proteins and guide almost every aspect of cells' construction, operation, and repair.

Genetic mutation—A permanent change in a gene that can be passed on to children. The rare, early-onset familial form of Alzheimer's disease is associated with mutations in genes on chromosomes 1, 14, or 21.

Genetic risk factor—A change in a gene that increases a person's risk of developing a disease.

Genetic variant—A change in a gene that may increase or decrease a person's risk of developing a disease or condition.

Genome-wide association study (GWAS)—A study approach that involves rapidly scanning complete sets of DNA, or genomes, of many individuals to find genetic variations associated with a particular disease.

Hippocampus—Hippocampus is a major component of the brain of human or other mammals. It belongs to the limbic system and plays a significant role in long-term memory. In AD, it is one of the first regions of brain to suffer damage, memory problems appear.

Protein—A substance that determines the physical and chemical characteristics of a cell and therefore of an organism. Proteins are essential to all cell functions and are created using genetic information.

APPENDICES

APPENDIX A - SNPs Related to Alzheimer's Disease

The tables present SNPs related to the AD obtained from previous association studies[26].

RS ID/Chromosome	Related Gene
rs1415985 at chr1	in AGBL4
rs11205641 at chr1	in AGBL4
rs4926831 at chr1	in AGBL4
rs9659092 at chr1	in AGBL4
rs11583200 at chr1	n ELAVL4
rs12725861 at chr1	
rs6428503 at chr1	in GBP2
rs10922573 at chr1	in GBP2
rs7537752 at chr1	in CSF1
rs12044355 at chr1	in DISC1
rs11683103 at chr2	
rs2119067 at chr2	in SCN2A, SCN2A2
rs10184275 at chr2	in SCN2A, SCN2A2
rs2681411 at chr3	in CD86
rs3846421 at chr4	in SORCS2
rs12639920 at chr4	in ATP8A1
rs1425967 at chr4	
rs4416533 at chr4	
rs12514426 at chr5	in WWC1
rs179943 at chr6	in ATXN1
rs3807031 at chr6	in NCRNA00171, PPP1R11, ZNRD1, ZNRD1AS
rs929156 at chr6	in TRIM15
rs13213247 at chr6	
rs11754661 at chr6	in MTHFD1L
rs9455973 at chr6	
rs6942930 at chr7	in INTS1
rs2039461 at chr9	

[26] http://www.pharmgkb.org/disease/PA443319?tabType=tabGenetics#tabview=tab0&subtab=

RS ID/Chromosome	Related Gene
rs7893928 at chr10	
rs16934131 at chr10	in KCNMA1
rs3740057 at chr10	in DNMBP
rs10883421 at chr10	in DNMBP
rs2986017 at chr10	in CALHM1, CALHM2
rs10786828 at chr10	in SORCS3
rs7894737 at chr10	in SORCS3
rs11244841 at chr10	in ADAM12
rs7946599 at chr11	in SORL1
rs2298814 at chr11	in SORL1
rs6589885 at chr11	in SORL1
rs720099 at chr11	in SORL1
rs11218342 at chr11	in SORL1
rs11218343 at chr11	in SORL1
rs1784919 at chr11	in SORL1
rs1792124 at chr11	in SORL1
rs3781835 at chr11	in SORL1
rs3781838 at chr11	in SORL1
rs11610206 at chr12	
rs2387100 at chr13	
rs6313 at chr13	in HTR2A
rs9544105 at chr13	
rs659628 at chr13	in KCTD12
rs12146962 at chr14	
rs11159647 at chr14	
rs4555132 at chr15	
rs1480090 at chr15	
rs1383139 at chr15	
rs5882 at chr16	in CETP, NLRC5
rs11653716 at chr17	in NOS2
rs4343 at chr17	in ACE
rs4351 at chr17	in ACE

RS ID/Chromosome	Related Gene
rs1402627 at chr18	in DLGAP1
rs4459653 at chr19	in ZNF224
rs4802207 at chr19	in ZNF224
rs3746319 at chr19	in ZNF224, ZNF225
rs2061332 at chr19	in ZNF224, ZNF225
rs2061333 at chr19	in ZNF224, ZNF225
rs10524523 at chr19	in APOE, TOMM40
rs429358 at chr19	in APOC1, APOE
rs7412 at chr19	in APOC1, APOE
rs4420638 at chr19	in APOC1, APOC1P1
rs3826656 at chr19	in CD33
rs2180566 at chr20	in DEFB122, DEFB123

The tables present SNPs related to the AD obtained from previous association studies[27].

SNPs on **TRPC4AP** gene on chromosome 20q11.22 seem to be associated with late-onset of Alzheimer disease
Multiple genetic variations in **SORL1** are associated with Alzheimer disease
rs5984894 is linked and associated with increased risk in females
SNPs in the **CLU, CR1** and **PICALM** genes, *rs11136000, rs3818361,* and *rs3851179,* which have been replicated in independent (European-descent) populations
rs10519262, an intergenic SNP on chromosome 15
rs908832 a SNP in **ABCA2**, associated with early-onset AD
rs1050283 in the **OLR1** gene may increase risk for both early-onset and late-onset Alzheimer disease
rs2227564 a SNP in **PAU** gene
rs2333227 in the **MPO** gene, and *rs669* in the **A2M** gene, and possible synergistic interaction between them
rs2373115 is one of several SNPs in the **GAB2** gene that are associated with higher risk of Alzheimer disease
rs2986017 a SNP in the **CALHM1** gene
rs3025786 which can decrease risk slightly among ApoE4 carriers
rs5963409 in the **OTC** gene promoter region
rs10868366, rs7019241, rs9886784 are associated with Alzheimer disease in a study of ~1100 Canadian and UK patients
A SNP in the **PON1** gene
A SNP in intron 9 of the **CHAT** gene
rs4878104 and *rs4877365* in the **DAPK1** gene
SNPs in the **DNMBP** gene
SNPs in the **MME** gene, most notably, *rs1836915*
A SNP in the **TLR4** gene, *rs4986790*, with many disease associations
A SNP in the **BACE1** gene
Numerous SNPs such as *rs4293*(risk allele appears to be A), *rs1799752* in the **ACE** gene are associated with susceptibility to Alzheimer's disease
rs1868402 in the **PPP3R1** gene is associated with progression of dementia

[27] http://snpedia.com/index.php/Alzheimer's_disease

APPENDIX B - RS IDs of Selected SNPs Based on AHP Scoring

Number	RS_ID	AHP SCORE
1	rs7161889	0,717559
2	rs2437357	0,692803
3	rs6494031	0,677721
4	rs2124459	0,677721
5	rs16951252	0,677721
6	rs1619631	0,654962
7	rs839511	0,654374
8	rs8029805	0,654374
9	rs4673644	0,654374
10	rs16969899	0,654374
11	rs7166325	0,654374
12	rs16940651	0,654374
13	rs6494030	0,654374
14	rs9526245	0,654374
15	rs4254542	0,654374
16	rs9651118	0,654374
17	rs7866199	0,631615
18	rs870695	0,631615
19	rs1806760	0,631615
20	rs811925	0,552691
21	rs7736650	0,551519
22	rs1060743	0,550694
23	rs2230721	0,550694
24	rs6937193	0,537417
25	rs7213894	0,529932
26	rs1653586	0,529863
27	rs8041254	0,529546
28	rs6426282	0,529546
29	rs982804	0,529546
30	rs17828120	0,529546
31	rs3793511	0,529546
32	rs7652762	0,529546
33	rs3802428	0,52876

34	rs12029094	0,52876
35	rs11100790	0,527935
36	rs914358	0,522721
37	rs5764698	0,507959
38	rs12327141	0,507104
39	rs4310078	0,506787
40	rs52911	0,506787
41	rs12191369	0,506787
42	rs16896641	0,506787
43	rs1572898	0,506787
44	rs7760666	0,506787
45	rs2747008	0,506787
46	rs6475236	0,506787
47	rs2225193	0,506787
48	rs672901	0,506787
49	rs1053495	0,506787
50	rs3798887	0,506787
51	rs13273088	0,506787
52	rs5764825	0,506787
53	rs6540910	0,506787
54	rs11055616	0,506787
55	rs10832613	0,506787
56	rs17109366	0,506787
57	rs4767550	0,506787
58	rs1662332	0,506787
59	rs17174714	0,506787
60	rs10951201	0,506787
61	rs2225176	0,506787
62	rs2052852	0,506787
63	rs2610201	0,506787
64	rs10790332	0,506194
65	rs7928477	0,506194
66	rs16966703	0,505962
67	rs1126828	0,505962
68	rs17114803	0,505962
69	rs2278812	0,505962
70	rs2820718	0,505962
71	rs796283	0,492561

72	rs2567982	0,492022
73	rs1050045	0,492022
74	rs3732923	0,491705
75	rs2121866	0,491705
76	rs2121867	0,491705
77	rs7950059	0,491705
78	rs2905967	0,491705
79	rs10911111	0,491705
80	rs10503404	0,491705
81	rs1638196	0,491705
82	rs2450130	0,491705
83	rs17154454	0,491705
84	rs1467051	0,491705
85	rs10882332	0,491705
86	rs12485273	0,491705
87	rs1867982	0,491705
88	rs10255956	0,491705
89	rs7141909	0,491705
90	rs10793302	0,491705
91	rs10732447	0,491705
92	rs9329334	0,491705
93	rs12901591	0,491705
94	rs612759	0,491705
95	rs11087626	0,491705
96	rs6596384	0,491705
97	rs3846133	0,491705
98	rs10892434	0,491705
99	rs10238586	0,491705
100	rs7089353	0,491705
101	rs16869739	0,491705
102	rs375241	0,491705
103	rs3776586	0,491705
104	rs341795	0,491705
105	rs4639844	0,491705
106	rs10084692	0,491705
107	rs10235838	0,491705
108	rs7431408	0,491705
109	rs3739390	0,491705

110	rs6793943	0,491705
111	rs8024206	0,491705
112	rs901104	0,491705
113	rs2773502	0,491705
114	rs4685765	0,491705
115	rs934854	0,491705
116	rs4316429	0,491705
117	rs12594561	0,491705
118	rs12592002	0,491705
119	rs2619911	0,491705
120	rs10797791	0,491705
121	rs17067931	0,491705
122	rs584828	0,49088
123	rs1385600	0,49088
124	rs2152183	0,48344
125	rs17114641	0,48344
126	rs882114	0,476229
127	rs4789786	0,476229
128	rs3786507	0,476229
129	rs12202209	0,476229
130	rs460262	0,476229
131	rs6464211	0,475404
132	rs1150782	0,470739
133	rs1381335	0,470588
134	rs4870723	0,470118
135	rs6540547	0,469597
136	rs10024098	0,469263
137	rs10491131	0,469009
138	rs6503633	0,469009
139	rs2252304	0,469009
140	rs10402361	0,469009
141	rs4687319	0,468946
142	rs602668	0,468946
143	rs2742424	0,468946
144	rs7602727	0,468946
145	rs606114	0,468946
146	rs2057116	0,468946
147	rs11742602	0,468946

148	rs17813652	0,468946
149	rs6803572	0,468946
150	rs6687672	0,468946
151	rs7842055	0,468946
152	rs6007009	0,468946
153	rs2038252	0,468946
154	rs2278844	0,468946
155	rs12881652	0,468946
156	rs7029570	0,468946
157	rs17817919	0,468946
158	rs6926560	0,468946
159	rs17214144	0,468946
160	rs17475367	0,468946
161	rs12090877	0,468946
162	rs1587607	0,468946
163	rs4706990	0,468946
164	rs7152571	0,468946
165	rs2172876	0,468946
166	rs2071272	0,468946
167	rs1551762	0,468946
168	rs10873172	0,468946
169	rs2278295	0,468946
170	rs17402830	0,468946
171	rs2034478	0,468946
172	rs788796	0,468946
173	rs12424516	0,468946
174	rs2277326	0,468946
175	rs934750	0,468946
176	rs705770	0,468946
177	rs2306207	0,468946
178	rs9368288	0,468946
179	rs831287	0,468946
180	rs2236345	0,468946
181	rs9513000	0,468946
182	rs7750443	0,468946
183	rs2915765	0,468946
184	rs7569357	0,468946
185	rs3776403	0,468946

186	rs1010788	0,468946
187	rs4513434	0,468946
188	rs2290324	0,468946
189	rs306783	0,468946
190	rs2180090	0,468946
191	rs12368653	0,468946
192	rs757428	0,468946
193	rs10519763	0,468946
194	rs2030533	0,468946
195	rs17408919	0,468946
196	rs7969488	0,468946
197	rs3817627	0,468946
198	rs363301	0,468946
199	rs3171980	0,468946
200	rs3829930	0,468946
201	rs17023934	0,468946
202	rs10150311	0,468946
203	rs10059683	0,468946
204	rs10489965	0,468946
205	rs11102040	0,468946
206	rs6086969	0,468946
207	rs16945692	0,468946
208	rs10495020	0,468946
209	rs136585	0,468946
210	rs4909259	0,468946
211	rs4608591	0,468946
212	rs41480152	0,468946
213	rs1149158	0,468946
214	rs13299632	0,468946
215	rs230966	0,468946
216	rs10489964	0,468946
217	rs9824931	0,468946
218	rs11134233	0,468946
219	rs462907	0,468946
220	rs12241747	0,468946
221	rs1556408	0,468946
222	rs10424282	0,468946
223	rs3793032	0,468946

224	rs962459	0,468946
225	rs7665654	0,468946
226	rs16914041	0,468946
227	rs10775471	0,468946
228	rs9839410	0,468946
229	rs2239764	0,468946
230	rs10931347	0,468946
231	rs468308	0,468946
232	rs217873	0,468946
233	rs12356533	0,468946
234	rs3783215	0,468946
235	rs6725519	0,468946
236	rs12353519	0,468946
237	rs2280391	0,468946
238	rs4380535	0,468946
239	rs1065035	0,468675
240	rs980989	0,468675
241	rs17709552	0,468675
242	rs1010169	0,468675
243	rs17804446	0,468358
244	rs6991221	0,468358
245	rs7004238	0,468358
246	rs9314604	0,468358
247	rs17624022	0,468358
248	rs17377379	0,468358
249	rs687766	0,468358
250	rs633297	0,468358
251	rs3020233	0,468358
252	rs150088	0,468358
253	rs8100239	0,468358
254	rs11061149	0,468358
255	rs3739392	0,468358
256	rs10924471	0,468358
257	rs3797443	0,468358
258	rs9456538	0,468358
259	rs2922894	0,468358
260	rs2922893	0,468358
261	rs7793197	0,468358

262	rs2181624	0,468358
263	rs646228	0,468358
264	rs10911125	0,468358
265	rs2238973	0,468358
266	rs980653	0,468358
267	rs1038919	0,468358
268	rs10050568	0,468358
269	rs1273349	0,468358
270	rs2306021	0,468358
271	rs4529465	0,468358
272	rs17203328	0,468358
273	rs10489298	0,468358
274	rs2784176	0,468358
275	rs1416086	0,468358
276	rs679449	0,468358
277	rs2241780	0,468358
278	rs13026243	0,468358
279	rs7648530	0,468358
280	rs13261597	0,468358
281	rs7012244	0,468358
282	rs1273369	0,468358
283	rs2325788	0,468358
284	rs3807306	0,468358
285	rs4497180	0,468358
286	rs448281	0,468358
287	rs9388981	0,468358
288	rs7204799	0,468358
289	rs9557765	0,468358
290	rs11810899	0,468358
291	rs754107	0,468358
292	rs11127582	0,468358
293	rs7990870	0,468358
294	rs17642119	0,468358
295	rs2421987	0,468358
296	rs10519864	0,468358
297	rs4683990	0,468358
298	rs1570355	0,468358
299	rs11562973	0,468358

300	rs12406164	0,468358
301	rs1441951	0,468358
302	rs6882032	0,468358
303	rs7241781	0,468358
304	rs706120	0,468358
305	rs4695256	0,468358
306	rs2288693	0,468358
307	rs7839119	0,468358
308	rs4143055	0,468358
309	rs8004481	0,468358
310	rs6001585	0,468358
311	rs10142154	0,468358
312	rs7526860	0,468358
313	rs4238137	0,468358
314	rs2985340	0,468358
315	rs2227607	0,468358
316	rs552191	0,468358
317	rs17101017	0,468358
318	rs17522183	0,468358
319	rs703261	0,468358
320	rs1108923	0,468358
321	rs11211654	0,468358
322	rs10797719	0,468358
323	rs1003873	0,468358
324	rs3791624	0,468358
325	rs1533469	0,468358
326	rs10910966	0,468358
327	rs1007837	0,468358
328	rs1535505	0,468358
329	rs10749832	0,468358
330	rs7179228	0,468358
331	rs1573045	0,468358
332	rs1544622	0,468358
333	rs373759	0,468358
334	rs17581311	0,468358
335	rs7945395	0,468358
336	rs7115850	0,468358
337	rs2833423	0,468358

338	rs11729081	0,468358
339	rs4742006	0,468358
340	rs2887202	0,468358
341	rs12902857	0,468358
342	rs10515489	0,468358
343	rs4742008	0,468358
344	rs2510038	0,468358
345	rs4589663	0,468358
346	rs1476359	0,468358
347	rs712839	0,468358
348	rs2759251	0,468358
349	rs761222	0,468358
350	rs9827237	0,468358
351	rs2378991	0,468358
352	rs638859	0,468358
353	rs11188342	0,468358
354	rs12592370	0,468358
355	rs2511170	0,468358
356	rs319751	0,468358
357	rs2373115	0,468358
358	rs9474576	0,468358
359	rs662999	0,468358
360	rs16917171	0,468358
361	rs6474387	0,468358
362	rs1466998	0,468358
363	rs9906088	0,468358
364	rs10860787	0,468358
365	rs17776503	0,468358
366	rs7988271	0,468358
367	rs2833426	0,468358
368	rs16951777	0,468358
369	rs8131958	0,468358
370	rs41498044	0,468358
371	rs16980706	0,468358
372	rs10491731	0,468358
373	rs1295741	0,468358
374	rs6850108	0,468358
375	rs12462609	0,468358

376	rs7280029	0,468358
377	rs9457252	0,468358
378	rs6981002	0,468358
379	rs8105903	0,468358
380	rs10487888	0,468358
381	rs9886720	0,468358
382	rs10974624	0,468358
383	rs1508411	0,468358
384	rs4901047	0,468358
385	rs535112	0,468358
386	rs12012995	0,468358
387	rs319760	0,468358
388	rs11602622	0,468358
389	rs2304717	0,468358
390	rs6592775	0,468358
391	rs2176283	0,468358
392	rs9295385	0,468358
393	rs6081611	0,468358
394	rs6920829	0,468358
395	rs2661810	0,468358
396	rs11624601	0,468358
397	rs1528972	0,468358
398	rs11762469	0,468358
399	rs4291702	0,468358
400	rs12035887	0,468358
401	rs1755609	0,468358
402	rs12623816	0,468358
403	rs4244192	0,468358
404	rs571701	0,468358
405	rs6548485	0,468358
406	rs1315130	0,468358
407	rs13143866	0,468358
408	rs10435360	0,468358
409	rs9867093	0,468358
410	rs2510054	0,468358
411	rs7629708	0,468358
412	rs16893166	0,468358
413	rs7945424	0,468358

414	rs6474388	0,468358
415	rs7543453	0,468358
416	rs11071548	0,468358
417	rs7084706	0,468358
418	rs7901883	0,468358
419	rs10494080	0,468358
420	rs10905930	0,468358
421	rs10162627	0,468358
422	rs849538	0,468358
423	rs3774862	0,468358
424	rs11206955	0,468358
425	rs1346944	0,468358
426	rs3135715	0,468358
427	rs672203	0,468358
428	rs2578269	0,468358
429	rs6822971	0,468358
430	rs17018731	0,468358
431	rs17685233	0,468358
432	rs3767364	0,468358
433	rs3904857	0,468358
434	rs3807918	0,468358
435	rs9309766	0,468358
436	rs7749278	0,468358
437	rs11022254	0,468358
438	rs1441952	0,468358
439	rs2253211	0,468358
440	rs6798616	0,468358
441	rs2444043	0,468358
442	rs4945261	0,468358
443	rs8187945	0,468358
444	rs12186105	0,468358
445	rs17154432	0,468358
446	rs2243283	0,468358
447	rs4636424	0,468358
448	rs17015201	0,468358
449	rs2788019	0,468358
450	rs376027	0,468358
451	rs1880084	0,468358

452	rs9383882	0,468358
453	rs16860440	0,468358
454	rs12201030	0,468358
455	rs2240492	0,468358
456	rs1894603	0,468358
457	rs816868	0,468358
458	rs6673646	0,468358
459	rs4732416	0,468358
460	rs4742009	0,468358
461	rs7872937	0,468358
462	rs1661444	0,468358
463	rs706119	0,468358
464	rs9544544	0,468358
465	rs16954106	0,468358
466	rs7032871	0,468358
467	rs9332471	0,468358
468	rs487865	0,468358
469	rs7679010	0,468358
470	rs2295050	0,468358
471	rs16906549	0,468358
472	rs232262	0,468121
473	rs572846	0,468121
474	rs2310312	0,467533
475	rs11830378	0,447686
476	rs4123837	0,447241
477	rs752662	0,447075
478	rs8192100	0,447075
479	rs4683139	0,44625
480	rs41381550	0,44625
481	rs6945447	0,44625
482	rs2883456	0,44625
483	rs2251021	0,44625
484	rs1919969	0,44625
485	rs3210458	0,44621
486	rs4930265	0,445916
487	rs17114808	0,445916
488	rs2537828	0,445599
489	rs41433548	0,445599

490	rs17793957	0,445599
491	rs654631	0,445599
492	rs7107498	0,445599
493	rs17381596	0,445599
494	rs17071628	0,445599
495	rs17752628	0,445599
496	rs10208185	0,445599
497	rs7730403	0,445599
498	rs10493173	0,445599
499	rs17752640	0,445599
500	rs10487780	0,445599
501	rs509556	0,445599
502	rs17494418	0,445599
503	rs1808529	0,445599
504	rs3807874	0,445599
505	rs17470122	0,445599
506	rs6039135	0,445599
507	rs10509709	0,445599
508	rs11682005	0,445599
509	rs6761956	0,445599
510	rs12507552	0,445599
511	rs7023041	0,445599
512	rs10197159	0,445599
513	rs5022059	0,445599
514	rs12125867	0,445599
515	rs647130	0,445599
516	rs17221034	0,445599
517	rs6006733	0,445599
518	rs9911460	0,445599
519	rs4944551	0,445599
520	rs6830624	0,445599
521	rs10137468	0,445599
522	rs17129159	0,445599
523	rs7802083	0,445599
524	rs12453085	0,445599
525	rs1527369	0,445599
526	rs776023	0,445599
527	rs2216386	0,445599

528	rs7516312	0,445599
529	rs2694643	0,445599
530	rs9303590	0,445599
531	rs1793284	0,445599
532	rs234914	0,445599
533	rs7245009	0,445599
534	rs4789189	0,445599
535	rs7387373	0,445599
536	rs11715222	0,445599
537	rs17817690	0,445599
538	rs12212067	0,445599
539	rs10851257	0,445599
540	rs10791894	0,445599
541	rs17107695	0,445599
542	rs746018	0,445599
543	rs10501603	0,445599
544	rs4430517	0,445599
545	rs5753336	0,445599
546	rs1569437	0,445599
547	rs12197213	0,445599
548	rs7518943	0,445599
549	rs7211994	0,445599
550	rs6006743	0,445599
551	rs1507357	0,445599
552	rs2277304	0,445599
553	rs746019	0,445599
554	rs2065922	0,445599
555	rs17293003	0,445599
556	rs2116390	0,445599
557	rs778143	0,445599
558	rs16951395	0,445599
559	rs12765825	0,445599
560	rs4571967	0,445599
561	rs2304195	0,445599
562	rs234915	0,445599
563	rs11162341	0,445599
564	rs17105095	0,445599
565	rs10924173	0,445599

566	rs6742774	0,445599
567	rs11651891	0,445599
568	rs1889522	0,445599
569	rs17744938	0,445599
570	rs936160	0,445599
571	rs6501786	0,445599
572	rs17310529	0,445599
573	rs17099883	0,445599
574	rs2513077	0,445599
575	rs41503946	0,445599
576	rs12974182	0,445599
577	rs2038915	0,445599
578	rs2600672	0,445599
579	rs1912401	0,445599
580	rs2934683	0,445599
581	rs10926756	0,445599
582	rs6988293	0,445599
583	rs12248642	0,445599
584	rs12622458	0,445599
585	rs10875286	0,445599
586	rs2889490	0,445599
587	rs4790406	0,445599
588	rs3807219	0,445599
589	rs4789782	0,445599
590	rs13217051	0,445599
591	rs17227580	0,445599
592	rs1546914	0,445599
593	rs2058469	0,445599
594	rs10416445	0,445599
595	rs17537634	0,445599
596	rs7418577	0,445599
597	rs10000185	0,445599
598	rs9840074	0,445599
599	rs1334600	0,445599
600	rs4789788	0,445599
601	rs12713431	0,445599
602	rs1145908	0,445599
603	rs773853	0,445599

604	rs4789188	0,445599
605	rs12196205	0,445599
606	rs6466819	0,445599
607	rs12155080	0,445599
608	rs6700589	0,445599
609	rs31505	0,445599
610	rs12153735	0,445599
611	rs12160956	0,445599
612	rs2166440	0,445599
613	rs17746778	0,445599
614	rs10494374	0,445599
615	rs9512986	0,445599
616	rs923711	0,445599
617	rs1073246	0,445599
618	rs6953295	0,445599
619	rs7160534	0,445599
620	rs7548780	0,445599
621	rs3781560	0,445599
622	rs6503275	0,445599
623	rs3801410	0,445599
624	rs12719718	0,445599
625	rs2407548	0,445599
626	rs4238390	0,445599
627	rs1339226	0,445599
628	rs716333	0,445599
629	rs11201804	0,445599
630	rs586284	0,445599
631	rs11963528	0,445599
632	rs2837970	0,445599
633	rs17009672	0,445599
634	rs6917851	0,445599
635	rs1005092	0,445599
636	rs17105164	0,445599
637	rs12600845	0,445599
638	rs2030080	0,445599
639	rs7334078	0,445599
640	rs10485147	0,445599
641	rs12406713	0,445599

642	rs9319410	0,445599
643	rs5986510	0,445599
644	rs6937379	0,445599
645	rs6677116	0,445599
646	rs889730	0,445599
647	rs17105914	0,445599
648	rs17107578	0,445599
649	rs17105907	0,445599
650	rs12783090	0,445599
651	rs2153612	0,445599
652	rs10205160	0,445599
653	rs788799	0,445599
654	rs10266006	0,445599
655	rs16851949	0,445599
656	rs6738344	0,445599
657	rs2074614	0,445599
658	rs2284293	0,445599
659	rs747996	0,445599
660	rs11615548	0,445599
661	rs514643	0,445599
662	rs2305206	0,445599
663	rs17096257	0,445599
664	rs10495022	0,445599
665	rs6115381	0,445599
666	rs6793635	0,445599
667	rs1591636	0,445599
668	rs492563	0,445599
669	rs2290519	0,445599
670	rs542600	0,445599
671	rs544398	0,445599
672	rs1173850	0,445599
673	rs6536392	0,445599
674	rs706133	0,445599
675	rs7130004	0,445599
676	rs16857272	0,445599
677	rs4919008	0,445599
678	rs38808	0,445599
679	rs17712300	0,445599

680	rs10038062	0,445599
681	rs4970722	0,445599
682	rs11070765	0,445599
683	rs6500288	0,445599
684	rs11671309	0,445599
685	rs142	0,445599
686	rs507326	0,445599
687	rs7656697	0,445599
688	rs17057475	0,445599
689	rs4771207	0,445599
690	rs7206735	0,445599
691	rs1437337	0,445599
692	rs1539439	0,445599
693	rs7672834	0,445599
694	rs17217647	0,445599
695	rs12080760	0,445599
696	rs3798889	0,445599
697	rs2273151	0,445599
698	rs2273152	0,445599
699	rs1911594	0,445599
700	rs10060763	0,445599
701	rs12568559	0,445599
702	rs2098781	0,445599
703	rs2473138	0,445599
704	rs136575	0,445599
705	rs7985565	0,445599
706	rs10462535	0,445599
707	rs2493766	0,445599
708	rs1323430	0,445599
709	rs2144317	0,445599
710	rs17792105	0,445599
711	rs2347946	0,445599
712	rs491730	0,445599
713	rs10937121	0,445599
714	rs10906451	0,445599
715	rs7578592	0,445599
716	rs17469935	0,445599
717	rs10894801	0,445599

718	rs10906224	0,445599
719	rs1621293	0,445599
720	rs10486950	0,445599
721	rs17101464	0,445599
722	rs4791287	0,445599
723	rs17686720	0,445599
724	rs9614462	0,445599
725	rs9583334	0,445599
726	rs9316153	0,445599
727	rs203050	0,445599
728	rs13353636	0,445599
729	rs16946506	0,445599
730	rs17250107	0,445599
731	rs7339274	0,445599
732	rs17804923	0,445599
733	rs17026635	0,445599
734	rs745639	0,445599
735	rs13298370	0,445599
736	rs6540614	0,445599
737	rs252545	0,445599
738	rs11466511	0,445599
739	rs491928	0,445599
740	rs7228240	0,445599
741	rs536054	0,445599
742	rs12529407	0,445599
743	rs2152066	0,445599
744	rs2191316	0,445599
745	rs3765098	0,445599
746	rs16952975	0,445599
747	rs10889155	0,445599
748	rs9316514	0,445599
749	rs2423464	0,445599
750	rs4254878	0,445599
751	rs12614102	0,445599
752	rs3793181	0,445599
753	rs2153000	0,445599
754	rs7934652	0,445599
755	rs4669573	0,445599

756	rs4491964	0,445599
757	rs10196146	0,445599
758	rs1994312	0,445599
759	rs1177234	0,445599
760	rs1483449	0,445599
761	rs7201414	0,445599
762	rs155387	0,445599
763	rs6935296	0,445599
764	rs11045970	0,445599
765	rs7731153	0,445599
766	rs12956638	0,445599
767	rs1082214	0,445599
768	rs3735487	0,445599
769	rs2498500	0,445599
770	rs10224793	0,445599
771	rs2753614	0,445599
772	rs2219250	0,445599
773	rs16939880	0,445599
774	rs10943930	0,445599
775	rs2990877	0,445599
776	rs12464067	0,445599
777	rs996379	0,445599
778	rs7315682	0,445599
779	rs13198062	0,445599
780	rs4362705	0,445599
781	rs6672561	0,445599
782	rs1431486	0,445599
783	rs978290	0,445599
784	rs7589790	0,445599
785	rs12625776	0,445599
786	rs500243	0,445599
787	rs1149160	0,445599
788	rs7946133	0,445599
789	rs2448246	0,445599
790	rs2498434	0,445599
791	rs3782837	0,445599
792	rs1629507	0,445599
793	rs1663584	0,445599

794	rs1690918	0,445599
795	rs492823	0,445599
796	rs593479	0,445599
797	rs9392325	0,445599
798	rs7527246	0,445599
799	rs10066756	0,445599
800	rs2660228	0,445599
801	rs10492629	0,445599
802	rs17355265	0,445599
803	rs3801712	0,445599
804	rs12415467	0,445599
805	rs1149155	0,445599
806	rs6086964	0,445599
807	rs17096091	0,445599
808	rs17066720	0,445599
809	rs4103380	0,445599
810	rs2466089	0,445599
811	rs6881888	0,445599
812	rs4611855	0,445599
813	rs10137465	0,445599
814	rs1149156	0,445599
815	rs1501253	0,445599
816	rs479640	0,445599
817	rs31669	0,445599
818	rs12995333	0,445599
819	rs11629324	0,445599
820	rs4466137	0,445599
821	rs7896883	0,445599
822	rs10487998	0,445599
823	rs6971925	0,445599
824	rs17133543	0,445599
825	rs1030110	0,445599
826	rs3770375	0,445599
827	rs149667	0,445599
828	rs11223351	0,445599
829	rs17234886	0,445599
830	rs1997865	0,445599
831	rs10926758	0,445599

832	rs1864744	0,445599
833	rs10949739	0,445599
834	rs12113120	0,445599
835	rs11854890	0,445599
836	rs17159833	0,445599
837	rs9442956	0,445599
838	rs13207114	0,445599
839	rs4707570	0,445599
840	rs3796632	0,445599
841	rs848086	0,445599
842	rs17714306	0,445599
843	rs567348	0,445599
844	rs11847417	0,445599
845	rs7044045	0,445599
846	rs4835774	0,445599
847	rs1277215	0,445599
848	rs177374	0,445599
849	rs4372296	0,445599
850	rs2399858	0,445599
851	rs2143988	0,445599
852	rs17105916	0,445599
853	rs9589986	0,445599
854	rs7573820	0,445599
855	rs13201188	0,445599
856	rs10141687	0,445599
857	rs154942	0,445599
858	rs7621841	0,445599
859	rs4433764	0,445599
860	rs2729682	0,445599
861	rs2439222	0,445599
862	rs13157965	0,445599
863	rs1784178	0,445599
864	rs6476078	0,445599
865	rs569376	0,445599
866	rs11688935	0,445599
867	rs11246340	0,445599
868	rs16910520	0,445599
869	rs6698474	0,445599

870	rs7118874	0,445599
871	rs17023936	0,445599
872	rs3807222	0,445599
873	rs10753688	0,445599
874	rs4652742	0,445599
875	rs746064	0,445599
876	rs4379706	0,445599
877	rs6694158	0,445599
878	rs17797801	0,445599
879	rs10747489	0,445599
880	rs6504077	0,445599
881	rs154631	0,445599
882	rs4486743	0,445599
883	rs11143589	0,445599
884	rs1587608	0,445599
885	rs10138002	0,445599
886	rs17135053	0,445599
887	rs11147040	0,445599
888	rs123241	0,445599
889	rs16887057	0,445599
890	rs192795	0,445599
891	rs2184723	0,445599
892	rs994424	0,445599
893	rs4904262	0,445599
894	rs12709653	0,445599
895	rs5986480	0,445599
896	rs10904831	0,445599
897	rs4655414	0,445599
898	rs6105115	0,445599
899	rs6759186	0,445599
900	rs7904146	0,445599
901	rs36133	0,445599
902	rs37332	0,445599
903	rs2813728	0,445599
904	rs10519344	0,445599
905	rs7621973	0,445599
906	rs7437956	0,445599
907	rs2768941	0,445599

908	rs7064104	0,445599
909	rs893179	0,445599
910	rs6042365	0,445599
911	rs9511941	0,445599
912	rs12719719	0,445599
913	rs152060	0,445599
914	rs735756	0,445599
915	rs7685063	0,445599
916	rs6546119	0,445599
917	rs3108630	0,445599
918	rs10510170	0,445599
919	rs13009588	0,445599
920	rs468821	0,445599
921	rs6134890	0,445599
922	rs9507991	0,445599
923	rs11639569	0,445599
924	rs17115411	0,445599
925	rs2165888	0,445599
926	rs4986122	0,445599
927	rs11773627	0,445599
928	rs16855155	0,445599
929	rs7548768	0,445599
930	rs2288270	0,445599
931	rs17455458	0,445599
932	rs3767910	0,445599
933	rs6994888	0,445599
934	rs17124895	0,445599
935	rs12455083	0,445599
936	rs6666724	0,445599
937	rs8131547	0,445599
938	rs177407	0,445599
939	rs4787040	0,445599
940	rs27139	0,445599
941	rs586037	0,445599
942	rs5762546	0,445599
943	rs7338172	0,445599
944	rs7872482	0,445599
945	rs1024681	0,445599

946	rs9863706	0,445599
947	rs7756062	0,445599
948	rs4796640	0,444774
949	rs12001326	0,444774
950	rs6565624	0,444774
951	rs604411	0,444774
952	rs12512157	0,444774
953	rs10503065	0,425062
954	rs3743473	0,424786
955	rs2286720	0,424237
956	rs999147	0,40998
957	rs8052055	0,40998
958	rs8468	0,409511

APPENDIX C - Electronic Format of Decision Trees

Attached CD below contains decision trees constructed using;

- Representative SNPs(only genotype data)
- Representative SNPs and Clinical data in .swf format.

Flash Player or any web browser with plugged in flash player may run the files.

APPENDIX D - Decision Tree Generated by Representative SNPs

SNP_A-4213932 = [C_C]
| SNP_A-2146889 = [G_G]
| | SNP_A-1896372 = [G_G]: CASE {CASE=7, CNTL=1}
| | SNP_A-1896372 = [T_G] / [G_T]: CNTL {CASE=0, CNTL=2}
| SNP_A-2146889 = [T_G] / [G_T]
| | SNP_A-1849082 = [A_G] / [G_A]: CASE {CASE=17, CNTL=0}
| | SNP_A-1849082 = [G_G]
| | | SNP_A-2118885 = [T_C] / [C_T]: CNTL {CASE=1, CNTL=9}
| | | SNP_A-2118885 = [T_T]
| | | | SNP_A-2125015 = [A_G] / [G_A]: CASE {CASE=12, CNTL=0}
| | | | SNP_A-2125015 = [G_G]
| | | | | SNP_A-2034790 = [A_C] / [C_A]: CASE {CASE=8, CNTL=0}
| | | | | SNP_A-2034790 = [C_C]
| | | | | | SNP_A-1865279 = [A_G] / [G_A]: CASE {CASE=8, CNTL=0}
| | | | | | SNP_A-1865279 = [G_G]
| | | | | | | SNP_A-1812298 = [G_G]
| | | | | | | | SNP_A-2083550 = [T_C] / [C_T]: CNTL {CASE=0, CNTL=4}
| | | | | | | | SNP_A-2083550 = [T_T]: CASE {CASE=97, CNTL=54}
| | | | | | | SNP_A-1812298 = [T_G] / [G_T]: CNTL {CASE=0, CNTL=4}
| SNP_A-2146889 = [T_T]
| | SNP_A-2114080 = [A_A]
| | | SNP_A-1838931 = [C_C]
| | | | SNP_A-2272567 = [C_C]
| | | | | SNP_A-1892259 = [G_G]: CASE {CASE=3, CNTL=0}
| | | | | SNP_A-1892259 = [T_G] / [G_T]
| | | | | | SNP_A-2034790 = [A_C] / [C_A]: CASE {CASE=2, CNTL=0}
| | | | | | SNP_A-2034790 = [C_C]
| | | | | | | SNP_A-2272456 = [G_G]
| | | | | | | | SNP_A-4261255 = [T_G] / [G_T]: CASE {CASE=2, CNTL=0}
| | | | | | | | SNP_A-4261255 = [T_T]
| | | | | | | | | SNP_A-1987132 = [C_G] / [G_C]: CASE {CASE=2, CNTL=0}
| | | | | | | | | SNP_A-1987132 = [G_G]: CNTL {CASE=15, CNTL=39}
| | | | | | SNP_A-2272456 = [T_G] / [G_T]: CASE {CASE=2, CNTL=0}
| | | | | SNP_A-1892259 = [T_T]
| | | | | | SNP_A-2296526 = [A_A]: CNTL {CASE=0, CNTL=4}
| | | | | | SNP_A-2296526 = [A_G] / [G_A]
| | | | | | | SNP_A-2125015 = [A_G] / [G_A]: CASE {CASE=4, CNTL=0}
| | | | | | | SNP_A-2125015 = [G_G]: CNTL {CASE=17, CNTL=31}
| | | | | | SNP_A-2296526 = [G_G]
| | | | | | | SNP_A-1980601 = [T_C] / [C_T]: CNTL {CASE=0, CNTL=5}

| | | | | | | SNP_A-1980601 = [T_T]
| | | | | | | | SNP_A-1902372 = [T_C] / [C_T]
| | | | | | | | SNP_A-1965422 = [T_C] / [C_T]: CASE {CASE=2, CNTL=0}
| | | | | | | | SNP_A-1965422 = [T_T]: CNTL {CASE=0, CNTL=10}
| | | | | | | | SNP_A-1902372 = [T_T]
| | | | | | | | | SNP_A-2113638 = [C_C]
| | | | | | | | | SNP_A-2030266 = [A_G] / [G_A]: CNTL {CASE=5, CNTL=12}
| | | | | | | | | SNP_A-2030266 = [G_G]: CASE {CASE=283, CNTL=176}
| | | | | | | | SNP_A-2113638 = [T_C] / [C_T]
| | | | | | | | | SNP_A-1872465 = [A_A]: CNTL {CASE=16, CNTL=31}
| | | | | | | | | SNP_A-1872465 = [A_G] / [G_A]: CASE {CASE=3, CNTL=0}
| | | | | | | | SNP_A-2113638 = [T_T]: CASE {CASE=1, CNTL=1}
| | | SNP_A-2272567 = [T_C] / [C_T]: CASE {CASE=9, CNTL=0}
| | | SNP_A-1838931 = [T_C] / [C_T]: CNTL {CASE=0, CNTL=9}
| | SNP_A-2114080 = [A_G] / [G_A]
| | | SNP_A-1987132 = [C_G] / [G_C]: CASE {CASE=7, CNTL=0}
| | | SNP_A-1987132 = [G_G]
| | | | SNP_A-1809622 = [A_G] / [G_A]: CASE {CASE=2, CNTL=0}
| | | | SNP_A-1809622 = [G_G]
| | | | | SNP_A-1997332 = [A_A]
| | | | | | SNP_A-4273581 = [T_C] / [C_T]: CASE {CASE=8, CNTL=2}
| | | | | | SNP_A-4273581 = [T_T]: CNTL {CASE=32, CNTL=92}
| | | | | SNP_A-1997332 = [A_G] / [G_A]: CASE {CASE=2, CNTL=0}
| | SNP_A-2114080 = [G_G]: CNTL {CASE=1, CNTL=8}
SNP_A-4213932 = [T_C] / [C_T]
| SNP_A-2258450 = [A_A]
| | SNP_A-1901559 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=6}
| | SNP_A-1901559 = [G_G]
| | | SNP_A-1930045 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=5}
| | | SNP_A-1930045 = [G_G]
| | | | SNP_A-4198010 = [A_A]
| | | | | SNP_A-2164852 = [C_C]
| | | | | | SNP_A-2118885 = [T_C] / [C_T]: CNTL {CASE=0, CNTL=10}
| | | | | | SNP_A-2118885 = [T_T]
| | | | | | | SNP_A-2268610 = [A_A]
| | | | | | | | SNP_A-1865279 = [A_G] / [G_A]: CASE {CASE=8, CNTL=1}
| | | | | | | | SNP_A-1865279 = [G_G]
| | | | | | | | | SNP_A-2021678 = [C_C]
| | | | | | | | | SNP_A-1929138 = [C_C]: CNTL {CASE=131, CNTL=162}
| | | | | | | | | SNP_A-1929138 = [C_G] / [G_C]: CASE {CASE=5, CNTL=0}
| | | | | | | | | SNP_A-2021678 = [T_C] / [C_T]
| | | | | | | | | SNP_A-2304918 = [A_A]: CNTL {CASE=0, CNTL=18}
| | | | | | | | | SNP_A-2304918 = [A_G] / [G_A]: CASE {CASE=2, CNTL=0}
| | | | | | | SNP_A-2268610 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=7}

| | | | | SNP_A-2164852 = [C_G] / [G_C]
| | | | | | SNP_A-2075360 = [A_C] / [C_A]: CNTL {CASE=0, CNTL=2}
| | | | | | SNP_A-2075360 = [C_C]: CASE {CASE=9, CNTL=0}
| | | | SNP_A-4198010 = [A_G] / [G_A]: CNTL {CASE=1, CNTL=11}
| SNP_A-2258450 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=8}
SNP_A-4213932 = [T_T]
| SNP_A-1849082 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=2}
| SNP_A-1849082 = [G_G]
| | SNP_A-1893931 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=2}
| | SNP_A-1893931 = [G_G]: CASE {CASE=20, CNTL=8}

APPENDIX E - Decision Rules to Predict Disease Status in Terms of AD Using Representative SNPs

No	***Decision Rules for CASES***	***Gene***	***Chr.***	***P.***
1	SNP_A-4213932 = [C_C] SNP_A-2146889 = [G_G] SNP_A-1896372 = [G_G]	DBT dihydrolipoamide branched chain transacylase E2 FOXO3 forkhead box O3 DISC1 disrupted in schizophrenia 1	1 6 1	87,50%
2	SNP_A-4213932 = [C_C] SNP_A-2146889 = [T_G] / [G_T] SNP_A-1849082 = [A_G] / [G_A]	DBT dihydrolipoamide branched chain transacylase E2 FOXO3 forkhead box O3 NBN nibrin	1 6 8	100,00%
3	SNP_A-4213932 = [C_C] SNP_A-2146889 = [T_G] / [G_T] SNP_A-1849082 = [G_G] SNP_A-2118885 = [T_T] SNP_A-2125015 = [A_G] / [G_A]	DBT dihydrolipoamide branched chain transacylase E2 FOXO3 forkhead box O3 NBN nibrin DDO D-aspartate oxidase PDZD8 PDZ domain containing 8	1 6 8 6 10	100,00%
4	SNP_A-4213932 = [C_C] SNP_A-2146889 = [T_G] / [G_T] SNP_A-1849082 = [G_G] SNP_A-2118885 = [T_T] SNP_A-2125015 = [G_G] SNP_A-2034790 = [A_C] / [C_A]	DBT dihydrolipoamide branched chain transacylase E2 FOXO3 forkhead box O3 NBN nibrin DDO D-aspartate oxidase PDZD8 PDZ domain containing 8 ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	1 6 8 6 10 4	100,00%
5	SNP_A-4213932 = [C_C] SNP_A-2146889 = [T_G] / [G_T] SNP_A-1849082 = [G_G] SNP_A-2118885 = [T_T] SNP_A-2125015 = [G_G]	DBT dihydrolipoamide branched chain transacylase E2 FOXO3 forkhead box O3 NBN nibrin DDO D-aspartate oxidase PDZD8 PDZ domain	1 6 8 6 10	100,00%

		containing 8		
	SNP_A-2034790 = [C_C]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
	SNP_A-1865279 = [A_G] / [G_A]	C9orf3 chromosome 9 open reading frame 3	9	
6	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	64,24%
	SNP_A-2146889 = [T_G] / [G_T]	FOXO3 forkhead box O3	6	
	SNP_A-1849082 = [G_G]	NBN nibrin	8	
	SNP_A-2118885 = [T_T]	DDO D-aspartate oxidase	6	
	SNP_A-2125015 = [G_G]	PDZD8 PDZ domain containing 8	10	
	SNP_A-2034790 = [C_C]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
	SNP_A-1865279 = [G_G]	C9orf3 chromosome 9 open reading frame 3	9	
	SNP_A-1812298 = [G_G]	ARHGAP26 Rho GTPase activating protein 26	5	
	SNP_A-2083550 = [T_T]	ANGPT2 angiopoietin 2	8	
7	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [G_G]	LIPH lipase, member H	3	
8	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig),	7	

		short basic domain, secreted, (semaphorin) 3C		
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	SNP_A-2034790 = [A_C] / [C_A]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
9	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	SNP_A-2034790 = [C_C]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
	SNP_A-2272456 = [G_G]	SLC35A3 solute carrier family 35 (UDP-N-acetylglucosamine (UDP-GlcNAc) transporter), member A3	1	
	SNP_A-4261255 = [T_G] / [G_T]	PTPRM protein tyrosine phosphatase, receptor type, M	18	
10	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	

	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	SNP_A-2034790 = [C_C]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
	SNP_A-2272456 = [G_G]	SLC35A3 solute carrier family 35 (UDP-N-acetylglucosamine (UDP-GlcNAc) transporter), member A3	1	
	SNP_A-4261255 = [T_T]	PTPRM protein tyrosine phosphatase, receptor type, M	18	
	SNP_A-1987132 = [C_G] / [G_C]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
11	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	SNP_A-2034790 = [C_C]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
	SNP_A-2272456 = [T_G] / [G_T]	SLC35A3 solute carrier family 35 (UDP-N-acetylglucosamine (UDP-GlcNAc) transporter), member A3	1	
12	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	

	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [A_G] / [G_A]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-2125015 = [A_G] / [G_A]	PDZD8 PDZ domain containing 8	10	
13	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [G_G]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-1980601 = [T_T]	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	
	SNP_A-1902372 = [T_C] / [C_T]	GSN gelsolin	9	
	SNP_A-1965422 = [T_C] / [C_T]	FMNL2 formin-like 2	2	

14	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	61,66%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [G_G]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-1980601 = [T_T]	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	
	SNP_A-1902372 = [T_T]	GSN gelsolin	9	
	SNP_A-2113638 = [C_C]	TPO thyroid peroxidase	2	
	SNP_A-2030266 = [G_G]	SYN3 synapsin III	22	
15	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [G_G]	KCNN3 potassium intermediate/small	1	

		conductance calcium-activated channel, subfamily N, member 3		
	SNP_A-1980601 = [T_T]	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	
	SNP_A-1902372 = [T_T]	GSN gelsolin	9	
	SNP_A-2113638 = [T_C] / [C_T]	TPO thyroid peroxidase	2	
	SNP_A-1872465 = [A_G] / [G_A]	C9orf3 chromosome 9 open reading frame 3	9	
	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
16	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	50,00%
	SNP_A-2296526 = [G_G]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-1980601 = [T_T]	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	
	SNP_A-1902372 = [T_T]	GSN gelsolin	9	
	SNP_A-2113638 = [T_T]	TPO thyroid peroxidase	2	
17	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	

	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [T_C] / [C_T]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
18	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [C_G] / [G_C]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
19	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [G_G]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
	SNP_A-1809622 = [A_G] / [G_A]	HMGA1 high mobility group AT-hook 1	6	
20	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	80,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [G_G]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
	SNP_A-1809622 = [G_G]	HMGA1 high mobility group AT-hook 1	6	
	SNP_A-1997332 = [A_A]	GABBR2 gamma-aminobutyric acid (GABA) B receptor, 2	9	
	SNP_A-4273581 = [T_C] / [C_T]	MAML3 mastermind-like 3	4	

		(Drosophila)		
21	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [G_G]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
	SNP_A-1809622 = [G_G]	HMGA1 high mobility group AT-hook 1	6	
	SNP_A-1997332 = [A_G] / [G_A]	GABBR2 gamma-aminobutyric acid (GABA) B receptor, 2	9	
22	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	88,89%
	SNP_A-2258450 = [A_A]	ANGPT2 angiopoietin 2	8	
	SNP_A-1901559 = [G_G]	STK39 serine threonine kinase 39	2	
	SNP_A-1930045 = [G_G]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-4198010 = [A_A]	FGD4 FYVE, RhoGEF and PH domain containing 4	4	
	SNP_A-2164852 = [C_C]	TLL2 tolloid-like 2	10	
	SNP_A-2118885 = [T_T]	DDO D-aspartate oxidase	6	
	SNP_A-2268610 = [A_A]	PIKFYVE phosphoinositide kinase, FYVE finger containing	2	
	SNP_A-1865279 = [A_G] / [G_A]	C9orf3 chromosome 9 open reading frame 3	9	
23	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2258450 = [A_A]	ANGPT2 angiopoietin 2	8	
	SNP_A-1901559 = [G_G]	STK39 serine threonine kinase 39	2	
	SNP_A-1930045 = [G_G]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-4198010 = [A_A]	FGD4 FYVE, RhoGEF and PH	4	

		domain containing 4		
	SNP_A-2164852 = [C_C]	TLL2 tolloid-like 2	10	
	SNP_A-2118885 = [T_T]	DDO D-aspartate oxidase	6	
	SNP_A-2268610 = [A_A]	PIKFYVE phosphoinositide kinase, FYVE finger containing	2	
	SNP_A-1865279 = [G_G]	C9orf3 chromosome 9 open reading frame 3	9	
	SNP_A-2021678 = [C_C]	DOK1 docking protein 1, 62kDa (downstream of tyrosine kinase 1)	2	
	SNP_A-1929138 = [C_G] / [G_C]	TRHDE thyrotropin-releasing hormone degrading enzyme	12	
24	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2258450 = [A_A]	ANGPT2 angiopoietin 2	8	
	SNP_A-1901559 = [G_G]	STK39 serine threonine kinase 39	2	
	SNP_A-1930045 = [G_G]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-4198010 = [A_A]	FGD4 FYVE, RhoGEF and PH domain containing 4	4	
	SNP_A-2164852 = [C_C]	TLL2 tolloid-like 2	10	
	SNP_A-2118885 = [T_T]	DDO D-aspartate oxidase	6	
	SNP_A-2268610 = [A_A]	PIKFYVE phosphoinositide kinase, FYVE finger containing	2	
	SNP_A-1865279 = [G_G]	C9orf3 chromosome 9 open reading frame 3	9	
	SNP_A-2021678 = [T_C] / [C_T]	DOK1 docking protein 1, 62kDa (downstream of tyrosine kinase 1)	2	
	SNP_A-2304918 = [A_G] / [G_A]	SNW1 SNW domain containing 1	14	
25	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2258450 = [A_A]	ANGPT2 angiopoietin 2	8	
	SNP_A-1901559 = [G_G]	STK39 serine threonine kinase 39	2	
	SNP_A-1930045 = [G_G]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted,	7	

		(semaphorin) 3C		
	SNP_A-4198010 = [A_A]	FGD4 FYVE, RhoGEF and PH domain containing 4	4	
	SNP_A-2164852 = [C_G] / [G_C]	TLL2 tolloid-like 2	10	
	SNP_A-2075360 = [C_C]	CAMKK2 calcium/calmodulin-dependent protein kinase kinase 2, beta	12	
26	SNP_A-4213932 = [T_T]	DBT dihydrolipoamide branched chain transacylase E2	1	71,43%
	SNP_A-1849082 = [G_G]	NBN nibrin	8	
	SNP_A-1893931 = [G_G]	PLCB1 phospholipase C, beta 1 (phosphoinositide-specific)	20	

APPENDIX F - Decision Tree Generated by Representative SNPs and Clinical Data

SNP_A-4213932 = [C_C]
| SNP_A-2146889 = [G_G]
| | SNP_A-1896372 = [G_G]: CASE {CASE=7, CNTL=1}
| | SNP_A-1896372 = [T_G] / [G_T]: CNTL {CASE=0, CNTL=2}
| SNP_A-2146889 = [T_G] / [G_T]
| | Body_Mass_Indx > 36.250: CNTL {CASE=0, CNTL=3}
| | Body_Mass_Indx ≤ 36.250
| | | TRIG_MMOL_L > 0.700
| | | | age > 90.200: CNTL {CASE=0, CNTL=2}
| | | | age ≤ 90.200
| | | | | SNP_A-2118885 = [T_C] / [C_T]
| | | | | | age > 64.550: CNTL {CASE=1, CNTL=9}
| | | | | | age ≤ 64.550: CASE {CASE=2, CNTL=0}
| | | | | SNP_A-2118885 = [T_T]: CASE {CASE=140, CNTL=54}
| | | TRIG_MMOL_L ≤ 0.700: CNTL {CASE=0, CNTL=3}
| SNP_A-2146889 = [T_T]
| | SNP_A-2114080 = [A_A]
| | | SNP_A-1838931 = [C_C]
| | | | SNP_A-2272567 = [C_C]
| | | | | SNP_A-1892259 = [G_G]: CASE {CASE=3, CNTL=0}
| | | | | SNP_A-1892259 = [T_G] / [G_T]
| | | | | | HBA1C_PCT > 0.062: CASE {CASE=5, CNTL=0}
| | | | | | HBA1C_PCT ≤ 0.062
| | | | | | | Body_Mass_Indx > 18.900
| | | | | | | | CHOL_MMOL_L > 3.505
| | | | | | | | | SNP_A-2034790 = [A_C] / [C_A]: CASE {CASE=2, CNTL=0}
| | | | | | | | | SNP_A-2034790 = [C_C]
| | | | | | | | | | SNP_A-2164852 = [C_C]: CNTL {CASE=10, CNTL=39}
| | | | | | | | | | SNP_A-2164852 = [C_G] / [G_C]: CASE {CASE=2, CNTL=0}
| | | | | | | | | CHOL_MMOL_L ≤ 3.505: CASE {CASE=2, CNTL=0}
| | | | | | | Body_Mass_Indx ≤ 18.900: CASE {CASE=2, CNTL=0}
| | | | | SNP_A-1892259 = [T_T]
| | | | | | SNP_A-2296526 = [A_A]: CNTL {CASE=0, CNTL=4}
| | | | | | SNP_A-2296526 = [A_G] / [G_A]
| | | | | | | SNP_A-2125015 = [A_G] / [G_A]: CASE {CASE=4, CNTL=0}
| | | | | | | SNP_A-2125015 = [G_G]
| | | | | | | | HB_G_L > 112

| | | | | | | | HDLCH_MMOL_L > 0.825
| | | | | | | | | SNP_A-1948390 = [A_A]: CNTL {CASE=4, CNTL=28}
| | | | | | | | | SNP_A-1948390 = [A_G] / [G_A]: CASE {CASE=7, CNTL=3}
| | | | | | | | | SNP_A-1948390 = [G_G]: CASE {CASE=2, CNTL=0}
| | | | | | | | HDLCH_MMOL_L ≤ 0.825: CASE {CASE=2, CNTL=0}
| | | | | | | | HB_G_L ≤ 112: CASE {CASE=2, CNTL=0}
| | | | | | SNP_A-2296526 = [G_G]
| | | | | | | WBC_GIGA_L > 11.800: CNTL {CASE=0, CNTL=5}
| | | | | | | WBC_GIGA_L ≤ 11.800
| | | | | | | | SNP_A-1980601 = [T_C] / [C_T]: CNTL {CASE=0, CNTL=5}
| | | | | | | | SNP_A-1980601 = [T_T]
| | | | | | | | | SNP_A-1902372 = [T_C] / [C_T]
| | | | | | | | | SNP_A-1965422 = [T_C] / [C_T]: CASE {CASE=2, CNTL=0}
| | | | | | | | | SNP_A-1965422 = [T_T]: CNTL {CASE=0, CNTL=9}
| | | | | | | | SNP_A-1902372 = [T_T]
| | | | | | | | | SNP_A-2113638 = [C_C]: CASE {CASE=288, CNTL=184}
| | | | | | | | | SNP_A-2113638 = [T_C] / [C_T]: CNTL {CASE=19, CNTL=31}
| | | | | | | | | SNP_A-2113638 = [T_T]: CASE {CASE=1, CNTL=1}
| | | SNP_A-2272567 = [T_C] / [C_T]: CASE {CASE=9, CNTL=0}
| | | SNP_A-1838931 = [T_C] / [C_T]: CNTL {CASE=0, CNTL=9}
| | SNP_A-2114080 = [A_G] / [G_A]
| | | SNP_A-1987132 = [C_G] / [G_C]: CASE {CASE=7, CNTL=0}
| | | SNP_A-1987132 = [G_G]
| | | | TRIG_MMOL_L > 0.695
| | | | | SNP_A-1809622 = [A_G] / [G_A]: CASE {CASE=2, CNTL=0}
| | | | | SNP_A-1809622 = [G_G]
| | | | | | SNP_A-1997332 = [A_A]: CNTL {CASE=38, CNTL=94}
| | | | | | SNP_A-1997332 = [A_G] / [G_A]: CASE {CASE=2, CNTL=0}
| | | | TRIG_MMOL_L ≤ 0.695: CASE {CASE=2, CNTL=0}
| | SNP_A-2114080 = [G_G]: CNTL {CASE=1, CNTL=8}
SNP_A-4213932 = [T_C] / [C_T]
| age > 53.050
| | Body_Mass_Indx > 17.150
| | | HBA1C_PCT > 0.091: CASE {CASE=3, CNTL=0}
| | | HBA1C_PCT ≤ 0.091
| | | | LDLCH_MMOL_L > 5.135: CASE {CASE=3, CNTL=0}
| | | | LDLCH_MMOL_L ≤ 5.135
| | | | | SNP_A-2258450 = [A_A]
| | | | | | SNP_A-1901559 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=6}
| | | | | | SNP_A-1901559 = [G_G]
| | | | | | | WBC_GIGA_L > 3.650
| | | | | | | | SNP_A-1930045 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=5}
| | | | | | | | SNP_A-1930045 = [G_G]
| | | | | | | | | SNP_A-1929138 = [C_C]

| | | | | | | | | SNP_A-1865279 = [A_G] / [G_A]: CASE {CASE=7, CNTL=2}
| | | | | | | | | SNP_A-1865279 = [G_G]: CNTL {CASE=129, CNTL=203}
| | | | | | | | SNP_A-1929138 = [C_G] / [G_C]: CASE {CASE=5, CNTL=1}
| | | | | | | WBC_GIGA_L ≤ 3.650: CNTL {CASE=0, CNTL=5}
| | | | | SNP_A-2258450 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=8}
| | Body_Mass_Indx ≤ 17.150: CASE {CASE=3, CNTL=0}
| age ≤ 53.050: CASE {CASE=6, CNTL=0}
SNP_A-4213932 = [T_T]
| CHOL_MMOL_L > 7.285: CNTL {CASE=0, CNTL=2}
| CHOL_MMOL_L ≤ 7.285
| | SNP_A-1849082 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=2}
| | SNP_A-1849082 = [G_G]
| | | SNP_A-1893931 = [A_G] / [G_A]: CNTL {CASE=0, CNTL=2}
| | | SNP_A-1893931 = [G_G]: CASE {CASE=20, CNTL=6}

APPENDIX G - Decision Rules to Predict Disease Status in Terms of AD Using Representative SNPs and Clinical Data

No	Decision Rules for CASES	Gene	Chr.	P.
1	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	87,50%
	SNP_A-2146889 = [G_G]	FOXO3 forkhead box O3	6	
	SNP_A-1896372 = [G_G]	DISC1 disrupted in schizophrenia 1	1	
2	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_G] / [G_T]	FOXO3 forkhead box O3	6	
	Body_Mass_Indx ≤ 36.250	#N/A	#N/A	
	TRIG_MMOL_L > 0.700	#N/A	#N/A	
	age ≤ 90.200	#N/A	#N/A	
	SNP_A-2118885 = [T_C] / [C_T]	DDO D-aspartate oxidase	6	
	age ≤ 64.550	#N/A	#N/A	
3	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	72,16%
	SNP_A-2146889 = [T_G] / [G_T]	FOXO3 forkhead box O3	6	
	Body_Mass_Indx ≤ 36.250	#N/A	#N/A	
	TRIG_MMOL_L > 0.700	#N/A	#N/A	
	age ≤ 90.200	#N/A	#N/A	
	SNP_A-2118885 = [T_T]	DDO D-aspartate oxidase	6	
4	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [G_G]	LIPH lipase, member H	3	
5	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched	1	100,00%

		chain transacylase E2		
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	HBA1C_PCT > 0.062	#N/A	#N/A	
6	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	HBA1C_PCT ≤ 0.062	#N/A	#N/A	
	Body_Mass_Indx > 18.900	#N/A	#N/A	
	CHOL_MMOL_L > 3.505	#N/A	#N/A	
	SNP_A-2034790 = [A_C] / [C_A]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
7	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-	13	

		family C (CFTR/MRP), member 4		
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	HBA1C_PCT ≤ 0.062	#N/A	#N/A	
	Body_Mass_Indx > 18.900	#N/A	#N/A	
	CHOL_MMOL_L > 3.505	#N/A	#N/A	
	SNP_A-2034790 = [C_C]	ENPP6 ectonucleotide pyrophosphatase/phosphodiesterase 6	4	
	SNP_A-2164852 = [C_G] / [G_C]	TLL2 tolloid-like 2	10	
8	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	HBA1C_PCT ≤ 0.062	#N/A	#N/A	
	Body_Mass_Indx > 18.900	#N/A	#N/A	
	CHOL_MMOL_L ≤ 3.505	#N/A	#N/A	
9	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_G] / [G_T]	LIPH lipase, member H	3	
	HBA1C_PCT ≤ 0.062	#N/A	#N/A	
	Body_Mass_Indx ≤ 18.900	#N/A	#N/A	
10	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched	1	100,00%

		chain transacylase E2		
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [A_G] / [G_A]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-2125015 = [A_G] / [G_A]	PDZD8 PDZ domain containing 8	10	
11	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	70,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [A_G] / [G_A]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-2125015 = [G_G]	PDZD8 PDZ domain containing 8	10	
	HB_G_L > 112	#N/A	#N/A	
	HDLCH_MMOL_L > 0.825	#N/A	#N/A	
	SNP_A-1948390 = [A_G] / [G_A]	LUM lumican	12	
12	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	

	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [A_G] / [G_A]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-2125015 = [G_G]	PDZD8 PDZ domain containing 8	10	
	HB_G_L > 112	#N/A	#N/A	
	HDLCH_MMOL_L > 0.825	#N/A	#N/A	
	SNP_A-1948390 = [G_G]	LUM lumican	12	
13	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [A_G] / [G_A]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-2125015 = [G_G]	PDZD8 PDZ domain containing 8	10	
	HB_G_L > 112	#N/A	#N/A	
	HDLCH_MMOL_L ≤ 0.825	#N/A	#N/A	
14	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	

	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [A_G] / [G_A]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	SNP_A-2125015 = [G_G]	PDZD8 PDZ domain containing 8	10	
	HB_G_L ≤ 112	#N/A	#N/A	
15	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [G_G]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	WBC_GIGA_L ≤ 11.800	#N/A	#N/A	
	SNP_A-1980601 = [T_T]	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	
	SNP_A-1902372 = [T_C] / [C_T]	GSN gelsolin	9	
	SNP_A-1965422 = [T_C] / [C_T]	FMNL2 formin-like 2	2	

16	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	61,02%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [G_G]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	
	WBC_GIGA_L ≤ 11.800	#N/A	#N/A	
	SNP_A-1980601 = [T_T]	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	
	SNP_A-1902372 = [T_T]	GSN gelsolin	9	
	SNP_A-2113638 = [C_C]	TPO thyroid peroxidase	2	
17	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	50,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [C_C]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
	SNP_A-1892259 = [T_T]	LIPH lipase, member H	3	
	SNP_A-2296526 = [G_G]	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily	1	

		N, member 3		
	WBC_GIGA_L ≤ 11.800	#N/A	#N/A	
	SNP_A-1980601 = [T_T]	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	
	SNP_A-1902372 = [T_T]	GSN gelsolin	9	
	SNP_A-2113638 = [T_T]	TPO thyroid peroxidase	2	
18	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1838931 = [C_C]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-2272567 = [T_C] / [C_T]	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	
19	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [C_G] / [G_C]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
20	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [G_G]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
	TRIG_MMOL_L > 0.695	#N/A	#N/A	
	SNP_A-1809622 = [A_G] / [G_A]	HMGA1 high mobility group AT-hook 1	6	
21	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%

	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [G_G]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
	TRIG_MMOL_L > 0.695	#N/A	#N/A	
	SNP_A-1809622 = [G_G]	HMGA1 high mobility group AT-hook 1	6	
	SNP_A-1997332 = [A_G] / [G_A]	GABBR2 gamma-aminobutyric acid (GABA) B receptor, 2	9	
22	SNP_A-4213932 = [C_C]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	SNP_A-2146889 = [T_T]	FOXO3 forkhead box O3	6	
	SNP_A-2114080 = [A_G] / [G_A]	KIF26B kinesin family member 26B	1	
	SNP_A-1987132 = [G_G]	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	
	TRIG_MMOL_L ≤ 0.695	#N/A	#N/A	
23	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	age > 53.050	#N/A	#N/A	
	Body_Mass_Indx > 17.150	#N/A	#N/A	
	HBA1C_PCT > 0.091	#N/A	#N/A	
24	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	age > 53.050	#N/A	#N/A	
	Body_Mass_Indx > 17.150	#N/A	#N/A	
	HBA1C_PCT ≤ 0.091	#N/A	#N/A	
	LDLCH_MMOL_L > 5.135	#N/A	#N/A	
25	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	77,78%
	age > 53.050	#N/A	#N/A	
	Body_Mass_Indx > 17.150	#N/A	#N/A	
	HBA1C_PCT ≤ 0.091	#N/A	#N/A	
	LDLCH_MMOL_L ≤ 5.135	#N/A	#N/A	
	SNP_A-2258450 = [A_A]	ANGPT2 angiopoietin 2	8	
	SNP_A-1901559 = [G_G]	STK39 serine threonine kinase 39	2	
	WBC_GIGA_L > 3.650	#N/A	#N/A	

	SNP_A-1930045 = [G_G]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-1929138 = [C_C]	TRHDE thyrotropin-releasing hormone degrading enzyme	12	
	SNP_A-1865279 = [A_G] / [G_A]	C9orf3 chromosome 9 open reading frame 3	9	
26	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	83,33%
	age > 53.050	#N/A	#N/A	
	Body_Mass_Indx > 17.150	#N/A	#N/A	
	HBA1C_PCT ≤ 0.091	#N/A	#N/A	
	LDLCH_MMOL_L ≤ 5.135	#N/A	#N/A	
	SNP_A-2258450 = [A_A]	ANGPT2 angiopoietin 2	8	
	SNP_A-1901559 = [G_G]	STK39 serine threonine kinase 39	2	
	WBC_GIGA_L > 3.650	#N/A	#N/A	
	SNP_A-1930045 = [G_G]	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	
	SNP_A-1929138 = [C_G] / [G_C]	TRHDE thyrotropin-releasing hormone degrading enzyme	12	
27	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	age > 53.050	#N/A	#N/A	
	Body_Mass_Indx ≤ 17.150	#N/A	#N/A	
28	SNP_A-4213932 = [T_C] / [C_T]	DBT dihydrolipoamide branched chain transacylase E2	1	100,00%
	age ≤ 53.050	#N/A	#N/A	
29	SNP_A-4213932 = [T_T]	DBT dihydrolipoamide branched chain transacylase E2	1	76,92%
	CHOL_MMOL_L ≤ 7.285	#N/A	#N/A	
	SNP_A-1849082 = [G_G]	NBN nibrin	8	
	SNP_A-1893931 = [G_G]	PLCB1 phospholipase C, beta 1 (phosphoinositide-specific)	20	

APPENDIX H - Relevant SNPs Information in the Tree Constructed Using Genotype Data (Representative SNPs)

SNP No	RefSNP Alleles	MAF/Minor Allele Count	Gene	Chr	Chr position
SNP_A-4213932	C/T	T=0.218/476	DBT dihydrolipoamide branched chain transacylase E2	1	100690122
SNP_A-2146889	G/T	G=0.133/291	FOXO3 forkhead box O3	6	108981196
SNP_A-1896372	G/T	T=0.148/324	DISC1 disrupted in schizophrenia 1	1	232176195
SNP_A-1849082	A/G	A=0.078/170	NBN nibrin	8	90952245
SNP_A-2118885	A/G	G=0.040/87	DDO D-aspartate oxidase	6	110733646
SNP_A-2125015	C/T	T=0.032/69	PDZD8 PDZ domain containing 8	10	119048752
SNP_A-2034790	A/C	A=0.096/209	ENPP6 ectonucleotide pyrophosphatase/ phosphodiesterase 6	4	185012101
SNP_A-1865279	A/G	T=0.046/100	C9orf3 chromosome 9 open reading frame 3	9	97766209
SNP_A-1812298	A/C	T=0.141/309	ARHGAP26 Rho GTPase activating protein 26	5	142305388
SNP_A-2083550	C/T	C=0.110/240	ANGPT2 angiopoietin 2	8	6411683
SNP_A-2114080	C/T	C=0.196/427	KIF26B kinesin family member 26B	1	245816347
SNP_A-1838931	A/G	A=0.034/74	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	80407703
SNP_A-2272567	C/T	T=0.037/82	ABCC4 ATP-binding cassette,	13	95809057

			sub-family C (CFTR/MRP), member 4		
SNP_A-1892259	A/C	C=0.068/149	LIPH lipase, member H	3	185246257
SNP_A-2272456	A/C	A=0.198/432	SLC35A3 solute carrier family 35 (UDP-N-acetylglucosamine (UDP-GlcNAc) transporter), member A3	1	100463422
SNP_A-4261255	A/C	C=0.037/80	PTPRM protein tyrosine phosphatase, receptor type, M	18	8165568
SNP_A-1987132	C/G	C=0.028/60	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	106750202
SNP_A-2296526	C/T	T=0.042/92	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	154717781
SNP_A-1980601	C/T	G=0.058/127	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	9130347
SNP_A-1902372	A/G	G=0.077/169	GSN gelsolin	9	124088240
SNP_A-1965422	C/T	C=0.136/296	FMNL2 formin-like 2	2	153497910
SNP_A-2113638	A/G	A=0.105/230	TPO thyroid peroxidase	2	1474131
SNP_A-2030266	C/T	T=0.052/113	SYN3 synapsin III	22	32934139
SNP_A-1872465	A/G	G=0.122/267	C9orf3 chromosome 9 open reading frame 3	9	97682788
SNP_A-1809622	C/T	A=0.015/33	HMGA1 high mobility group AT-hook 1	6	34213868

SNP_A-1997332	C/T	C=0.082/178	GABBR2 gamma-aminobutyric acid (GABA) B receptor, 2	9	101327048
SNP_A-4273581	A/G	C=0.147/322	MAML3 mastermind-like 3 (Drosophila)	4	140677870
SNP_A-2258450	C/T	C=0.074/161	ANGPT2 angiopoietin 2	8	6409944
SNP_A-1901559	A/G	A=0.161/351	STK39 serine threonine kinase 39	2	169087232
SNP_A-1930045	A/G	A=0.061/134	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	80418123
SNP_A-4198010	A/G	G=0.022/48	FGD4 FYVE, RhoGEF and PH domain containing 4	4	32687352
SNP_A-2164852	C/G	C=0.063/137	TLL2 tolloid-like 2	10	98148557
SNP_A-2268610	C/T	C=0.063/137	PIKFYVE phosphoinositide kinase, FYVE finger containing	2	209160659
SNP_A-2021678	C/T	T=0.073/159	DOK1 docking protein 1, 62kDa (downstream of tyrosine kinase 1)	2	74778644
SNP_A-1929138	C/G	C=0.063/137	TRHDE thyrotropin -releasing hormone degrading enzyme	12	72678994
SNP_A-2304918	C/T	C=0.057/125	SNW1 SNW domain containing 1	14	78198786
SNP_A-2075360	G/T	T=0.077/169	CAMKK2 calcium/calmodulin-dependent protein kinase kinase 2, beta	12	121675575
SNP_A-1893931	A/G	A=0.049/107	PLCB1 phospholipase C, beta 1 (phosphoinositide-specific)	20	8292683

APPENDIX I - Relevant SNPs Information in the Tree Constructed Using Genotype (Representative SNPs) and Clinical Data

SNP No	RefSNP Alleles	MAF/Minor Allele Count	Gene	Chr	Chr position
SNP_A-4213932	C/T	T=0.218/476	DBT dihydrolipoamide branched chain transacylase E2	1	100690122
SNP_A-2146889	G/T	G=0.133/291	FOXO3 forkhead box O3	6	108981196
SNP_A-1896372	G/T	T=0.148/324	DISC1 disrupted in schizophrenia 1	1	232176195
SNP_A-2118885	A/G	G=0.040/87	DDO D-aspartate oxidase	6	110733646
SNP_A-2114080	C/T	C=0.196/427	KIF26B kinesin family member 26B	1	245816347
SNP_A-1838931	A/G	A=0.034/74	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	80407703
SNP_A-2272567	C/T	T=0.037/82	ABCC4 ATP-binding cassette, sub-family C (CFTR/MRP), member 4	13	95809057
SNP_A-1892259	A/C	C=0.068/149	LIPH lipase, member H	3	185246257
SNP_A-2034790	A/C	A=0.096/209	ENPP6 ectonucleotide pyrophosphatase/ phosphodiesterase 6	4	185012101
SNP_A-2164852	C/G	C=0.063/137	TLL2 tolloid-like 2	10	98148557
SNP_A-2296526	C/T	T=0.042/92	KCNN3 potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	1	154717781
SNP_A-2125015	C/T	T=0.032/69	PDZD8 PDZ domain containing 8	10	119048752
SNP_A-1948390	C/T	C=0.127/277	LUM lumican	12	91499367

SNP_A-1980601	C/T	G=0.058/127	SEMA5A sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A	5	9130347
SNP_A-1902372	A/G	G=0.077/169	GSN gelsolin	9	124088240
SNP_A-1965422	C/T	C=0.136/296	FMNL2 formin-like 2	2	153497910
SNP_A-2113638	A/G	A=0.105/230	TPO thyroid peroxidase	2	1474131
SNP_A-1987132	C/G	C=0.028/60	ATG5 ATG5 autophagy related 5 homolog (S. cerevisiae)	6	106750202
SNP_A-1809622	C/T	A=0.015/33	HMGA1 high mobility group AT-hook 1	6	34213868
SNP_A-1997332	C/T	C=0.082/178	GABBR2 gamma-aminobutyric acid (GABA) B receptor, 2	9	101327048
SNP_A-2258450	C/T	C=0.074/161	ANGPT2 angiopoietin 2	8	6409944
SNP_A-1901559	A/G	A=0.161/351	STK39 serine threonine kinase 39	2	169087232
SNP_A-1930045	A/G	A=0.061/134	SEMA3C sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C	7	80418123
SNP_A-1929138	C/G	C=0.063/137	TRHDE thyrotropin-releasing hormone degrading enzyme	12	72678994
SNP_A-1865279	A/G	T=0.046/100	C9orf3 chromosome 9 open reading frame 3	9	97766209
SNP_A-1849082	A/G	A=0.078/170	NBN nibrin	8	90952245
SNP_A-1893931	A/G	A=0.049/107	PLCB1 phospholipase C, beta 1 (phosphoinositide-specific)	20	8292683

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Erdoğan, Onur

Nationality: Turkish (TC)

Date and Place of Birth: 12 February 1985 , Ankara

Marital Status: Single

email: onurer007@hotmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU, Medical Informatics	2012
BS	Hacettepe University, Statistics	2008
High School	75. Yıl Lisesi(Yabancı Dil Ağırlıklı)	2003

WORK EXPERIENCE

Year	Place	Enrollment
2011-Present	TÜBİTAK	Measurement&Analysis Specialist
2010 May	Yargıtay	Computer Programmer
2008 October	SAD Marketing	Mediterranean Area Manager

FOREIGN LANGUAGES

Advanced English, Beginner German

PUBLICATIONS

Günel B.D., Erdoğan O., Doğan G.H., Çağlayan Özbaş K., Tümay A. "Sistem Testlerinde İstatistiksel Süreç Kontrolünün Kullanımı", 6. Ulusal Yazılım Mühendisliği Sempozyumu, 2012

HOBBIES

Flight Simulator, Free Diving, Computer Technologies, Puzzle

METU LIBRARY

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ■

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : Erdoğan
Adı : Onur
Bölümü : Sağlık Bilişimi

TEZİN ADI (İngilizce) : Pedicting the Disease of Alzheimer (AD) with SNP Biomarkers and Clinical Data Based Decision Support System Using data Mining Classification Approaches

TEZİN TÜRÜ : Yüksek Lisans ■ Doktora ☐

1. Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. ■
2. Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullancılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐
3. Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

Yazarın imzası ..... Tarih .....