OPTIMIZATION OF WEIGHTS AND FEATURES
IN USE OF AHP FOR SNP PRIORITIZATION

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

BY

ARİF YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

JANUARY 2018

Approval of the thesis:

OPTIMIZATION OF WEIGHTS AND FEATURES IN USE OF AHP FOR SNP PRIORITIZATION

Submitted by ARİF YILMAZ in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** ____________________

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** ____________________

Assoc. Prof. Dr. Yeşim Aydın Son
Supervisor, **Health Informatics Dept., METU** ____________________

Examining Committee Members:

Prof. Dr. Hasan Oğul
Computer Engineering Dept., Başkent University ____________________

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU ____________________

Assist. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU ____________________

Assist. Prof. Dr. Can Alkan
Computer Engineering Dept., Bilkent University ____________________

Assoc. Prof. Dr. Cem İyigün
Industrial Engineering Dept., METU ____________________

Date: *16.01.2018*

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, Last name :	**Arif Yılmaz**
Signature :	________________

ABSTRACT

OPTIMIZATION OF WEIGHTS AND FEATURES IN USE OF AHP FOR SNP PRIORITIZATION

Yılmaz, Arif
Ph.D., Department of Health Informatics
Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

January 2018, 113 pages

Single Nucleotide Polymorphisms (SNP) holds a promise in identification of genomic footprints of complex diseases such as cancer and diabetes. However, identification of SNPs associated to complex diseases is a challenging problem due to the high number and variety of SNPs present in individual genomes. Analysis of genome wide studies of SNP datasets mainly focus on statistical evidence. As there are close to hundred million SNPs in human genome, incorporating biological and functional knowledge about statistically significant SNPs provides valuable features for further selection of SNPs. Analytical Hierarchy Process (AHP) based SNP prioritization approach is a method developed for this purpose. However, AHP requires expert knowledge, which results in subjective decisions. In this work, we propose a novel approach for AHP design and optimization by utilizing Random Forest based AHP (RF-AHP) assessment on categories. We utilized the results of previously developed genomic model on Prostate Cancer. Proposed RF-AHP approach was compared with Delphi-AHP based method on Schizophrenia, Prostate Cancer, Type 2 Diabetes and Alzheimer's disease genomic datasets and same performance was achieved. Additionally, RegulomeDB database was integrated to RF-AHP. While similar performance was obtained in most of the datasets better prioritization scoring is achieved for Schizophrenia disease.

Keywords: SNP Prioritization, Analytic Hierarchy Processing, Random Forest, Prostate Cancer, Type 2 Diabetes

ÖZ

SNP ÖNCELİKLENDİRME AMAÇLI AHP KULLANIMINDA AĞIRLIKLARIN VE ÖZNİTELİKLERİN ENİYİLENMESİ

Yılmaz, Arif
Doktora Tıp Bilişimi Bölümü
Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son

Ocak 2018, 113 sayfa

Tekil Nükleotid Polimorfizmleri (SNP), kanser ya da tip 2 diyabet gibi karmaşık hastalıkların tespitinde umut vadetmektedir. Bununla birlikte karmaşık hastalıklarla ilişkili SNP'lerin tespit edilmesi bireylerin genomlarındaki çok sayıdaki ve değişkenlikteki SNP'ler nedeniyle zorlayıcı bir problemdir. SNP veri setlerinin genom çapında ilişkilendirme çalışmalarında çoğunlukla istatistiksel bulgular üzerinde odaklanılmaktadır. Bununla birlikte, bir insan genomunda yaklaşık yüz milyon SNP bulunmaktadır. İstatistiksel olarak anlamlı SNP'lerle ilgili biyolojik ve işlevsel bilgilerin eklenmesi daha ileri SNP seçimi için önemli özellikler sağlamaktadır. Analitik Hiyerarşi İşleme (AHİ) temelli SNP önceliklendirme tekniği bu görevi yerine getirmek amacıyla geliştirilmiştir. Fakat AHİ'nin uzmanların deneyimlerine ihtiyaç duyması özniteliklerin seçiminde ve ağırlıklarında öznel kararlara neden olmaktadır. Bu çalışmada AHİ tasarımı ve eniyilemesi için Rastgele Orman tabanlı AHİ (RO-AHİ) kategorilerinin ağırlık ve öznitelik belirleme yaklaşımı önerilmektedir. Bu amaçla Prostat Kanseri üzerinde daha önceden yapılmış olan çalışmalar sonucunda geliştirilmiş olan genomik model kullanılmıştır. Geliştirilen yöntem, Şizofreni, Prostat kanseri, Tip 2 Diyabet ve Alzheimer Hastalığı genetik veri setlerinde Delphi AHİ tabanlı bir yöntem ile karşılaştırılmış ve aynı başarıma ulaşılabilmiştir. Ek olarak, RegulomeDB veritabanı da RO-AHİ ye eklendiğinde Şizofreni hastalığı ile ilgili daha iyi sonuçlara ulaşılmış, diğer hastalıklar ile ilgili aynı başarım sonuçlarına ulaşılmıştır.

Anahtar Sözcükler: SNP Önceliklendirme, Analitik Hiyerarşi İşleme, Rastgele Orman, Prostat Kanseri, Tip 2 Diyabet

To My Family

ACKNOWLEDGMENTS

I would like to thank and express sincere gratitude to my supervisor Assoc. Prof. Yeşim Aydın Son. She has guided this work with her knowledge, professionalism, positivity with a lot of patience and excellent support.

I also would like to thank to my follow-up committee members Assoc. Prof. Cem İyigün and Assist. Prof. Can Alkan for their guidance and accurate feedbacks.

I am grateful to Dr. Gürkan Üstünkar for his work during his Ph. D. thesis results of which helped me during my work.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Alzheimer’s Disease
AHP	Analytic Hierarchy Process
BMI	Body Mass Index
CA	Cancer
CART	Classification and Regression Tree
CDS	Coding Sequence
dbGaP	Database of Genotypes and Phenotypes
DNA	Deoxyribo-Nucleic Acid
FDM	Fuzzy Delphi Method
GAD	Database of Genetic Association of Complex Diseases
GWAS	Genome Wide Association Study
HGP	Human Genome Project
MAF	Minor Allele Frequency
MCDM	Multi Criteria Decision Making
MDA	Mean Decrease in Accuracy
MDI	Mean Decrease in Impurity (Gini)
NCBI	National Center of Biotechnology Information
NHRI	National Human Genome Research Institute
OMIM	Online Mendelian Inheritance in Man
OOB	Out Of Bag Data
PCa	Prostate Cancer
PSA	Prostate Specific Antigen
SVM	Support Vector Machine
QTL	Quantitative Trait Locus
Sz	Schizophrenia
RF	Random Forest
RF-AHP	Random Forest based Analytic Hierarchy Process
RNA	Ribo-Nucleic Acid
SNP	Single Nucleotide Polymorphism
T2D	Type 2 Diabetes
UTR	Untranslated Region
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

1.1 Motivation

Human genome may be represented as a sequence, which consists of 3.3 billion letters, each representing a single nucleotide. The nucleotides are biomolecules, which are symbolized by one of the A, C, G, T letters [1]. In human genome, 99% of the nucleotide composition is identical. The remaining 1% consists of the variations which are basis for the differences between individuals. If a variation at a nucleotide locus is observable in at least 1% of the population, this variation is called a Single Nucleotide Polymorphism (SNP). SNPs are very efficient biomarkers, as they can be associated with many complex diseases by using statistical or intelligent methods.

Variant data enables research of complex diseases such as diabetes, cardiovascular diseases, neuro-degenerative diseases and cancers [2][3]. Genome Wide Association Studies (GWAS) can be designed as case-control, cohort or trio study. It is used to identify the statistically associated SNPs with complex diseases by investigating millions of SNPs in a single experimental set-up. In case-control studies, which is the most widely used type, statistically significant variations differentiating between case and control samples are found.

Nonetheless, the number of statistically significant SNPs that should be inspected still reaches over tens of thousands. Hence, after GWAS, prioritization of significant SNPs according to its biological relevance, and other prior information is required.

Decision making techniques are shown to appropriately provide a solution to the prioritization problem of the candidate SNPs according to the complex criteria [4].

The objective of this thesis is to propose Random Forest based AHP (RF-AHP) method to address expert judgment uncertainty in decision making with AHP. It is accomplished by training the analytical hierarchy process input data using Random Forest machine learning method [5]. So, the AHP categories are evaluated according to calculated Variable Importances in the trained Random Forest model.

The RF-AHP method offers pairwise comparisons of categories without any requirement of expert knowledge. Consequently, all of the criticisms of AHP related to judgement, subjectivity, and uncertainty and imprecision is avoided. A case study related to Single Nucleotide Polymorphism (SNP) Prioritization was performed to demonstrate the proposed approach. As a supervised learning algorithm, Random Forest (RF) method was employed to evaluate the importances of AHP categories. In evaluation of AHP categories using RF-AHP for SNP prioritization, three types of data source were used. First type of data source is the database existing in METU-SNP software developed previously in METU-BIN laboratory [4]. Second type of data source is the genomic model for a complex disease. Results of a previously published study [6] on Prostate Cancer was used as the genomic model in this study. Third type of data source is the GWAS disease datasets from the literature. Four disease datasets were used in analysis and comparison tests; Schizophrenia, Prostate Cancer, Type 2 Diabetes, and Alzheimer's disease. Prostate Cancer was used in training of Random Forest machine learning algorithm for AHP Category Evaluation. However, for performance evaluation all of the four disease datasets were used. Additionally, we have integrated RegulomeDB scores into RF-AHP based prioritization in the latest version of METU-SNP.

Findings: We have compared our results with Delphi AHP and found out that some categories were uninformative and may be removed from AHP hierarchy. Consequently, a much simpler AHP tree was obtained providing same or better performance without any requirement for expert judgment. After incorporation of RegulomeDB, better results obtained for Schizophrenia.

Originality: The proposed Random Forest based category evaluation method may be used in calculation of weights of AHP categories without requiring experts. To the best of our knowledge, at the time of writing the thesis, random forest based method for AHP category evaluation was not introduced in the literature.

1.2 Thesis Organization

This thesis consists of six chapters. In Chapter 1, i.e. this chapter, motivation and brief introduction to basic concepts is presented. Single Nucleotide Polymorphisms

(SNP) and Genome Wide Association Studies (GWAS) and SNP prioritization are briefly explained. Analytic Hierarchy Process decision making technique for SNP prioritization, its advantages and disadvantages are listed. The proposed Random Forest based Analytic Hierarchy Process to overcome subjectivity problem is mentioned.

In Chapter 2, theoretical background for the concepts referenced in this thesis is provided. Firstly, basics of molecular biology science and genomic variations is presented. Later, the GWAS and SNP prioritization is explained. Theory of decision making with Analytic Hierarchy Processing is provided. Then theory of Random Forest machine learning algorithm is described.

In Chapter 3, materials and methods employed to realize the proposed algorithms are presented. Detailed explanation of utilized data sources and their use is provided. Then software environment for developing the methodologies is outlined.

In Chapter 4, theory of proposed Random Forest based Analytic Hierarch Process method using variable importances is presented. Its application for SNP prioritization is provided in detail. Incorporation of RegulomeDB database into RF-AHP is described. Integration principles and transformation of RegulomeDB to RF-AHP based SNP prioritization is detailed.

In Chapter 5, results on realized RF-AHP and RegulomeDB integrated RF-AHP are presented. Comparison of Delphi AHP based method to RF-AHP method and RF-AHP with RegulomeDB is explained. Then the performance of three methods are compared with respect to various performance plots. Later, discussion of the results is presented.

In Chapter 6, conclusions of presented RF-AHP method and possible future improvements are discussed.

CHAPTER 2

BACKGROUND

Bioinformatics is a research and technology domain which aims to interpret molecular biology via computational techniques such as machine learning, statistics and high performance computing. It is an interdisciplinary field that relates biology and biomedical sciences to statistics and computer science [7]. Recent developments in the bioinformatics field drives revolutionary approaches to problems in various other domains such as health [8][9][10], agriculture [11], genetic engineering [12], biology [13], medicine [2], pharmacology [14], sociology [15][16] i.e. all dimensions in the universe where life exists. Bioinformatics is subdivided to many “omics” fields such as genomics, proteomics, and transcriptomics. These research are enabled by employing molecular biology and computational analysis algorithms on smallest components such as nucleotides, genes, amino acids, proteins and metabolites. Bioinformatics itself is also being revolutionized by new developments in biotechnology and data sciences [17]. This has paved the way for higher throughput and cheaper sequencing technologies. With emerging big data processing technologies such as cluster computing, artificial intelligence and data mining techniques, it is possible to make detailed investigation on omics data.

Recently, one of the mostly focused fields in Bioinformatics is on understanding the meaning of the vast amount of data, identify molecular basis of genetic diseases and develop personalized medicine approaches [3][18][19][20]. The purpose of this thesis is such an effort in that research field. For this purpose, related background knowledge on basics of molecular biology, etiology of genetic diseases and genome wide association studies and related computational techniques are presented as follows.

2.1. Molecular Biology, DNA Replication, Transcription and Translation

DNA is made of nucleic acids which are formed as a double stranded and twisted to the shape of double helix made of nucleotide pairs. It is constructed with four different types of nucleotides namely Adenine, Cytosine, Thymine and Guanine. These nucleotides represented by letters A, C, G, and T [1]. There are about 3 billion base pairs in human genome. A genome is made of chromosomes. Diploid organisms have pairs of chromosomes that are paternal (inherited from father) and maternal (inherited from mother) chromosomes. For instance, human genome consists of 23 maternal and 23 paternal chromosomes. The central dogma of Molecular Biology explains replication, transcription and translation [21] as shown in Figure 1. Replication is the process of duplication of a DNA during cell division as shown in Figure 2. It begins with untwisting the chromosome [22]. Then DNA is unzipped by enzyme Helicase by breaking of hydrogen bonds between nucleotides and opening a replication fork. In elongation phase, the two separated strands work as a template for the nucleotides. An enzyme called DNA polymerase bonds the complementary nucleotides to two strands. Here, A is the complementary for T, similarly, C is complementary for G. DNA polymerase can operate only in 5’ to 3; direction. Therefore, the polymerization and duplication process always continues in 5’ to 3’ direction on DNA. During replication, the leading strand nucleotides are bond in forward direction continuously. However, in lagging strand, nucleotides are bonded in reverse direction in small sequences which are called Okazaki frames. Each small fragment is joined to previous fragment with enzyme DNA Ligase (Figure 2).

During DNA replication, enzyme called Exonuclease checks and proofreads the bases and if there is an error in A-T or C-G matches, finds and corrects them. After replication complete telomere sequences are bonded the both ends of DNA by enzyme Telomerase. 99% of chromosome is non-coding regions with mainly regulative function. Remaining regions are named exome, where genes are located. Gene is a sequence of DNA that produces a specific protein for a specific function. A gene maps to specific genetic locus on a chromosome as shown in Figure 3.

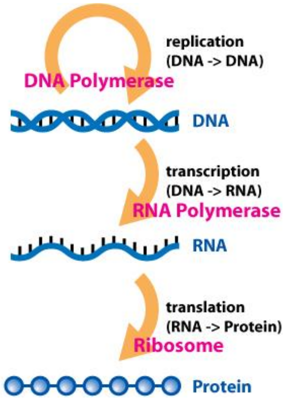


Figure 1 Central Dogma of Molecular Biology: Replication, Transcription, Translation (Credit: Creative Commons License).

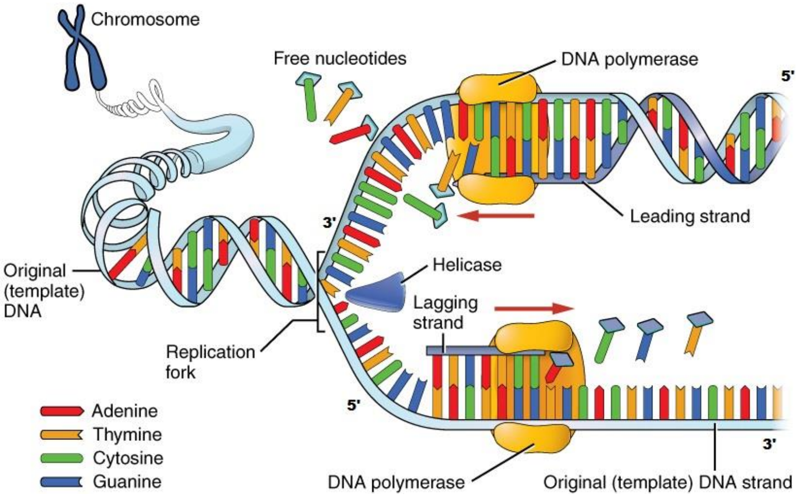


Figure 2 DNA Replication process (Credit: Creative Commons License).

In **transcription**, the Deoxyribo-Nucleic Acid (DNA) corresponding to a gene is rewritten as Ribo-Nucleic Acid (RNA). Later RNA is processed by splicing. In splicing, non-coding parts (introns) are excised, and remaining coding regions (exons) are combined as the messenger RNA (mRNA). **Translation** is the coding of polypeptides as amino acids according to messenger RNAs as shown in Figure 4. Amino acids make up of polypeptides as shown in Figure 5. Proteins are composed of polypeptides. They perform various functions in different biological processes, and define the characteristics of organisms.

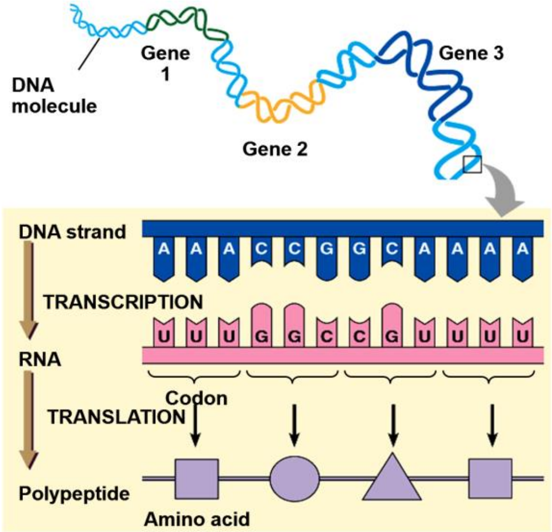
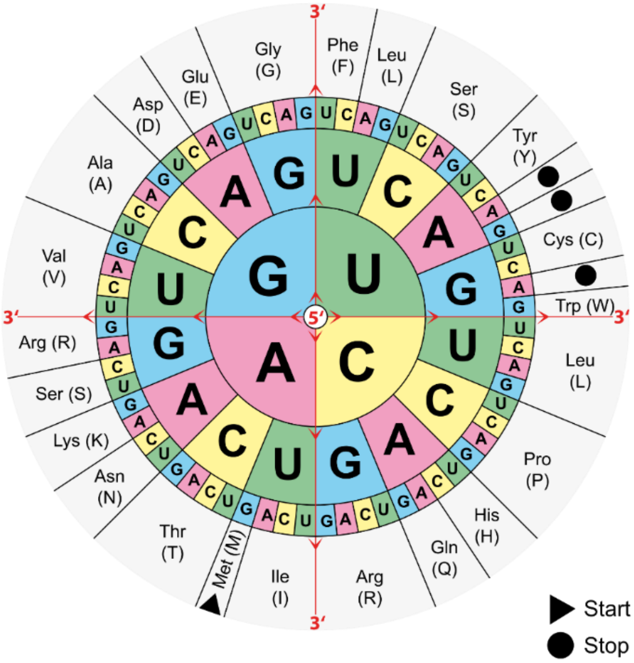


Figure 3 Transcription and Translation of DNA nucleotides on a chromosome (Credit: Creative Commons License).



e

Figure 4 Amino Acid Triplet Codes (Credit: Creative Commons License).

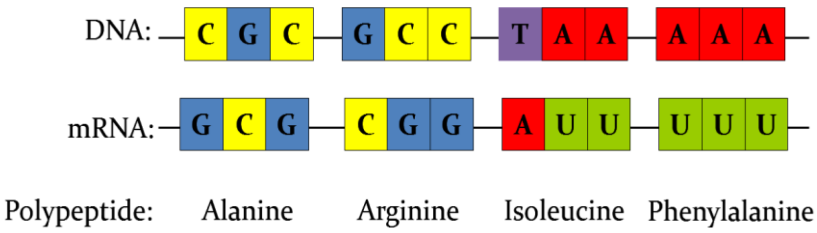


Figure 5 DNA-mRNA-Polypeptide relation. (Credit: Creative Commons License)

2.2. Mutations

Changes that occur in DNA sequence is called mutations. This may occur as a result of natural processes such as binding an incorrect base during DNA replication or due to external factors. The external factors causing mutations are called mutagens[23]. A mutagen may be a chemical material such as poison, tobacco, pollution or a physical event such as UV light or X-Ray radiation that causes a genetic change in DNA [24][25]. A mutation may affect small locus in a gene or large portion of a chromosome. Some of the most frequently seen mutations include [26]:

Substitution: When a nucleotide is replaced by another nucleotide, it is called substitution.

Deletion: A nucleotide is removed from the sequence. This may cause a shift in transcription of mRNA, therefore in amino acid synthesis all amino acids may be different.

Insertion: A nucleotide is inserted into a sequence. Result of insertion is similar to deletion and may cause large difference in synthesized amino acids.

If a mutation affects single amino acid it is called point mutation. One or a few nucleotides are changed in a DNA sequence [27]. Insertions and deletions may cause frameshift mutations if amount of change is not multiple of three bases. Drastic changes may occur after the transcription and translation of a sequence. Codons are translated to amino acids in triplets, therefore, one shift in the sequence may result in completely different amino acids. Chromosomal mutations on the other hand are those who affect all chromosome for instance loss or gain of chromosome duplication, deletion or insertion in structure of chromosome. According to result of the mutation in amino acid, they are organized under following groups [28][29]:

Silent or Synonymous Mutation: A mutation that does not alter the produced amino acid. The resulting protein is the same as that produced before mutation.

Nonsynonymous Mutation: A mutation that results in alteration of produced amino acid. Then resulting polypeptide and protein is going to be different. Types of non-synonymous mutations are as follows:

Missense: A single base substitution (non-conservative substitution) causes different coding in amino acid production resulting in different protein synthesis.

Nonsense: A substitution causes introduction of a stop codon. This results in early termination in polypeptide string. The shorter polypeptide string causes loss of normal biological activity.

Frameshift: As mentioned above, this may be result of insertion or deletion of nucleotides in DNA sequence.

If mutations occur in a gamete, they are passed to the next generation i.e. offsprings. Otherwise, the mutations remain in the individual, which are called somatic mutations.

2.3. Polymorphism

If a variation is observed in at least 1% of a population, it is called as polymorphism [30][31]. Variations, in which only one nucleotide is different from the population is called Single Nucleotide Polymorphisms (SNP) [32]. SNP is very common type of polymorphism. SNPs are the most common genomic polymorphisms; in average a SNP can be observed in every 300 nucleotide of coding regions and in every 1000 nucleotide of non-coding genome [33][34]. The most frequently observed nucleotide form of the SNP is called as major allele. Likewise, the allele with lowest frequency is called as minor allele.

Inheritance of closely mapped alleles are linked to each other as a group. These genetically linked alleles that are on closely linked locations on genes or chromosomes are called haplotypes. Haplotypes are likely to be inherited together. This non-random linkage between the alleles is called Linkage Disequilibrium (LD).

As of 2018, 10 million common SNPs have been identified in human genome. As shown in Figure 6, according to location and nature of change, SNPs cause different outcomes at biological and phenotypic level. Over 60% of all SNPs reported in dbSNP are exonic variations, where majority are missense, indicating a change in the amino acid code. These changes can have varying level of effect from individuals’ phenotype such as eye color, hair type etc. to susceptibility to diseases such as cancer, or complex diseases. Therefore, SNPs are highly promising biological markers in research of molecular genetic origins of disease risk, susceptibility, or response to treatment. Through understanding of disease causing changes in genome, molecular etiology of diseases can be revealed, by identifying genes, biological pathways, and other biological interactions.

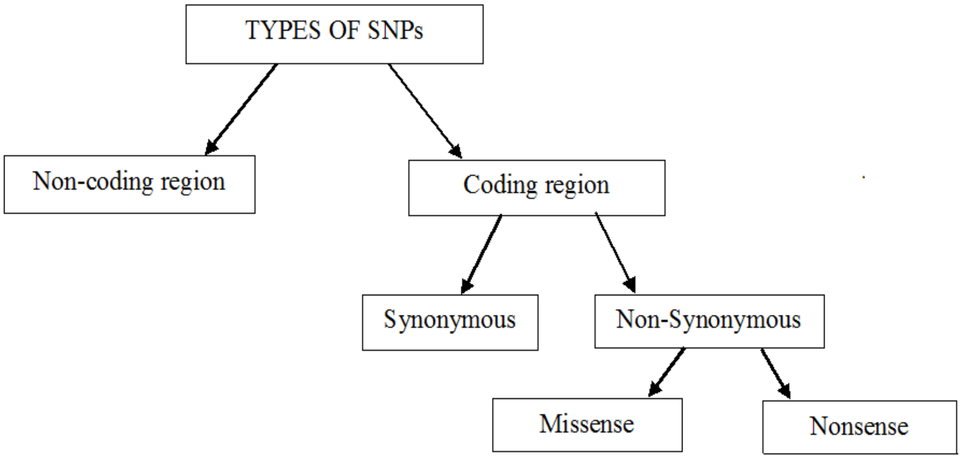


Figure 6 Types of SNPs (Credit: Creative Commons License).

2.4. Regulation of Gene Expression

On DNA molecule, genes are the loci that code proteins [35]. However, just before the coding sequence of gene on DNA, a region called the promoter region exists as shown .in Figure 7. The promoter region may be as short as a few nucleotides or as long as a few hundred nucleotides. The proteins bind to these regions to regulate the transcription [36]. Therefore, if the promoter region is long, more proteins bind to the region, and gene expression is more controlled. General transcription factors such as TFIID, assemble the transcription initiation complex on the promoter [37]. Then RNA polymerase binds to transcription initiation complex in promoter to initiate transcription. Other special transcription factors may also bind to promoter to regulate the gene expression for specific genes.

As seen in Figure 7. There are enhancer regions outside promoter that help enhance the transcription process [38]. Transcription factors also bind to enhancers. When DNA bending proteins bind, the DNA bends, the activators bind to enhancers and transcription factors, to increase gene expression. There are also transcription repressors that prevent transcription. They bind to promoter and stop to RNA polymerase effectively blocking the transcription of gene. Activators and repressors respond to external effects to prevent the binding of transcription factors.

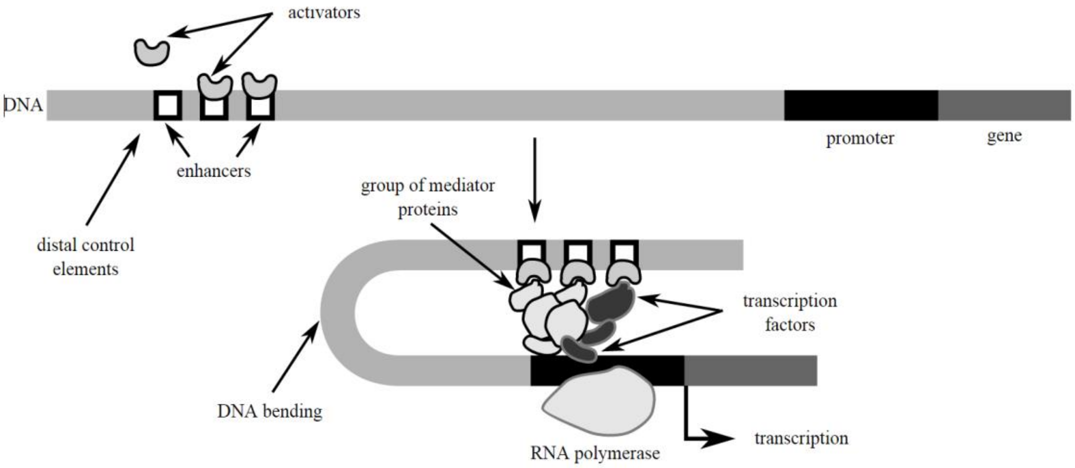


Figure 7 Operation of Transcription Factors in Gene Transcription Process (Credit: Creative Commons License).

This process is regulated according to cell type and many involved proteins [39]. Variation in these regulatory processes has an important role on activation, deactivation and expression of a gene [40][41]. For instance, in [42], Gobbi et al. identified a SNP that introduces a new promoter element which interferes with activation of alpha-like globin genes that results in a form of a blood disorder Alpha Thalassemia. Similarly, in [43], Zhou et al. analyzed the effects of a SNP which is on the promoter of GRK3 that causes Bipolar Disorder.

The Encyclopedia of DNA Elements project aims to systematically maps the regions in human DNA [44]. These regions include transcriptions, transcription factors, and chromatin structure and histone modifications. These mappings assign 80% of genome to biochemical functions. These do not include protein coding regions, and correspond to regulatory functions for 147 different cell types. The datasets available from ENCODE project is compiled and annotated according to various requirements. RegulomeDB is such a database that aims to annotate variants in gene regulating loci [45].

2.5. SNP Profiling

In order to analyze the genetic foundation or characteristic of a trait or a disease, various variants are determined through genotyping. This process is called as variant profiling. For instance, in SNP profiling variations at specific SNP loci are measured. SNPs are found to be involved in etiology of many complex diseases. Therefore, they are used as markers in disease association studies. Additionally, a combination of specific SNPs is useful as a genetic method for identity testing. There are various SNP profiling techniques.

Hybridization (microarray) based Techniques: It is based on binding a primer to a DNA target sequence. A RNA or cDNA is labeled with fluorescent. The labeled targets are then hybridized on microarray surface which contains hundreds of probe sequences at different position. When a target hybridized at a probe position, fluorescent intensity at that location is high meaning the labeled target is expressed [46].

Polymerase Chain Reaction (PCR): It is an easy to use method. A template DNA molecule is amplified by using DNA primers and enzyme DNA polymerase and DNA nucleotides a mixture is prepared [47]. By cycling the temperature between two temperatures, hydrogen bonds between strands are broken and restored. During this DNA polymerase hybridizes the primer to template strand. It is more accurate then NGS or microarray based techniques and mostly used type for DNA sequencing.

Next Generation Sequencing (NGS): It enables massively multiplexed sequence processing of more than ten million nucleotides at once [48].

2.6. Genetic Diseases

Genetic diseases may be classified as Chromosomal diseases, Mendelian diseases and complex diseases [49].

Chromosomal Diseases: They are caused because of chromosomal abnormalities such as lack of a chromosome, a non-disjunction of chromosome etc.

Mendelian diseases: These are also called monogenic diseases. They are caused by one gene mutations inherited from parents. Hemophilia and color blindness, cystic fibrosis are such diseases.

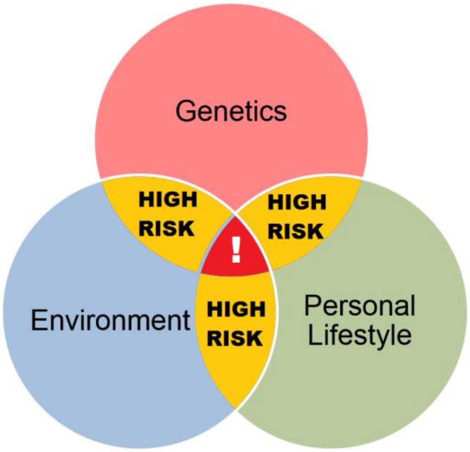


Figure 8 Risk factors for complex diseases [50].

Complex Diseases: These diseases are caused by many genetic factors with small effects as well as environmental factors and personal lifestyle [50][51]. The relation between these factors may be seen in Figure 8. Genetics, Environment and Personal lifestyle pose a risk of occurrence [52]. When any two factors come together, risk is high. If three factors come together the risk level is critical. Genetic factors define the risk of occurrence of a disease because of genetic variations in the individual's genome. Personal life style and environmental factors affect the prevalence of the disease in time, earlier or later. For instance, in [53], Prostate Cancer disease model was developed. As a result, 108 SNPs were found to be associated with the disease. Additionally, BMI exercise and smoking was identified to be associated life style phenotypes.

Complex Traits: Environmental and genetic factors may affect a phenotype occurrence [54]. This type of trait is called as complex trait [55]. Large percent of diseases that damage health status of human are complex traits [56]. In [57], genetics only associations of complex traits to loci is performed using disease data for 42 traits. As a result, 392 loci were identified. Moreover, finding the factors such as genetic loci that affect a complex trait as well as the environmental factors were studied [58]. For instance, in eye related disease called Age Related Macular Degeneration (AMD) disease, age is a major influence. However in [59], researchers identified BMI and smoking affect probability prevalence of AMD disease.

Epistasis: In complex traits, effect of a polymorphism may be dependent on another polymorphism elsewhere in the genome [60][61][62]. When the mutational effect size is large, the complex trait or disease consist of many variations with small interactions [63]. These interactions between polymorphisms are called epistasis [64]. In epistasis, effect of one gene or locus is dependent on other genes or loci.

Pleiotropy: If a genetic locus is associated with multiple trait, it is called pleiotropy [65][66]. Genetic variants may be related to multiple traits and distinct traits [67]. In literature there are reports that identified loci that are associated with multiple traits [68][69]. In [57], 42 complex traits were identified. These are called as cross-phenotype associations. Cross-phenotype associations may be due to pleiotropy. Pleiotropy may be caused by single locus or single region. If a phenotype is related to another then a change in first phenotype causes change in the other phenotype. This is called mediated pleiotropy. However, sometimes the interaction due to other factors may be associated falsely as cross-phenotypes causing spurious pleiotropy [70]. Some variants causing cross-phenotypes are protein coding variants, splice site variants, mutations in the protein coding genes and intergenic regulatory elements. They may have a major role in pleiotropy because they may cause deregulation of hundreds of target proteins.

2.7. Genome-Wide Association Study

Searching for statistically significant variations by analyzing whole genome and identification of the genetic differences that result in differences between individuals is called Genome Wide Association Study (GWAS) [71]. The resulting differences may be a phenotype variation, disease or response to a drug e.g., traits like height, blood pressure, complex diseases like cancer, penicillin intolerance. The common variations for a particular disease may be identified by statistically analyzing the occurrences of genetic markers in the disease [72]. For this purpose, the variations related to a trait may be checked in the genome of an individual. If certain group of variations causes the disease in an individual, then it must not exist in healthy person.

The number of variations related to other traits prevents easy identification by just comparing all genomic locations. In order to perform GWAS to identify significant variations one of following types of study configuration are performed [73]:

Trio Study: Data of the subject with the disease is studied against the genome of his/her mother and father. The differences in genotypes between the parents who do not have the disease is compared to the child with the disease. The variation between parents and child is expected to be responsible of the disease.

Cohort Study: For the research of a phenomenon, a population with specific trait and another without that trait is observed for a long time then the result and underlying reasons are analyzed.

Case-Control Study: The population with the disease or the trait are called cases. Similarly, population without the case i.e. health individuals are called as controls. Populations are collected to discover common non-random differences between all cases and all controls to identify the cause of the disease. Because the data is collected from readily performed genotyping of cases and controls, the analysis may be completed in short time. About thousands of samples may be used in case-control studies, therefore the genotyping errors in individuals are not important.

In GWAS, most used type of variation is SNP because of the high information content, ease of genotyping and lower cost. Most preferred type of GWAS is Case-Control Study due to use of already existing data from cases and controls. The amount of case and control subjects are usually more than thousand to obtain enough statistical power. This results in large amount of data to be processed. Statistical tests are applied to SNPs to identify SNPs that have differences in allele frequencies between cases and controls. After these SNPs known, the disease may be easily identified later in other subjects by checking the identified SNPs. Later drugs may be developed targeting the disease before it advances.

However, in all methods of GWAS, a large amount of data in order of terabytes has to be processed. This requires complex computations and algorithm along with high performance processors as well as large storage resources. Fortunately, for GWAS there is no need to process all the genomic data for sequencing and extraction of SNPs for subjects. In the Human Genome Project all of sequencing and extraction of SNPs are performed and these findings are publicly available in various databases. Some information about these databases are as follows:

HapMap Project: This project aims to identify and map nearby SNPs as blocks that are inherited together. These blocks are accessible by representative tag SNPs. 1 million SNPs in human is mapped to 500 thousand tag SNPs. By using tag SNPs, the project aims to catalog the similarities and dissimilarities in human. The catalog variants are used to link the relations between variants, genes and diseases. Then it would be used to produce medications targeting the disease or vaccines that will be most effective in individuals.

The dbSNP Project: This database keeps polymorphisms such as SNPs insertions and deletions. Additionally, SNP-gene, chromosome and SNP-disease relations are available in the database. It is publicly available at NCBI. In dbSNP about 45 million SNPs are identified.

The dbGaP Project: Also developed by NCBI, this database contains the relations between genotypes and phenotype information. The dbGaP database is essentially a large repository keeping ordered data from many GWAS studies, genetic data of the samples in the studies as well as their phenotypes. In the database, previously performed studies of various types such as cohort, trio or case-control studies are presented in well-organized form for other researchers.

RegulomeDB Project: This database stores SNP–regulatory element relations in non-coding genetic locations of Human Genome Project. It incorporates many datasets such as ENCODE and eQTL.

Online Mendelian Inheritance in Man (OMIM) Project: This project keeps a catalog of Mendelian diseases. Mendelian Disease can be described by genomic errors versus gene functions as a complex disease. NCBI hosts the OMIM database that contains Mendelian disorders for about 12,000 genes.

Genetic Association Database (GAD) Project: In GAD, data about genes and diseases are collected from academic literature and presented in gene based format. GAD is also available from NCBI.

Using the databases listed above, biomarkers such as SNPs related to diseases or traits may be identified using statistical or functional approaches. In statistical analysis, statistical significance of each variation i.e. non-randomness between control and case subjects is calculated. For instance, in a population with 1000 cases and 1000 controls, the significance of variations are calculated. This analysis includes millions of SNPs obtained from the databases above. Then, most significant SNPs are obtained by ranking them according to significance. In significance calculations, the difference of allele frequencies between cases and controls are used for association.

2.8. Statistical Analysis using PLINK

In calculation of statistical significance there are various tools. Plink is a well-known open source software developed by Shaun Purcell [74] for genome wide association and statistical identification of significantly differential SNPs between cases and controls. It consists of following features:

1. Data management,
2. Summary statistics,
3. Population stratification,
4. Association testing,
5. Haplotype testing,
6. Meta-Analysis

In significance analysis, association testing feature is used. In association testing, plink calculates the frequencies of alleles separately for cases and controls. Here the calculations are performed according to Fisher's Exact Test and T Test. In order to perform association test, .bed, .bim and .fam files are required. After plink analysis, the calculated significances are obtained as a specially formatted ".assoc.adjusted" file. The file is a tab separated file that contains fields such as chromosome number, SNP rsID, unadjusted asymptotic significance as p-value, Genomic Control (GC) adjusted significance, Bonferroni adjusted significance, Sidak single-step adjusted significance, Sidak step-down adjusted significance, step-up Benjamini-Hochberg False Discovery Rate (FDR) control, and step-up Benjamini-Yekutieli FDR control. Basically, as a result of association analysis, unadjusted asymptotic significance i.e. p-value is used. After obtaining association test results as unadjusted significance values, the prioritization of SNPs is performed. In biostatistics, SNPs with p-value<0.05 are usually considered significant. In genome-wide studies a multiple test correction for up to 1M samples should be considered, so a p-value $< 10^{-5}$ is usual threshold for GWAS [75]. However, the SNPs should be considered according to their genetic function and biological features. Therefore, considering p-value alone is not enough and SNP prioritization step should be executed after GWAS analysis [76].

2.9. SNP Prioritization and Candidate SNP Selection

In spite of being scored with high statistical significance value, not all small effects causing a complex disease may be biologically relevant variations. Most of the SNPs in the SNP databases have no known disease -related results. Prioritization of SNPs is essential in accurate disease SNP detection [77]. Prioritization is performed according to additional knowledge such as meta-data or annotations about SNPs that provide functional information [72]. SNP prioritization process consists of three stages as shown in Figure 9.

At first stage, SNP data quality control is performed according to minor allele frequency, missing values and Hardy-Weinberg equilibrium criteria. In the second stage, each SNP's multiple testing adjusted p-values of association are calculated. A significance threshold is set and only statistically significant SNPs are inspected for prioritization instead of inspecting millions of SNPs. Otherwise processing all of the SNPs to discover the ones associated with a disease is a very demanding task requiring large processing power and appropriate selection technique. In the third stage, statistically significant SNPs are prioritized according to additional important features such as SNP location, associated gene, associated disease, pathway etc. as well as p-value. There are various tools in the literature for SNP prioritization such as SPOT [78], SNPLogic [79] or Fast SNP [80] and METU-SNP [4].

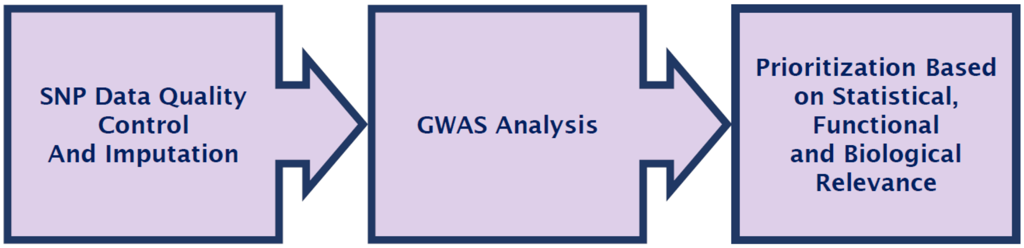

SNP Data Quality Control And Imputation
GWAS Analysis
Prioritization Based on Statistical, Functional and Biological Relevance

Figure 9 Outline of generic SNP prioritization process.

In prioritization, statistically significant SNPs below a threshold value, for example SNPs with significance p-value 0.05 or lower, are further analyzed with respect to biological facts [81]. In this stage various features from domain knowledge is integrated into associations to discover the SNPs that are really associated to investigated trait or disease. For instance, linkage disequilibrium is a strong association between SNPs at neighboring loci and may be given importance. It is also possible to weight the genes according to their significance by using all SNPs on the genes [82]. It may be done by calculating combined p-values for SNPs on a gene using methods such as Fisher's combination test as follows:

$$p_{gene(combined)} = -2 \sum ln\, p_i \qquad \textbf{(EQUATION 1)}$$

Where $p_{gene(combined)}$ is the combined p-value of the gene and p_i is the p-value for i_{th} SNP on the gene for which combined p-value is calculated. Similarly, it is also possible to obtain significance of pathways in which genes exists by calculating Fisher's exact test as follows:

$$p_{pathway(combined)} = 1 - \sum_{i=1}^{K} \frac{\binom{s}{i}\binom{N-s}{m-i}}{\binom{N}{m}} \qquad \textbf{(EQUATION 2)}$$

Here, $p_{pathway(combined)}$ is the combined p-value of the pathway, N is the total number of genes, s is the number of genes associated to disease, m is the number of genes on the pathway and K is the number of significantly associated genes in the pathway. In addition to statistical evidences, biological and functional facts about knowledge such as genomic location, disease association are considered in prioritization of SNPs. The resulting list of SNPs from prioritization is called Candidate SNP list.

2.10. Decision Making and Analytic Hierarchy Process

Multi Criteria Decision Making (MCDM) is the process of making decisions, which requires consideration of number of subjective factors. Analytic Hierarchy Process (AHP) is one of the frequently used MCDM methods [83]. It has gained popularity over the years especially in the fields of management, engineering and medicine [84][85][86][87]. The outline of AHP is shown in Figure 10. One of its advantages is its ease of use. Simple pairwise comparisons based upon the judgments of experts are required to derive priority scales [88][89]. Performing pairwise comparisons allows decision makers to assign weights to coefficients and compare alternatives. Because it provides hierarchical view to problem structure, additions and improvements to a model is easy. Since its introduction in 1980's it was employed in many areas such as economics, resource management, corporate policy and strategy, public policy, political strategy, planning [90][91].

AHP has been extensively used in decision making problems in many fields literature. For instance, some researchers studied AHP in contracting company selection [24][92], product quality evaluation [93], plant or facility location selection[94], game design factor evaluation [95] and search engine evaluation [96], Risk Assessment Modeling for security of cross border gas pipeline [97] and Single Nucleotide Polymorphism Prioritization after Genome Wide Association Studies in Bioinformatics [4].

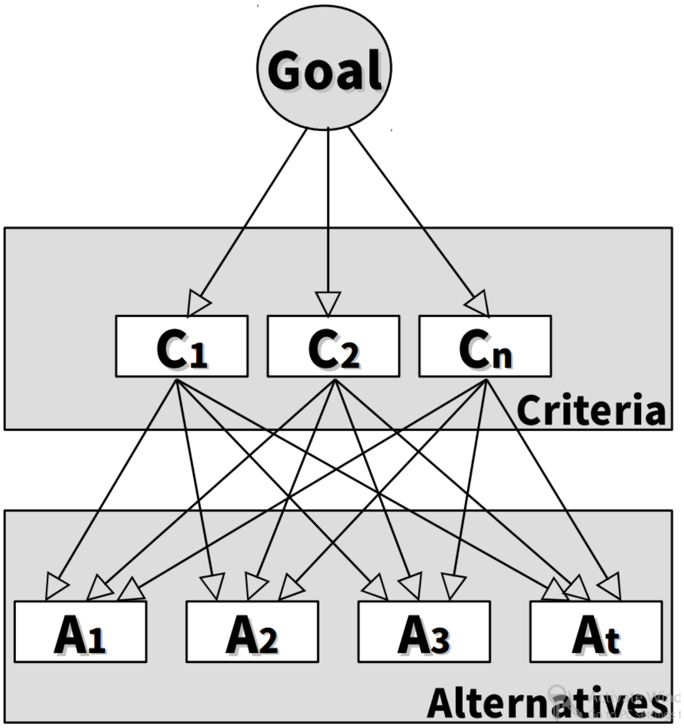


Figure 10 Outline of generic AHP Hierarchy Model.

AHP requires assigning weights to the investigated categories. In order to set these weights expert opinion is required. However, the judgement based on the expert evaluations may be subjective, incorrect or not precise. Improper weighting of the AHP categories may result in complete failure in the decision making process. Moreover, the detection of related categories for the problem is another subjective task. Literature search and expert knowledge is also required to obtain the relevant categories to provide proper decision making capability for the AHP about the problem. For this purpose questionnaires are provided to the experts to select between important categories or add new ones [98][99][100]. As these methods do not provide a complete solution to the problem, as long as the weighting is based on expert evaluations, the problems stated above are yet unsolved [101].

These drawbacks i.e. subjectivity, uncertainty, imprecision in expert judgements for pairwise comparisons have been mentioned and solutions have been proposed in literature. An expert’s approach to a problem may not be the same. The results of the decision may differ according to the expert Delphi method which was introduced by Dalkey [102][103][104] is one of most widely used methods. In Delphi method, in order to reduce subjectivity, inputs from multiple experts are obtained and the collected data is consolidated. It has been used to obtain single weight for each pairwise comparison in the AHP model. Another approach is Interval AHP in which categories are assigned as intervals instead of giving single value in pairwise comparisons [105].

Researchers also proposed use of Fuzzy Set Theory to handle experts' inputs [101][106]. Fuzzy Logic Theory was introduced by Lotfi A. Zadeh [107]. It has been extensively used in many studies and systems successfully. In real systems, experts performs actions according to their preferences and customizations [108]. In many cases, converting a real world parameter to a precise mathematical model is not always possible because a human senses and decides using previous experience more than making calculations. For instance, a person can feel if water is cold or warm. But cannot measure exact temperature of water. The felt measurements are not crisp but fuzzy values. However, he or she does not know its temperature quantitatively. Fuzzy logic theory takes this reality into account. The weights assigned by experts during pairwise comparisons are evaluated as fuzzy values instead of crisp values [109]. Using this method, the inadequate use of crisp values is avoided for modeling of imprecise real life problems. Instead, experts provide their evaluations as fuzzy assessments. In order to benefit from both, combination of Fuzzy Theory and Delphi process was also proposed as Fuzzy Delphi Method (FDM). FDM was used to mitigate risks of subjective evaluations in pairwise comparisons in literature [99][106][109]. Although it has been used widely, the efficiency of using Fuzzy Logic in AHP is also criticized in [101].

Here, we have proposed a novel "Random Forest based Analytic Hierarchy Process" (RF-AHP) method to address the expert judgment uncertainty in AHP decision making. The dependency to the expert opinion is eliminated by training the AHP input data using Random Forest machine learning method. Evaluation of AHP categories i.e. criteria is made according to the assigned importances by the trained Random Forest model.

Although the methods above provide successful approaches to reduce the effect of experts in the loop, to the best of our knowledge, there was no study that removes the necessity of experts.

2.11. Analytic Hierarchy Process Workflow

Analytic Hierarchy Process is essentially a decision making approach based on the priorities which were obtained from pairwise comparison of criteria and alternatives. The application steps of AHP is outlined in Figure 11.

The steps AHP may be explained as follows:

1. Problem Definition: In the first step of AHP, the problem to be solved is defined. Therefore, the goal of the decision making is defined.
2. Criteria Definition: In the second step, the criteria that should be considered in solving the problem are defined. Criteria are subjective based on the requirements of experts. Also their validity should be checked.
3. Listing of Alternatives: In this step, possible decision alternatives are identified.
4. Construction of Hierarchy: In order to solve the problem, the goal, criteria and alternatives are organized as a hierarchy as shown in Figure 10. The goal is in the top level, criteria are in the second level, and the alternatives are at the bottom. In the hierarchy, the elements which are at the same level are independent (Independence axiom of AHP). The complexity of the hierarchy may change according to number of levels in the hierarchy and complexity of the problem.
5. Scaling of the relative importances: In this step the range of relative importances to be used in pairwise comparisons are defined. Although different scales may be used, most commonly used scale is 1-9 scale as shown in Table 1 [88][89]. This scale contains five major scores namely 1, 3, 5, 7, 9. However, if the expert is not certain about these values, for instance, a comparison may require giving more than 2 but less than 3, then inter-values 2,4,6,8 may also be used.

Table 1 Available Options for Pairwise Comparisons in Criteria Evaluation.

Option	**Numerical Value**
Equally Important	1
Weakly Important	3
More Important	5
Very Important	7
Extremely Important	9
Intermediate values (if required)	2 ,4,6, 8

6. Receiving Expert Selections: During application of AHP, one or more experts are interviewed or made questionnaires about the relative importances. In these interviews experts make pairwise comparisons of criteria. The consistency of results and decision performance of AHP is very much dependent on these decisions. Therefore, expertise and appropriate knowledge of the selected people have critical importance in AHP method. If there is only one expert, acquiring preferences and making pairwise comparisons are easy. Otherwise multiple user evaluations are averaged by using arithmetic or geometric mean to obtain single comparison result.

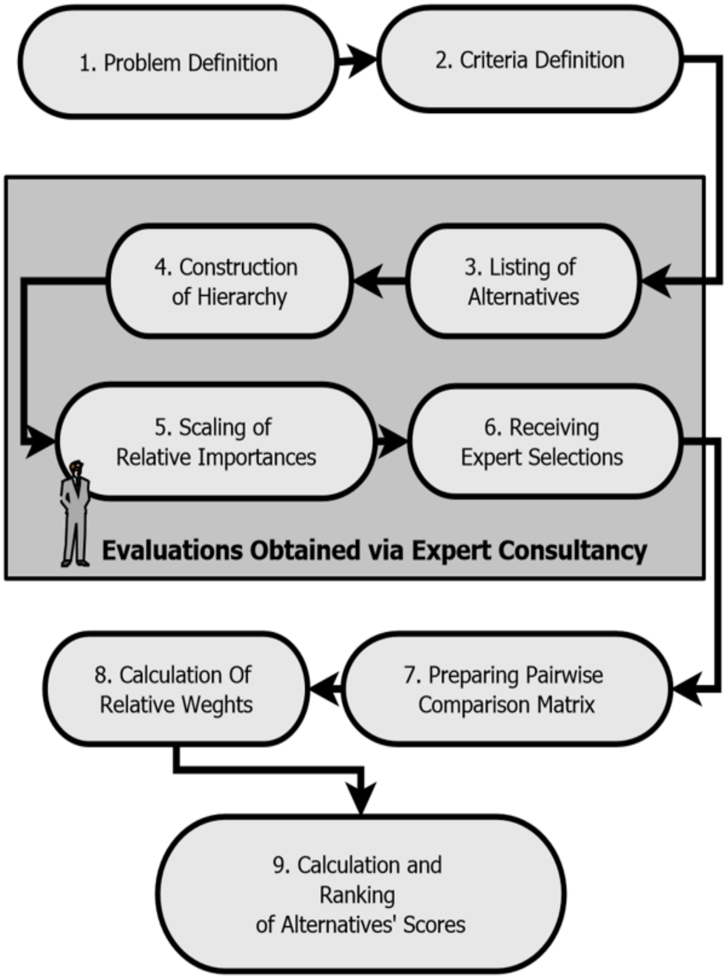

1. Problem Definition
2. Criteria Definition
4. Construction of Hierarchy
3. Listing of Alternatives
5. Scaling of Relative Importances
6. Receiving Expert Selections
Evaluations Obtained via Expert Consultancy
8. Calculation Of Relative Weghts
7. Preparing Pairwise Comparison Matrix
9. Calculation and Ranking of Alternatives' Scores

Figure 11 Implementation steps of Analytic Hierarchy Process.

7. Preparing Pairwise Comparison Matrix: Using the pairwise comparisons from Step 6, a pairwise comparison matrix is populated. If there are n criteria in the hierarchy then $n.(n\text{-}1)/2$ comparisons are performed. Therefore the size of comparison matrix is $nxn.$

Although, relative comparisons are used mostly, use of absolute scales such as weight, height i.e. are also possible. In this case, the absolute values are written to matrix directly. At the end of this step, the relative or absolute importances i.e. preferences are obtained in matrix form. In this matrix, $a_{ij} = 1/a_{ij}$.

$$A = \begin{bmatrix} a_{11} = 1 & a_{12} & \dots & a_{1n} \\ 1/a_{21} & a_{22} = 1 & \dots & a_{2n} \\ . & . & \dots & . \\ 1/a_{1n} & 1/a_{2n} & \dots & a_{nn} = 1 \end{bmatrix} \qquad \textbf{(EQUATION 3)}$$

8. Calculation of Relative Weights of Criteria: After developing pairwise comparison matrix, these values should be normalized. For normalization firstly sum of cells for each column is obtained as shown in (4).

$$b_j = \sum_{j=1}^{n} a_{ij} \qquad \textbf{(EQUATION 4)}$$

Then, each cell value is divided to sum of its column value as follows:

$$c_{ij} = \frac{a_{ij}}{b_j} \qquad \textbf{(EQUATION 5)}$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ . & . & \cdots & . \\ c_{1n} & c_{2n} & \cdots & c_{nn} \end{bmatrix} \qquad \textbf{(EQUATION 6)}$$

Having obtained matrix C which consists of c_{ij} values, relative weights of each category is obtained using arithmetic mean as shown in (6):

$$w_i = \frac{\sum_{j=1}^{n} c_{ij}}{n}$$ **(EQUATION 7)**

$$W = \begin{bmatrix} w_1 \\ w_2 \\ . \\ w_n \end{bmatrix}$$ **(EQUATION 8)**

9. Calculation of Scores and Ranking of Alternatives: When AHP hierarchy with weights is available, all of the alternatives' AHP scores are calculated finally. Then the alternatives are ranked according to their AHP score. According to decision requirement most ranking alternatives are identified and decision making process is accomplished.

2.12. Random Forest Machine Learning Algorithm

Random Forest (RF) is a supervised machine learning model developed by Leo Breiman in 2001 [5][110][111]. It is called as forest because it consists of many randomly produced decision trees. An RF is aggregation of thousands different randomly produced decision trees. Each tree consist of randomly selected features i.e. variables in the feature space. Also the training data is selected randomly using bootstrap method. The trees are trained according to classification and regression tree (CART) algorithm, which was also invented by Leo Breiman. Because each tree is constructed randomly and using different variable set, each is trained uniquely to classify the input space. Each tree classifies which class should be predicted in their output. For an input instance, some trees may classify correctly or some may do incorrectly. However the outcome of the random forest is obtained by combining results from all of the trees based on voting. Overall output is the most voted value by trees. The success of random forest relies on the fact that the major vote should be the right selection for the whole forest.

Random Forest method has become a popular algorithm in machine learning. It is successful in problems where the number of features is large. Being an ensemble method, it is able to deal with complex interactions [5]. It is also able to provide a measure about variable importances. Random forest is a classification and regression method based on the ensemble of large number of trees. The trees are constructed from a training dataset. The generalized flow of Random Forest algorithm is shown in Figure 12.

Step 1. The available data about the problem is sampled for each tree training using bootstrapping. Sampled data is divided to 2 parts. The 2/3 portion of sampled data is used as training data and 1/3 portion as Out-of Bag (OOB) data. OOB data is used to estimate the error for each trained tree.

Step 2. For each sampled data a standard classification and regression tree (CART) is grown. The parameters for the CART training is as follows:

$a_n \in \{1,\ldots,n\}$: bootstrap data size in sampled data;

$mtry \in \{1,\ldots,p\}$: random number of features to be used at splitting at each node of each tree;

$nodesize \in \{1,\ldots,n\}$: the number of instances in each cell before split.

Step 2.a In RF training of CART is controlled by $mtry$ parameter. In each split (i.e. branching in tree), randomly, $mtry$ number of prediction feature is selected from all feature set.

Step 2.b Identify the best splitting feature in this feature set. In CART, decrease of Gini impurity is used for splitting a cell. According to Gini index a cell is divided to two branches.

Step 2.c Repeat until all of the feature splits are completed. Then a tree is grown.

Step 3. After growing a tree, remaining data portion i.e. OOB is used to validate prediction performance. Hence each tree is trained and evaluated by uniquely bootstrapped data set. After a tree is grown, it is aggregated to the forest.

Step 4. Repeat the same steps to train new tree until number of trees in settings for the forest is obtained.

Step 5. By averaging the OOB error for each tree, OOB error prediction for the forest is obtained.

After the forest is obtained, the prediction by the forest is obtained by aggregation of predictions of all of the trees in the forest. For random classification forest, most voted response by the trees is the output class. In random regression forest the average of the voted values is the response of the forest.

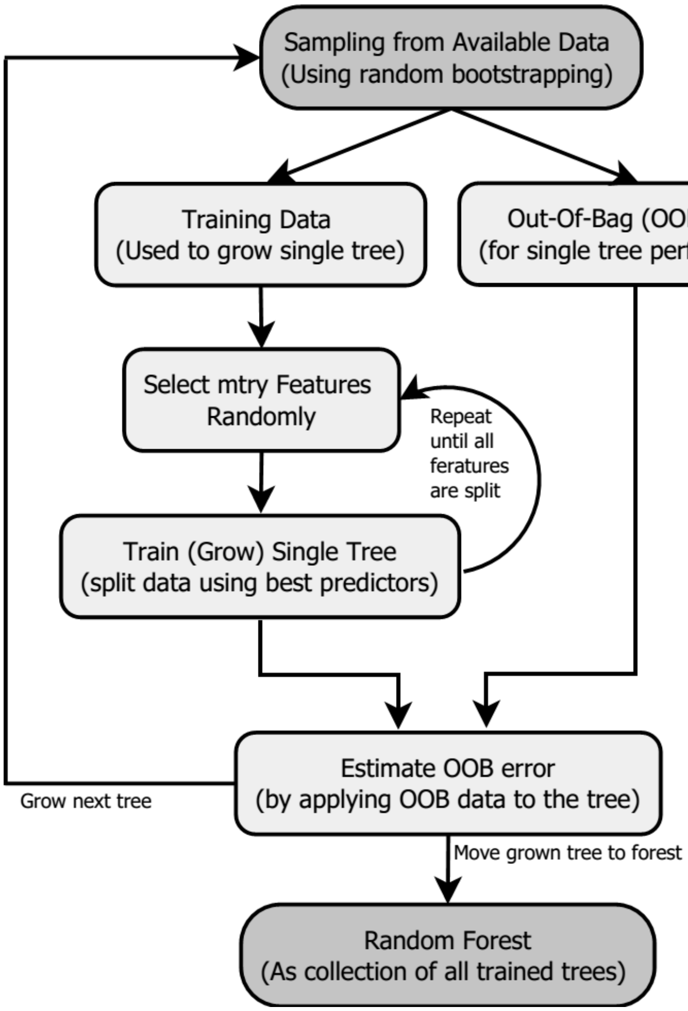

Sampling from Available Data
(Using random bootstrapping)
Training Data
(Used to grow single tree)
Out-Of-Bag (OOB) Data
(for single tree performance)
Select mtry Features
Randomly
Repeat
until all
feratures
are split
Train (Grow) Single Tree
(split data using best predictors)
Estimate OOB error
(by applying OOB data to the tree)
Grow next tree
Move grown tree to forest
Random Forest
(As collection of all trained trees)

Figure 12 Flowchart of Random Forest model.

2.13. Advantages of Random Forest

- Random Forest method may be used for categorical or numerical data for classification and regression, respectively.

- It is successful for predictions for both large and small data sets. There is no need for dividing data for training and testing because it uses bootstraps from the full data set to grow individual trees. Hence, cross validation is not necessary.

- It is able to classify when the classes in the data set is imbalanced.
- Random Forest is relatively easy to use for the expert because it requires setting only number of trees in the forest and selecting number of features in each split by the user.

2.14. Pitfalls of Random Forest

In decision tree, the result of splits may be easily visualized to understand the nature of the problem. However, in RF, there are thousands of trees and it is not possible to understand any information by visualizing the trees. Hence RF model is a black box method to the data analyst.

Additionally, in application software of the RF on a computer, all of the trees in the forest should be kept in memory. Therefore, it requires larger memory and greater processor resources during both training and prediction. However, as stated in [112], it is possible to remove a tree instance in the forest when keeping all of the trees in memory is not necessary.

2.15. Random Forest Use Cases: Prediction and Ranking

Random forest method may be used for prediction purpose in a classification or regression problem or ranking purpose of the features according to their performance in prediction used in that problem. Ranking is accomplished by calculating the variable importances of features within the RF method. Here in the presented work, we are using RF for ranking purpose. The ranking is made by evaluation of the variable importances that are inherently calculated during the training of RF model. The details of Ranking via variable importances are detailed as follows.

2.16. RF Variable Importances

After training an RF model, it is possible to rank the feature importances. It is an essential property of RF. There are two methods for calculation of variable i.e. feature importances. The First is Mean Decrease in Gini Impurity (MDI) and the second is Mean Decrease in Accuracy (MDA) [5]. MDI is based on calculating the total amount of decrease in node impurities i.e. information gain in all trees when splits made on that variable [113] [114]. The amount of decrease of impurity (ΔI) of variable $X^{(j)}$ in a split at a node for only one tree (t) may be calculated as follows:

$$\Delta I_{(t)}(X^{(j)}) = I_{parent} - (p_{left}.I_{left} + p_{right}.I_{right}) \qquad \textbf{(EQUATION 9)}$$

Here, the proportion of samples in left and right nodes is given by p_{left} and p_{right}, and the impurities for parent, left and right nodes are I_{parent}, I_{left} and I_{right} respectively. The impurities for nodes are calculated with:

$$I = 1 - \sum_j (p(j))^2 \qquad \textbf{(EQUATION 10)}$$

Where $p(j)$ is the proportion of samples with label (j) in that node. Then MDI of $X^{(j)}$ for all trees may be calculated as:

$$MDI(X^{(j)}) = \frac{1}{ntree} \sum_{t=1}^{ntree} \Delta I^{(t)}(X^{(j)}) \qquad \textbf{(EQUATION 11)}$$

Similarly, MDA may be calculated by averaging decrease of accuracies for all tree [115]. If variable is not important, its MDA should not degrade prediction accuracy. The amount of decrease of accuracy (ΔA) of variable $X^{(j)}$ for only one tree (t) may be calculated according to following equation:

$$\Delta A^{(t)}(X^{(j)}) = \frac{\sum_{i=1}^{m} \overline{y}_i^{(t)}}{m} - \frac{\sum_{i=1}^{m} \overline{y}_{i,\pi j}^{(t)}}{m} \qquad \textbf{(EQUATION 12)}$$

Here, m is the no of samples in the Out of Bag portion of training data.$\overline{y}_i^{(t)}$ is 1 if y_i is correctly classified by tree t, 0 otherwise. Similarly, When $X^{(j)}$ is permuted, $y_{i\pi j(t)}$ is 1 if y_i is correctly classified by tree (t) after X_j is permuted , 0 otherwise. MDA of $X^{(j)}$ for all trees may be calculated as:

$$MDA(X^{(j)}) = \frac{1}{ntree} \sum_{t=1}^{ntree} \Delta A^{(t)}(X^{(j)}) \qquad \textbf{(EQUATION 13)}$$

On this equation, $ntree$ is the number of trees in the random forest. MDA is obtained by calculating the arithmetic mean of DA all trees in the random forest.

CHAPTER 3

MATERIALS AND METHODS

3.1. Utilized Data Sources

In order to evaluate AHP categories using RF-AHP for SNP prioritization, four types of data source were used: 1. Candidate SNP Lists, 2. Statistically Significant SNP List, 3. SNP Feature Database, 4. dbGaP Disease Datasets. For evaluation of regulatory contents RegulomeDB dataset is used. The details of these datasets are presented as follows:

3.1.1 Candidate SNP List from Genomic Model

First type of data source is the candidate SNPs modeled in previous SNP Prioritization studies. In these studies various data-mining approaches were used on disease specific case-control data to identify SNPs associated with these diseases. For PCa model, significant SNPs set was obtained from [6], where SVM+ID3 methods were used successively to detect PCa related SNPs. As a result, 108 SNPs were selected as candidate. The list of these SNPs is available in APPENDIX A. We have analyzed the previously published model, and used disease associated candidate SNPs as Response. It should be noted that, the datasets were selected from a study, in which, techniques other than AHP were used. This ensures that the candidate SNPs identified by these models do not cause any bias in the training of proposed model.

3.1.2 Statistically Significant SNP List

Second type of data source is statistically significant SNPs list. In order to obtain statistically significant SNPs, Prostate Cancer (PCa) genotyping dataset from dbGaP collection was used. The dataset obtained from "Multi Ethnic Genome Wide Scan of Prostate Cancer" Study. It consists of 4650 cases and 4795 controls with 600.000 SNPs. [116][117]. In this study we used the SNP list from the genomic model results already completed in [6]. It consisted of three stages. At first stage, SNP data quality control is performed according to minor allele frequency, missing values and Hardy-Weinberg equilibrium criteria. In the second stage, each SNP's multiple testing adjusted p-values of association were calculated. The analysis was made using PLINK and results were obtained in the form of "association.assoc.adjusted" file. Finally, a significance threshold of p-value=0.05 is set in the "association.assoc.adjusted" file and only statistically significant SNPs are obtained.

3.1.3 SNP Features Database

Third type of data source is the SNP features database. This is the most essential component of data oriented training of AHP and contains vast amount of information accumulated from clinical bioinformatics domain. The database contains detailed annotations of SNPs according to their genomic location, disease association based on literature, prediction of consequences, and conservation across species [4]. These features were collected from public dbSNP, EntrezGene, KEGG and Gene Ontology databases. The database included following categories that may be used for SNP prioritization:

1. GWAS Results:
    - SNP p value
    - SNPs related with significant genes according to combined p-value with respect to Linkage Disequilibrium (i.e. non-random association of alleles at two or more loci)
    - SNP is on significant gene,
    - SNP is on a significant gene which is on a significant pathway.
2. Biological Facts:
    - SNP with evolutionary conserved regions.
    - SNPs on a gene.
    - SNPs that are associated with a gene which is related to a complex disease.
    - SNP is proved to be associated to a Disease gene via either directly or LD with another SNP, or a Pathway.
    - SNP is associated to a Disease gene (but not proved) via either directly or LD with another SNP, or a Pathway.
3. Genomic Location and Functional Effects:
    - Non-Coding- UTR-3
    - Non-Coding- UTR-5
    - Non-Coding Intronic
    - Non-Coding - Near Gene 3
    - Non-Coding - Near Gene 5

- Non-Coding - Splice3
- Non-Coding Splice5
- Coding- Frameshift
- Coding - CDS Non Synonymous

3.1.4 dbGaP Disease Datasets

First type of data source is the genotyping data from the following dbGaP collections. Four disease datasets used in the analysis namely were 1.Prostate Cancer (PCa), 2.Type 2 Diabetes Mellitus (T2DM), 3. Alzheimer's Disease (AD) and 4. Schizophrenia (Sz). Details of datasets are as follows:

Prostate Cancer (PCa) Dataset: The PCa dataset for tests is the same as that used in training. As explained above, it was obtained from “Multi Ethnic Genome Wide Scan of Prostate Cancer”. This dataset consists of 4650 cases and 4795 controls with 600.000 SNPs [116].

Type 2 Diabetes Mellitus (T2DM) Dataset: T2DM dataset was obtained from “Nurses' Health Study” (NHS, all female 1,769 controls and 1,479 cases) and the Health Professionals Follow-up Study (HPFS - male 1,277 controls and 1,114 cases) on Type 2 Diabetes Mellitus" and includes 642,576 SNPs [118][119].

Alzheimer’s Disease (AD) Dataset: The AD dataset was obtained from GenADA dataset. Genotyping data included 806 AD cases and 782 controls. It consists of 500,000 SNPs [120][121].

Schizophrenia (Sz) Dataset: Sz disease dataset was available from dbGaP public database as “Molecular Genetics of Schizophrenia (MGS) study”. It consisted of 3,972 cases and 3,629 controls [122].

PCa data was used in training of Random Forest machine learning algorithm for AHP Category Evaluation as shown in this table. However, for the performance evaluation between Delphi-AHP and RF-AHP all of the four disease datasets were used.

The imported version of RegulomeDB is 1.1 and is based on the data from dbSNP141 version. We downloaded all of the RegulomeDB dataset on a server machine due to its size. RegulomeDB 1.1 contains about 60 million SNPs. However, the database reused from METU-SNP contained 11 million SNPs. About 10 million of these SNPs existed in RegulomeDB as well.

3.1.5 RegulomeDB Regulatory Variant Datasets

The imported version of RegulomeDB is 1.1 and is based on the data from dbSNP141 version. All of the RegulomeDB dataset was downloaded and adapted to the study.

3.2 Software Environment for the Analysis

RF-AHP development, analysis, performance comparisons are completed in R. In the training process of Random Forest algorithm “randomForest” package in R data mining and statistical analysis software [112] was used. Using the permutation test facility of this package, Mean Decrease in Accuracies (MDA) and Mean Decrease in Impurity (MDI) importance measures for the categories were calculated. METU-SNP source code was tailored or reused when possible for development of RF-AHP and performance comparison on GWAS Disease Datasets. MySQL Database used to store SNP Annotation Datasets.

CHAPTER 4

IMPLEMENTATION OF RANDOM FOREST AHP AND INCORPORATION OF REGULOMEDB

4.1. Objectives

Random Forest based Analytic Hierarchy Process (RF-AHP) method has been developed to address the expert judgment uncertainty in AHP decision making. The dependency to the expert opinion is eliminated by training the AHP input data using machine learning method called Random Forest. Evaluation of AHP categories is made according to the calculated variable importances of Random Forest model.

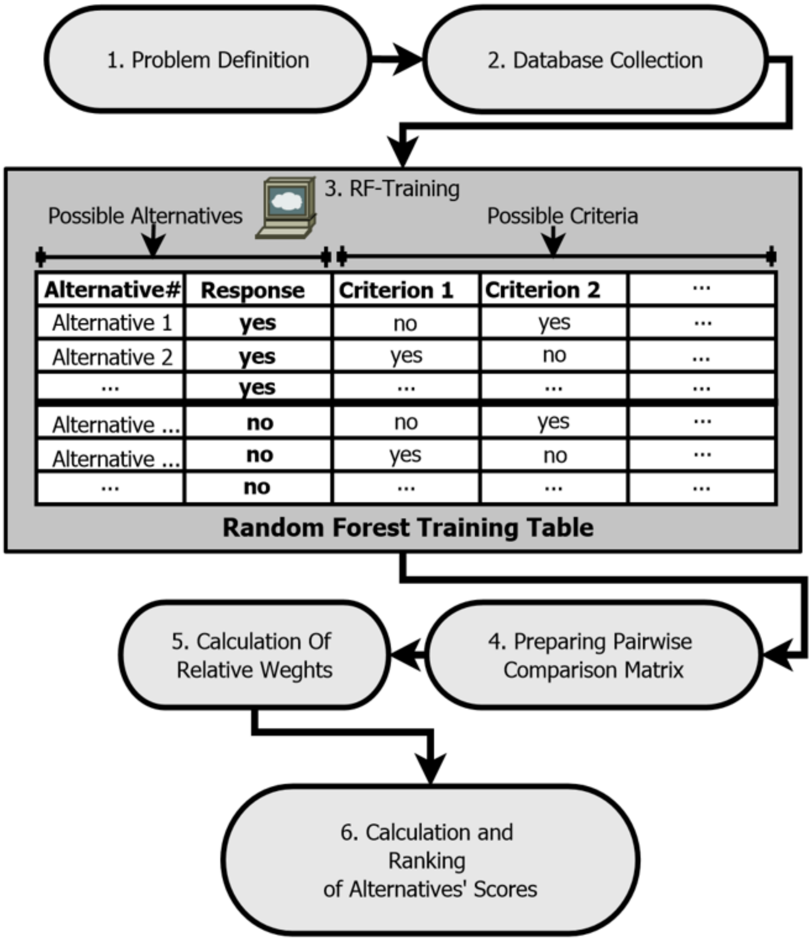
The proposed RF-AHP method requires data preparation, application of Random Forest machine learning method, inspecting RF model and evaluation steps. The explanation of these steps is presented in following section. Sample application of each step is detailed as a case study in Application of RF-AHP to SNP Prioritization section of this chapter.

Finally, incorporation of RegulomeDB as an additional feature is presented at the end of the chapter.

4.2. Implementation of RF-AHP Methodology

The RF step is implemented to eliminate the tasks that require an expert’s consultancy. The proposed RF-AHP requires an initial database construction, thus the availability of available datasets is essential for this method. The steps of the proposed RF-AHP are explained below, and as shown in Figure 13.

1. Problem Definition: Similar to Delphi AHP, the first step is to define the decision problem to be solved.


1. Problem Definition
2. Database Collection
3. RF-Training
Possible Alternatives
Possible Criteria
Random Forest Training Table
4. Preparing Pairwise Comparison Matrix
5. Calculation Of Relative Weghts
6. Calculation and Ranking of Alternatives' Scores

Alternative#	Response	Criterion 1	Criterion 2	...
Alternative 1	**yes**	no	yes	...
Alternative 2	**yes**	yes	no	...
...	**yes**	...	...	...
Alternative ...	**no**	no	yes	...
Alternative ...	**no**	yes	no	...
...	**no**	...	...	...

Figure 13 Implementation steps of Random Forest based Analytic Hierarchy Process.

2. Database Collection: RF-AHP, being a data driven methodology, assumes the availability of extensive data, and requires analysis of these data sources. The data may be obtained from large databases, transaction systems, management information systems, web, cloud, literature etc. The possible alternatives for the decision problem should be researched from the structured or unstructured data and transformed into rows as shown in Figure 13. Similarly, the criteria for making the decision should be digested or consolidated from resources. The criteria are assigned as features for the random forest method. The response variable should also be populated according to each alternative whether it is selected or not selected in the decision. In the end a classification (or regression) table should be populated in which rows are the alternatives and columns are criteria for the decision. The table is called as Random Forest Training Table as shown in Figure 13.

3. RF Training: After obtaining the training table for the classification, a Random Forest model is trained. The details of Random Forest training process are presented Section 2.12 Random Forest Machine Learning Algorithm, therefore it will not be repeated here. In RF-AHP, the variable importances from the trained RF model are used. As mentioned in Section 2.12, two types of variable importances namely, Mean Decrease in Impurity (MDI) and Mean Decrease in Accuracy (MDA) calculations for each category in the RF training table are made on the trained RF model. The user of the RF-AHP method may decide which importance to use according to problem context.

4. Preparing Pairwise Comparison Matrix: After obtaining Criteria importances i.e. Variable Importances as MDA or MDI, pairwise comparison matrix is calculated as follows:

$$A = \begin{bmatrix} a_{11} = 1 & VI_1/VI_2 & ... & VI_1/VI_n \\ VI_2/VI_1 & a_{22} = 1 & ... & VI_2/VI_n \\ . & . & ... & . \\ VI_n/VI_1 & VI_n/VI_2 & ... & a_{nn} = 1 \end{bmatrix} \qquad \textbf{(EQUATION 14)}$$

Here, *VI_i* stands for *MDA_i* if MDA importances are used in pairwise comparisons, or *$V\,I_i$* stands for *MDI_i* if MDI importances are used in pairwise comparisons. Moreover, user provided evaluation of combination of MDI and MDA is also possible. Following precautions and alternatives may also be considered in construction of pairwise comparison matrix:

1. Criteria which have VIs as zero should be removed. These criteria are not effective in decision making. Additionally, it is not possible to construct pairwise comparison matrix because it causes “division by zero” in elements as explained by Saaty [123].

2. Likewise, the user may choose to apply the scale presented in Table 1 with respect to MDA or MDI importances.

3. Although in normal process it is possible to use the importance values directly according to MDA or MDI values, the decision maker may choose to use manual scales by simply looking at VI values. For instance one variable may have very high VI value that saturates the scaling, therefore very high values may be clamped at a limit.

5. Calculation of Relative Weights of Criteria for RF-AHP: After developing pairwise comparison matrix above, its values should be normalized. For normalization firstly sum of cells for each column is obtained as showed in (15) previously. This step is covered in detail in Section 2.11 as Step 8 in the Analytic Hierarchy Process Workflow. For the sake of completeness we summarize the calculation of criteria weights again. The comparison matrix elements are normalized as follows:

$$c_{ij} = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}}$$ **(EQUATION 15)**

Having obtained normalized matrix C which consists of c_{ij} values, relative weights of each criteria is obtained as W vector using arithmetic mean as shown in (16):

$$W = \begin{bmatrix} \frac{\sum_{j=1}^{n} c_{1j}}{n} \\ \frac{\sum_{j=1}^{n} c_{2j}}{n} \\ . \\ \frac{\sum_{j=1}^{n} c_{nj}}{n} \end{bmatrix}$$ **(EQUATION 16)**

6. Calculation of RF-AHP Scores for Alternatives and Ranking: After RF-AHP hierarchy with criteria weights is obtained, RF-AHP scores are calculated for the alternatives and ranked. The details of these calculations are covered in Section 2.11 under Step 9. By applying a threshold to scores of alternatives prioritized alternatives are obtained decision making process is completed.

4.3. Application of RF-AHP to SNP Prioritization

In order to validate proposed RF-AHP method, a case study was conducted for Single Nucleotide Polymorphism (SNP) Prioritization [124]. Details of SNP prioritization was also presented in Section 2.9 SNP Prioritization and Candidate SNP Selection.

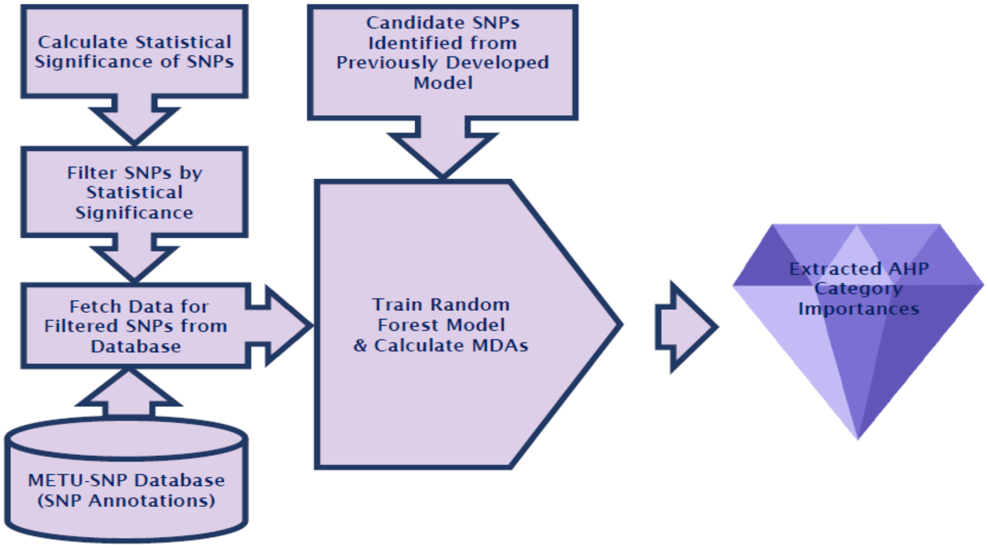
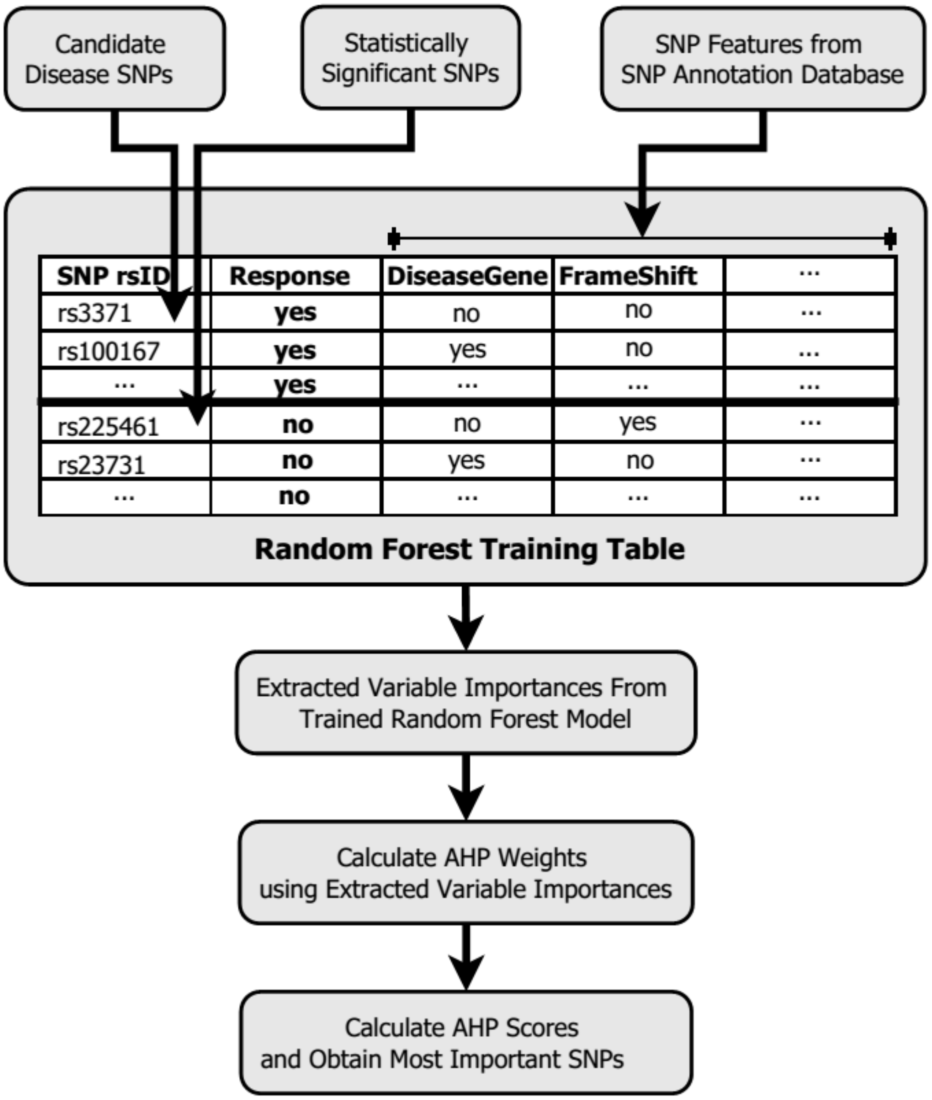

Calculate Statistical Significance of SNPs
Filter SNPs by Statistical Significance
Fetch Data for Filtered SNPs from Database
METU-SNP Database (SNP Annotations)
Candidate SNPs Identified from Previously Developed Model
Train Random Forest Model & Calculate MDAs
Extracted AHP Category Importances

Figure 14 Order of activities in application of RF-AHP method in SNP Prioritization.

The objective of RF-AHP in SNP Prioritization process is to give score to each SNP in the dataset such that, the SNPs those responsible of subject disease get highest score. Firstly, statistical analysis results were obtained as "association.assoc.adjusted" files from PLINK and significance threshold p-value was selected. The SNPs which have smaller p-value than threshold were queried from SNP features database to collect their accompanying data. After obtaining SNP features, an input table was populated as shown in Figure 15.

On this table, if a SNP was supplied with information that, for instance, if the field in ‘‘Frameshift” in the SNP Features database is valued as "1", then it is recorded as ‘‘yes", otherwise it is recorded as ‘‘no". Similarly, if there was no data available for the SNP for that category, it is recorded as ‘‘no" by default, meaning that it has no effect for the prioritization. If the SNP is one of the disease SNPs in the candidate SNP then OUTPUT class is recorded as ‘‘yes", otherwise it is recorded as ‘‘no". The steps to implement RF-AHP was explained in Section 4.2. Here we apply each step to SNP prioritization as follows:

Step 1. Problem Definition: The goal of decision making problem in SNP prioritization is to find the most relevant SNPs for a given SNP list. In order to achieve this objective, the supplied SNPs are analyzed for various features from SNP annotation database.


Candidate Disease SNPs
Statistically Significant SNPs
SNP Features from SNP Annotation Database
Random Forest Training Table
Extracted Variable Importances From Trained Random Forest Model
Calculate AHP Weights using Extracted Variable Importances
Calculate AHP Scores and Obtain Most Important SNPs

SNP rsID	Response	DiseaseGene	FrameShift	...
rs3371	yes	no	no	...
rs100167	yes	yes	no	...
...	yes	...	...	...
rs225461	no	no	yes	...
rs23731	no	yes	no	...
...	no	...	...	...

Figure 15 Application of RF-AHP method in SNP Prioritization.

For instance, one may consider DiseaseGene_ViaDirect feature in SNP annotation database. If the SNP (i.e. an alternative) is already associated with any other disease gene, it may be highly a candidate for the subject disease as well. Therefore, the decision making method should appropriately handle it for SNP prioritization.

Step 2. Database Collection: In this step, the datasets mentioned in Section 3.1are used to populate the RF database table shown in Figure 15.

First type of dataset is the candidate SNP list mentioned in Section 3.1.1. This SNP list contains 108 SNPs associated to prostate cancer obtained from [53]. These SNPs are inserted to the RF training table as Response = “yes” rows as shown in Figure 15.

Secondly, statistically significant SNPs explained in Section 3.1.2 are inserted to RF training table as Response = “no” as shown in Figure 15. The significance threshold p-value=0.05 is set, 26367 SNPs whose significance lower (i.e. better) than threshold is selected.

The third type of dataset is SNP features dataset as explained in Section 3.1.3. The features are all categorical data. It is used to populate the criteria columns in the RF database table. If a SNP has an annotation for a criterion in the SNP features dataset, the annotated value is recorded to the table as “yes”. If there is no annotation found for a SNP, the feature is recorded as “no” for imputation.

Step 3. RF Training for SNP list and SNP features database: The training of RF model is performed after obtaining the populated RF database table.

RandomForest Library Training Parameters: The RandomForest library accepts the RF database training table as training data. The Response field was set as the predictor output explicitly. Variable Importance Calculation feature was set to true. Number of features to train for each tree was set to *mtry*=10 and *number of trees* in the forest was set to *N*=5000. With these settings the training code was run and the trained RF model was obtained.

Step 4. Calculation of Pairwise Comparison Matrix

In our model, the Variable Importances are the criteria in AHP. Random Forest package in R calculates Mean Decrease in Accuracy i.e. MDA and mean decrease in Gini (i.e. Impurity) MDI as importances.

The calculated MDA and MDI variable importances from RF model is plotted and shown in Figure 16. On the figure, most of the annotated features for criteria are found to be zero importance. When the MDA and MDI plot is inspected carefully, it may be seen that MDA value for some criteria are negative valued. It means that permuting those variables affect the model accuracy. For this purpose of keeping all information possible, they are saved for criteria list. The MDA and MDI plot is not similar because of negative values in MDA plot. For this reason, it is not easy to follow “importance based weight calculation” approach. Instead, we opted for using

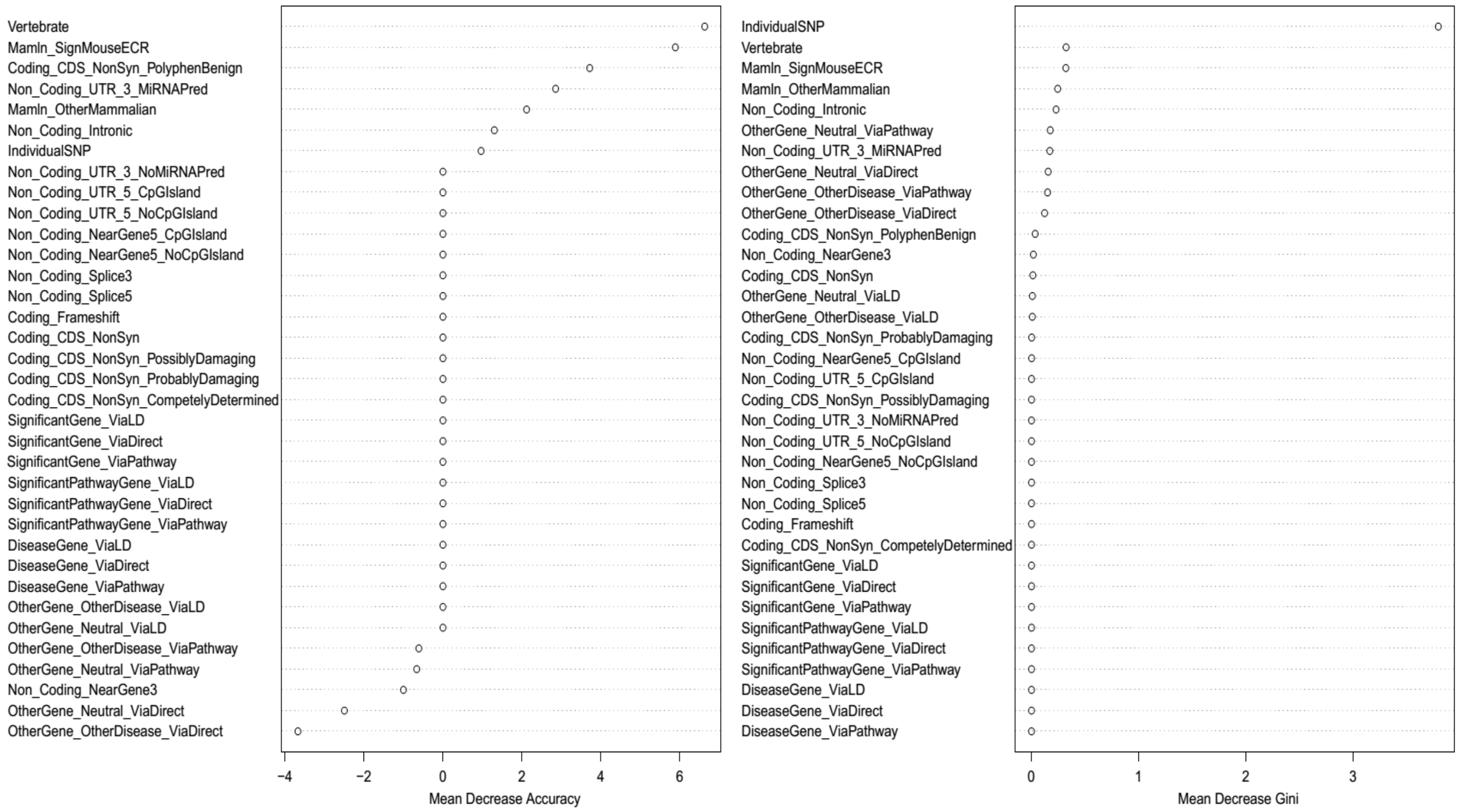


Figure 16 AHP category importance evaluation based on MDA and MDI plot

RF selected categories, then used pairwise comparison weights from [4] based on genomic, statistical and biological features.

Step 5. Calculation of Criteria Weights: The relative weights are calculated by considering the importances, pairwise comparison weights and using equations (13), (14) and (15) respectively. The results of calculated weights are shown in Figure 17.

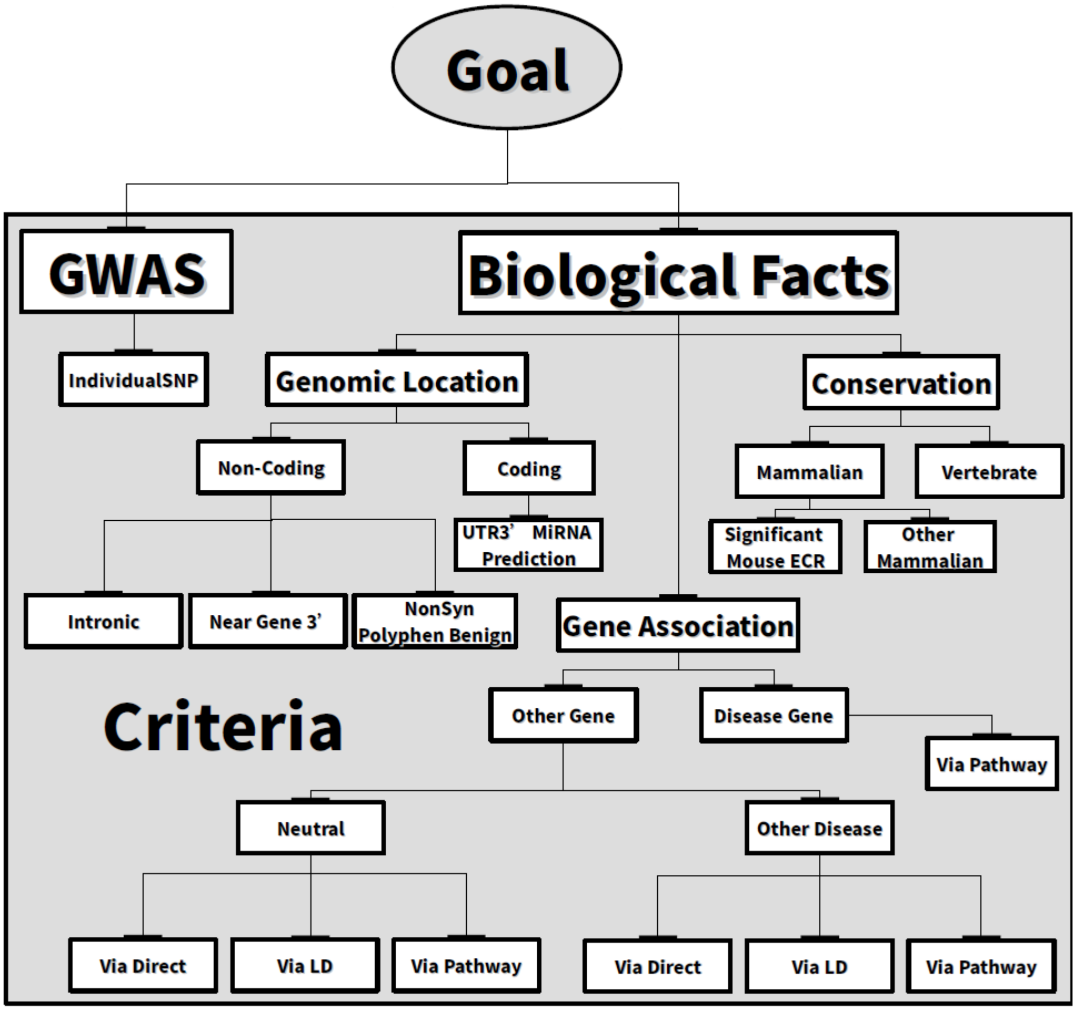


Figure 17 Resulting RF-AHP Tree for SNP Prioritization.

Step 6. Calculation of RF-AHP Scores for Alternatives and Ranking:

After obtaining the criteria with weights the resulting AHP model is for calculation of score for each SNP and ranked by their AHP scores. RF-AHP Score for each SNP is calculated according to the following equation:

$$S_{(SNP_t)} = \sum_{i=1}^{n} I_{i(SNP_t)} W_i : t = 1, ...m \qquad \textbf{(EQUATION 17)}$$

Here, m is the number of SNPs, W_i is the normalized weight of a category in AHP tree obtained by RF-AHP method and I_i is the activation indicator for SNP_t for criteria i. For instance, if *SNPt* is related to disease gene on the same pathway (DiseaseGene ViaPathway), its indicator value is 1, otherwise 0. Therefore, the score of a SNP (i.e. alternative) is calculated by sum of activated weights. The resulting RF-AHP Tree is shown in Figure 17. Application results for SNP prioritization using obtained RF-AHP model is presented in Chapter 5.

4.4. Incorporation of RegulomeDB: RF-AHP-R

A large percentage of GWAS hits for common traits and common diseases or phenotypes fall outside of exome i.e. non-coding regions [125][126][127]. RegulomeDB provides annotations for SNPs with known and predicted regulatory DNA elements such as regions of DNase hypersensitivity, binding sites of transcription factors, and promoter regions that have been biochemically characterized to regulation transcription.

In RF-AHP, most of the SNP annotations currently existing in the database is mostly on coding regions. Non-coding SNPs, without having any associated gene for a annotation, it is difficult to effectively annotate a variant. One of the best features of integrative approach with RF-AHP is ability to add new databases easily. Thus, we have incorporated RegulomeDB dataset, which combines sources from public ENCODE project [44], into RF-AHP scoring is done to improve its prioritization performance. Details of incorporation of RegulomeDB to system is as follows.

The RegulomeDB incorporated version of RF-AHP method i.e. RF-AHP-R is able to score and prioritize SNPs in both coding and regulatory regions. In importing the data, the categorical RegulomeDB score converted to numeric score in RF-AHP-R. As explained in Section 3.1.5, the imported version of RegulomeDB is version 1.1. The lookup table for conversion of RegulomeDB scores to RF-AHP-R scoring is shown in Table 2.

Table 2 RegulomeDB Scoring to RF-AHP-R Scoring for Computations.

RegulomeDB Score	**Supporting Data**	**RF-AHP-R Score**
1a	eQTL + TF binding + Matched TF Motif + Matched DNase Footprint + DNase Peak	6
1b	eQTL + TF Binding + Any Motif + DNase Footprint + DNase Peak	6
1c	eQTL + TF Binding + Matched TF Motif + DNase Peak	5
1d	eQTL + TF Binding + Any Motif + DNase peak	5
1e	eQTL + TF Binding + Matched TF Motif	4
1f	eQTL + TF Binding / DNase Peak	3
2a	TF binding + Matched TF Motif + Matched DNase Footprint + DNase Peak	4
2b	TF Binding + Any Motif + DNase Footprint + DNase Peak	4
2c	TF Binding + Matched TF Motif + DNase Peak	3
3a	TF Binding + Any Motif + DNase peak	3
3b	TF Binding + Matched TF Motif	2
4	TF Binding + DNase Peak	2
5	TF Binding or DNase peak	1
6	Other	0

CHAPTER 5

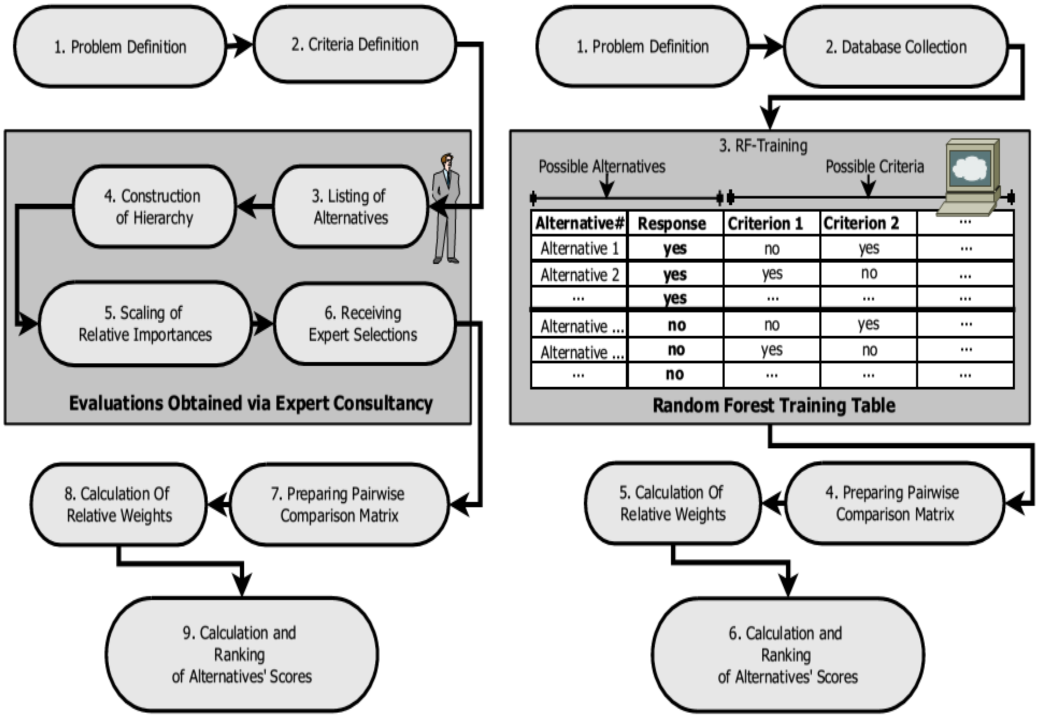
RESULTS AND DISCUSSIONS

In this study, variable importance property of Random Forest algorithm was used to identify informative and uninformative categories in construction of RF-AHP. Our results showed that, using the RF variable importances, AHP based SNP prioritization can be performed without any subjectivity to obtain better decision performance. Here, we have presented our results in two sections; First, RF-AHP performance is demonstrated on the Prostate Cancer and the Alzheimer's Disease datasets. Next, comparison of RF-AHP and RF-AHP-R methods to Delphi AHP is presented for all four disease datasets described in Section 3.1.4.

5.1 Comparison of Delphi AHP Categories to RF-AHP Categories

METU-SNP was developed in 2011 in METU-BIN Bioinformatics Laboratory. It was constructed using Delphi-AHP methodology [4]. The database of the Delphi AHP contains SNP annotations with features such as gene, disease etc. It employs a novel approach to SNP prioritization using Analytic Hierarchy Process based decision making technique for SNP prioritization. It analyzes and calculates a score for each SNP according to different categories in the dataset. The weights of categories were scored by six molecular biology experts using the Delphi method. A brief view of Delphi AHP tree is shown in APPENDIX B. Consequently, the category evaluations are based on judgement. Moreover, it is not clear that all of the inspected features in AHP tree are really necessary. A subset of questions which were asked to experts is available in APPENDIX C.

In this step comparative case study, the categories used in RF-AHP was compared to categories in Delphi AHP. The Delphi AHP tree was analyzed in order to detect uninformative and uncertain category weights based on RF-AHP results. The main objective of RF-AHP is to select categories to evaluate from the list of categories that exists in METU-SNP database. The application steps of Delphi AHP and RF-AHP are shown in Figure 18 (a) and (b). When compared to Delphi AHP categories, 19 categories do not exist in RF-AHP categories shown in Table 3. They are found to be uninformative by RF variable importances.


1. Problem Definition
2. Criteria Definition
4. Construction of Hierarchy
3. Listing of Alternatives
5. Scaling of Relative Importances
6. Receiving Expert Selections
Evaluations Obtained via Expert Consultancy
8. Calculation Of Relative Weights
7. Preparing Pairwise Comparison Matrix
9. Calculation and Ranking of Alternatives' Scores
1. Problem Definition
2. Database Collection
3. RF-Training
Possible Alternatives
Possible Criteria
Random Forest Training Table
5. Calculation Of Relative Weights
4. Preparing Pairwise Comparison Matrix
6. Calculation and Ranking of Alternatives' Scores

Alternative#	Response	Criterion 1	Criterion 2	...
Alternative 1	yes	no	yes	...
Alternative 2	yes	yes	no	...
...	yes	...	...	...
Alternative ...	no	no	yes	...
Alternative ...	no	yes	no	...
...	no	...	...	...

(a) (b)

Figure 18 Comparison of Delphi AHP and RF-AHP implementation steps.

Table 3 List of Uninformative Categories According to Calculated Importances.

Non Coding UTR 3 NoMiRNAPred
Non Coding UTR 5 CpGIsland
Non Coding UTR 5 NoCpGIsland
Non Coding NearGene5 CpGIsland
Non Coding NearGene5 NoCpGIslnd
Non Coding Splice3
Non Coding Splice5
Coding Frameshift
Coding CDS NonSyn PolyphenBenign
Coding CDS NSyn PossiblyDamaging
Coding CDS NSyn ProbablyDamaging
Coding CDS NSyn CompletelyDetermined
Significant Gene ViaLD
Significant Gene ViaDirect
Significant Gene ViaPathway
Significant Pathway Gene ViaLD
Significant Pathway Gene ViaDirect
Significant Pathway Gene ViaPathway
Disease Gene ViaLD
Disease Gene ViaDirect

5.2 Application of RF-AHP for SNP Prioritization

In order to validate the performance of trained RF-AHP model, analyses for Prostate Cancer and Alzheimer's Disease is performed. Details of Prostate Cancer and Alzheimer's Disease datasets used in the analyses were described in Section 3.1.4.

For the first analysis, statistically significant SNPs are selected from Prostate Cancer GWAS analysis was used in the AHP prioritization. The trained RF-AHP model described in the Section 4.3 was used to calculate the RF-AHP scores. In order to select informative SNP set to analyze with SNPNexus, we filtered the SNPs with highest scores. The number of SNPs having RF-AHP score greater than 0.1 was found as 121 SNPs. For the second analysis, a completely different dataset i.e. Alzheimer's Disease dataset was selected. Following the same work-flow with the Prostate Cancer variant analysis, statistically significant SNPs obtained from Alzheimer's Disease GWAS analysis was selected for RF-AHP analysis. Then RF-

AHP model was used to calculate RF-AHP scores. The number of SNPs, associated with AD, having RF-AHP score greater than 0.1 was 56.

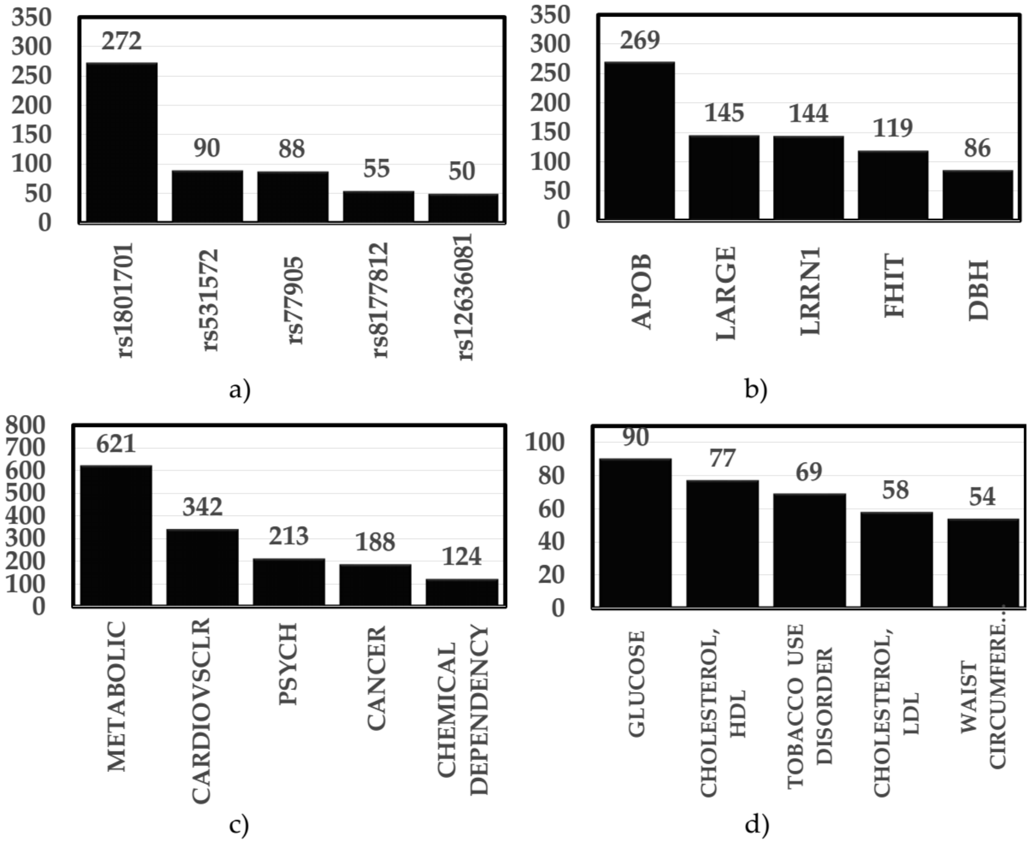
5.3 Analysis of RF-AHP Based SNP Prioritization Results

Prostate Cancer Analysis Results: Prostate Cancer associated 121 SNPs selected through RF-AHP, returned rs1801701, rs531572, rs77905, rs8177812, rs12636081 SNP ids as the most frequently referenced SNPs by GAD. As shown in Table 4, all the SNPs was successfully placed in Top 20 by RF-AHP except rs8177812. When inspected, in the RF-AHP results, rs8177812 was ranked as 49th.

Table 4 Top 20 SNPs, calculated RF-AHP scores and GAD Rank for PCa.

Rank	SNP ID	RF-AHP Score	GAD Rank
1	rs3912492	0.338088	8
2	rs12636081	0.338088	5
3	rs17061864	0.338088	6
4	rs6803449	0.338088	7
5	rs1801701	0.215257	1
6	rs77905	0.21326	3
7	rs12948056	0.212474	121
8	rs4794488	0.212474	66
9	rs1433369	0.197392	50
10	rs16930396	0.190501	35
11	rs1608114	0.190501	91
12	rs1915940	0.190501	95
13	rs2574824	0.182625	26
14	rs7249230	0.177416	109
15	rs11563056	0.175419	118
16	rs8064691	0.175419	75
17	rs12592981	0.175419	87
18	rs531572	0.175419	2
19	rs965560	0.159278	26
20	rs138726	0.159278	22

SNPNexus results provided APOB, LARGE, LRRN1, FHIT, DBH as most referenced genes. These genes were also associated with Metabolic, Cardiovascular, Psychiatric, Cancer, Chemical Dependency disease classes. As most referenced phenotypes Glucose, Cholesterol HDL, Tobacco Use Disorder, Cholesterol LDL and Waist Circumference categories were obtained (Figure 19).



a

Figure 19 SNPNexus Frequency Results of AHP results for Prostate Cancer Analysis: a) Most Referenced SNPs b) Most Referenced Genes c) Most Referenced Disease Classes d) Most Referenced Phenotypes.

Alzheimer’s Disease Analysis Results: For Alzheimer’s Disease, the results obtained from the SNPNexus provided the references for 56 SNPs having RF-AHP Score > 0.1. SNPNexus returned rs6023, rs5897, rs4972, rs132954, rs6160 SNP ids as the most referenced SNPs. When it is checked in Table 5, all of the SNPs were successfully found in Top 20 by RF-AHP. SNPNexus results provided F5, F2, LARGE, ADD1, LRRN1 as most referenced genes. SNPNexus associated Cardiovascular, Metabolic, Reproduction, Hematological, Cancer as disease classes for the submitted SNPs. Glucose, Venous Thrombosis, Hypertension, Thrombo-Embolism, Cholesterol HDL categories was obtained as the most associated phenotype classes (Figure 20).

Table 5 Top 20 SNPs and calculated RF-AHP scores and GAD Rank for AD.

Rank	SNP ID	RF-AHP Score	GAD Rank
1	rs3084	0.474415	41
2	rs11523	0.460004	36
3	rs879	0.45345	52
4	rs897530	0.451068	21
5	rs7384	0.432316	N/A
6	rs2668	0.421324	26
7	rs6160	0.412137	5
8	rs7769	0.408969	46
9	rs5897	0.400453	2
10	rs1237	0.397055	25
11	rs14810	0.393636	14
12	rs11522	0.393636	32
13	rs783305	0.385102	20
14	rs4161	0.375043	34
15	rs9511	0.370289	18
16	rs1615	0.370289	N/A
17	rs17032	0.368507	10
18	rs42019	0.350949	23
19	rs4972	0.339956	3
20	rs138222	0.329827	9

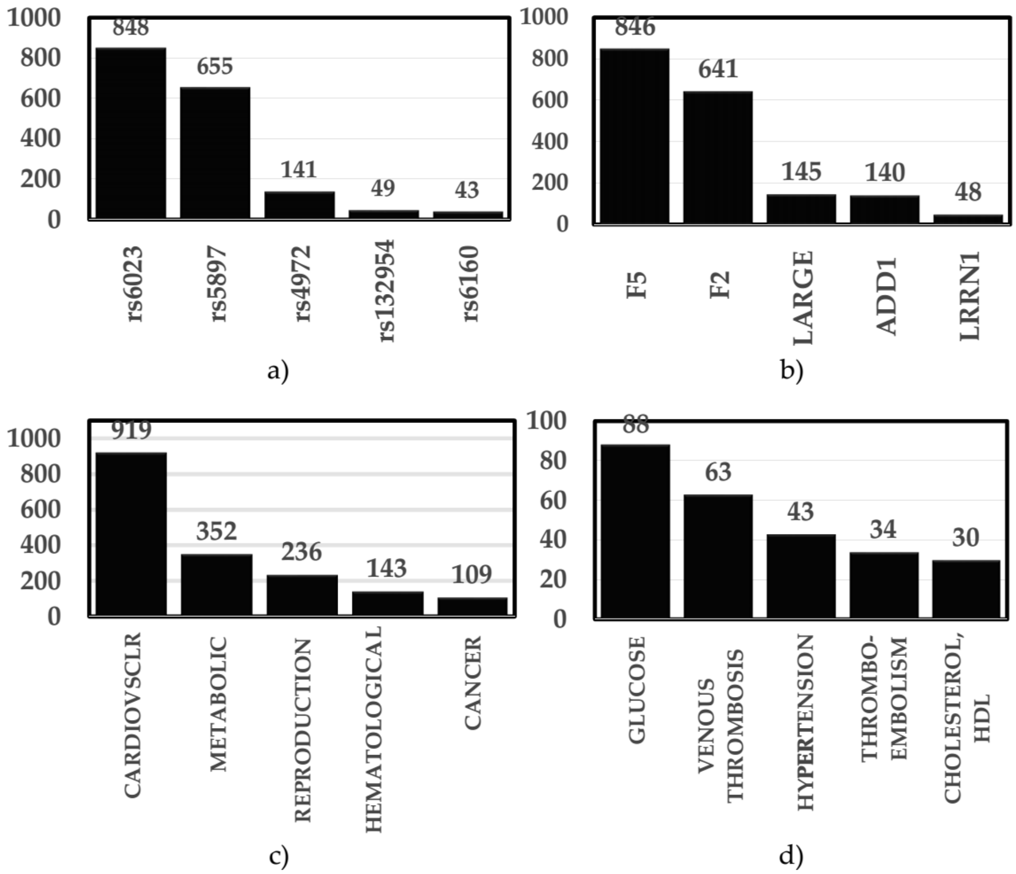


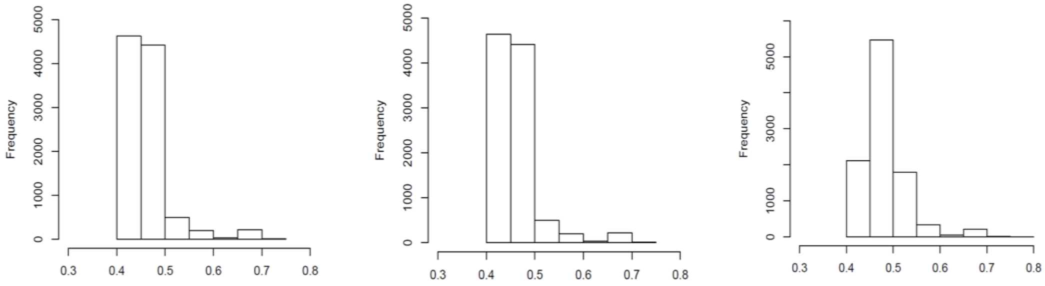
Figure 20 SNPNexus Frequency Results of AHP results for Alzheimer’s Disease Analysis: a) Most Referenced SNPs, b) Most Referenced Genes, c) Most Referenced Disease Classes, d) Most Referenced Phenotypes

5.4 Performance Comparison between Delphi-AHP and RF-AHP and RegulomeDB Incorporated RF-AHP-R versions

In order to compare the prioritization performance of the three models, GWAS Disease Datasets presented in Section 3.1.4 were used. There are four disease datasets namely, Schizophrenia, Type 2 Diabetes Mellitus, Alzheimer’s Disease and Prostate Cancer. RF-AHP and RegulomeDB incorporated version RF-AHP-R are compared to Delphi AHP according to SNP prioritization results of these four datasets. In these analyses, SNP scores for each method are calculated according to Equation (17). Performance of the RF-AHP and RF-AHP-R are compared to the expert evaluated Delphi-AHP by using the methods explained in Chapter 4. In order to obtain statistically significant SNPs, p-value threshold was selected as 0.05. Then following analyses are repeated for each model for all disease datasets. The top 20 SNPs according to all methods for various disease datasets are presented in APPENDIX D as tables. Number of available SNPs is also counted and presented in APPENDIX E. Additionally, results of references calculated from SNPNexus GAD results are presented in APPENDIX F.

5.4.1 Comparative Analysis of Schizophrenia SNP Prioritization Results for Delphi AHP RF-AHP and RF-AHP-R

AHP Score distribution for Schizophrenia analyses are shown in Figure 21. Amount of available SNPs which have AHP scores higher than threshold are shown in Table 6. For Schizophrenia disease, general distribution of AHP scores were higher. Therefore, threshold was set to 0.5.



a) Sz scoring by Delphi AHP

b) Sz scoring by RF- AHP

c) Sz scoring by RF-AHP-R

Figure 21 Distribution of computed for Sz scores in Delphi AHP, RF-AHP and RF-AHP-R

According to Table 6, it may be seen that the number of available SNPs having AHP scores above 0.1 was the same at 961 SNPs for Delphi AHP and 959 SNPs for RF-AHP. This shows that designing AHP process using RF was able to provide almost same response without requiring any expert and no loss of information occurred. Sz analysis using RF-AHP-R provided 2424 SNPs which scored over threshold value.

Table 6 No of Available SNPs as a Result of Schizophrenia Analysis

Analysis Name	**No of SNPs whose AHP score > 0.5**
Sz analysis using Delphi AHP	961
Sz analysis using RF-AHP	959
Sz analysis using RF-AHP-R	2424

For each analysis result a separate SNPNexus job was submitted. Result of each job analyzed and GAD results are shown in Figure 22.

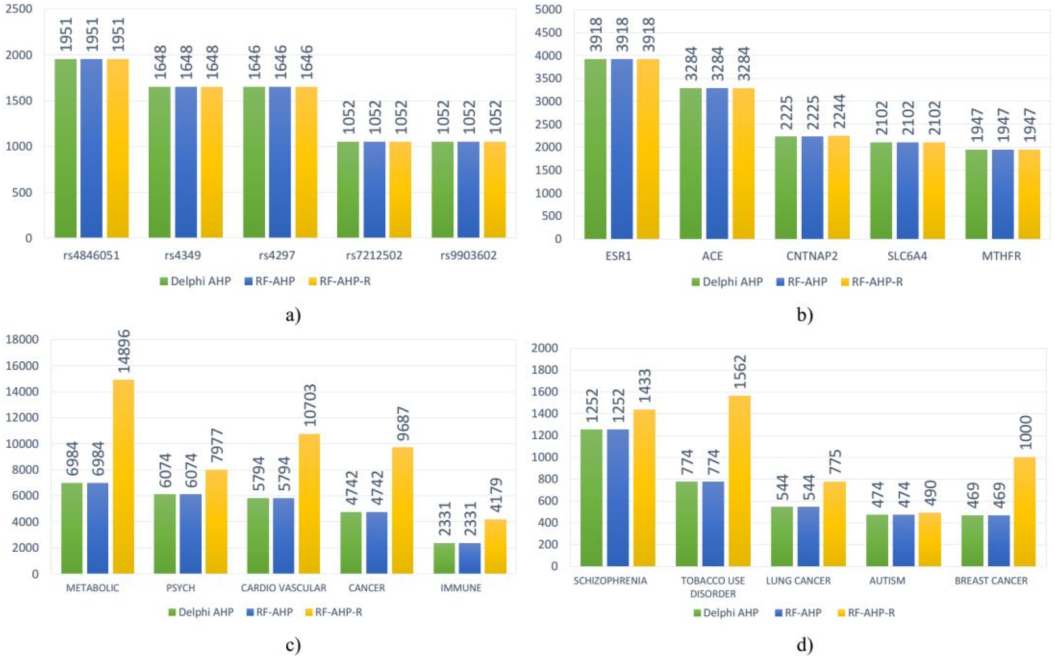
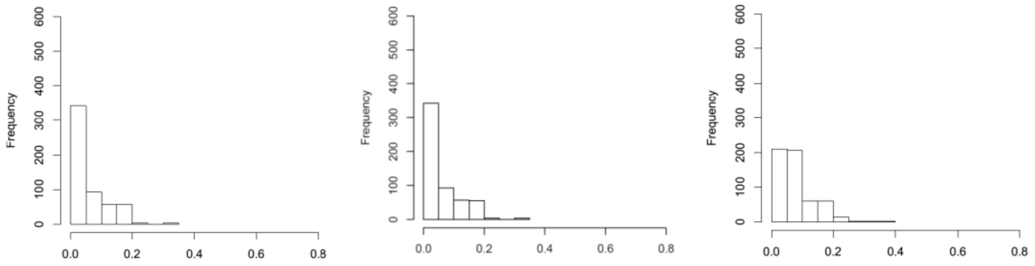


Figure 22 SNPNexus-GAD Frequency Results for Schizophrenia Analysis: a) Most Referenced SNPs b) Most Referenced Genes c) Most Referenced Disease Classes d) Most Referenced Phenotypes

5.4.2 *Comparative Analysis of Prostate Cancer SNP Prioritization Results for Delphi AHP RF-AHP and RF-AHP-R*

AHP Score distribution for Prostate Cancer analyses are shown in Figure 23. For Prostate Cancer according to general distribution of AHP scores threshold was set to 0.1. Amount of available SNPs which have AHP score higher than threshold are shown in Table 7.



a) PCa scoring by Delphi AHP b) PCa scoring by RF- AHP c) PCa scoring by RF-AHP-R

Figure 23 Distribution of computed AHP Scores for PCa for Delphi AHP, RF-AHP and RF-AHP-R

According to Table 7, it may be seen that the number of available SNPs having AHP scores above 0.1 was the same at 121 SNPs for Delphi AHP and RF-AHP.

Table 7 No of Available SNPs as a Result of Prostate Cancer Analysis

Analysis Name	**No of SNPs whose AHP score > 0.1**
PCa analysis using Delphi AHP	121
PCa analysis using RF-AHP	121
PCa analysis using RF-AHP-R	140

This shows that designing AHP process using RF was able to provide same response without requiring any expert and no loss of information occurred. By incorporating RegulomeDB data, in RF-AHP-R, number of available SNPs increased to 140. For each analysis result a separate SNPNexus job was submitted. Result of each job analyzed and GAD results are shown in Figure 24.

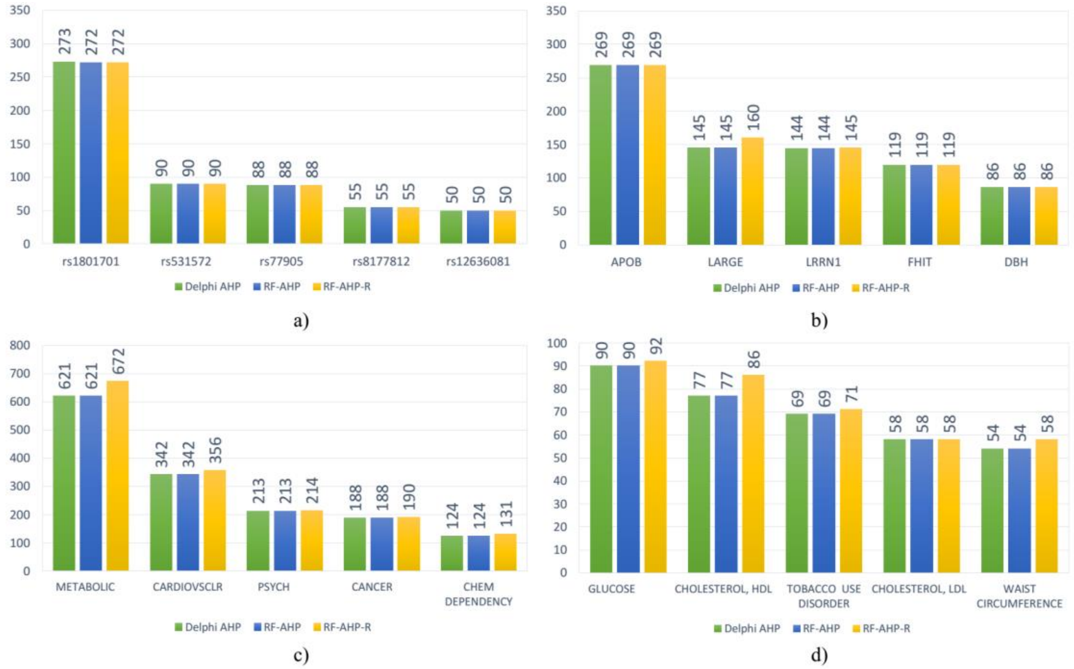
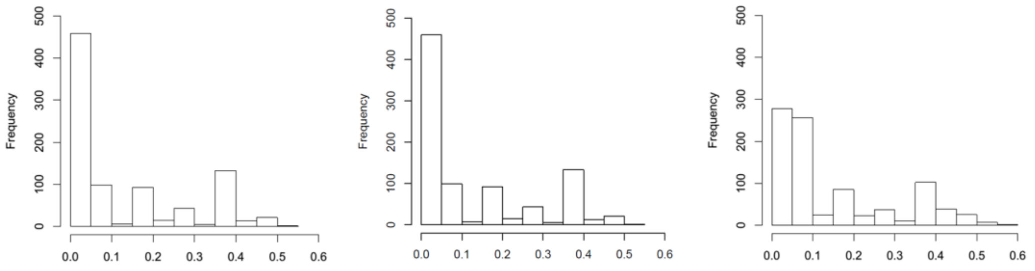


Figure 24 SNPNexus-GAD Frequency Results for Prostate Cancer Analysis: a) Most Referenced SNPs b) Most Referenced Genes c) Most Referenced Disease Classes d) Most Referenced Phenotypes

5.4.3 Comparative Analysis of Type 2 Diabetes Mellitus SNP Prioritization Results for Delphi AHP RF-AHP and RF-AHP-R

AHP Score distribution for Type 2 Diabetes Mellitus analyses are shown in Figure 25. For Type 2 Diabetes Mellitus disease, according to general distribution of AHP scores threshold was set to 0.1.



a) T2DM scoring by Delphi AHP

b) T2DM scoring by RF- AHP

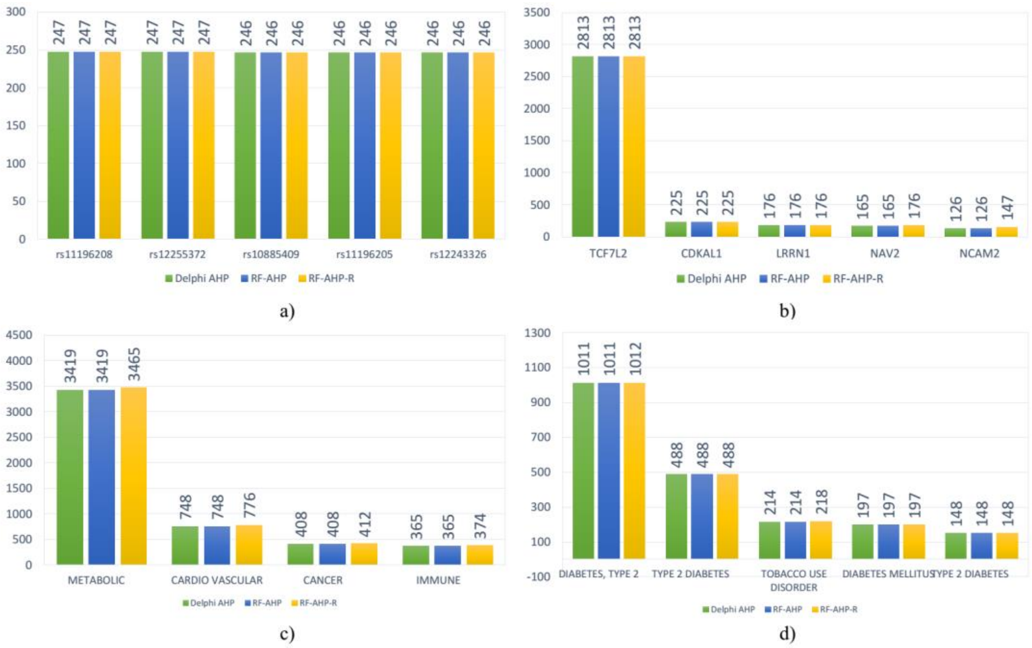
c) T2DM scoring by RF-AHP-R

Figure 25 Distribution of computed AHP Scores for T2DM Delphi AHP, RF-AHP and RF-AHP-R

Amount of available SNPs which have AHP score higher than threshold are shown in Table 8. According to Table 8, it may be seen that the number of available SNPs having AHP scores above 0.1 was the same at 330 SNPs for Delphi AHP and RF-AHP. This shows that designing AHP process using RF was able to provide same response without requiring any expert and no loss of information occurred. By incorporating RegulomeDB data, in RF-AHP-R, number of available SNPs increased to 353 SNPs. For each analysis result a separate SNPNexus job was submitted. Result of each job analyzed and GAD results are shown in Figure 26.

Table 8 No of Available SNPs as a Result of Type 2 Diabetes Analysis

Analysis Name	**No of SNPs whose AHP score > 0.1**
T2DM analysis using Delphi AHP	330
T2DM analysis using RF-AHP	330
T2DM analysis using RF-AHP-R	353



a)

b)

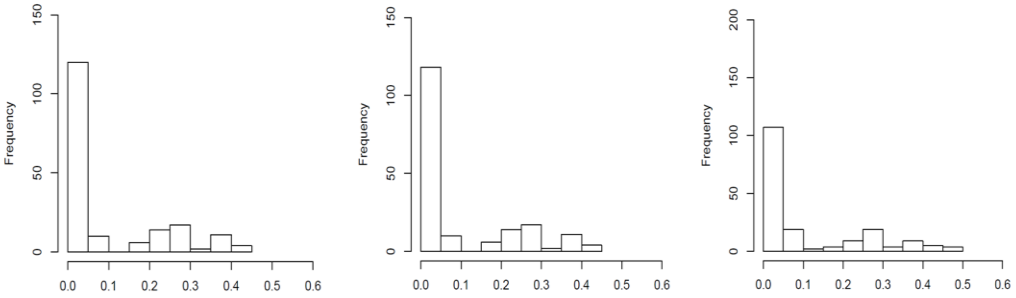
c)

d)

Figure 26 SNPNexus-GAD Frequency Results for T2DM Analysis: a) Most Referenced SNPs b) Most Referenced Genes c) Most Referenced Disease Classes d) Most Referenced Phenotypes

5.4.4 Comparative Analysis of Alzheimer's Disease SNP Prioritization Results for Delphi AHP RF-AHP and RF-AHP-R

AHP Score distribution for Alzheimer's Disease analyses are shown in Figure 27. For Alzheimer's Disease according to general distribution of AHP scores threshold was set to 0.1.



a) AD scoring by Delphi AHP

b) AD scoring by RF- AHP

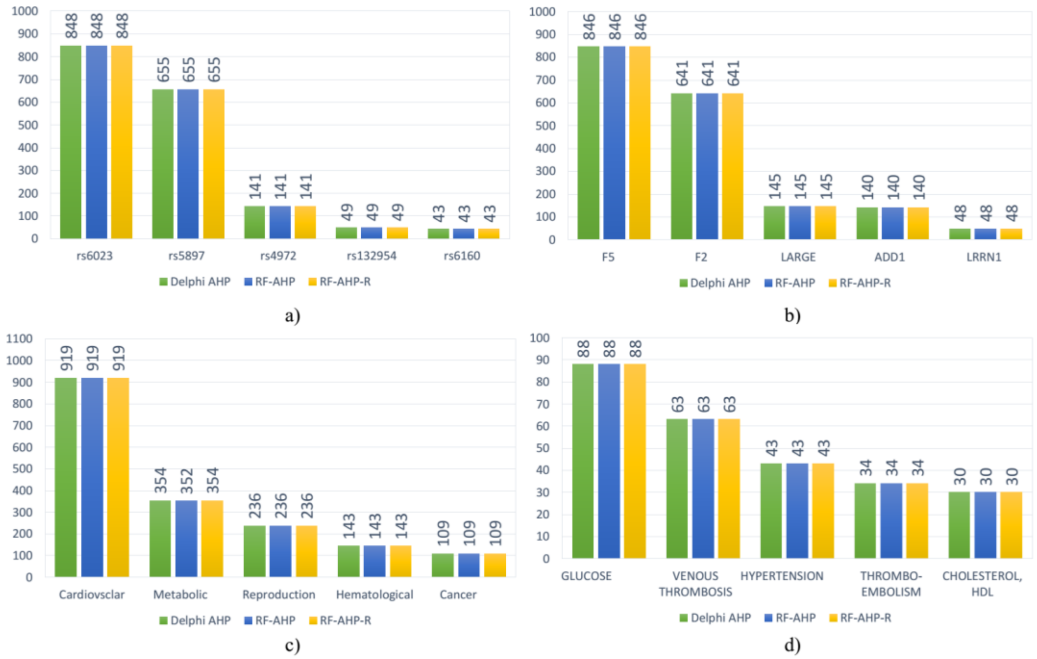
c) AD scoring by RF-AHP-R

Figure 27 Distribution of computed scores for AD in Delphi AHP, RF-AHP and RF-AHP-R

As a result of Alzheimer's Disease GWAS analysis number of significant SNPs above p-value=0.05 was comparably lower than other diseases. Therefore prioritization results contained less SNPs. Amount of available SNPs which have AHP score higher than threshold are shown in Table 9. It may be seen that the number of available SNPs having AHP scores above 0.1 was the same at 54 SNPs for Delphi AHP and RF-AHP. Therefore, RF-AHP was able to provide same response without requiring any expert and no information loss occurred. By incorporating RegulomeDB, in RF-AHP-R, SNPs increased to 56 SNPs. GAD results are shown in Figure 28.

Table 9 No of Available SNPs as a Result of Alzheimer's Disease Analysis

Analysis Name	**No of SNPs whose AHP score > 0.1**
AD analysis using Delphi AHP	54
AD analysis using RF-AHP	54
AD analysis using RF-AHP-R	56



a)

b)

c)

d)

Figure 28 SNPNexus-GAD Frequency Results for Alzheimer’s Disease Analysis: a) Most Referenced SNPs b) Most Referenced Genes c) Most Referenced Disease Classes d) Most Referenced Phenotypes

5.5 Discussion on the Analysis Results

Overall results of this thesis provided a proof-of-concept for Random Forest based AHP (RF-AHP) method to address expert judgment uncertainty in decision making with AHP. The AHP categories are evaluated according to calculated Variable Importances in the trained Random Forest model providing automatic identification of necessary set of features and weights.

Remarkably, according to Figure 16, categories such as Coding_Frameshift and Other_Coding_Nonsynonymous have zero importance and they are omitted in RF-AHP and RF-AHP-R It is logical because if these mutations occur then result is not a complex disease but a monogenic disease with major effects.

When the number of SNPs over an AHP score threshold are compared between Delphi AHP and RF-AHP analyses (Table 6 through Table 9) no change was observed. This proved that designing AHP based on RF importances was successful in representing the same decision making performance.

Use of RegulomeDB incorporated RF-AHP-R method for Sz analysis nearly doubled the number of available SNPs as it may be seen in Table 6. RF-AHP-R scoring also provided insight into Schizophrenia disease. These results supported that, Schizophrenia associated regions of genome are mainly located in regulatory regions. Additionally, in PCa, T2D and AD analyses, the number of available SNPs increased consistently in small amounts for RF-AHP-R because of additional knowledge provided by RegulomeDB table.

The Top 5 most referenced SNPs were the same for all three of the methods for each disease (Figure 22, 24, 26, and 28). Number of most referenced disease classes increased proving that the optimized model is able to link higher number of SNPs to diseases. One interesting result is that most referenced disease class is METABOLIC although its phenotype is a cancer. In the T2DM results, METABOLIC disease class results increased mostly meaning that the RF-AHP model is able to perform better for T2D which is also a metabolic disease. In Schizophrenia, most referenced disease types are METABOLIC and PSYCHOLOGICAL class which are also logical. For Alzheimer CARDIOVASCULAR type is mostly referenced in GAD database. The number of most referenced phenotypes in analyses are consistent between Delphi AHP and RF-AHP models. Proving that the RF-AHP model is able to link same or higher number of SNPs to phenotypes. Moreover both the Delphi AHP and RF-AHP models were able to detect most Diabetes Mellitus type phenotypes successfully. When the most referenced number of genes in SNPNexus query is checked, it was found that they are the same in all methods for each disease.

In summary, pruning the Delphi AHP tree categories did not cause any loss of information. However, when these categories were asked to experts, mistakenly these categories were given high importance. Therefore, data driven approach avoided mistakes due to subjective weighing of categories by experts.

CHAPTER 6

CONCLUSIONS AND FUTURE RESEARCH

6.1 Conclusions

The objective of this thesis work is to provide a solution to subjectivity and uncertainty problem in Analytic Hierarchy Process based decision making caused by the expert evaluations. In particular, a solution for Single Nucleotide Polymorphism (SNP) prioritization problem for disease associated SNP biomarker detection problem has been proposed. The introduction of the problem is provided in the first chapter of this thesis. In the second chapter, the molecular biology background concepts for complex diseases and SNP prioritization are presented to establish a base of knowledge. Related literature for complex disease biomarker discovery by use of SNPs is reviewed. Genome wide association study which is widely used in prior to SNP prioritization is presented. Additionally, background for Analytic Hierarchy Process based decision making and Random Forest based machine learning method, their advantages and disadvantages are reviewed. In the third chapter, materials and methods employed to realize the proposed algorithms are presented. Detailed explanation of utilized data sources such as Schizophrenia, Prostate Cancer, Type 2 Diabetes Mellitus and Alzheimer's Disease is presented. Additionally, software environment in development and testing of the methodologies is outlined. In the fourth chapter, proposed Random Forest based Analytic Hierarchy Process RF-AHP is described in detail. Its advantages for eliminating subjective decisions is explained. A case study for application of RF-AHP to SNP prioritization is presented as a step by step procedure. Moreover, implementation details of RF-AHP-R method is described. With RF-AHP-R, prioritization using SNPs with regulatory functions are provided by incorporation of RegulomeDB.

Results and Discussion of tests performed on implemented methods are presented in the fifth chapter. Performance of the realized RF-AHP is tested through two disease datasets namely Prostate Cancer and Alzheimer's Disease. Comparison of Delphi AHP based method to RF-AHP method and RF-AHP with RegulomeDB i.e. RF-AHP-R is presented. The comparisons of performance for the three methods are shown as various plots. Later, discussion of the results is presented. Finally in this chapter, in the following sections, the contributions and possible future search alternatives are provided.

6.2 Contributions

Proposed Random Forest based Analytic Hierarchy Process (RF-AHP) method to address the expert judgment uncertainty problem in Analytic Hierarchy Process design was showed to be viable for decision making in SNP prioritization. Both important and uninformative AHP categories are identified by using the Random Forest machine learning method using Prostate Cancer dataset [116][117] .

There are studies in the literature, where Random Forest analysis was applied to discover candidate SNPs [128][129]. However, to the best of our knowledge, incorporating Random Forest to the AHP approach in order to provide RF-AHP prioritization was not reported in previous academic studies. In this study, variable importance property of Random Forest method was shown to be useful to identify informative and uninformative categories in a previously developed expert designed Delphi-AHP. Consequently, using the RF-AHP method, AHP based SNP prioritization can be performed without the need for experts, therefore subjectivity in decisions may be eliminated.

Four disease datasets namely Schizophrenia [122], Prostate Cancer [116][117], Type 2 Diabetes Mellitus [118][119] and Alzheimer's Disease[120][121] were used in analyses. RF-AHP and RF-AHP-R compared to Delphi AHP according to SNP prioritization results of these four datasets. As a consequence of comparative analyses, it is concluded that there is no loss of information in the results with respect to the number of prioritized SNPs between three approaches. For instance, Coding_Frameshift and Other_Coding_Nonsynonymous categories were not used in RF-AHP as they were not found to have any impact on the decision. As coding SNPs would have serious consequences on the biological outcome, they are more likely to be the causative changes in single gene disorders. However, complex genetic disorders such as, cancer, diabetes or neurological diseases studied here, are the focus of GWAS. In a complex genetic disorder, many SNPs, spread throughout the genome, contributes to increase disease susceptibility as small effects [55]. Furthermore, the genetic factors associated with the disease are not only the variations between these groups, but there are additional factors such as demographical and clinical findings, lifestyle, and other environmental factors [54][66].

The focus in the applied method was identification of valuable information in each category in the Delphi-AHP tree. Therefore, building a more objective and more efficient AHP tree was possible without use of any experts by Random Forest based evaluation of uninformative categories of the AHP tree.

6.3 Future Research

As a future research on this study, firstly, the proposed method may be used to analyze other complex diseases such as bipolar disorder or other types of cancers. By this way, etiology of other diseases may be discovered by employing RF-AHP.

Secondly, considered SNP annotations include genetic biological and functional annotations. In the analysis of complex diseases there are other factors that should be considered such as clinical phenotypes, real time sensor data and lifestyle information e.g. BMI, smoking and environmental conditions such as air pollution. After these datasets are obtained and incorporated into RF-AHP, we believe the prioritization performance may be further improved.

Thirdly, in the current database, only SNPs are evaluated in RF-AHP. However other types of polymorphisms such as STRs and CNVs which were mentioned in Section 2.3 may be analyzed for finding complex disease related biomarkers.

Finally, performance comparison of RF-AHP may be made to other machine learning methods such as naïve-Bayes or neural networks.

REFERENCES

[1] K. F. Tipton, “Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions.,” *Eur. J. Biochem.*, vol. 223, no. 1, p. 1, 1994.

[2] J. M. Fernández and A. Valencia, *Bioinformatics for personalized medicine*, vol. 6620. 2012.

[3] N. Risch, “Genetic linkage and complex diseases: A response,” *Genet. Epidemiol.*, vol. 7, no. 1, pp. 41–45, 1990.

[4] G. Üstünkar and Y. Aydın Son, “METU-SNP: an integrated software system for SNP-complex disease association analysis.,” *J. Integr. Bioinform.*, vol. 8, p. 187, 2011.

[5] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[6] S. C. Yücebaş and Y. Aydın Son, “A Prostate Cancer Model Build by a Novel SVM-ID3 Hybrid Feature Selection Method Using Both Genotyping and Phenotype Data from dbGaP.,” *PLoS One*, vol. 9, no. 3, p. e91404, 2014.

[7] J. H. Moore, “Bioinformatics,” *Journal of Cellular Physiology*, vol. 213, no. 2. pp. 365–369, 2007.

[8] V. Maojo and B. I. Group, “Bioinformatics : Towards New Directions for Public Health * Bioinformatics – Current Issues,” *Methods Inf Med*, vol. 3, no. 43, pp. 208–214, 2004.

[9] A. J. Butte, “Translational bioinformatics applications in genome medicine,” *Genome Medicine*, vol. 1, no. 6. 2009.

[10] M. Chen and R. Hofestädt, “A medical bioinformatics approach for metabolic disorders: Biomedical data prediction, modeling, and systematic analysis,” *J. Biomed. Inform.*, vol. 39, no. 2, pp. 147–159, 2006.

[11] S. Y. Rhee, J. Dickerson, and D. Xu, "Bioinformatics and Its Applications in Plant Biology," *Annu. Rev. Plant Biol*, vol. 57, pp. 335–60, 2006.

[12] L. Yount, "Biotechnology and genetic engineering," *Libr. a B.*, p. 316 p., 2004.

[13] H. Akbar, F. C. Cardoso, S. Meier, C. Burke, S. Mcdougall, M. Mitchell, C. Walker, S. L. Rodriguez-zas, R. E. Everts, H. a Lewin, J. R. Roche, and J. J. Loor, "Bioinformatics and Biology Insights," *Bioinform. Biol. Insights*, vol. 8, no. i, pp. 45–63, 2014.

[14] C. F. Thorn, T. E. Klein, and R. B. Altman, "Pharmacogenomics and bioinformatics: PharmGKB.," *Pharmacogenomics*, vol. 11, no. 4, pp. 501–505, 2010.

[15] P. A. Martin and R. Dingwall, "Medical Sociology and Genetics," in *The New Blackwell Companion to Medical Sociology*, 2009, pp. 511–529.

[16] C. Hauskeller, S. Sturdy, and R. Tutton, "Genetics and the Sociology of Identity," *Sociology*, vol. 47, no. 5, pp. 875–886, 2013.

[17] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5. pp. 774–781, 2013.

[18] X. Wang and L. Liotta, "Clinical bioinformatics: A new emerging science," *Journal of Clinical Bioinformatics*, vol. 1, no. 1, 2011.

[19] R. J. R. Jiang, F. Z. F. Zeng, W. Z. W. Zhang, X. W. X. Wu, and Z. Y. Z. Yu, "Accelerating Genome-Wide Association Studies Using CUDA Compatible Graphics Processing Units," *2009 Int. Jt. Conf. Bioinformatics, Syst. Biol. Intell. Comput.*, pp. 92–98, 2009.

[20] Q. Z. Li and E. K. Wakeland, "Autoimmune Diseases in the Bioinformatics Paradigm," *Genomics, Proteomics and Bioinformatics*, vol. 13, no. 4, pp. 205–207, 2015.

[21] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[22] D. A. Payne, "Basics of molecular biology," in *Molecular Pathology in*

Clinical Practice:Second Edition, 2016, pp. 1–17.

[23] X. Wu, M. R. Spitz, C. I. Amos, J. Lin, L. Shao, J. Gu, M. De Andrade, N. L. Benowitz, P. G. Shields, and G. E. Swan, “Mutagen sensitivity has high heritability: Evidence from a twin study,” *Cancer Res.*, vol. 66, no. 12, pp. 5993–5996, 2006.

[24] P. S.-W. Fong and S. K.-Y. Choi, “Final contractor selection using the analytical hierarchy process,” *Constr. Manag. Econ.*, vol. 18, no. 5, pp. 547–557, 2000.

[25] K. -hoon. Ng, “Non-Ionizing Radiations - Sources, Biological Effects, Emissions and Exposures,” *Proc. Int. Conf. Non-Ionizing Radiat. UNITEN*, no. October, pp. 1–16, 2003.

[26] F. W. Nussbaum, Robert L; McInnes, Roderick R; Huntington, *Thompson & Thompson Genetics in Medicine*. 2016.

[27] J. M. Mullaney, R. E. Mills, W. Stephen Pittard, and S. E. Devine, “Small insertions and deletions (INDELs) in human genomes,” *Hum. Mol. Genet.*, vol. 19, no. R2, 2010.

[28] R. Nielsen, “Mutations as missing data: Inferences on the ages and distributions of nonsynonymous and synonymous mutations,” *Genetics*, vol. 159, no. 1, pp. 401–411, 2001.

[29] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. 2002.

[30] A. J. Brookes, “The essence of SNPs,” *Gene*, vol. 234, no. 2. pp. 177–186, 1999.

[31] R. Karki, D. Pandya, R. C. Elston, and C. Ferlini, “Defining ‘mutation’ and ‘polymorphism’ in the era of personal genomics,” *BMC Medical Genomics*, vol. 8, no. 1. 2015.

[32] J. E. Seeb, G. Carvalho, L. Hauser, K. Naish, S. Roberts, and L. W. Seeb, “Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms,” *Molecular Ecology Resources*, vol. 11, no. SUPPL. 1, pp. 1–8, 2011.

[33] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–933, 2001.

[34] E. Tuzun, A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M. V. Olson, and E. E. Eichler, "Fine-scale structural variation of the human genome," *Nat. Genet.*, vol. 37, no. 7, pp. 727–732, 2005.

[35] F. Villarroya, M. Peyrou, and M. Giralt, "Transcriptional regulation of the uncoupling protein-1 gene," *Biochimie*, vol. 134. pp. 86–92, 2017.

[36] O. Hobert, "Gene regulation by transcription factors and microRNAs.," *Science*, vol. 319, no. 5871, pp. 1785–6, 2008.

[37] P. J. Farnham, "Insights from genomic profiling of transcription factors," *Nature Reviews Genetics*, vol. 10, no. 9. pp. 605–616, 2009.

[38] J. Dekker, "Gene regulation in the third dimension," *Science*, vol. 319, no. 5871. pp. 1793–1794, 2008.

[39] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz, "Gene regulation at the single-cell level.," *Science*, vol. 307, no. 5717, pp. 1962–5, 2005.

[40] S. M. Park, E. Y. Choi, M. Bae, S. Kim, J. B. Park, H. Yoo, J. K. Choi, Y. J. Kim, S. H. Lee, and I. H. Kim, "Histone variant H3F3A promotes lung cancer cell migration through intronic regulation," *Nat. Commun.*, vol. 7, 2016.

[41] K. A. Beaumont, R. A. Newton, D. J. Smit, J. H. Leonard, J. L. Stow, and R. A. Sturm, "Altered cell surface expression of human MC1R variant receptor alleles associated with red hair and skin cancer risk," *Hum. Mol. Genet.*, vol. 14, no. 15, pp. 2145–2154, 2005.

[42] M. De Gobbi, V. Viprakasit, J. R. Hughes, C. Fisher, V. J. Buckle, H. Ayyub, R. J. Gibbons, D. Vernimmen, Y. Yoshinaga, P. De Jong, J. F. Cheng, E. M. Rubin, W. G. Wood, D. Bowden, and D. R. Higgs, “A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter,” *Science (80-.).*, vol. 312, no. 5777, pp. 1215–1217, 2006.

[43] X. Zhou, T. B. Barrett, and J. R. Kelsoe, “Promoter Variant in the GRK3 Gene Associated with Bipolar Disorder Alters Gene Expression,” *Biol. Psychiatry*, vol. 64, no. 2, pp. 104–110, 2008.

[44] ENCODE Project Consortium, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[45] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder, “Annotation of functional variation in personal genomes using RegulomeDB,” *Genome Res.*, vol. 22, no. 9, pp. 1790–1797, 2012.

[46] A. Pemov, “DNA analysis with multiplex microarray-enhanced PCR,” *Nucleic Acids Res.*, vol. 33, no. 2, pp. e11–e11, 2005.

[47] L. S. Kristensen and L. L. Hansen, “PCR-based methods for detecting single-locus DNA methylation biomarkers in cancer diagnostics, prognostics, and response to treatment,” *Clinical Chemistry*, vol. 55, no. 8. pp. 1471–1483, 2009.

[48] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, “Ten years of next-generation sequencing technology,” *Trends Genet.*, vol. 30, no. 9, pp. 418–426, 2014.

[49] M. Oti and H. G. Brunner, “The modular nature of genetic diseases,” *Clinical Genetics*, vol. 71, no. 1. pp. 1–11, 2007.

[50] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, “Human genetic variation and its contribution to complex traits,” *Nature Reviews Genetics*, vol. 10, no. 4. pp. 241–251, 2009.

[51] T. A. Manolio, L. L. Rodriguez, L. Brooks, G. Abecasis, D. Ballinger, M. Daly, P. Donnelly, S. V Faraone, K. Frazer, S. Gabriel, and others, “New models of collaboration in genome-wide association studies: the Genetic Association Information Network,” *Nat. Genet.*, vol. 39, no. 9, pp. 1045–

1051, 2007.

[52] N. Risch and K. Merikangas, “The Future of Genetic Studies of Complex Human Diseases,” *Science (80-.).*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[53] S. C. Yu, “A Prostate Cancer Model Build by a Novel SVM-ID3 Hybrid Feature Selection Method Using Both Genotyping and Phenotype Data from dbGaP,” vol. 9, no. 3, pp. 1–8, 2014.

[54] W. H. Wei, G. Hemani, and C. S. Haley, “Detecting epistasis in human complex traits,” *Nat. Rev. Genet.*, vol. 15, no. 11, pp. 722–733, 2014.

[55] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn, “Genome-wide association studies for complex traits: Consensus, uncertainty and challenges,” *Nature Reviews Genetics*, vol. 9, no. 5. pp. 356–369, 2008.

[56] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, “GCTA: A tool for genome-wide complex trait analysis,” *Am. J. Hum. Genet.*, vol. 88, no. 1, pp. 76–82, 2011.

[57] P. M. Visscher and J. Yang, “A plethora of pleiotropy across complex traits,” *Nature Genetics*, vol. 48, no. 7. pp. 707–708, 2016.

[58] N. R. Wray, J. Yang, B. J. Hayes, A. L. Price, M. E. Goddard, and P. M. Visscher, “Pitfalls of predicting complex traits from SNPs,” *Nat. Rev. Genet.*, vol. 14, no. 7, pp. 507–515, 2013.

[59] J. M. Seddon, R. Reynolds, J. Maller, J. A. Fagerness, M. J. Daly, and B. Rosner, “Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables,” *Investig. Ophthalmol. Vis. Sci.*, vol. 50, no. 5, pp. 2044–2053, 2009.

[60] J. A. G. M. de Visser, T. F. Cooper, and S. F. Elena, “The causes of epistasis,” *Proc. R. Soc. B Biol. Sci.*, vol. 278, no. 1725, pp. 3617–3624, 2011.

[61] R. Sanjuan and S. F. Elena, “Epistasis correlates to genomic complexity,” *Proc. Natl. Acad. Sci.*, vol. 103, no. 39, pp. 14402–14405, 2006.

[62] T. R. Hughes, “Universal epistasis analysis,” *Nature Genetics*, vol. 37, no. 5.

pp. 457–458, 2005.

[63] M. F. Schenk, I. G. Szendro, M. L. M. Salverda, J. Krug, and J. A. G. M. De Visser, “Patterns of epistasis between beneficial mutations in an antibiotic resistance gene,” *Mol. Biol. Evol.*, vol. 30, no. 8, pp. 1779–1787, 2013.

[64] H. J. Cordell, “Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans,” *Hum. Mol. Genet.*, vol. 11, no. 20, pp. 2463–2468, 2002.

[65] J. Zhang and G. P. Wagner, “On the definition and measurement of pleiotropy,” *Trends in Genetics*, vol. 29, no. 7. pp. 383–384, 2013.

[66] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, “Pleiotropy in complex traits: Challenges and strategies,” *Nat. Rev. Genet.*, vol. 14, no. 7, pp. 483–495, 2013.

[67] A. B. Paaby and M. V. Rockman, “The many faces of pleiotropy,” *Trends in Genetics*, vol. 29, no. 2. pp. 66–73, 2013.

[68] O. A. Andreassen, W. K. Thompson, A. J. Schork, S. Ripke, M. Mattingsdal, J. R. Kelsoe, K. S. Kendler, M. C. O’Donovan, D. Rujescu, T. Werge, P. Sklar, J. C. Roddey, C.-H. Chen, L. McEvoy, R. S. Desikan, S. Djurovic, A. M. Dale, C. The Psychiatric Genomics, D. Bipolar, and G. Schizophrenia Working, “Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate,” *PLoS Genet*, vol. 9, no. 4, p. e1003455, 2013.

[69] J. Hodgkin, “Seven types of pleiotropy,” *International Journal of Developmental Biology*, vol. 42, no. 3. pp. 501–505, 1998.

[70] P. W. HEDRICK, “Antagonistic pleiotropy and genetic polymorphism: a perspective,” *Heredity (Edinb).*, vol. 82, no. 2, pp. 126–133, 1999.

[71] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, “Five years of GWAS discovery,” *American Journal of Human Genetics*, vol. 90, no. 1. pp. 7–24, 2012.

[72] R. M. Cantor, K. Lange, and J. S. Sinsheimer, “Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application,” *Am. J. Hum. Genet.*, vol. 86, no. 1, pp. 6–22, 2010.

[73] J. Flint, “GWAS,” *Current Biology*, vol. 23, no. 7. pp. R265–R266, 2013.

[74] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. De Bakker, M. J. Daly, and others, “PLINK: a tool set for whole-genome association and population-based linkage analyses,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, 2007.

[75] Y. Aydın Son, Ş. Tüzmen, and C. Hizel, “Designing and implementing pharmacogenomics study,” in *Omics for Personalized Medicine*, 2013, pp. 97–122.

[76] P. Marjoram, A. Zubair, and S. V. Nuzhdin, “Post-GWAS: Where next More samples, more SNPs or more biology,” *Heredity*, vol. 112, no. 1. pp. 79–88, 2014.

[77] C. Liu and Z. Xuan, “Prioritization of cancer-related genomic variants by SNP association network,” *Cancer Inform.*, vol. 14, pp. 57–70, 2015.

[78] S. F. Saccone, R. Bolze, P. Thomas, J. Quan, G. Mehta, E. Deelman, J. a. Tischfield, and J. P. Rice, “SPOT: A web-based tool for using biological databases to prioritize SNPs after a genome-wide association study,” *Nucleic Acids Res.*, vol. 38, no. June, pp. 201–209, 2010.

[79] A. R. Pico, I. V. Smirnov, J. S. Chang, R. F. Yeh, J. I. Wiemels, J. K. Wiencke, T. Tihan, B. R. Conklin, and M. Wrensch, “SNPLogic: An interactive single nucleotide polymorphism selection, annotation, and prioritization system,” *Nucleic Acids Res.*, vol. 37, no. SUPPL. 1, 2009.

[80] H. Y. Yuan, J. J. Chiou, W. H. Tseng, C. H. Liu, C. K. Liu, Y. J. Lin, H. H. Wang, A. Yao, Y. T. Chen, and C. N. Hsu, “FASTSNP: An always up-to-date and extendable service for SNP function analysis and prioritization,” *Nucleic Acids Res.*, vol. 34, no. WEB. SERV. ISS., 2006.

[81] B. L. Fridley, E. Iversen, Y. Y. Tsai, G. D. Jenkins, E. L. Goode, and T. A. Sellers, “A latent model for prioritization of SNPs for functional studies,” *PLoS One*, vol. 6, no. 6, 2011.

[82] J. Che and M. Shin, “A meta-analysis strategy for gene prioritization using gene expression, SNP genotype, and eQTL data,” *Biomed Res. Int.*, vol. 2015, 2015.

[83] M. Velasquez and P. T. Hester, “An Analysis of Multi-Criteria Decision Making Methods,” vol. 10, no. 2, pp. 56–66, 2013.

[84] J. R. Yu, Y. Hsiao, and H. Sheu, “A Multiplicative Approach to Derive Weights in the Interval Analytic Hierarchy Process,” vol. 13, no. 3, pp. 225–231, 2011.

[85] H. Chen-yi, C. Ke-ting, and T. Gwo-hshiung, “FMCDM with Fuzzy DEMATEL Approach for Customers ’ Choice Behavior Model,” vol. 9, no. 4, pp. 236–246, 2007.

[86] J.-P. Brans, “The management of the future: Ethics in OR: Respect, multicriteria management, happiness,” *Eur. J. Oper. Res.*, vol. 153, no. 2, pp. 466–467, 2004.

[87] C.-H. Yeh, H. Deng, S. Wibowo, and Y. Xu, “Fuzzy multicriteria decision support for information systems project selection,” *Int. J. Fuzzy Syst.*, vol. 12, no. 2, pp. 170–174, 2010.

[88] L. Saaty Thomas, “The analytic Hierarchy process,” *New York McGrow-Hill*, 1980.

[89] T. L. S. Ã and J. S. Shang, “Group decision-making : Head-count versus intensity of preference,” vol. 41, pp. 22–37, 2007.

[90] K. M. A.-S. Al-Harbi, “Application of the AHP in project management,” *Int. J. Proj. Manag.*, vol. 19, no. 1, pp. 19–27, 2001.

[91] A. H. S. Chan, W. Y. Kwok, and V. G. Duffy, “Using AHP for determining priority in a safety management system,” *Ind. Manag. Data Syst.*, vol. 104, no. 5, pp. 430–445, 2004.

[92] E. W. L. Cheng and H. Li, “Contractor selection using the analytic network process,” *Constr. Manag. Econ.*, vol. 22, no. 10, pp. 1021–1032, 2004.

[93] “Product quality evaluation system based on AHP fuzzy comprehensive evaluation _ Xi _ Journal of Industrial Engineering and Management.” .

[94] S. M. Mousavi, R. Tavakkoli-Moghaddam, M. Heydar, and S. Ebrahimnejad, “Multi-Criteria Decision Making for Plant Location Selection: An Integrated Delphi-AHP-PROMETHEE Methodology,” *Arab. J. Sci. Eng.*, vol. 38, no.

5, pp. 1255–1268, 2013.

[95] B. Hwang, N. Pai, and C. Wu, "Fuzzy AHP for determining the key features and cognitive differences of mobile game development among designer and game player," *Multimed. Tools Appl.*, no. 140, 2016.

[96] D. Lewandowski, Q. Zhu, J. Tina Du, F. Meng, K. Wu, and X. Sun, "Using a Delphi method and the analytic hierarchy process to evaluate Chinese search engines: A case study on Chinese search engines," *Online Inf. Rev.*, vol. 35, no. 6, pp. 942–956, 2011.

[97] A. Hasan, "Security of Cross-Country Oil and Gas Pipelines : A Risk-Based Model," vol. 7, no. 3, pp. 1–8, 2016.

[98] T.-Y. Hsieh, S.-T. Lu, and G.-H. Tzeng, "Fuzzy MCDM approach for planning and design tenders selection in public office buildings," *Int. J. Proj. Manag.*, vol. 22, no. 7, pp. 573–584, 2004.

[99] A. Jindal and K. S. Sangwan, "An integrated fuzzy multi-criteria evaluation of sustainable reverse logistics network models," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, 2013, pp. 1–7.

[100] M. Bouzon, K. Govindan, C. M. Taboada, and L. M. S. Campos, "Resources , Conservation and Recycling Identification and analysis of reverse logistics barriers using fuzzy Delphi method and AHP," *"Resources, Conserv. Recycl.*, vol. 108, pp. 182–197, 2016.

[101] L. Lun and P. K. Leng, "Does ' Fuzzifying ' AHP Improve the Quality of Multi-Attribute Decision Making ?," pp. 1–14.

[102] O. I. Review, "Using a Delphi method and the Analytic Hierarchy Process to evaluate the search engines : A case study on Chinese search," no. November 2016, 2011.

[103] S. M. Mousavi, R. Tavakkoli-Moghaddam, M. Heydar, and S. Ebrahimnejad, "Multi-criteria decision making for plant location selection: an integrated Delphi--AHP--PROMETHEE methodology," *Arab. J. Sci. Eng.*, vol. 38, no. 5, pp. 1255–1268, 2013.

[104] N. Dalkey and O. Helmer, "An experimental application of the Delphi method to the use of experts," *Manage. Sci.*, vol. 9, no. 3, pp. 458–467, 1963.

[105] Z. Nong, P. Dong-jiang, and Z. Yi-ming, “Evaluation of Relative Mining Intensity in Western China Based on Interval Analytic Hierarchy Process,” pp. 2941–2953.

[106] G. Tajik, A. H. Azadnia, S. A. H. S. Hassan, and others, “A hybrid fuzzy MCDM approach for sustainable third-party reverse logistics provider selection,” in *Advanced Materials Research*, 2014, vol. 845, pp. 521–526.

[107] L. A. Zadeh, “Fuzzy sets,” *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[108] D. Kannan, A. B. L. de Sousa Jabbour, and C. J. C. Jabbour, “Selecting green suppliers based on GSCM practices: Using Fuzzy TOPSIS applied to a Brazilian electronics company,” *Eur. J. Oper. Res.*, vol. 233, no. 2, pp. 432–447, 2014.

[109] M. Ghorbani, S. Mohammad Arabzad, and A. Shahin, “A novel approach for supplier selection based on the Kano model and fuzzy MCDM,” *Int. J. Prod. Res.*, vol. 51, no. 18, pp. 5469–5484, 2013.

[110] G. Biau and E. Scornet, “A Random Forest Guided Tour,” *arXiv.org*, vol. math.ST, pp. 1–35, 2015.

[111] B. Fröhlich, E. Rodner, M. Kemmler, and J. Denzler, “Large-scale Gaussian Process Classification Using Random Decision Forests,” *Pattern Recognit. Image Anal.*, vol. 22, no. 1, pp. 113–120, 2012.

[112] a Liaw and M. Wiener, “Classification and Regression by randomForest,” *R news*, vol. 2, no. December, pp. 18–22, 2002.

[113] D. F. Schwarz, I. R. König, and A. Ziegler, “On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data.,” *Bioinformatics*, vol. 26, no. 14, pp. 1752–8, Jul. 2010.

[114] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, “Conditional variable importance for random forests.,” *BMC Bioinformatics*, vol. 9, no. 1, p. 307, Jan. 2008.

[115] A.-L. Boulesteix, A. Bender, J. Lorenzo Bermejo, and C. Strobl, “Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations,” *Brief. Bioinform.*, vol. 13, no. 3, pp. 292–304, Sep. 2011.

[116] L. N. Kolonel, B. E. Henderson, J. H. Hankin, A. M. Y. Nomura, L. R. Wilkens, M. C. Pike, D. O. Stram, K. R. Monroe, M. E. Earle, and F. S. Nagamine, “A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics,” *Am. J. Epidemiol.*, vol. 151, no. 4, pp. 346–357, 2000.

[117] C. A. Haiman, N. Patterson, M. L. Freedman, S. R. Myers, M. C. Pike, A. Waliszewska, J. Neubauer, A. Tandon, C. Schirmer, G. J. McDonald, and others, “Multiple regions within 8q24 independently affect risk for prostate cancer,” *Nat. Genet.*, vol. 39, no. 5, pp. 638–644, 2007.

[118] F. B. Hu, J. E. Manson, M. J. Stampfer, G. Colditz, S. Liu, C. G. Solomon, and W. C. Willett, “Diet, lifestyle, and the risk of type 2 diabetes mellitus in women,” *N. Engl. J. Med.*, vol. 345, no. 11, pp. 790–797, 2001.

[119] R. M. van Dam, E. B. Rimm, W. C. Willett, M. J. Stampfer, and F. B. Hu, “Dietary patterns and risk for type 2 diabetes mellitus in US men,” *Ann. Intern. Med.*, vol. 136, no. 3, pp. 201–209, 2002.

[120] N. Filippini, A. Rao, S. Wetten, R. A. Gibson, M. Borrie, D. Guzman, A. Kertesz, I. Loy-English, J. Williams, T. Nichols, and others, “Anatomically-distinct genetic associations of APOE ε4 allele load with regional cortical atrophy in Alzheimer’s disease,” *Neuroimage*, vol. 44, no. 3, pp. 724–728, 2009.

[121] H. Li, S. Wetten, L. Li, P. L. S. Jean, R. Upmanyu, L. Surh, D. Hosford, M. R. Barnes, J. D. Briley, M. Borrie, and others, “Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease,” *Arch. Neurol.*, vol. 65, no. 1, pp. 45–53, 2008.

[122] B. K. Suarez, J. Duan, A. R. Sanders, A. L. Hinrichs, C. H. Jin, C. Hou, N. G. Buccola, N. Hale, A. N. Weilbaecher, D. A. Nertney, and others, “Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23. 3-p21. 2 and 11p13. 1-q14. 1 in the combined sample,” *Am. J. Hum. Genet.*, vol. 78, no. 2, pp. 315–333, 2006.

[123] T. L. Saaty, “Decision making with the analytic hierarchy process,” *Int. J. Serv. Sci.*, vol. 1, no. 1, p. 83, 2008.

[124] X. Ding, J. Wang, A. Zelikovsky, X. Guo, M. Xie, and Y. Pan, “Searching high-order SNP combinations for complex diseases based on energy

distribution difference," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 12, no. 3, pp. 695–704, 2015.

[125] B. C. Kim, W. Y. Kim, D. Park, W. H. Chung, K. Shin, and J. Bhak, "SNP@Promoter: A database of human SNPs (Single Nucleotide Polymorphisms) within the putative promoter regions," in *Asia Pacific Bioinformatics Network (APBioNet) 6th International Conference on Bioinformatics, InCoB 2007 - Proceedings*, 2007, vol. 9, no. SUPPL. 1.

[126] D. Menendez, O. Krysiak, A. Inga, B. Krysiak, M. A. Resnick, and G. Schönfelder, "A SNP in the flt-1 promoter integrates the VEGF system into the p53 transcriptional network.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 5, pp. 1406–11, 2006.

[127] J. Lamba, V. Lamba, S. Strom, R. Venkataramanan, and E. Schuetz, "Novel single nucleotide polymorphisms in the promoter and intron 1 of human pregnane X receptor/NR1I2 and their association with CYP3A4 expression," *Drug Metab. Dispos.*, vol. 36, no. 1, pp. 169–181, 2008.

[128] Y. A. Meng, Y. Yu, L. A. Cupples, L. A. Farrer, and K. L. Lunetta, "Performance of random forest when SNPs are in linkage disequilibrium," *BMC Bioinformatics*, vol. 10, no. 1, p. 78, 2009.

[129] L. Zou, Q. Huang, A. Li, and M. Wang, "A genome-wide association study of Alzheimer's disease using random forests and enrichment analysis," *Science China Life Sciences*, vol. 55, no. 7. pp. 618–625, 2012.

APPENDICES

APPENDIX A

LIST OF SNPs FROM PROSTATE CANCER GENOMIC MODEL [6]

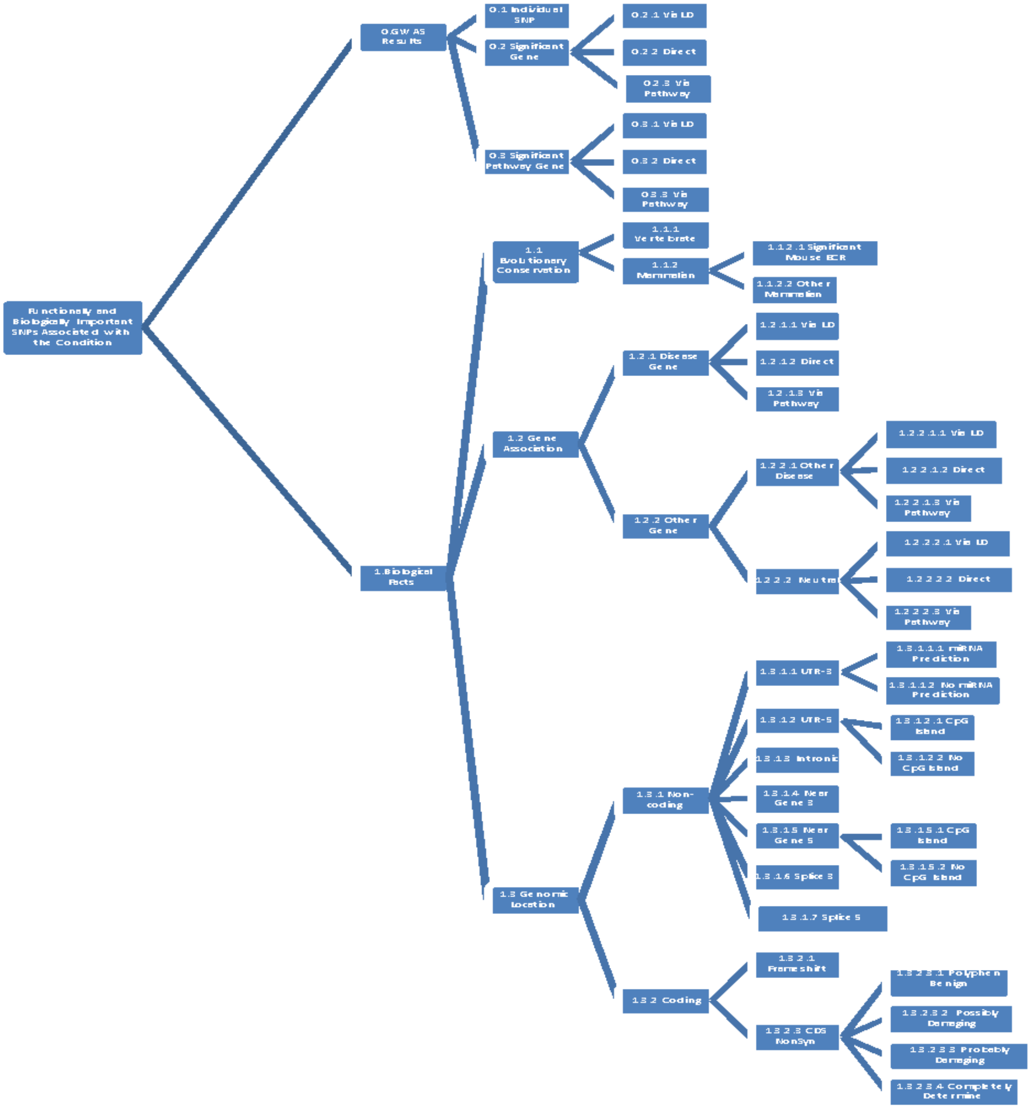
SNPs: 1-60	**SNPs: 61-108**
rs2442602	rs16863955
rs11729739	rs504207
rs17363393	rs17152800
rs7562894	rs12980509
rs17701543	rs12119983
rs3093679	rs9963110
rs280986	rs10068915
rs17595858	rs2296370
rs9848588	rs6708126
rs9347691	rs960278
rs11790106	rs1020235
rs5972169	rs7843255
rs964130	rs2853668
rs6851444	rs2115101
rs11126869	rs10106027
rs4782945	rs2194505
rs10195113	rs524534

rs11086671	rs2602296
rs7775829	rs17111584
rs12243805	rs2120806
rs1433369	rs17799219
rs6887293	rs17400029
rs9401290	rs17178580
rs1454186	rs10517581
rs12733054	rs7183502
rs3812906	rs2948268
rs17284653	rs3760903
rs4827384	rs2103869
rs17375010	rs13011951
rs6549458	rs11685549
rs1379015	rs11584032
rs1122170	rs10788555
rs766045	rs12266639
rs2666205	rs6676372
rs1965340	rs4562278
rs501700	rs7067548
rs12201462	rs2826802
rs7010457	rs4793790
rs6704731	rs11885120
rs17432165	rs17001078
rs4908656	rs7024840
rs10854395	rs2711134
rs6475584	rs7584223
rs1470494	rs918285
rs9462806	rs197265
rs12644498	rs4517938
rs7876199	rs7152946
rs744346	rs7034430
rs1974562	rs517036
rs12247568	rs340542

rs17673975	rs1074525
rs6774902	
rs10954845	
rs6686571	
rs6779266	
rs6747704	

APPENDIX B

DELPHI ANALYTIC HIERARCHY PROCESS TREE (Adapted from [4])



APPENDIX C

SAMPLE QUESTIONNAIRE FOR DELPHI-AHP CATEGORY WEIGHTS BY EXPERTS [4]

Priority vectors	**Description**	**Expert 1**	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**
0	Gwas Results	0.33	0.25	0.83	0.14	0.36
1	Biological Facts	0.67	0.75	0.17	0.86	0.64
0.1	Individual SNP	0.07	0.07	0.08	0.16	0.06
0.2	Significant -Gene	0.64	0.64	0.19	0.75	0.27
0.2.1	Significant Gene - Via LD	0.06	0.11	0.07	0.07	0.11
0.2.2	Significant Gene - Via Direct	0.66	0.63	0.75	0.81	0.33
0.2.3	Significant Gene - Via Pathway	0.28	0.26	0.18	0.12	0.56
0.3	Significant Pathway Gene	0.28	0.28	0.72	0.09	0.67
0.3.1	Significant Pathway Gene - Via LD	0.06	0.11	0.08	0.11	0.11
0.3.2	Significant Pathway Gene - Via Direct	0.66	0.63	0.69	0.7	0.33
0.3.3	Significant Pathway Gene - Via Pathway	0.28	0.26	0.23	0.19	0.56
1.1	Evolutionary Conservation	0.06	0.12	0.26	0.31	0.11
1.1.1	Vertebrate	0.33	0.17	0.9	0.13	0.25
1.1.2	Mammalian	0.67	0.83	0.1	0.88	0.75
1.1.2.1	Mammalian - Significant Mouse ECR	0.67	0.83	0.25	0.75	0.8
1.1.2.2	Mammalian - Other Mammalian	0.33	0.17	0.75	0.25	0.2
1.2	Gene Association	0.66	0.32	0.63	0.62	0.58
1.2.1	Disease Gene	0.9	0.83	0.88	0.9	0.88
1.2.1.1	Disease Gene - Via LD	0.06	0.11	0.09	0.23	0.11

APPENDIX D

DELPHI-AHP, RF-AHP AND RF-AHP-R SCORES FOR ANALYZED DISEASES

TOP 20 DELPHI-AHP, RF-AHP AND RF-AHP-R SCORES FOR ALZHEIMER’S DISEASE

AD DELPHI-AHP Scores:

rs879	0.435252
rs2668	0.422149
rs11523	0.421324
rs6160	0.412137
rs3084	0.398197
rs1237	0.398197
rs7384	0.376914
rs14810	0.375438
rs11522	0.375438
rs897530	0.374533
rs783305	0.366587
rs7769	0.353567
rs9511	0.352091
rs1615	0.351774
rs42019	0.351774
rs5897	0.323093
rs4972	0.320616
rs4161	0.298508
rs17032	0.292289
rs138222	0.291147

AD RF-AHP Scores:

rs879	0.43411
rs2668	0.421324
rs11523	0.421324
rs6160	0.412137
rs3084	0.397055
rs1237	0.397055
rs14810	0.374296
rs7384	0.374296
rs11522	0.374296
rs897530	0.373708
rs783305	0.365762
rs9511	0.350949
rs1615	0.350949
rs7769	0.350949
rs42019	0.350949
rs5897	0.323093
rs4972	0.320616
rs4161	0.297683
rs17032	0.291147
rs138222	0.291147

AD RF-AHP-R Scores:

rs3084	0.474415
rs11523	0.460004
rs879	0.45345
rs897530	0.451068
rs7384	0.432316
rs2668	0.421324
rs6160	0.412137
rs7769	0.408969
rs5897	0.400453
rs1237	0.397055
rs14810	0.393636
rs11522	0.393636
rs783305	0.385102
rs4161	0.375043
rs9511	0.370289
rs1615	0.370289
rs17032	0.368507
rs42019	0.350949
rs4972	0.339956
rs138222	0.329827

TOP20 DELPHI-AHP, RF-AHP AND RF-AHP-R SCORES FOR PROSTATE CANCER DISEASE

PCa DELPHI-AHP Scores:

rs3912492	0.338913
rs12636081	0.338913
rs17061864	0.338913
rs6803449	0.338913
rs1801701	0.215257
rs4794488	0.213299
rs77905	0.21326
rs12948056	0.212474
rs1433369	0.198217
rs1608114	0.191643
rs16930396	0.191326
rs1915940	0.191326
rs2574824	0.18345
rs7249230	0.177416
rs11563056	0.176244
rs8064691	0.176244
rs12592981	0.176244
rs531572	0.176244
rs965560	0.160103
rs138726	0.159278

PCa RF-AHP Scores:

rs3912492	0.338088
rs12636081	0.338088
rs17061864	0.338088
rs6803449	0.338088
rs1801701	0.215257
rs77905	0.21326
rs12948056	0.212474
rs4794488	0.212474
rs1433369	0.197392
rs16930396	0.190501
rs1608114	0.190501
rs1915940	0.190501
rs2574824	0.182625
rs7249230	0.177416
rs11563056	0.175419
rs8064691	0.175419
rs12592981	0.175419
rs531572	0.175419
rs965560	0.159278
rs138726	0.159278

PCa RF-AHP-R Scores:

rs3912492	0.357428
rs12636081	0.357428
rs17061864	0.338088
rs6803449	0.338088
rs1433369	0.274752
rs1801701	0.253937
rs1608114	0.248521
rs77905	0.2326
rs12948056	0.231814
rs666721	0.23002
rs16930396	0.229181
rs4794488	0.212474
rs3782851	0.21068
rs11695247	0.210092
rs138000	0.210092
rs1915940	0.209841
rs680949	0.206673
rs11253552	0.206673
rs6949101	0.206673
rs2574824	0.201965

TOP20 DELPHI-AHP, RF-AHP AND RF-AHP-R SCORES FOR SCHIZOPHRENIA DISEASE

Sz DELPHI-AHP Scores:

rs17115004	0.737535
rs3793504	0.737535
rs2229163	0.73671
rs7009117	0.71945
rs6589360	0.715562
rs7128875	0.715562
rs6475523	0.715562
rs16895119	0.714737
rs17011998	0.692803
rs2982712	0.692215
rs12295969	0.677721
rs11819808	0.677721
rs11568942	0.677721
rs720024	0.677721
rs7074934	0.677721
rs7111410	0.677721
rs16848098	0.677721
rs5030351	0.677721
rs17021884	0.677721
rs6669695	0.677721

Sz RF-AHP Scores:

rs17115004	0.73671
rs2229163	0.73671
rs3793504	0.73671
rs7009117	0.71945
rs6589360	0.714737
rs7128875	0.714737
rs6475523	0.714737
rs16895119	0.714737
rs17011998	0.691978
rs2982712	0.69139
rs12295969	0.676896
rs11819808	0.676896
rs11568942	0.676896
rs720024	0.676896
rs7074934	0.676896
rs7111410	0.676896
rs16848098	0.676896
rs5030351	0.676896
rs17021884	0.676896
rs6669695	0.676896

Sz RF-AHP-R Scores:

rs7009117	0.77747
rs2229163	0.77539
rs3793504	0.77539
rs7541690	0.754256
rs6589360	0.753417
rs17115004	0.73671
rs4655836	0.734916
rs7128875	0.734077
rs16895119	0.734077
rs10790976	0.730909
rs2227284	0.730909
rs7019331	0.730909
rs10757185	0.715576
rs7875344	0.715576
rs41368546	0.715576
rs447	0.715576
rs3213219	0.715576
rs951240	0.715576
rs6475523	0.714737
rs7722406	0.712157

TOP20 DELPHI-AHP, RF-AHP AND RF-AHP-R SCORES FOR TYPE 2 DIABETES MELLITUS DISEASE

T2DM DELPHI-AHP Scores:

rs12592542	0.506199
rs3935795	0.491705
rs3935794	0.491705
rs3935796	0.491705
rs11593943	0.491705
rs16886364	0.491705
rs16886448	0.491705
rs17109221	0.491705
rs10841843	0.468946
rs190092	0.468358
rs12907278	0.468358
rs17764096	0.468358
rs7153625	0.468358
rs7154599	0.468358
rs6866823	0.468358
rs6871286	0.468358
rs6886001	0.468358
rs1979398	0.468358
rs4685598	0.468358
rs7649544	0.468358

T2DM RF-AHP Scores:

rs12592542	0.505374
rs3935795	0.49088
rs3935794	0.49088
rs3935796	0.49088
rs11593943	0.49088
rs16886364	0.49088
rs16886448	0.49088
rs17109221	0.49088
rs10841843	0.468121
rs190092	0.467533
rs12907278	0.467533
rs17764096	0.467533
rs7153625	0.467533
rs7154599	0.467533
rs6866823	0.467533
rs6871286	0.467533
rs6886001	0.467533
rs1979398	0.467533
rs4685598	0.467533
rs7649544	0.467533

T2DM RF-AHP-R Scores:

rs3935795	0.56824
rs3935794	0.56824
rs3935796	0.52956
rs7144011	0.525553
rs12592542	0.524714
rs11593943	0.51022
rs16886364	0.51022
rs17109221	0.51022
rs17764096	0.506213
rs10518694	0.492916
rs228768	0.492916
rs16886448	0.49088
rs11603383	0.489748
rs1402002	0.489748
rs1979398	0.486873
rs4685598	0.486873
rs11693602	0.470996
rs4077463	0.470996
rs10841843	0.468121
rs190092	0.467533

APPENDIX E

AVAILABLE SNPS FOR AHP, RF-AHP RF-AHP-R METHODS FOR PROSTATE CANCER, TYPE 2 DIABETES, SCHIZOPHRENIA AND ALZHEIMER'S DISEASE ANALYSES

Table E.1 No of Available SNPs as a Result of Prostate Cancer Analysis

Analysis Name	**No of SNPs whose AHP score > 0.1**
PCa analysis with Delphi-AHP	121
PCa analysis with RF-AHP	121
PCa analysis with RF-AHP-R	140

Table E.2 No of Available SNPs as a Result of Type 2 Diabetes Analysis

Analysis Name	**No of SNPs whose AHP score > 0.1**
T2D analysis with Delphi-AHP	330
T2D analysis with RF-AHP	330
T2D analysis with RF-AHP-R	353

Table E.3 No of Available SNPs as a Result of Schizophrenia Analysis

Analysis Name	**No of SNPs whose AHP score > 0.5**
Sz analysis with Delphi-AHP	961
Sz analysis with RF-AHP	959
Sz analysis with RF-AHP-R	2424

Table E.4 No of Available SNPs as a Result of Alzheimer’s Disease Analysis

Analysis Name	**No of SNPs whose AHP score > 0.1**
AD analysis with Delphi-AHP	54
AD analysis with RF-AHP	54
AD analysis with RF-AHP-R	56

APPENDIX F

GAD RESULTS FOR PROSTATE CANCER DISEASE ANALYSIS IN AHP, RF-AHP RF-AHP-REGULOME METHODS

Table F.1 No of most referenced SNPs in the results of GAD query

SNP#	Delphi-AHP	RF-AHP	RF-AHP-R
rs1801701	272	272	272
rs531572	90	90	90
rs77905	88	88	88
rs8177812	55	55	55
rs12636081	50	50	50
rs17061864	50	50	50
(Other)	1437	1437	1531

Table F.2 No of most referenced disease classes in the results of GAD query

Disease Class	Delphi-AHP	RF-AHP	RF-AHP-R
METABOLIC	621	621	672
CARDIOVASCULAR	342	342	356
PSYCH	213	213	214
CANCER	188	188	190
CHEMDEPENDENCY	124	124	131
NEUROLOGICAL	101	101	105
(Other)	453	453	468

Table F.3 No of most referenced phenotypes in the results of GAD query

Phenotype	Delphi-AHP	RF-AHP	RF-AHP-R
Glucose	90	90	92
Cholesterol, HDL	77	77	86
Tobacco Use Disorder	69	69	71
Cholesterol, LDL	58	58	58
Waist Circumference	54	54	58
Menopause	46	46	51
(Other)	1648	1648	1720

Table F.4 No of most referenced genes in the results of GAD query

Gene	Delphi-AHP	RF-AHP	RF-AHP-R
APOB	269	269	269
LARGE	145	145	160
LRRN1	144	144	145
FHIT	119	119	119
DBH	86	86	86
MGMT	84	84	84
(Other)	1195	1195	1273

GAD RESULTS FOR TYPE 2 DIABETES DISEASE ANALYSIS IN AHP, RF-AHP RF-AHP-REGULOME METHODS

Table F.5 No of most referenced SNPs in the results of GAD query

SNP#	Delphi-AHP	RF-AHP	RF-AHP-R
rs11196208	247	247	247
rs12255372	247	247	247
rs10885409	246	246	246
rs11196205	246	246	246
rs12243326	246	246	246
rs7077039	246	246	246
(Other)	5204	5204	5345

Table F.6 No of most referenced disease classes in the results of GAD query

Disease Class	Delphi-AHP	RF-AHP	RF-AHP-R
METABOLIC	3419	3419	3465
CARDIOVASCULAR	748	748	776
CANCER	408	408	412
IMMUNE	365	365	374
CHEMDEPENDENCY	296	296	309
UNKNOWN	273	273	274
(Other)	1173	1173	1213

Table F.7 No of most referenced phenotypes in the results of GAD query

Phenotype	Delphi-AHP	RF-AHP	RF-AHP-R
Phenotype	1011	1011	1012
diabetes, type 2	488	488	488
Type 2 diabetes	214	214	218
Tobacco Use Disorder	197	197	197
Diabetes Mellitus, Type 2	148	148	148
type 2 diabetes	123	123	123
Waist Circumference	4501	4501	4637

Table F.8 No of most referenced genes in the results of GAD query

Gene	Delphi-AHP	RF-AHP	RF-AHP-R
TCF7L2	2813	2813	2813
CDKAL1	225	225	225
LRRN1	176	176	176
NAV2	165	165	176
NCAM2	126	126	147
NRXN3	123	123	123
(Other)	3054	3054	3163

GAD RESULTS FOR SCHIZOPHRENIA DISEASE ANALYSIS IN AHP, RF-AHP RF-AHP-REGULOME METHODS

Table F.9 No of most referenced SNPs in the results of GAD query

SNP#	Delphi-AHP	RF-AHP	RF-AHP-R
rs4846051	1951	1951	1951
rs4349	1648	1648	1648
rs4297	1646	1646	1646
rs7212502	1052	1052	1052
rs9903602	1052	1052	1052
rs8191446	794	794	794
(Other)	29989	29989	63874

Table F.10 No of most referenced disease classes for Schizophrenia in the results of GAD query

Disease Class	Delphi-AHP	RF-AHP	RF-AHP-R
Metabolic	6984	6984	14896
Psych	6074	6074	10703
Cardiovascular	5794	5794	9687
Cancer	4742	4742	7977
Immune	2331	2331	6230
Neurological	2320	2320	4179
(Other)	9887	9887	18345

Table F.11 No of most referenced phenotypes for Schizophrenia in the results of GAD query

Phenotype	Delphi-AHP	RF-AHP	RF-AHP-R
Schizophrenia	1252	1252	1562
Undefined	1011	1011	1433
Tobacco Use Disorder	774	774	1425
Lung Cancer	544	544	1039
Autism	474	474	1000
Breast Cancer	469	469	881
(Other)	33608	33608	64677

Table F.12 No of most referenced genes in the results of GAD query

Gene	Delphi-AHP	RF-AHP	RF-AHP-R
ESR1	3918	3918	3918
ACE	3284	3284	3284
CNTNAP2	2225	2225	2906
SLC6A4	2102	2102	2244
MTHFR	1947	1947	2102
IL4	1804	1804	1947
(Other)	22852	22852	55616

GAD RESULTS FOR ALZHEIMER'S DISEASE ANALYSIS IN AHP ,RF-AHP RF-AHP-REGULOME METHODS

Table F.13 No of most referenced SNPs in the results of GAD query

SNP#	Delphi-AHP	RF-AHP	RF-AHP-R
rs6023	848	848	848
rs5897	655	655	655
rs4972	141	141	141
rs132954	49	49	49
rs6160	43	43	43
rs35627	37	37	37
(Other)	520	520	523

Table F.14 No of most referenced disease classes for Alzheimer's Disease in the results of GAD query

Disease Class	Delphi-AHP	RF-AHP	RF-AHP-R
Cardiovascular	919	919	919
Metabolic	354	352	354
Reproduction	236	236	236
Unknown	196	196	196
Hematological	143	143	143
Cancer	109	109	109
(Other)	338	338	339

Table F.15 No of most referenced phenotypes for Alzheimer's Disease in the results of GAD query

Phenotype	Delphi-AHP	RF-AHP	RF-AHP-R
Glucose	88	88	88
Venous Thrombosis	63	63	63
Hypertension	43	43	43
Undefined	41	41	41
Thromboembolism, Venous	34	34	34
Cholesterol, HDL	30	30	30
(Other)	1994	1994	1997

Table F.16 No of most referenced genes in the results of GAD query

Gene	Delphi-AHP	RF-AHP	RF-AHP-R
F5	846	846	846
F2	641	641	641
LARGE	145	145	145
ADD1	140	140	140
LRRN1	48	48	48
CYP11A1	43	43	43
(Other)	430	430	433

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Yılmaz, Arif

Nationality: Turkish

Date and Place of Birth: Kaman, KIRŞEHİR, 1974

Marital Status: Married

Phone: +905055052743

email: arif.yilmaz@tubitak.gov.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	Middle East Technical University, Department of Electrical and Electronics Engineering	2002
BS	Hacettepe University, Department of Electrical and Electronics Engineering	1999

WORK EXPERIENCE

Year	Place	Enrollment
2013-Present	TUBITAK-UZAY Space Technologies Research Institute-Big Data and Computing Systems Development Group	Chief Researcher - Group Leader
2012-2013	TUBITAK-UZAY Space Technologies Research Institute-Software Technologies Development Group	Senior Researcher - Group Leader

2008-2013	TUBITAK-UZAY Space Technologies Research Institute-Software Technologies Development Group	Senior Researcher
2005-2008	TUBITAK-BILTEN Information Technologies Research Institute- E-Commerce, E-Signature and Software Development Group	Senior Researcher
2004-2005	Turkish War Academies Command-War Games Simulation Center, Military Service	Software Developer
2003-2004	TUBITAK-BILTEN Information Technologies Research Institute-Intelligent Energy Conversion Systems Group	Senior Researcher
2000-2003	TUBITAK-BILTEN Information Technologies Research Institute-Intelligent Energy Conversion Systems Group	Researcher
1999-2000	Hacettepe University, Department of Electrical and Electronics Engineering	Research Assistant

FOREIGN LANGUAGES

English: Fluent, French: Basic