A HYBRID FEATURE SELECTION MODEL FOR GENOME WIDE ASSOCIATION STUDIES

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

BY

SAİT CAN YÜCEBAŞ

IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY
IN
HEALTH INFORMATICS

SEPTEMBER 2013

A HYBRID FEATURE SELECTION MODEL FOR GENOME WIDE ASSOCIATION STUDIES

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

Prof. Dr. Nazife Baykal _____
Director, Graduate School of Informatics

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

Prof. Dr. Nazife Baykal _____
Supervisor, Department of Information Systems, METU

Assist. Prof. Dr. Yeşim Aydın Son _____
Co-Supervisor, Department of Health Informatics, METU

Examining Committee Members

Prof. Dr. Hayri Sever _____
Department of Computer Engineering, Hacettepe University

Prof. Dr. Nazife Baykal _____
Department of Information Systems, METU

Assist. Prof. Dr. Aybar Can Acar _____
Department of Health Informatics, METU

Assoc. Prof. Dr. Levent Çarkacıoğlu _____
T.C. Merkez Bankası

Assoc. Prof. Dr. Hasan Oğul _____
Department of Computer Engineering, Başkent University

Date: 04.09.2013

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: Sait Can Yücebaş

Signature :

ABSTRACT

A HYBRID FEATURE SELECTION MODEL FOR GENOME WIDE ASSOCIATION STUDIES

Yücebaş, Sait Can
Ph.D., Department of Health Informatics
Supervisor: Prof. Dr. Nazife Baykal
Co-supervisor: Assist. Prof. Dr. Yesim Aydın Son

September 2013, 166 pages

Through Genome Wide Association Studies (GWAS) many SNP-complex disease relations have been investigated so far. GWAS presents high amount – high dimensional data and relations between SNPs, phenotypes and diseases are most likely to be nonlinear. In order to handle high volume-high dimensional data and to be able to find the nonlinear relations, data mining approaches are needed. A hybrid feature selection model of support vector machine and decision tree has been designed. This model also combines the genotype and phenotype information to increase the diagnostic performance. The model is tested on prostate cancer and melanoma data that have been downloaded from NCBI's dbGaP database. On prostate cancer data the hybrid system performed 71.67% accuracy on data set consists of only genotypes, 84.23% accuracy on data set consists of only phenotypes and when genotyping and phenotypes are integrated accuracy increased to 93.81%. On melanoma data, the hybrid system performed 57.12% accuracy for only genotypes, 75.48% accuracy for only phenotypes and when genotyping and phenotypes are integrated accuracy increased to 86.35%.

For prostate cancer case the hybrid system's has performance indicators of 90.92% of sensitivity and 0.91 AUC, which outperforms Prostate Specific Antigen (PSA) test. In melanoma case selected phenotypic and genotypic features were also examined by previous studies that shows the ability of the system to select most predictive features so the hybrid system on melanoma case has a potential to be used for identifying the risk groups.

Keywords: Genome Wide Association Studies, Decision Tree, Support Vector Machine, Prostate Cancer, Melanoma

ÖZ

BÜTÜNSEL GENOM İLİŞKİLENDİRME ÇALIŞMALARI İÇİN HİBRİD ÖZNİTELİK SEÇME MODELİ

Yücebaş, Sait Can
Doktora, Sağlık Bilişimi Bölümü
Tez Danışmanı: Prof. Dr. Nazife Baykal
Eş-Danışman: Assist. Prof. Dr. Yesim Aydın Son

Eylül 2013, 166 sayfa

Bütünsel genom ilişkilendirme çalışmaları karmaşık hastalıklar ve SNP'ler arasındaki ilişkileri keşfetmektedir. Bu çalışmalar yüksek miktarda çok boyutlu veri sunmaktadır. Ayrıca SNP'ler, hastalıklar ve fenotipler arasındaki ilişkiler genellikle doğrusal değildir. Yüksek miktarlı çok boyutlu verilerle çalışmak ve aralarındaki ve doğrusal olmayan ilişkileri bulabilmek için veri madenciliğine ihtiyaç duyulmaktadır. Genotip ve fenotip bilgilerini birleştirerek bunlar üzerinde çıkarsama yapan, karar ağacı ve destekçi vektör makinasından oluşan bir hibrid sistem tasarlanmıştır. Tasarlanan model NCBI'ın dbGaP veritabanından indirilmiş olan prostat ve melanoma veri kümeleri üzerinde denenmiştir. Prostat veri kümesinde sadece genotip bilgileri kullanıldığında %71,67'lik kesinlik sonucu, sadece fenotip bilgileri kullanıldığında %84,3 kesinlik sonucu; genotip ve fenotip bilgileri birleştirildiğinde ise %93,81'e yükselen kesinlik sonucu elde edilmiştir. Melanoma veri kümesinde sadece genotipler kullanıldığında %57,12, sadece fenotip bilgileri kullanıldığında %75,18 ve fenotip ile genotip bilgileri birleştirildiğinde %86,35'lik bir kesinlik sonucu elde edilmiştir.

Prostat üzerinde çalıştırılan sistem, %90.92'lik duyarlılık ve 0.91'lik ROC eğrisi altında kalan alan ile Prostata Özel Antijen Testi'ne üstünlük göstermiştir. Melanoma için ise hibrid modelin seçmiş olduğu fenotip ve genotip özellikleri bundan önceki çalışmalarda incelenmiş olup bu durum kurulan hibrid sistemin melanoma için en ayırt edici özellikleri seçme yetisi olduğunu göstermektedir. Bu sayede melanoma için kurulan hibrid sistemin risk gruplarını ayırt etme potansiyeli bulunmaktadır.

Anahtar Kelimeler: Bütünsel Genom İlişkilendirme Çalışmaları, Karar Ağacı, Destek Vektör Makinası, Prostat Kanseri, Melanoma

dedicated to my dear wife Burcu Yücebaş, my daughter Su Bade Yücebaş, my dear father Tevfik F. Yücebaş and my dear mother M. Gül Yücebaş

ACKNOWLEDGEMENTS

I express sincere appreciation to **Prof. Dr. Nazife Baykal** and **Assist. Prof. Dr. Yeşim Aydın Son** for their guidance and insight throughout the research. Thanks to other faculty members **Prof. Dr. Hayri Sever, Assoc. Prof. Dr. Hasan Oğul, Assoc**., **Assist. Prof. Dr. Aybar C. Acar and Dr. Levent Çarkacıoğlu** for their suggestions and comments. Valuable contribution of **Remzi Çelebi** is greatfully acknowledged. Thanks to Sibel Gülnar, M. Hakan Güler, Çetin İnci and Yaşar Sayın for their support in administrative issues throughout my studentship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

AC: Allergic Conjunctivitis
AD: Alzheimer's Disease / Atopic Dermatitis
ADT: Alternating Decision Tree
AHP: Analytical Hierarchical Process
AMD: Age Related Macular Degeneration
ANN: Artificial Neural Network
AR: Allergic Rhinitis
ASD: Autisim Spectrum Disorder
AUC: Area Under Curve
BA: Bronchial Asthma
BBN: Bayesian Belief Network
BEBT: Backward Elimination with Back Tracking
BMI: Body Mass Index
CA: Cancer
CAA: Childhood Allergic Asthma
CAD: Coronary Artery Disease
CART: Classification Regression Tree
CHAID: Chi Squared Automatic Interaction Detection
CPM: Combinational Partitioning Method
CSM: Cancer Somatic Variation
dbGaP: Database of Genotypes and Phenotypes
DBSCAN: Density Based Spatial Clustering of Applications
DENCLUE: Density Based Clustering
DNA: Deoxyribonucleic Acid
ELSI: Ethical Legal Social Implications
FSBT: Forward Selection and Back Tracking
GA: Genetic Algorithm
GAD: Genetic Association of Complex diseases and Disorders
GE: Grammatical Evolution
GENN: Genetic Evaluation Neural Network
GEDT: Genetic Evolution Decision Tree
GP: Genetic Programming
GPDT: Genetic Programming Decision Tree
GPNN: Genetic Programming Neural Network
GWAS: Genome Wide Association Studies
HGP: Human Genome Project
IIG: Incremental Information Gain
KNN: K Nearest Neighbour
MI: Myocardial Infraction
MDR: Multi Factor Dimensionality Reduction
MRA: Mıltiple Regression Analysis
MS: Multiple Sclerosis
NCBI: National Center of Biotechnology Information
NHRI: The National Human Genome Researc Institute
OPTICS: Ordering Points to Identify Clustering structure
PCA: Principal Conponent Analysis
PDM: Parameter Decreasing Method
PSA: Prostate Specific Antigen
QUEST: Quick Unbiased Efficient Statistical Tree

RBF: Radial Basis Function
RBFS: Rule Based Feature Selection
RF: Random Forest
RNA: Ribonucleic Acid
SNP: Single Nucleotide Polymorphism
SVM: Support Vector Machine
WHO: World Health Organization

CHAPTER 1

INTRODUCTION

The differences among living organisms have been the interest of researchers for a long time. Underlying reasons for these differences have been studied for centuries. These researches go back to mid-1800s. The first work about this topic was profounded by Charles Darwin, The Origin of Species, in 1859. His work tries to explain the evolution of species through natural selection phenomena. In 1866 Gregor Mendel published his work on inheritance of plants. He had a suspicion that some factors play an important role in inheriting phenotypic features in living organisms. With his thoughts and works on inheritance he is believed to be the founder of modern genetics. Mendel's work and the discovery of DNA in 1869 paved the way of genetic studies.

In these days genetic studies gained up speed more than ever with the completion of the Human Genome Project. A research field raised by the advancements explained above is the Genome Wide Association Studies (GWAS). In basic, GWAS search for a genetic variation and try to find if that variation is related with a certain phenotype such as vulnerability to a specific complex disease. Through GWAS genetic variants based on Single Nucleotide Polymorphism (SNP), which is a single nucleotide change in genome, can be identified. In GWAS many SNP-complex disease relations are searched such as, age related macular degeneration [1], heart diseases [2], diabetes [3], rheumatoid arthritis [4], Crohn's Disease [5], Bipolar Disorder [6], Hypertension [7], Multiple Sclerosis [8] and cancer types [9-10-11].

Methods used in GWAS can be grouped under two main categories which are parametric and non-parametric [12]. Parametric methods are based on a genetic model and these models are constructed by statistical calculations such as regression based models. On the other hand non-parametric methods do not require a genetic model given beforehand; instead they build their own models based on given data by using data mining and machine learning methods [12]. The reason behind choosing non-parametric methods over parametric ones is the high dimensionality of the genetic data which make traditional statistics not sufficient enough for the analysis [13]. Almost all known machine learning algorithms have been used in GWAS, some of the foremost methods are decision trees [14-15-16-17-18-19-20-21], artificial neural networks [22-23-24-25-26-27-28], Bayesian belief networks [29-30-31], support vector machines [15-32-33-34-35-36-37-38-39-40-41] and genetic algorithms [42-43-44]. There is no clear evidence that one of the methods perform best among others [12].

All methods have their own advantages and disadvantages and selection of the appropriate method is mostly based on the given problem, data type, study design and aim of the work. Some studies in literature combine these methods to eliminate the disadvantages and strengthen the advantages of the methods used.

In such studies, genetic algorithm based approaches are combined with a master method and genetic algorithm is used to optimize the parameters of the main method. Motsinger's Genetic Evaluation Neural Network structure [45] and Grammatical Evaluation Decision Tree [46] are good examples for such approaches.

Our aim in this thesis is to build a hybrid feature selection model that will be used in the selection of most important and relevant phenotype and genotype features related with the selected cancer diseases. Our motivation behind this work can be gathered around three topics. These are:

To our knowledge there is no work that combines genotyping with multiple number of phenotypic features representing clinical findings, life style habits and demographic information of the subjects. Almost all works concentrate on the relation of genotypes with one specific phenotype attribute which is generally the disease itself. Some other works search for the relations between a known specific phenotype attribute and SNPs in a given disease such as effect of genetic factors on body mass index in type-2 diabetes. As we believe that the genetic background and the subject's life style along with clinical findings is equally important in detection of the complex diseases, our approach aims to combine all phenotypic features, including the clinical findings, and genotypes in the feature selection step. And in this study we designed our experiments to compare the performances of combined genotype-phenotypes information against only genotype and phenotype in disease diagnosis with different data mining approaches.

Combining different machine learning methods strengthens the overall performance. As distinct from many works in the literature, we have selected to use both methods individually rather than using one method as a master and other to optimize the master method. The first method selected for the proposed hybrid system is the ID3 decision tree. ID3 trees have a strong performance on classifying the discrete valued datasets like the genotyping and phenotype data used in this study. Its easy construction, easy application, good performance on dealing with data with noise and missing values, easy interpretation with its visual features are other reasons of this selection [47-48]. The other method selected is the SVMs. This method is selected because of its proven high performance in GWAS [49-50], and its ability to classify non separable problems while providing minimal overfitting, with lower chance to stuck at local minima, and easy interpretation of the attribute weights [37].

Based on the previous studies, it is known that in order to get more significant results, the genotype set must be reduced to a representative subset after GWAS. This is done by using statistical calculations, feature selection, feature and dimension reduction techniques [32-33-34-35-39-41].

But this also limits the work done into a statistical perspective and biological importance is often ignored. So, there is a need to prioritize the statistically significant SNPs associated with the given disease, according to their biological relevance, which also provides filtering of the data and reduction of the feature size. In order to prioritize and filter the data we have used METU-SNP's tool based on Analytical Hierarchical Process (AHP) which creates a SNP prioritization list that is based on both biological and statistical significance by combining P value statistics, genomic locations, functional consequences, evolutionary conservation and gene disease relations[1].

[1] METU-SNP: an integrated software system for SNP-complex disease association analysis. Ustünkar G, Aydın Son Y. J Integr Bioinform. 2011 Dec 12;8(1):187

The proposed SVM-ID3 hybrid system is tested on three different data sources (only genotype, only phenotype and combined genotype-phenotype data) for two different cancer studies multi ethnic prostate cancer and melanoma downloaded from NCBI's dbGAP database.

1.1 Data Mining Process

Due to the exponential growth in computing technology, computational power has increased to a level where thousands of instructions can be processed in a millisecond and their results can be stored in high capacity storage units. These results in generation of vast amount of raw data in almost all domains such as medicine, biology, engineering, finance, education, production etc. The specialists, decision makers and white collars in different field needs valuable information sieved from the raw data in order to give critical decisions such as a high financial investment or a risky health operation.

Data mining is a process to find important and hidden relations among vast amount of data and present them as information and/or knowledge. This process is a multi-step method that includes data collection, data management, data cleaning, data integration, data transformation, data mining and knowledge presentation [48]. This process can be depicted as in the Figure-1 below.

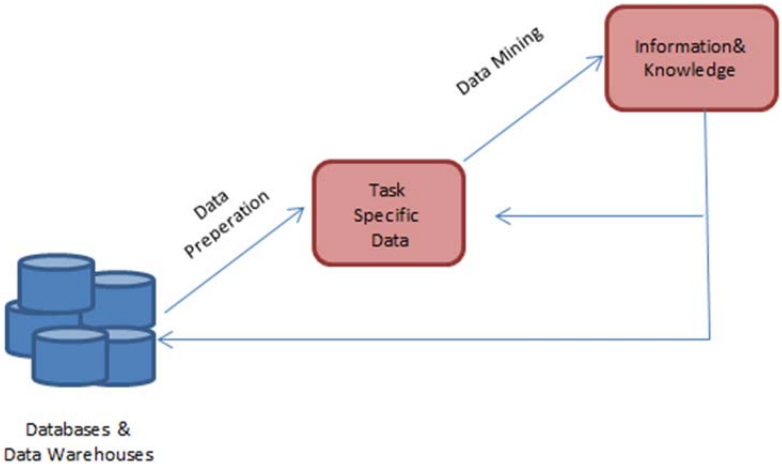


Figure 1: Data Mining Process. (Adapted from Data Mining: Concepts and Techniques. Jiawei Han, M. Kamber. 2001)

Generally data cleaning, integration and transformation are held in one step which is called Data Preparation. After the data is ready the main concerns remaining is the data-mining itself. The details of data preparation and data mining steps will be covered in the following subsections.

1.1.1 Data Preparation

Data preparation requires collection and management of the data. There are many various databases and database management systems that can be used at this step. Relational databases, data warehouses, transactional databases, object oriented databases form some examples.

The second step of the preparation is data cleaning. The raw data stored usually is not cleaned, and most of the data could suffer from missing values. The data can be present as an attribute field in the database but the value of the data may be missing due to mis-entries. Other problems that make the data unclean are noise and inconsistency. So in data cleaning one have to deal with missing values, noisy and inconsistent data. The basic approaches for data cleaning are:

- Tuple Erasing
- Manuel Fill
- Constant Value

Tuple erasing and manuel fill are easy and simple methods. In tuple erasing, the tuple containing the missing values simply deleted. But the other non-missing attributes in the tuple erased will lead to loss of information. Additionally, if the rate of missing values is very high deleting tuples could result in very few data point at the hand.

Manuel fill is another simple method. But this can be applied only when missing values are known for the tuples from other data sources. Unfortunately this is not the case most of the times.

A constant value can be used for the missing values. In a simple scheme a constant dummy value can be chosen and used to fill all missing values. But in this case decision model can interpret this constant as valuable information, especially for cases where missing value rate is high. To avoid such problems, instead of a dummy constant the mean of the attribute for the missing values can be used. This is a better application when compared to solutions above. Still, as the class information is ignored the importance of the attribute is somewhat smoothed and this can cause the model neglect some important relations based on this attribute. The best approach could be using the means according to their class information. In this method data is divided into classes according to given problem and the missing value is replaced by the attribute's mean according to the class that it belongs.
Noise *"is a random error or variance in a measured variable"* [2] and mostly appear as outliers in the data mining process. To clean a noisy data different smoothing methods can be used. Binning, clustering, regression are the more frequently selected for smoothing.

Inconsistency is another reason for unclean data. It may rise from improper data entry, miss use, faults in data integration from different data sources etc. A data record that has age attribute set to "3 years old" and marital status set to "married" is an example for inconsistency. In our case a data record that has sex attribute set to "female" and diagnosis attribute set to "prostate cancer" is another good example of inconsistency. Because prostate cancer is sex specific disease only occur in males.

[2] Data Mining: Concepts and Techniques. Jiawei Han, M. Kamber. 2001 Academic Press. ISBN 1-55860-489-8

Inconsistencies can be avoided at the time of data entry, data transformation and data integration phases. Even though these steps carried on carefully there still may be some inconsistencies. Thus a careful data examination with descriptive statistics such as frequency analysis, chi square test, odd's ratio, Phi and Cramer's V test can be helpful before applying any data mining methods to dataset.

1.1.2 Data Mining Methods

After data collection and preparation of data by cleaning is the third step in data mining process, where a suitable data mining method for the data and problem at the hand is applied. Data mining methods can be generalized under two main categories. These are Supervised Learning and Non Supervised Learning methods [51].

1.1.2.1 Supervised Learning Methods

In supervised learning, the labels of data which indicates the class information is known beforehand. Because the labels are known priori, the problem is reduced to class assignment for the new data, named as the Classification Problem. If data labels are divided into two groups such as "1" and "-1" or "benign" and "malignant" this is called "binary classification" and if there are more than two class labels it is called "multiclass classification" [52].

Before proceeding to classification methods one should know how to use the data sets effectively. The data at the hand used for the learning purpose. In order to learn from data, the data set is divided into training and test subsets. The model build learns from the training data according to model's specific features and its classification performance is evaluated by using the test subset. Separating the data set into training and test subsets correctly is very important for the performance of the model being developed. If this separation is not done properly the classification model could suffer from over-fitting or under-fitting. In over-fitting the model constructed memorizes the data instead of learning from it [53]. This mostly caused by "*noise and random error in the data*"[3]. So instead of generalization of the problem, the model becomes too specific by the training data and can not perform well on other datasets for a similar problem. This condition is depicted in the Figure-2 below:

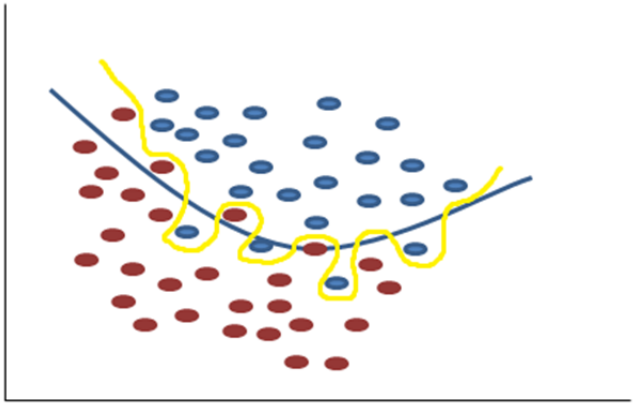


Figure 2: Overfitting Problem: The model memorizes the training dataset and so unable to make a generalization on different dataset of the same problem. (Adapted From: Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach. Burnham

[3] http://en.wikipedia.org/wiki/Overfitting accessed on 28.092012

The exact opposite of the over-fitting problem is the under-fitting of the data. In this scheme, the model constructed can not learn from the training data and have a poor performance on generalizing the problem. This situation is depicted in the Figure-3 below:

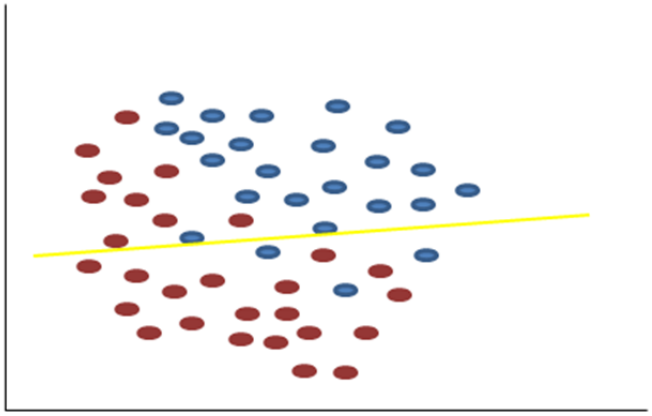


Figure 3: Underfitting Problem: Model can not learn from training data and has poor generalization performance. (Adapted From: Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach. Burnham, Kenneth P., Anderson, David R. 2nd ed. 2nd ed. 2002, XXVI, 448p.)

To avoid such situations training and testing data must be assigned carefully. There are different methods used in literature for healthy separation of dataset. The most widely used methods are:

- **Straight dividing:** This is a very basic method which divides the dataset as 70% training and 30% testing. The dataset is divided from the first row and training subset is formed until 70% of data is reached. And remaining data used as the test subset. Different percentages of division can be applied to adjust the model performance.
- **Random Sampling:** In straight dividing there is a risk that the problem at the hand can not be well generalized by the training data. As it is chosen from the first half of the dataset biased representation of a specific subset can result in a model representing only that data. To avoid such situations random sampling is suggested. In this scheme data elements are assigned randomly to the training and the test subsets [54]. Equal class rates and 30%-70% scheme could be applied to make random sampling stronger in performance.
- **Cross Validation** [55-56] Cross validation is designed for using most of the data samples both as training and testing data in different runs. By this way the chance of the model to learn almost all characteristics of the problem increases. In a basic approach the data set is divided into two equal sized subsets and data elements are randomly assigned to them. Each subset is used once for training and once for testing the model. This basic approach is called 2-Fold Cross Validation. Another method for cross validation is called K-Fold Cross Validation. In this scheme data is divided to equal size K folds. One of the folds is used for testing and the remaining folds (K-1) are used for training. This continues until all folds are used once as a test set. Leave One Out scheme is very similar to K-Fold Cross Validation but here the number of folds, K, is set to "*the number of observations in the original sample*[4]".

[4] http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29 accessed on 29.09.2012

After selecting the method to divide the dataset into training and testing once the turn comes to the method to be applied. There are many decision support methods each of have its own advantages and disadvantages. Some methods perform well on specific kind of problems and/or data. In the following section most widely used and well known methods will be summarized.

1.1.2.1.1 K Nearest Neighbors (KNN)

In this method, all data are represented as points in *n* dimensional space [48]. When a new data comes, the nearest of K neighbors are calculated and this new data is assigned to class of its neighbors. Because all data points are stored and classification is held on until arrival of a new data point, KNN learners are known as lazy learners.

Euclidean Distance is used to measure the distance between two points, X and Y, in *n* dimensional space. The calculation is done as follows:

$X=(x_1,x_2, \dots ,x_n)$ and $Y=(y_1,y_2, \dots ,y_n)$ **(EQUATION 1)**

$D(X,Y) = D(Y,X) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$ **(EQUATION 2)**

$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ **(EQUATION 3)**

The problem arises when k nearest neighbors belong to different classes. There are different approaches to solve this problem. The most straight forward method is to calculate the class rates of k neighbors and assign the new data point to a class with the highest rate.

In some conditions class rates can be equal. In this situation the distance between new data point and k neighbors calculated, and then new data point is assigned to the class of its nearest neighbor [57]. Or as a more reliable solution distant based voting [57] formulated below can be preferred.

$$\text{Vote}(y_i) = \sum_{c=1}^{k} \frac{1}{d(q,x_c)^n} 1(y_i, y_c)$$ **(EQUATION 4)**

> *"The vote assigned to class y_i to determine the class q, by neighbor x_c is 1 divided by the distance to that neighbor, i.e. $1(y_i , y_c)$ returns 1 if the class labels match and 0 otherwise. In equation above n would normally be 1 but values greater than 1 can be used to further reduce the influence of more distant neighbors*"[5].

Another important aspect in KNN algorithm is to decide the number *k*. Because smaller values of *k* tend to introduce variance and higher values of *k* tend to introduce bias. Generally a cross validation scheme is used among different k values, and best performing one is set to be the *k*.

[5] K-Nearest Neighbour Classifiers. P´adraig Cunningham and Sarah Jane Delany. University College Dublin, Dublin Institute of Technology. Technical Report UCD-CSI-2007-4. March 27, 2007

1.1.2.1.2 Decision Trees

Decision Trees are widely used for classification problems [58]. They are preferred over other methods due to some advantages such as noise handling, low computational requirements [59] and easy interpretation.

As the name indicates they form a structure that resembles a tree. The nodes in the structure represent the attributes and connections between the nodes represent the test conditions of the nodes. Nodes are grouped under root node, inner nodes and leaf nodes. The root node is the beginning of the tree and constructed by the attribute that has the best ability to divide the data set into equal pieces. The inner nodes are the other attributes that has a strong relation with the given problem and finally the nodes at the bottom of the tree which are called leaf nodes represent the classes of the given problem. A typical decision tree structure is given in the Figure-4 below:

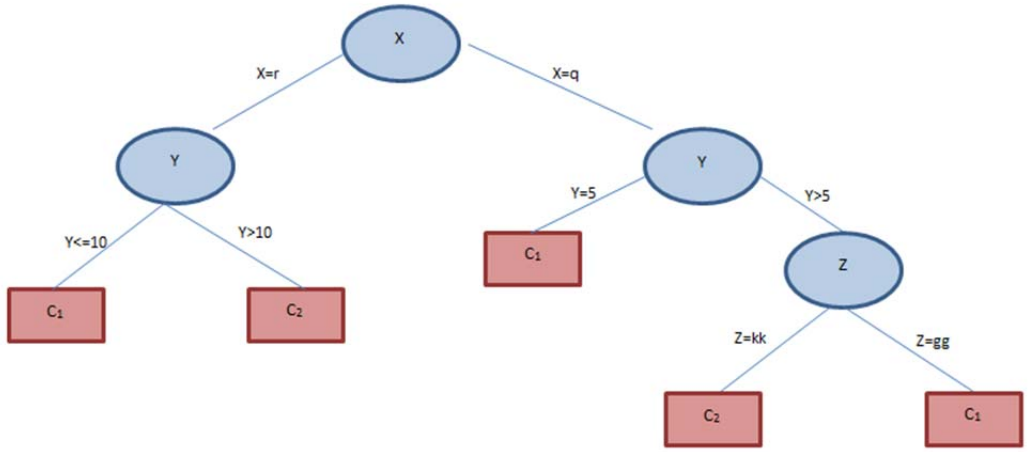


Figure 4: A typical decision tree structure. (Adapted from Data Mining: Concepts and Techniques. Jiawei Han, M. Kamber. 2001)

Although some differences can be discernible according to data type and performance modifications, the learning schemes of decision trees are generally based on Quinlan's ID3 Tree [60]. So the learning algorithm given in this section is based on Quinlan's method.

In decision tree induction each attribute is tested according to its information on the given classification problem. This information is calculated by some measures such as Entropy, Information Gain, Gain Ratio and Gini Index [59]. In a classical approach Information Gain and Entropy is used for the induction. Each attribute is tested and the attribute with highest information gain is chosen for the node of the tree [48].

The following explanations, notations and formulations of attribute selection are taken from [48].

S: set of s data samples
C_i: Class label with i=1 to m
S_i: number of samples of S in class C_i
P_i: probability of a data element i belongs to class C_i
A: Attribute with $A=\{a_1,a_2, \ldots ,a_v\}$

Information for classification of the samples is calculated by:

$$I(s_{1,}s_{2,}\dots,s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad \textbf{(EQUATION 5)}$$

To choose attribute A to divide the data samples into v subsets such {S_1,S_2, ... ,S_v} and S_j have the samples that have value a_j A, Entropy of A must be calculated as follows:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\cdots+s_{mj}}{s} I(s_{1j}, \dots, s_{mj}) \quad \textbf{(EQUATION 6)}$$

Where s_{ij} is the number of samples of class C_i in a subset S_j. Information of this subset S_j is calculated by:

$$I(s_{1j},s_{2j}, \dots ,s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad \textbf{(EQUATION 7)}$$

In this situation the Information Gain is the information minus entropy:

$$\text{Gain (A)} = I(s_{1,}s_{2,}\dots,s_m) - E(A) \quad \textbf{(EQUATION 8)}$$

The branching attribute is chosen as the attribute with highest information gain.

Information gain tends to introduce some bias when the range of the chosen attribute is high, so in order to balance such situation Gain Ratio can be used [61]. Gain Ratio is the normalization [47] of Information Gain and calculated as [47-62]:

$$\text{Gain Ratio}(a_i, S) = \frac{\text{Information Gain}(a_i,S)}{\text{Entropy}(a_i,S)} \quad \textbf{(EQUATION 9)}$$

Although this is the basic idea of decision trees, there are different types of induction measures such as Gini Index, Likelihood Ratio-Chi Square, Distance Measure, Orthogonal Criterion, AUC Splitting etc. can be used for tree branching [47].
The basic ID3 Decision Tree also copes with variety of real world problems such as data sets with continuous values. The most widely used trees can be summarized as [47]:

- C4.5: Developed by Quinlan [62]. Uses gain ratio for splitting criteria and can handle numeric valued attributes.
- CART: CART stands for Classification Regression Tree and developed by Breiman [63]. The most important feature of CART is that it can predict real numbers for the given problem [47].
- CHAID: Developed by Kass [64] a CHAID Tree uses statistical methods such as chi-square and F test, depending on the type of classification problem. If classes have nominal values the chi-square test is used, but if the classes have continuous values than F test is applied [47].
- QUEST: QUEST stands for Quick Unbiased Efficient Statistical Tree and developed by Loah [65]. Like CHAID it uses different statistical methods such as ANOVA, Levene's Test and chi square for different types of data [47].

Beside these conventional methods new decision tree methods are also proposed in the literature which mostly uses combination of different methodologies like evolutionary algorithms [59], fuzzy sets[6] [66], grammatical evolution [46] and random forests [67].

1.1.2.1.3 Bayesian Classification

It is a classification method that is based on a statistical approach called Bayes Theorem. This theorem shows the probability of an event to occur under the given conditions.
Suppose that the classification problem is about deciding the patients with or without hypertension and in our data set we have blood pressure measurements of the patents. In this condition:

P(H): is the probability that a patient has hypertension
P(V): is the probability that blood pressure measurement is high

So classification problem is:

P(H|V) = Probability of a patient to have hypertension given the evidence that patient has a high category blood pressure measurement.

This probability can be calculated by the Bayes Theorem as follows:

$$P(H|V) = \frac{P(V|H)P(H)}{P(V)} \qquad \textbf{(EQUATION 10)}$$

Where:

P(V|H)= Probability that a patient has high category blood pressure measurement given the evidence patient has hypertension.
P(H) = Probability of hypertension
P(V) = Probability of high category blood pressure measurement

A Naïve Bayes Classifier simply uses this rule, Bayes Theorem, to predict a class for a new data point. If the dataset consists of *n* distinct classes with $C_1, C_2, \ldots, C_n$ and D is the new data point all probabilities are calculated first with

$$P(C_i|D) = \frac{P(D|C_i)P(C_i)}{P(D)}, \quad i = 1 \text{ to } n \qquad \textbf{(EQUATION 11)}$$

And the data is assigned to highest probability class calculated with$P(C_i|D)$. At this point a computational complexity rises when dataset consists of many attributes. So an assumption is made and all attributes accepted to be independent [48]. By this assumption $P(D|C_i)$ can be calculated by:

$$P(D|C_i) = \prod_{k=1}^{n} P(d_k|C_i) \qquad \textbf{(EQUATION 12)}$$

[6] A complete fuzzy decision tree technique. Cristina Olaru, Louis Wehenkel. University of Liege, Department of Electrical Engineering and Computer Science, Montefiore Institute. Fuzzy Sets and Systems 138 (2003) 221–254. Received 5 February 2002; received in revised form 29 January 2003; accepted 18 February 2003

Naive Bayes method based on conditional independence but in real world problems mostly that is not the case. There can be relations between attributes and these relations can affect the data point to be assigned to one class or the other.

To model such conditions, attributes are not independent, Bayesian Belief Networks are developed. Bayesian Belief Networks as depicted in the Figure-5 below, are graph structures, can be directed or undirected, and with nodes representing the attributes and connections between nodes representing the relations [68].

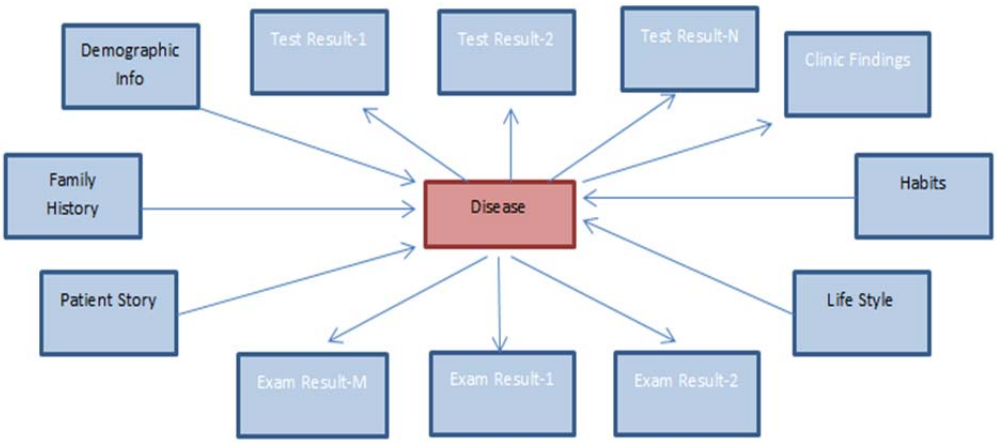


Figure 5: A simple Bayesian Network Structure

In the paper of Jie Cheng [69] learning scheme of Bayesian networks analyzed in two main categories. One of the learning schemes is based on dependencies of attributes and their joint distributions and uses scoring such as Kullback-Leibler (KL) entropy scoring function [70-71]. At the opposite, other scheme is based on independencies of the attributes and uses CI Based Algorithms [72-73]. According to the very same paper there are different works in literature that shows the performance of these two learning schemes.

Bayesian Belief Networks can also be categorized according to their network structure. Jie Cheng [59] groups those under four groups, which are:

- Tree Augmented Networks: The network forms a tree structure and learning schemes based on Chow-Liu algorithms [74].
- BN Augmented Networks: This networks form an *"arbitrary graph structure and learning structure is less efficient"* [59]
- Multi Networks: "*Bayesian multi-net allows the relations among the features to be different – i.e., for different values the class node takes, the features can form different local networks with different structures*" [59].
- General Bayesian Networks: All the network structures itemized above behave the class node as a separate special node but in General Bayesian Network class node is just an ordinary node like all the nodes presenting the attributes [59].

1.1.2.1.4 Artificial Neural Networks (ANN)

As the name indicates Artificial Neural Networks (ANN's) are built to mimic the nervous system of the living organisms. In this aim the first system developed by McCulloch and Pitts in 1943 [74]. They tried to mimic a single neuron and how it transmits the signals through the nervous system.

A biological neuron consists of Nucleus, Axon, Soma and Dendrite. Nucleus is found in the soma structure and responsible from cell's management. Dendrites are the extensions from soma and responsible from taking the signals from other cells. Axon is the single extension from soma that is responsible from transmitting the signals.

The basic idea they've introduced was a simple artificial neuron based on a mathematical function that adds all the inputs and calculates if the summation value is above a threshold or not. This approach is named as their founders and called McCulloch-Pitts neuron. The resemblance of artificial and biological neurons is depicted in the Figure-6 and Figure-7 below:

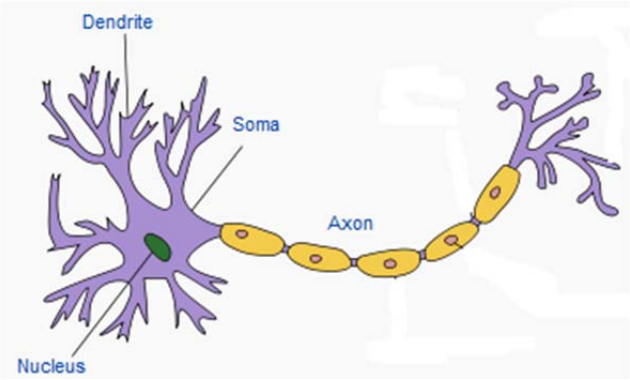


Figure 6: Structure of a neuron in living organisms (taken from http://en.wikipedia.org/wiki/Neuron accessed on 08.10.2010)

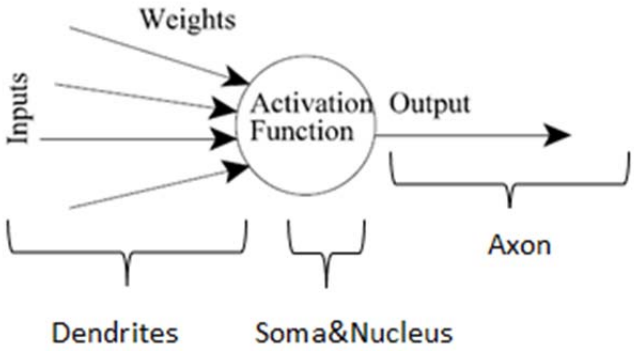


Figure 7: Structure of an artificial neuron (Adapted from Data Mining: Concepts and Techniques. Jiawei Han, M. Kamber. 2001)

The summation of inputs generally called Net Input [75] and the threshold value is determined by using an activation function. This can be formulated as:

$$\text{Net Input} = \sum_{i=1}^{m} W_i X_i \qquad \textbf{(EQUATION 13)}$$

Where *w* is the weight of input vector *x* and there are *m* distinct input vectors.
A stepwise or Linear combination function can be used for linearly separable problems as the activation function. The stepwise function is:

$$C = \begin{cases} 1 \text{ if Net Input } \geq \emptyset \\ 0 \text{ if Net Input } < \emptyset \end{cases} \qquad \textbf{(EQUATION 14)}$$

Where *C* indicates class and ϕ indicates the threshold value.

In the linear combination, a bias input is added to the system and the value of this input is set to 1. In this situation net input calculated by:

$$\text{Net Input} = 1 * w_b + \sum_{i=1}^{m} W_i X_i \qquad \textbf{(EQUATION 15)}$$

Where w_b indicates the weight of bias term.

To adapt perceptrons to non-linear problems different activation functions must be used. In such conditions different types of Sigmoid function is used.

- Logistic Function: $S(t) = \frac{1}{1+e^{-t}}$ **(EQUATION 16)**
- Hyperbolic Tangent: $\tanh x = \frac{e^{2x}-1}{e^{2x}+1}$ **(EQUATION 17)**
- Algebraic Function: $f(x) = \frac{x}{\sqrt{1+x^2}}$ **(EQUATION 18)**

In topological point of view an ANN can be divided to two groups which are single layer and multi-layer networks. Single layer ANN consists of only input and output neurons where multi-layered ANNs consist of more than two layers. The layers between the input and output are known as the hidden layers. A typical multi-layered ANN is given in the Figure-8 below:

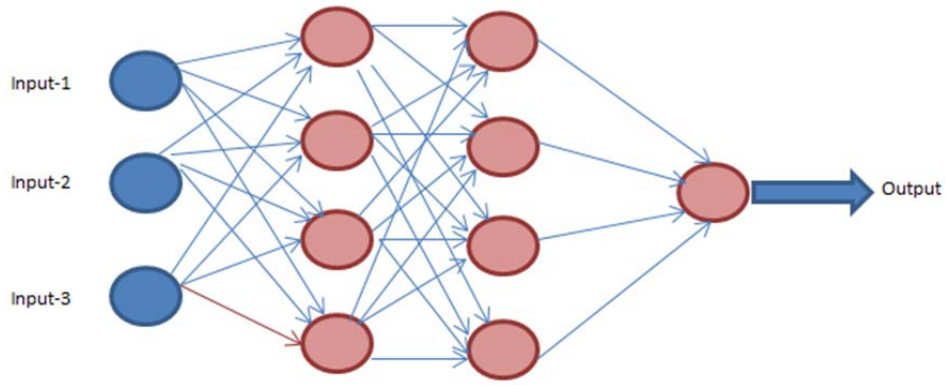


Figure 8: Structure of an artificial neural network. (Adapted from Data Mining: Concepts and Techniques. Jiawei Han, M. Kamber. 2001)

There are two main learning schemes for the ANN which are feed forward and the back propagation. In the feed forward approach a network is constructed by using one of the topologies and the activation functions described above. But such kind of ANNs could have some learning difficulties and could introduce a large margin of error between the actual and desired output. This is because no adjustments are made on the network's learning scheme. But in the back propagation algorithm the error between the actual and desired output is calculated than all weights are adjusted in all layers in order to minimize the error. This adjustment requires tracing back from output to input layers that is why the method is called back propagation.

The following explanations, notations and formulations to explain back propagation is taken from [76]:

The error between desired and actual value is calculated by:

$$E = \frac{1}{2\sum_{i=1}^{n}(D_i - A_i)^2} \qquad \textbf{(EQUATION 19)}$$

Where D_i indicates the desired output and A_i indicates the actual output. The weight adjustment is done according to:

$$\omega_{ki}(\tau) = \eta\, \delta_k\, x_i \qquad \textbf{(EQUATION 20)}$$

Where w_{ki} indicates the weight in τ iteration, xi indicates input vector, η indicates the learning rate and δ_k indicates the error of output. The learning rate here must be adjusted very carefully. Higher rates can introduce fast training but there is a risk that the model can not find the optimum decision point. Smaller rates sometimes better at finding the optimal decision point with a disadvantage of very long training time and they have the risk of sticking into local optimal points. To avoid such situations "*momentum term*" can be used as follows:

$$\omega_{ki}(\tau) = \eta\, \delta_k\, x_i + \alpha\, \omega_{ki}(\tau\text{-}1) \qquad \textbf{(EQUATION 21)}$$

By this way weight adjustments are depended to the previous adjustments.

Feed forward back propagation artificial neural networks are used in many different classification domains ranging from finance to medicine. Although their performance is sufficient enough for classification problems, new methods for ANNs are being developed. Nowadays one of the popular topics on ANNs is combining it with different algorithms such as Genetic Algorithms [77]. This approach is mostly used for finding the optimal parameters such as learning rates, number of hidden layers etc., for ANN.

1.1.2.1.5 Support Vector Machines (SVM)

Support Vector Machines (SVMs) developed by Vladimir N. Vapnik [78] are supervised classifiers that try to find a hyperplane which separates the data points belonging to different classes in the problem space. This separating vector must be at the furthest point to both classes which is known as Maximum-Margin.
Suppose that:

$$D = \{(x_i, y_i) | x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n}$$

Where:

D= Data Set
Yi = class indicator
Xi = p dimensional real vector that indicates the data point

A hyperplane in the given feature space can be written as:

$w \cdot x - b = 0$ **(EQUATION 22)**

Where :

W=normal vector
. = dot product

If the problem is linearly separable we can find two hyperplanes closest to the classes and they can be written as:

$w \cdot x - b = 1$ **(EQUATION 23)**
$w \cdot x - b = -1$ **(EQUATION 24)**

So the distance between these two hyperplanes can be stated as:

$\frac{2}{||w||}$ **(EQUATION 25)**

By this way optimization problem is to minimize ||w|| as in the Figure-9 below:

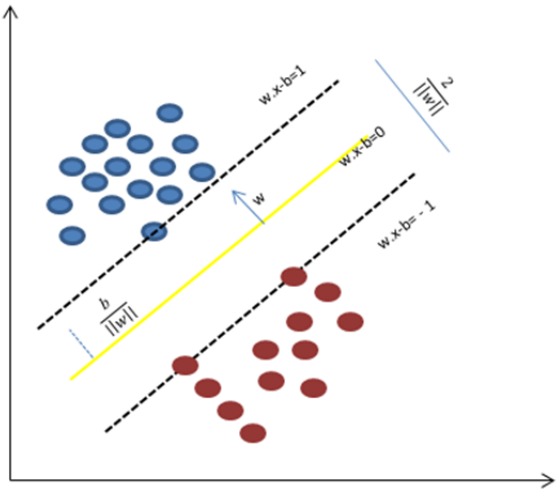


Figure 9: SVM optimization problem taken from
http://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

This model is not suitable for non-linearly separable problems. In non-linear problems, classes can not be separable by the hyperplane in the same dimension space because data points form a hackly structure as in the Figure-10 below.

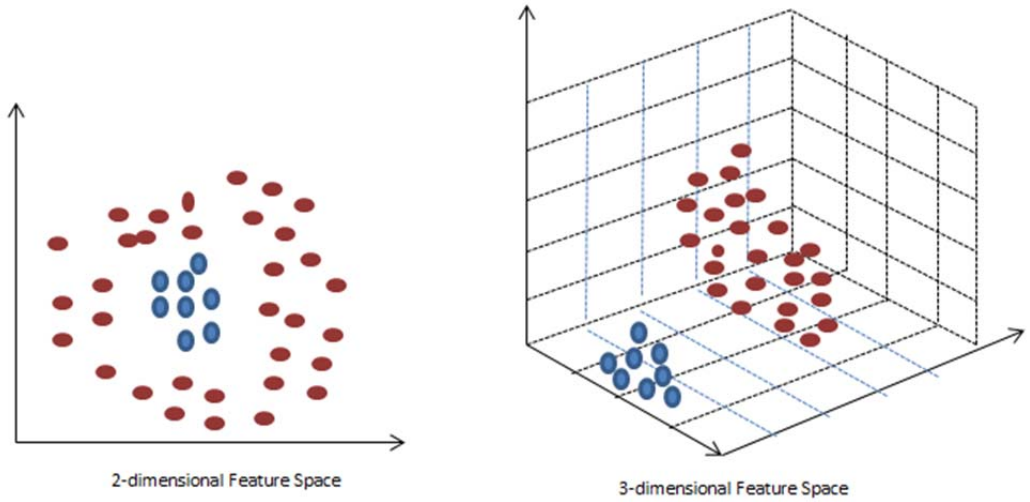


Figure 10: Matching two dimensional space onto three dimensional by Kernel Trick. (Adapted from Support Vector Machines Explained by Tristan Fletcher. UCL 2009)

In such cases it's sometimes useful to convert the feature space to a much higher dimensional space in order to separate the classes. This is done by using a Kernel Function and mapping the problem to a higher dimensional space is known as the Kernel Trick [79].
Let **x** be a vector in the *n*-dimensional input space and $\phi(\cdot)$ be a nonlinear mapping function from the input space to the high-dimensional feature space.

The hyperplane representing the decision boundary in the feature space can be defined as follows.

$$\mathbf{w} \cdot \phi(\mathbf{x}) - b = 0 \qquad \textbf{(EQUATION 26)}$$

where **w** denotes a weight vector that can map the training data into the high dimensional space. But calculating the dot product could introduce too much computational power. So instead of dot product Kernel Function can be used which eliminates the need for conversion the input vector to high dimensional space.

$$K(\mathbf{u},\mathbf{v}) = \phi(\mathbf{u}) \cdot \phi(\mathbf{v}) \qquad \textbf{(EQUATION 27)}$$

There are many different kernel functions [80] can be used according to given problem. These are:

- Polynomial: It's directional kernel function, mainly suitable for problems where all data points are normalized. Homogenous and in homogenous polynomial kernel functions are given in the formula below.

$$k(x_i,x_j) = (x_i.x_j)^d \qquad \textbf{(EQUATION 28)}$$
$$k(x_i,x_j) = (x_i.x_{j+1})^d \qquad \textbf{(EQUATION 29)}$$

- Sigmoid: This type of kernel originated from neural network field [81]. "*In fact, a SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network.*"[7] Sigmoid kernel function formula is given below

$$k(x_i,x_j) = \tanh(kx_i.x_j + c) \qquad \textbf{(EQUATION 30)}$$

- Radial Based (RBF): This kernel is also closely related with artificial neural networks [82] and known with its faster learning speed [83]. Its formulation is given below

$$k(x,y) = \exp\left(-\gamma||x-y|\right|^2\Big) \qquad \textbf{(EQUATION 31)}$$

- ANOVA: Analysis of Variance is a variance analysis used to discover the relations between the independent variables and effect of these relations on dependent variable. This kernel is a version of RBF which performs well on high dimensional regression problems [84].

$$k(x,y) = \sum_{k=1}^{n} \exp\left(-\sigma\left(x^k - y^k\right)^2\right)^d \qquad \textbf{(EQUATION 32)}$$

In SVM applications choosing the right kernel type that will give better classification performance than other kernels is the main problem. That problem arises because there are many different kernel functions can be used and even a specific type of kernel function has its own variations. For example RBF can be modified into Gaussian or ANOVA and polynomial kernels differ according to homogeneity. In this situation choosing the right type of kernel and its parameters discussed in the literature [85-86-87].

Our literature search so far showed us that there is no agreed on standard to choose the best kernel function. Choosing the kernel function mainly depends on the problem at the hand, feature space and data type. The work in [55] suggest to conduct a wide research on the problem and to find previously used methods in the field in order to determine which kernel function to use.

Once the appropriate function is selected the problem changes into finding the right parameters for the kernel. If a linear kernel is used the penalty factor has to be decided, or if a RBF is used penalty factor and epsilon have to be decided to find the best performing RBF. Generally, grid search [88-89-90] and evolutionary algorithms[8] [91] such as genetic [92] algorithms are used to find the optimum set of parameters. But details of such algorithms are beyond the scope of this thesis. Details of the SVM model used will be given in the Results Section.

[7] http://www.dtreg.com/svm.htm accessed on 27.09.2012

[8] Evolutionary tuning of multiple SVM parameters. Frauke Friedrichs, Christian Igel. Institut fur Neuroinformatik, Ruhr-Universitat Bochum, 44780 Bochum, Germany. Elsevier Neurocomputing 64 (2005) 107–117 Available online 7 January 2005

1.1.2.1.6 Genetic Algorithms

It is another algorithm that is influenced from nature. Like ANNs genetic algorithms try to imitate the evolution process. This idea is implemented by J. Holland in 1970's.

A genetic algorithm (GA) mainly consists of a population base, mutation and crossover operations for the generation evolution and a selection criterion to choose reproduction candidates from the population base.

The real world objects are encoded in binary strings which are called genotypes form the population base. The initial population base is the first generation. The second step is choosing the parents from the base in order to produce the next generations. To mimic the real world evolution theory only the best parents must be selected. Parents are evaluated by a fitness function. Each individual in the population base is assigned to a fitness value that represents how good they are for solving the problem in the hand. This fitness value is calculated by the fitness function. Different functions purposed are; Comprehensibility Metric, Predictive Accuracy and Interestingness based on information gain [93].

After each parent in population base is given a fitness value than the selection can be done by using different selection schemes some of which listed below:

- Roulette Wheel: Each parent is assigned to a place proportional to its fitness value on a circle that is called roulette wheel. Then a random selection is done over the wheel. But the parent with the highest fitness value has more chance to be chosen because the area that it occupies on the wheel is larger than others.
- Rank Based Selection: In roulette based approach if there is a big gap between the best and worse fitness values there could be a problem. That always the same parents selected because they occupy a very large area. To avoid that ranked based selection [94] can be used. In this method parents are ranked according to their fitness values. Each rank has a predefined place on the wheel. So parents with lower fitness values can also be selected. Comparison of roulette wheel and rank based selection is given in the Figure-11 below.

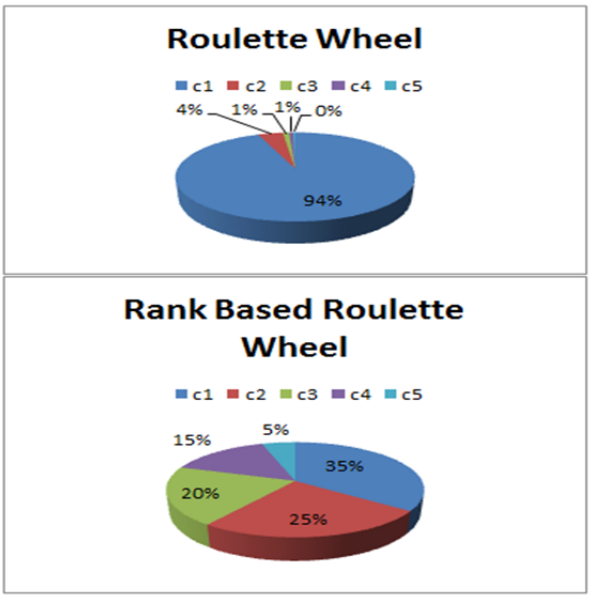


Figure 11: Roulette and Rank Based Selection. (Adapted from: A Comparative Analysis of Selection Scheme. Sonali Gandhi, Deeba Khan, Vikram Singh Solanki. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-4, September2012)

- Tournament Selection: In this approach a *Tour* number [95] is determined. The lowest tour number is 2 and the highest is *n* which is sample size. Then a random selection is done over the population base and a subset, sized by tour number, of individuals is formed. From this subset the best individuals chosen as a parent.
- Genitor: In this algorithm, parents are selected by linear regression and then the worst parent is deleted and replaced by the offspring [94].

The second important topic in GA is the reproduction of the next generations. Crossover and mutation operations are used for this purpose. Crossover is done between two parents but in mutation only a single parent is used.

There are mainly four crossover methods used by Gas. These are:

- One Point Crossover: In this scheme a random point is defined to cut the parents then children is formed by adding the cut points to each parent in a cross over scheme. This is depicted in Figure-12.

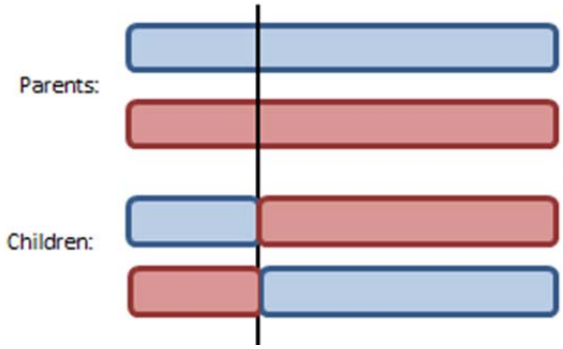


Figure 12: One Point Crossover. (Adapted from: A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. David E. Goldberg , Kalyanmoy Deb. Foundations of Genetic Algorithms Conference. 1991)

In this approach a positional bias can arise because places of the successive genes never change.

- Two Point Crossover: In this method two random points are selected for splitting the parents. By this way each parent is divided into three parts. Then each part is crossed over as in the Figure-13.

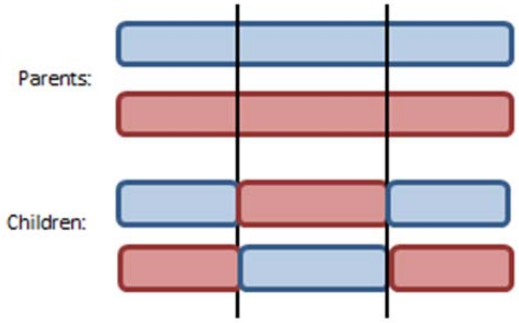


Figure 13: Two Point Crossover. (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

- Uniform Cross Over: In order to solve the positional bias, uniform cross over handles the bits of parents individually. And with a probability of 0.5 bits in the current position can be swapped between parents [96]. This scheme is depicted in the Figure-14 below.

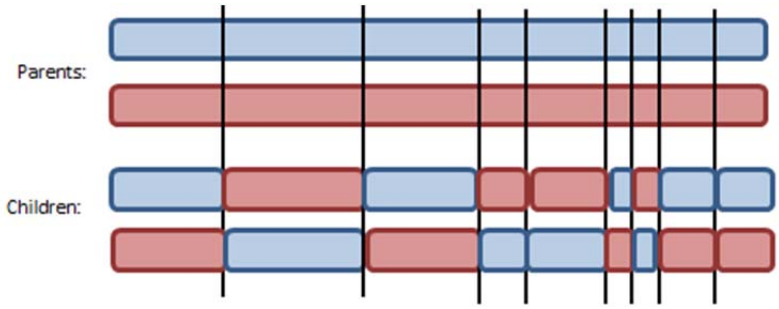


Figure 14: Uniform Cross Over (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

- Cut and Splice: The idea behind cut and splice is to divide the parents from different positions and cross over the divided parts to form different length children as described in the Figure-15.

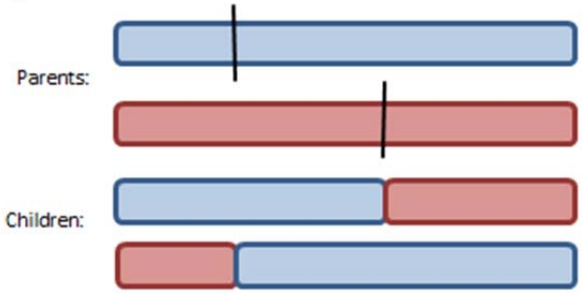


Figure 15: Cut and Splice (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

The crossover operation produces children with the information inherited from parents. By this way there is no new information produced. But in mutation the bits in a single parent is changed which cause the new information can be produced. There are five basic mutation operations can be used. These are:

- Bit Flip: In a single parent a bit in a position is changed to its opposite value.

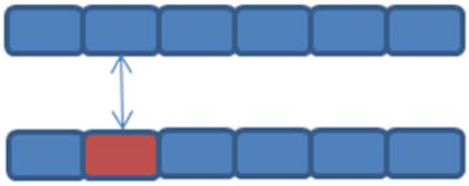


Figure 16: Bit-Flip (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

- Insert Mutation: Two bits are randomly selected. Then the second bit is placed adjacent to the first one as in the Figure-17.

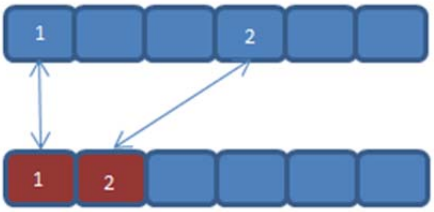


Figure 17: Insert Mutation (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

- Swap Mutation: As in insert mutation two bits are randomly selected. Then the positions of these bits are interchanged as in Figure-18.

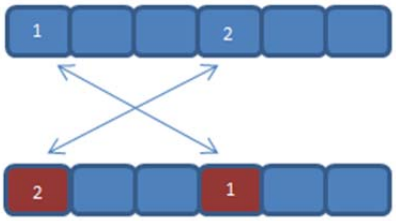


Figure 18: Swap Mutation (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

- Inversion Mutation: Two random points are selected over the parent string. Then the bit sequence between these two points is inverted as in Figure-19.

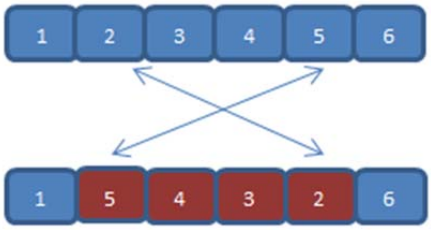


Figure 19:Insert Mutation (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

- Scramble Mutation: There are n bits selected over the parent string and then they are randomly distributed to their new positions over the parent as in figure-20.

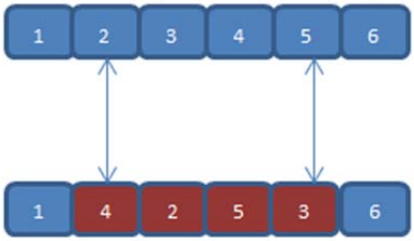


Figure 20: Scramble Mutation (Adapted from: Search Methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Edmund K. Burke, Graham Kendall. ISBN: 978-0-387-23460-1. Chapter4)

Genetic algorithms are generally used in optimization problems to find the best approach or setting the best parameters of the used methodology. Some of the recent uses in the literature are; prediction of traffic accidents [97], input and output estimation [98], layout problem solving [99]. Recent uses of GA in medical informatics field are; in medical imagining [100], cancer research such as breast [101] and lung [102], disease diagnosis and remote patient care [103].

1.1.2.2 Un-Supervised Learning Methods

In un-supervised learning, the labels of data which indicates the class information are not known beforehand. Because the labels are not known priori the problem becomes to which data points are similar to form cluster. So un-supervised learning is referred as clustering.

In order to group data points into clusters similarity of data points must be calculated. This is done by using similarity functions or distance measures which defines the distance between data points so that closest ones could form a cluster.

Distance measure is generally used if the data points are numeric valued. The most widely used distance measures are Euclidean and Manhattan distance. Formulation for these two distance measures are given below:

Euclidean Distance: $D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (**EQUATION 33**)

Manhattan distance: $D(X,Y) = \sum_{i=1}^{n}|X_i - Y_i|$ (**EQUATION 34**)

Similarity measure is generally used for categorical data bur can also be applied to numerical domain. For categorical data correlation and Jacckard's Coefficient [48] is used for similarity measure. Suppose that we have two binary data points namely *i* and *j* as given in the table below. According to this table formulation of the methods, Correlation and Jacckard, is given below:

Table 1: Four by Four Table matching exact and measured outcomes

		J	
		1	**0**
I	**1**	q	r
	0	s	t

Correlation: $\frac{q+t}{q+r+s+t}$ (**EQUATION 35**)

Jaccard's Coefficient: $\frac{q}{q+r+s+t}$ (**EQUATION 36**)

Clustering methods can be grouped under five categories [48]. These are:

- Partitioning Methods
- Hierarchical Methods
- Density Based Methods
- Grid Based Methods
- Model Based Methods

The following subsections briefly summarize these methods as a part of decision support techniques. But no further details will be given because clustering algorithms are not used in this thesis.

1.1.2.2.1 Partitioning Methods

As the name indicates these algorithms try to partition the data points and at the end each partition represents a cluster [48]. Among partition algorithms to most widely used ones are the K-Means and K-Medoids [104].

The K-Means [105] algorithm works in three main steps. In first step k numbers of initial cluster means are selected randomly. In the second step the distance between these means and data points are calculated. The data points that are closer to particular means form the initial clusters. In third step for each cluster the cluster means are re calculated. In this step data points may change their clusters according to their distances to new cluster means. This step is repeated until there is no new assignment is done or the desired error rate is reached. This error rate [48] can be calculated by:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \qquad \textbf{(EQUATION 37)}$$

Where p is data point, m_i is the cluster mean and C_i is the cluster.

K-Medoids algorithm is very similar to K-Means. The initial step is almost identical but in repeating step instead of changing the means the cost of changing the medoids is calculated [106]. If this cost is higher than the previous cost then no change is made.

1.1.2.2.2 Hierarchical Methods

Data points are clustered to form a hierarchical tree structure [48]. This structure is named as Dendogram [48-107]. In hierarchical methods the importance is based on how close is the cluster and according to this closeness can clusters be merged or not. In merging strategy three measures can be used, which are:

- Single Link: The distance between closest points of two different clusters measured. And these clusters are joined if the distance is under a predefined threshold.
- Complete Link: The distance between furthest points of two different clusters measured. Clusters are joined if the distance is under a predefined threshold.
- Average Link: The distance between center points of two different clusters measured. Clusters are joined if the distance is under a predefined threshold.

There are other hierarchical based methods such as BIRCH, CURE and Chamelon [48].

1.1.2.2.3 Density Based Methods

In density based clustering, clusters are formed according to dense points in data space. By this way arbitrary shaped cluster can be detected [48] and "*clusters are separated from each other by contiguous regions of low density of objects and low density regions are considered as noise or outliers*"[9]. There are different algorithms used in literature to compute the densities and forming clusters [48-108]. These are:

- DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSACN) is a very popular method in density based clustering. The main idea is based on a threshold distance called *neighborhood distance* [108] and another threshold that determines the number of points in cluster called *minPts* [108]. In this algorithm a point in data space is randomly selected and its neighborhood is formed. If this neighborhood contains data points above minPts then a new cluster is constructed.
- OPTICS: Ordering Points to Identify the Clustering Structure (OPTICS) is a form of DBSCAN that can handle varying density data space. In addition to DBSCAN's parameters OPTICS define two concepts, *core distance* and *reachable distance* [108]. The core distance is the distance that satisfies minPts around a chosen point. And reachable distance can be defined as [109]:

$$\begin{cases} \text{Undefined}, & |N_{\epsilon}(o)| < \text{MinPts} \\ \max\left(\text{core distance}_{e,\text{MinPts}}(o), \text{distance}(o,p)\right) & \text{otherwise} \end{cases} \quad \textbf{(EQUATION 38)}$$

  Where $N_{\epsilon}(o)$ is the neighborhood of point *o*.

- DENCLUE: Density Based Clustering (DENCLUE) is based on density distribution functions [48]. According to [48] DENCLUE has three important parameters which are Influence Function, Overall Density and Density Attractor. Influence function shows the effect of a data point in its cluster, overall density is the sum of all influence functions and density attractor is the "*local maxima of density function*" that is used to determine the cluster[1]. Formulation for these parameters are [110]:

  $$\text{Influence Function: } F_B^Y(x) = f_b(x,y) \quad \textbf{(EQUATION 39)}$$
  $$\text{Overall Density Function: } F_B^D(x) = \sum_{i=1}^{n} f_B^{x_i}(x) \quad \textbf{(EQUATION 40)}$$

  Where F^d is d dimensional feature space and y is a data point $y \in F^d$. Gaussian functions can be used as influence function then the equations above become[68] [110]:

  $$\text{Influence function: } f(x,y) = e^{-\frac{d(x,y)^2}{2\sigma^2}} \quad \textbf{(EQUATION 41)}$$
  $$\text{Overall Density Function: } f_G^d(x) = \sum_{i=1}^{n} -\frac{d(x,x_i)^2}{2\sigma^2} \quad \textbf{(EQUATION 42)}$$

Because DENCLUE is based on more strict mathematical bases it can perform well on noisy data sets and data space with arbitrary cluster shapes [48].

[9] Density Based Clustering. Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek. Wiley Interdisciplinary reviews: Data Mining and Knowledge Discovery **1** (3): 231–240. June 2011.

1.1.2.2.4 Grid Based Methods

In this approach the data space is divided into cells to form a uniform grid structure and in each cell clustering operations are carried out that makes the operation time independent from data size but depended to the size of the gird [48-111-112]. The most widely used grid based methods are the STING, Wavecluster and CLIQUE.

- STING: The grid structure is formed by rectangular shaped cells, each cell is further divided into child cells and generally a cell contains four child cell, then for each cell a number of statistical computations are done over the data points in cell such as maximum and minimum values, mean, standard deviation and chi-square test [48-111]. When a query is given to the grid, the most relevant top level cell is searched. In each level of this cell relevancy is calculated by statistical methods and by this way irrelevant cells are eliminated [48]. This continues until the bottom cell and if the requirements are met relevant cells found are returned but if the cells do not satisfy the requirements then other top level relevant cells are searched [48].
- Wave Cluster: Data is summarized by multi-dimensional grid structure [48-111]. Each cell in the grid contains the summary of its data points [48-111] then clustering is done by using wavelet transformation [48].
- CLIQUE: In this method the data space is divided to rectangular cells that do not overlap [48]. For a given query, each cell is calculated to see if it is dense for given query parameters. The intersection of these dense cells forms the candidate search space [48].

1.1.2.2.5 Model Based Clustering Methods

These methods are based on the assumption that data points show some probability distribution[1]. For this reason model based approaches try to find these distributions by different methods. In such approaches clusters resemble classes because each cluster has its characteristic descriptions [48]. Such clustering methods are known as conceptual clustering [113]. Because the clustering scheme is conceptual, like classification, some methods in classification can be used such as COBWEB, a classification tree structure [48] and Neural Network.

1.1.3 Conclusion on Data Mining Methods

In this section the advantages and disadvantages of the data mining methods will be given.

The firs method we've discussed was the KNN algorithm. KNN is known as its basic application. But it has optimization problems for the distance measure used and determining the number *k*. This approach does not give weights to attributes [48] so it is hard to interpret which attributes are more relevant to the classification problem than the others. Another disadvantage is that it can perform poor over very large high dimensional data so it can suffer from low accuracy [114].

Like KNN, Decision Trees area also very easy to implement. Because of their visual representation, the interpretations of the results are very easy too. This visual representation also enables rule extraction in the form of if-else structures. Another advantage of the decision tree is that they are somewhat independent of the data type. That is there are different types of trees that can handle numeric and/or categorical data.

If the attributes in data space is categorical then ID3 Tree can be used, in contrary if the attributes have numeric values then C4.5 or CART can be used. Decision trees can also work well on data with missing values and noise [115]. Although decision trees are good learners this requires data attributes to be closely relevant to the subject. If the relevancies of the attributes are not high, then decision trees can have some difficulties in proper classification [115]. Another disadvantage is if there is vast amount of data then the tree structure could become very complex and the same sub-trees could appear more than once in the tree [115].

Bayesian Belief Networks (BBN) is another method that presents visual representation of the given problem. By this way BBNs are easy to interpret [116]. Because BBN methods are based on probability distributions they are said to be robust to irrelevant data attributes [117]. In theoretical aspect BBNs should have the lowest error rate when compared to other classification methods such as ANN or Decision Tree, but conditional independence the absence of the prior probabilities hinders this performance [48]. Although they work well with categorical and numeric data, computational complexity on numeric data and categorical data with wide ranges can be given as a disadvantage for BBNs, because it becomes very hard to calculate the probability distributions [117].

Unlike the methods above, ANNs can perform well on large and complex datasets. They are generally in top methods according to classification performance because they can generalize both linear and non-linear classification problems and can handle data with noise and missing values [76]. The main disadvantage of ANNs is the determining the best model. That requires the optimization of the parameters such as number of neurons to be used in each layer, number of layer to be used and the learning rate to avoid overfitting. Other disadvantages stated are the long training times and the black box concept [76]. Although the decision performance is very good, the real decision mechanism of the ANNs is somewhat very hard to understand and interpret so this situation is named as the black box. But there are some works in the literature [118-119-120] try to formalize methods for interpreting the decision mechanism of ANNs by using weight analysis of neurons in each layer, sensitivity analysis and such [76].

Like ANNs, SVMs can handle large amount of high dimensional data. Because of its high accuracy rate on such kind of data space, SVMs are very useful for bioinformatics classification problems [49]. The main advantage of the SVMs is the flexibility to separate class regions [50]. SVMs are mostly compared to ANNs in the literature. This is because both methods have the ability to work on linear and nonlinear problems. In such comparison SVMs come forward by handling its convex structured problems so it is hard for an SVM to stuck at a local minima as in ANNs, this also introduce a less overfitting [50]. But SVMs can not perform well on categorical data as ANNs. In such conditions a data transformation must be done. The basic method offered is to form an *n* bit string for *n* value categorical attribute and set the bit to *1* to corresponding attribute [86]. SVMs are good at binary classification but they can not show good performance on multi class problems as multilayer ANNs [86]. As in some methods SVMs also need parameter optimization in order to find which of the kernel methods to be used and optimum values for the selected kernel function.

Genetic Algorithms are highly parallel in their structure [121]. This makes GA very suitable for optimization problems where the underlying problem is "*discontinuous, non-differentiable, stochastic, and highly non-linear*"[10].

[10] Prediction for Traffic Accident Severity: Comparing The Artificial Neural Network, Genetic Algorithm, Combined Genetic Algorithm And Pattern Search Methods: Mehmet Metin Kunt, Iman Aghayan, Nima Noii. Taylor & Francis Group. TRANSPORT ISSN 1648-4142 print 2011 Volume 26(4): 353–366

Generally speaking the main advantages of GA are; its parallelism, they do not tend to stuck at local minima so reaching the global minimum is easy and can handle noise. Because of its optimization ability GA is used to optimize the parameters of the methods discussed above and/or evolution of these methods. By this way GAs appears as one of the methods in hybrid data mining approaches. For example in Laura Diosan's [122] work, GA is used for SVM kernel evolution. Some examples for hybrid methods that use GA in literature are; ANN and GA hybrid [123], Decision Tree and GA hybrid for cancer disease [124]. Like the methods discussed above GA suffers from optimal model construction. In order to get the best results one must identify the fitness functions, mutation and crossover rates and selection functions very carefully and in a way that suits the given problem.

1.2 Genomic Analysis

In this section of thesis, brief information about genes, human genome project, single nucleotide polymorphism and its association with diseases such as cancer will be given.

1.2.1 Basic Concepts

Long before the Human Genome Project (HGP), the main concern of the biological sciences was the understanding of the genomic information [125].

The flow of genetic material starts with the Nucleotides which are the building blocks of the genetic material such as DNA and RNA. The main content of Nucleotide is the organic base. It is a special nitrogen structure and a nucleotide is named according to its organic base such as Adenine (A), Thymine (T), Guanine (G), Cytosine (C), and Uracil (U). A nucleotide also contains phosphate, carbon and sugar structures in it [126].

A single strand of nucleotides composed with ribose sugar forms the Ribonucleic Acid (RNA). Organic bases that RNA contains can be adenine, cytosine, guanine and uracil [127]. RNA is basically responsible from encoding the genetic information but there are other duties taken by different types of RNA. As its name indicates messenger RNA (mRNA) is responsible from carrying the genetic information during protein synthesis. The job of Transfer RNA (tRNA) is to transfer a specific amino acid structure in protein synthesis. Ribosomal RNA (rRNA) is the structural member of the ribosomes. The general RNA structure is given in the Figure-21.

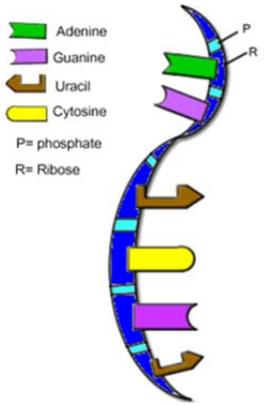


Figure 21: RNA Structure (Taken from http://www.biologycorner.com/bio1/DNA.htmlaccessed on 20.07.2012)

Deoxyribonucleic Acid (DNA) is formed by two strands of nucleotides which they form a double helix structure. To form such a structure only specific base pairs must be paired. These pairs are Adenine and Thymine, Cytosine and Guanine.

These pairs are connected with hydrogen bonds; in A-T connection there are two and in C-G connection there are three hydrogen bonds [128]. This whole structure, DNA, is responsible from carrying genetic instructions for biological evolution and living. DNA structure is given in the Figure-22 below:

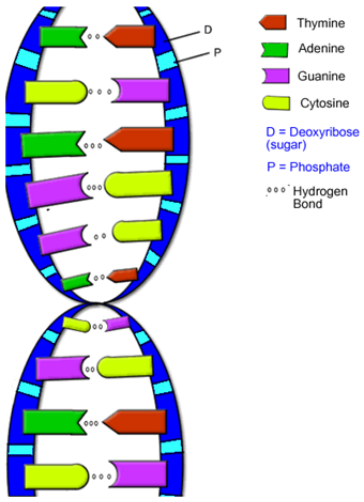


Figure 22: DNA Structure. (Taken from http://www.biologycorner.com/bio1/DNA.htmlaccessed on 20.07.2012)

The specific part of DNA that codes biological and chemical information of the living organism is called as Gene. There are two important parts of gene. One of them is the Exon [129] and responsible from transferring the DNA base pairs to mRNA in protein synthesis and the other is Intron [130] which occupies places between exons but not transferred. Gene structure is given in the Figure-23:

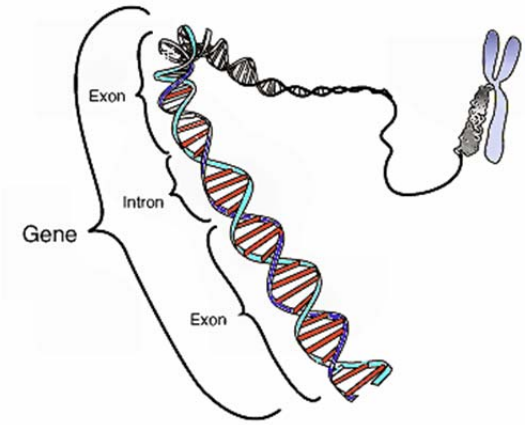


Figure 23: Structure of a gene. (Taken from http://en.wikipedia.org/wiki/File:Gene.png accessed on 20.07.2012)

The whole genetic information of the organism is called Genome and it is stored in the chromosome structure. Chromosome is formed by a single, long DNA and responsible from cell division and reproduction.

The genetic information is transferred by the protein synthesis. This is a three step process which consists of Transcription, Splicing and Translation. In transcription step, hydrogen bonds of the DNA are unzipped, one of the strands is chosen and its complementary strand is transferred to RNA. After that the introns of the RNA are dropped and exons are joined in splicing step. In the last step, translation, amino acid structure, amino acid chains and the specific protein is formed according to the genetic information carried. This synthesis process is given in the Figure-24.

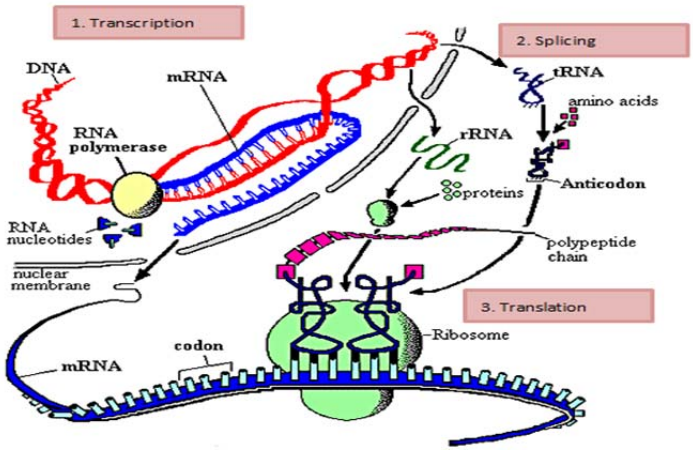


Figure 24: Process of protein synthesis in three steps. Transcription, splicing and translation. (Taken from http://www.accessexcellence.org/RC/VL/GG/images/protein_synthesis.gif accessed on 21.07.2012)

1.2.2 Human Genome Project

Human Genome Project (HGP) started in 1990 and completed in 2003, initiated by the Office of Science of the US Department of Energy and continued as an international research project with the aim of physiological and functional mapping of all human genes which are called genome. The National Human Genome Research Institute (NHRI) was established to conduct this international research. Participant countries were China, France, Germany, United Kingdom and Japan [131].

The main goals of the project are [132]:

- Identifying all human genes
- Sequencing of base pairs
- Developing new storage schemes and sequencing techniques

In a more detailed way, the milestones can be chronologically given as:

Table 2: Human Genome Project milestones taken from [131]

Goal	Achieved	Date
2- to 5-cMresolution map (600-1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
30,000 STSs	52,000 STSs	October 1998
95% of gene containing part of human sequence finished with 99% accuracy	99% of gene containing part of human sequence finished with 99% accuracy	April 2003
Sequence 500Mb/year at <0.25$	Sequence > 1,400Mb/year at <0.09$	November 2002
100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Full length human cDNAs	15,000 full length human cDNAs	March 2003
Complete genome sequences of E. Coli, S. Cerevisiae, C. Elegans, D. Melanogaster	Finished genome sequences of E. Coli, S. Cerevisiae, C. Elegans, D. Melanogaster plus whole genome drafts of others such as mouse and rat	April 2003
Develop genomic scale technologies	High throughput oligonucleotide synthesis DNA microarrays. Eukaryotic, whole genome knockouts (yeast). Scale up two hybrid systems for protein-protein interaction	1994-1996 1999-2002

As a byproduct of this research, genome mappings for other living organisms such as bacteria, worm, fruit fly and mouse [133] are done and ethical, legal, and social implications (ELSI) arising from this research examined.

Genetic mapping, physical mapping and DNA sequencing was the important methodologies used in HGP to achieve the goals stated above.

DNA sequencing is the main goal which in fact contains genetic mapping and physical mapping. That is because DNA is very long so it must be divided to smaller fragments for sequencing [134]. These smaller fragments are researched with genetic and physical mapping methods. This scheme, dividing the DNA into smaller parts, examining these smaller parts and searching for overlapping areas for sequencing is known as the Shotgun Method and depicted in the Figure-25.

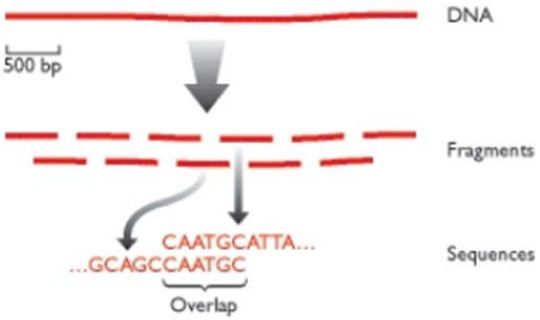


Figure 25: DNA sequencing by Shotgun Method (Taken from [134])

Genetic mapping uses genetic methods such as cross-breeding and try to find the locations of genes on a genome [133]. On the other hand physical mapping uses molecular biology methods [135] and "*maps the chemical characteristics of DNA molecule itself*"[11].

Conductors of this research believed that the sequencing and mapping the human genome will introduce new methods for diagnosis and treatment of the diseases. Further it will have a huge impact on early determination and/or even the prevention from diseases which are mostly caused by genetic variations. And give an opportunity to closely look the underlying effects of the diseases.

1.2.3 Genome Wide Association Studies and SNPs

A single nucleotide change in genome that cause variations in DNA sequence is called Single Nucleotide Polymorphism (SNP) and changed nucleotide pair is called Allele. The Figure-26 below depicts SNP and Allele structures.

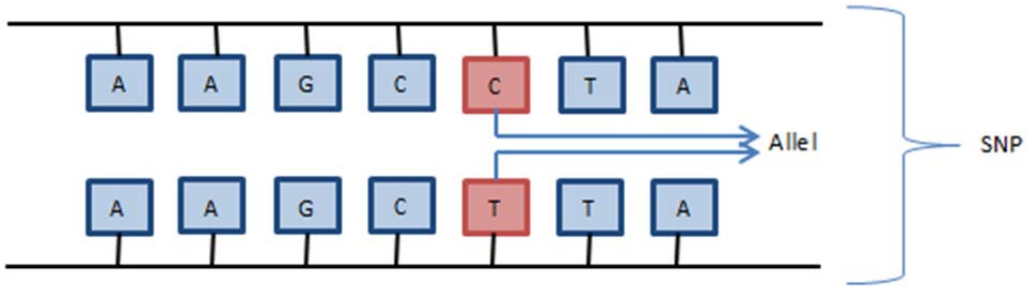


Figure 26: Single Nucleotide Polymorphism and Alleles

Such variations sometimes can occur in only one individual of the population. So in order to call a variation as a SNP, it must be observed among at least 1% of the population [136]. In general, there are two main variations which are called Synonymous and Non-Synonymous Polymorphism. As the name indicates the variations in synonymous polymorphism does not cause a different amino acid chain to be produced. That's because a single amino acid chain can be coded with more than one structure. On the other hand if the variation causes different amino acid chain then it is called Non-Synonymous Polymorphism. Non-Synonymous Polymorphism can be further divided into two categories. The variation can cause a premature stop codon then this scheme is called as nonsense. If a new amino acid chain is produced this is called missense.

SNPs are important due the variations they cause. The different amino acid chain caused by SNP can affect the living organism such as the presence or absence of that protein can make the living organism vulnerable to some diseases. With this property SNPs are treated as key point for complex diseases such as cancer, diabetes, cardio-vascular and mental diseases [137-138-139]. They are also examined for personalized medicine and pharmacogenomics [140] where drugs or therapies specially designed for the individual according to his/her genomic information.

In a broader term searching the gene variations among one species to identify the difference between its individuals is known as Genome Wide Association Studies (GWAS). With GWAS many SNPs can be identified and searched for their relations with complex diseases.

[11] Bioinformatics: A Primer. P. Narayanan. New Age Internationla (P) Limited Publishers. ISBN: 81-224-1610-1. P 28.

Today approximately 12 million SNPs identified and with GWAS nearly 40 complex diseases are found to be related with particular SNPs [141].

In order to conduct GWAS, three main epidemiologic study designs are generally preferred. These are Case-Control, Trio and Cohort Studies [141]. In case control studies two groups are formed with one having the disease, Case, and other healthy. Then differences between these two groups are examined. In cohort studies a large population is chosen and observed for a long period of time. Then the differences between individuals who developed the disease and healthy ones are examined. Trio studies include the information of the subject's parents. Genotyping is done over offspring and its parents to discover the transmission frequency of alleles.

Whatever study design is used, there are some general points that GWAS must handle. These are high volume of data, computational complexity and overfitting [12]. The number of identified SNPs, their allele frequencies and locus information forms a high volume high dimensional data. This high volume, high dimensional data requires great computational resources such as storage and computational power. Also analysis of such kind of data can introduce computational complexity like NP-hard problems. These analysis methods can suffer from overfitting according to parameters they use. The Table-3 is taken from table Joan Hardy's [138] work and defines the advantages, limitations and misconceptions of GWAS.

Table 3: Benefits, misconceptions and limitations of GWAS

Benefits
Does not require an initial hypothesis
Uses digital and additive data that can be mined and augmented without degradation
Encourages the formation of collaborative consortia which tend to continue for subsequent analysis
Rules out specific genetic associations
Provides data on ancestry of each subject
Provides data on both sequence and copy number variations
Misconceptions
Thought to provide data on all genetic variability associated with disease
Thought to screen out alleles with small effect size
Limitations
Requires samples from a large number of cases and controls, can be challenging to organize
Finds loci, not genes
Detects only alleles that are common in population
Requires replication in similar samples

Methods used in GWAS can be grouped under two main topics which are parametric and non-parametric methods [12]. Parametric methods are based on statistical approaches such as linear regression, multiple linear regression and logistic regression [12]. These models suffer from high volume of data and correspondence issues between different statistical methods [12-142-143]. On the other hand data reduction methods such as Combinational Partitioning Method (CPM) [144], Multifactor Dimensionality Reduction (MDR) [145] and pattern recognition methods such as neural networks and genetic algorithms are grouped under non-parametric methods.

As HGP, most of the results of GWAS and individual SNP findings are publicly available. These results are shared by the project's sites and some public databases. Among them most important ones are:

- HapMap Project
- dbSNP
- Regulome
- dbGAP

International HapMap Project tries to "*find the differences and similarities in human beings*[12]" by searching the genetic variations. The identified information is cataloged and used for worldwide researches to identify the relations between genes and diseases and furthermore to design medications and therapies that will target these genes [146]. This project was organized in three phases. In phase-1 approximately 1 million SNPs are identified, in phase-2 2.1 million SNPs are identified and in the last phase nearly 1.6 million SNPs are identified [147-148-149].

dbSNP is another database for storing SNPs as well as the deletion and insertion polymorphisms [150]. This database is founded by the collaboration of NHRI and The National Center of Biotechnology Information (NCBI). SNPs, SNP association with diseases, SNP locations on genes and chromosomes are the main information can be searched over this database.

Regulome database stores information about SNPs and other regulatory elements in gene locations of Human Genome which are mainly taken from datasets such as GEO and ENCODE projects [151]. This database is based on MySQL and MyISAM engines and stored information is divided into parts over grouped tables such as PubMed articles are stored in Author DB and ArticleDB, genetic information is stored in GeneDB, SpeciesDB etc. [152].

dbGaP is founded by NCBI and as its name indicates, database of genotypes and phenotypes (dbGaP), it stores phenotype and genotype information. It contains the results of GWAS and phenotype information of many studies that use different kinds of designs such as cohort, case-control and trio. Each research is organized as study, variables, documents, analysis and datasets. Study gives information about the work done while variable gives the basic description and statistical summary of the variables used. Documents give more detailed information about the study and dataset gives brief information about phenotype, genotype and variables they contain. Each research or work has a unique identification number according to dbGAP ID format [153]; this number is used to identify the subsections (study, phenotype, genotype, document and association) of the related work. dbGaP provides two types of access which is private and public. In order to gain private access one should have NIH eRA Commons account or accepted as principal investigator by NIH and NCBI [153]. The number formatting, public and private access scheme is depicted in the Figure-27 below:

[12] International HapMap Project web site. http://hapmap.ncbi.nlm.nih.gov/thehapmap.html.en accessed on 20.10.2012

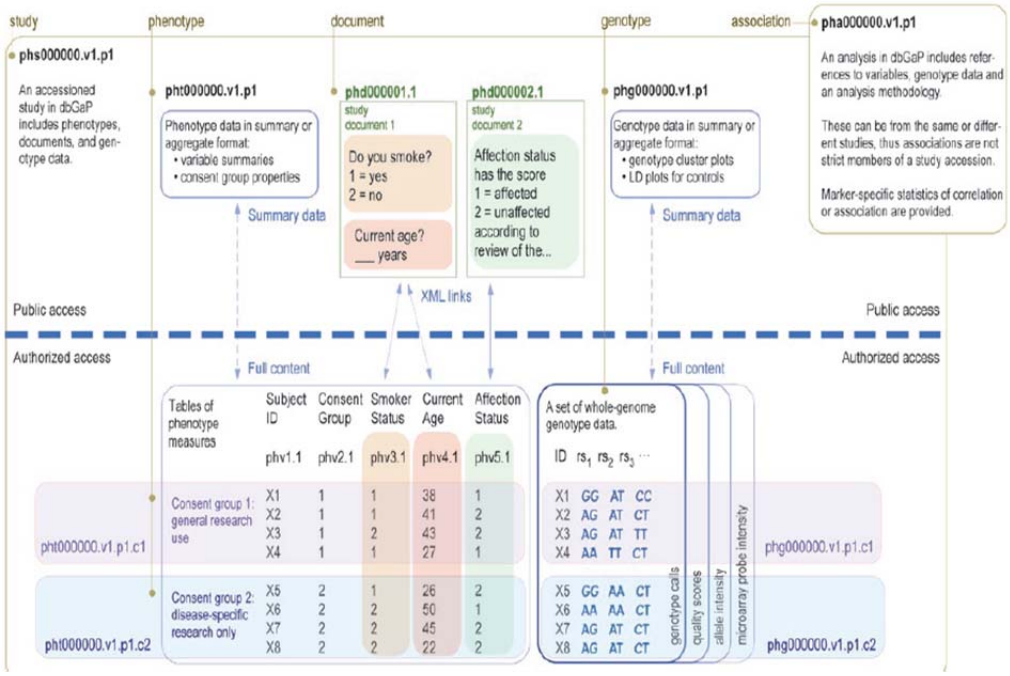


Figure 27: dbGaP file organization structure. (Taken from [153])

1.2.4 GWAS and Complex Diseases

GWAS focuses on SNPs and identify many relationships between SNPs and complex diseases. In this section a brief literature summary will be given on GWAS and related complex diseases.

Work of M. Martin and W. M. Brown searched the relation between genetic variations and ischemic stroke. They conduct a cohort study with 249 cases, 268 controls, 200million genotypes and exactly 408,803 SNPs [154]. But they could not find any important relations that are statistically significant.

Another study focuses on the relation between specific gene location and serum uric acid which is a strong indicator of Hyperuricemia [155]. With the help of specific loci in European population, they tried to map similar loci in African American population. As a result they found those strongly associated loci are almost the same for both populations.

Work of Alexander V. Alekseyenko [156] is one the researches done on GWAS and Rheumatoid Arthritis. They offer a graph based method over anti-CCP positive Rheumatoid Arthritis which performed well for finding the genetic factors behind the given disease with 0.81 ROC curve performance.

Another complex disease, Parkinson's disease, is studied by Hon Chung Fung and friends [157]. They try to find the gene variations behind this disease. They worked on a Parkinson's disease cohort with 267 cases, 270 controls and nearly 500,000 SNPs; at the result they published a publicly available data for the given disease with 220 million genotypes [157].

One of the complex diseases subject to GWAS is Alzheimer's disease (AD). Underlying clinical reasons of this disease are not so clear that makes this disease a good candidate for GWAS.

One of the studies on AD is Guiyou Liu's research [158]. In their work they used two datasets of French AD subjects one of them has 2031 cases and 5328 controls and other has 1034 cases and 118 controls. They used pathway analysis with statistical methods such as Z-statistics, Fisher's Exact Test, and Binomial Test for evaluating the results. In this work they found statistically significant pathways for AD.

Like AD, one of the diseases with unclear etiology is the Behcet's Disease. A research conducted by E. F. Remmers [159] on Behcet's disease with a dataset of 1215 cases ad 1278 controls. They worked on 311,459 SNPs and identify the associated locations. They checked their findings over a second dataset consist of different nationalities (Turkey, Middle East, Europe, Asia) with 2430 cases and 2660 controls. The detected locations were also the same in the second data set.

A popular disease topic for GWAS is the auto immune diseases. H. Hakonarson and S. F. A. Grant [160] made GWAS on two of such diseases which are Type-1 diabetes and inflammatory bowel disease. They found 16 loci for type-1 diabetes and 40 for the bowel disease.

Another form of inflammatory bowel disease is Ulcerative colitis. Two stage GWAS on ulcerative colitis over Japanese population is done by Kouichi Asano [161]. They used a dataset with 1384 cases, 3057 controls and identified three new susceptibility loci for the given disease.

In personalized medicine perspective one of the important aspects is the prevention of the disease by its early detection. SNPs that cause the disease are searched. If they found significant for the disease they are treated as one of the underlying factors. So an individual with that genetic variation is likely to have the disease. This makes very important to find the SNPs associated with the given disease for early detection. There are many works in the literature with this aim and one of them is the Mousheng Xu's [162] work for asthma disease, which is also an autoimmune disease, among children. They used random forests for selecting the associated SNPs. They related SNPs they found with the early detection of exacerbation of asthma disease.

Many skin related diseases are known as auto immune diseases. A survey like study is conducted by Xuejun Zhang [163] that explains the results of GWAS on various skin diseases. These diseases were Psoriasis, Systemic Lupus Erythematous, Leprosy, Virtiligo, Atopic Dermatitis, Male Baldness, Keloid Disease, Sarcoidosis, Cutaneous Basal Cell Carcinoma, and Cutaneous Melanoma. According to this survey many GWAS studies conducted on the given diseases and many genes and locations are identified as related with these diseases.

Age-related macular degeneration (AMD) is a disease that observed in higher percentages in elderly population has a risk for blindness. There are many GWAS done in literature for AMD [1-161-164-165]. One of the works is by Satoshi Arakawa et al. Their work is done over Japanese population with 1536 cases, 18,853 controls and they found new susceptibility loci for the given disease.

Multiple Sclerosis (MS) is an inflammatory disease which affects the functionalities of the nerve cells in brain and spinal cord. According to Anu Kemppinen, Stephen Sawcer and Alastair Compston's work there are 16 loci identified in association with MS [166]. These locations are close to the immunological functioning genes so they are also closely related with immune diseases. Because of this situation they offer to conduct much deeper and wider GWAS research on MS in order to identify the exact underlying genetic variants.

There are many GWAS conducted on cardio vascular diseases such as Myocardial Infraction (MI) and Coronary Artery Disease (CAD). A review made in 2010 by Pier Mannuccio Mannucci, Luca A Lotta, Flora Peyvandi [167] stated that there were seven GWAS made for MI-CAD relation up to that date. There were 13 loci found as a result of these studies and founded loci are also associated with some diseases such as Hypercholesterolemia, Abdominal Aortic Aneurism and Intracranial Aneurism.

A mental disease Schizophrenia is focused in the work of Peilin Jia [168]. They construct a network based pipeline structured with candidate module search, module assessment and module selection layers. They tested their model on three Schizophrenia datasets and identified 205 genes associated with the disease.

Type-2 Diabetes is another important complex disease. This disease is closely related to other complex diseases such as hypertension and cardio vascular diseases. A survey like work of Eleanor Wheeler and Ines Barroso made in 2011 states that there are 44 variant identified, new susceptibility loci are continued to be identified for therapeutic applications and further researches must be done in order to find the exact genetic variants [169].

1.2.5 GWAS and Oncological Diseases

Oncological diseases are the formal definition of the disease cancer. It is also a complex disease because it is affected by environmental and genetic factors. The reason they are given as a separate subsection in this thesis is that one of our aim is to identify underlying SNPs for different types of cancer. In this section a literature survey of GWAS on different types of cancer will be given. The mostly focused cancer types are:

- Prostate
- Melanoma
- Breast
- Lung
- Ovarian
- Pancreas

GLOBOCAN statistics state that prostate cancer is "*the second most frequently diagnosed cancer of men, the fifth most common cancer and sixth leading cause of death from cancer in men*"[13].

S. Tao and friends [170] conducted a research in order to find new SNPs additional to thirty known SNPs related with prostate cancer. They worked on a dataset consists of Swedish, Johns Hopkins Hospital and Cancer Genetics Markers of Susceptibility populations with a total of 4723 cases and 4792 controls. They identified six SNPs related with the disease.

Single population based studies also done for prostate cancer. One of the studies is the work of C. A. Haiman [171]. They worked 1,047,986 SNPs over African population with 3425 cases and 3290 controls. They identified seventeen new relations and one strongly related risk variant on chromosome 17q21. Related alleles found in this study are rare in other populations so they suggest further researches to be done.

[13] http://globocan.iarc.fr/factsheets/cancers/prostate.asp accessed on 01.10.2012

Another single population work is done over Japanese population by R. Takata [172] and friends. They made GWAS and a replication study on 4854 cases and 8801 controls and found five new loci for the disease. They also identified nine related SNPs out of thirty one SNPs that are known to be related with prostate cancer in European population.

L. M. FitzGerald [173] and friends analyzed aggressive prostate cancer in order to find biomarkers that could be used for early detection of the disease. They used a dataset of 202 cases ad 100 controls with 387,384 SNPs examined. They found two SNPs associated with the diseases which are rs3774315, very strongly related rs6497287.

The difference between aggressive and local prostate cancer is researched both in clinical and genetic aspects. The work of K. L. Penny [174] and friends conduct a search with this aim and compare the subjects of prostate cancer survivors and lethal cases. They searched 500,000 SNPs over 196 cases and 368 controls. They found fourteen copy number variants and three SNPs related with the aggressive disease.

J. Ciampa [175] conducted a two phase research for prostate cancer genetics. In first stage they worked on 523,841 SNP over 1175 cases and 1100 controls. In second phase they worked on 27,383 SNPs over 3941 cases and 3964 controls. Empirical Bayes and Logistic Regression methods are used. Although they found some related SNPs they could not reach genome wide significance levels.

Melanoma is caused by cancerous cells of melanocytes which give color to skin [176]. It is the top skin cancer type that causes death [177]. There are many GWAS made on melanoma in the literature. Some of them are summarized below.

A work by X. Wei and friends made whole exome sequencing over fourteen DNA samples [178]. The identified 68 related genes. By this way they were able to reveal alteration map for melanoma.

In order to find methylation markers in melanoma, Y. Koga [179] conducted a hybrid analysis combining gene expression and promoter methylation. They identified 68 hyper and 8 hypo-methylation markers.

S. MacGergor [180] conducted a comprehensive research on melanoma. They used 2168 cases in Australian population. The susceptibility variants they found are tested on three datasets that consists of European and United States populations. Total number of cases was 5193 and controls were 15,144. They identified three loci and SNPs over them related with melanoma.

Work of H. Nan [181] made a GWAS on European population dataset with 9136 cases and 3581 controls. They identified a gene related to disease and a SNP associated with low risk of the given disease.

Three new susceptibility loci were found by J. H. Barret [182]. Their research conducted on European population with 2981 cases and 1982 controls. They also used another dataset consists of British and French population of 6426 subjects as controls. They worked on 610,000 SNPs and found seven new regions. For replication studies they used two data sets which were Australia-United States and United Kingdom-Netherlands populations. As a result three new loci are found.

Nils Schoof and friends suspected that pathways for sunburn immunosuppression could be related to melanoma [183]. So they worked on a dataset with 1539 cases, 3917 controls and 43 genes with 1113 SNPs on which are them. In their first phase they found a relation between melanoma and immunosuppression. But they used different datasets which did not give the same results. Although they establish an evidence for the relation, further work must be done in the field.

Bladder cancer takes place in urinary bladder and caused by abnormal division of epithelial cells. Rothman [184] conducted a multistage GWAS on bladder cancer. In first phase they used 3532 cases, 5120 controls of European population and in second phase, which is a replication study, they used 8382 cases and 48,275 controls. They identified three new chromosomes associated with bladder cancer.

The same European population of 3532 cases and 5210 controls is used in another study to identify the pathways of bladder cancer. This study was conducted by I. Menashe [185]. They accessed public databases to use the known exact and candidate pathways in their study. With this approach they used 1421 pathways, 5647 genes and 90,000 SNPs. As a result they achieved to find 18 pathways related with bladder cancer.

The work of W. Obara [186] is a good example how GWAS can be used for the purposes rather than identifying the pathways or SNPs. They used the results of GWAS for therapeutic applications. They developed cancer peptide vaccine for bladder cancer. Nowadays caner vaccination is known as another therapy method like chemotherapy or radiotherapy after surgery [171].

According to World Health Organization (WHO) breast cancer is the most common cancer in worldwide women population by 16% [14].

An early GWAS on Chinese women population on breast cancer identified a SNP, rs2046210, strongly associated with the disease [187]. But in studies over European population like Stacey SN's [188] could not find a strong relation with that SNP. So another work is conducted by Rebecca Hein [189] to find the SNPs associated with breast cancer in Asian and European populations. In their work they studied 61,689 cases and 58,822 controls from both populations and SNPs with id's rs2046210 and rs12662670 are found to be strongly related with the disease.

To widen the studies above and to find the genetic risk factors of breast cancer in all ethnic and racial groups Fang Chen and friends [190] conducted study over European, African American, Native Hawaiian, Japanese and Latino populations. They worked over 2224 cases, 2827 controls and identified summary and overall risks based on alleles but they could not get significant results for African American population. Although their risk analysis could be used for preventive medicine as they stated a much wider study must be carried on over all ethnic and racial groups.

According to data gathered in GLOBOCAN project the number of worldwide deaths caused by lung cancer is 1.38 million [191].

[14] Taken from WHO web site about breast cancer.
http://www.who.int/cancer/detection/breastcancer/en/index1.html accessed on 21.10.2012

In addition to well-known three genomic regions, T. Rafnar [192] found seven additional variants related with lung cancer. They worked on a dataset with 1447 cases and 36.256 controls. They found rs748404 as a most significant SNP and checked this result on another dataset consists of multiple populations with 1299 cases and 4102 controls. As a result this variant also found to be related in the second dataset.

Work of E. Y. Bae [193] and friends tested the chromosomal regions earlier found for European population on Korean population. They found two of the regions are also related in the Korean population. They also found that these two regions are associated with adenocarcinoma and squamous cell carcinoma.

Another single population based research is the work of Z. Hu [194]. They worked on Chinese population with a dataset of 2331 cases, 3077 controls and identified two new susceptibility loci. The results were tested over a second dataset of 6313 cases and 6409 controls.

Although lung cancer is mostly seen in smokers, nonsmokers can suffer from this disease. A work by Y.Li [195] investigates the genetic variations related with lung cancer among nonsmokers. They conducted a four phase research. In the first phase they worked on 754 cases, 377 controls and identified 44 SNPs. In second phase they worked on a dataset with 735 cases and 253 controls to correlate the results of first phase and find the top SNP. Then this SNP is tested on the third phase over 530 subjects. In the last phase they searched for gene expression differences in order to establish the relations between SNPs and lung cancer.

Ovarian cancer is the seventh most common cancer type among women and third most common in gynecological diseases; in 2008 4.2% of deaths among women are caused by ovarian cancer [196].

A GWAS is conducted by E.L. Goode [197] on dataset of 1768 cases and 2354 controls. A total of 507,054 SNPs are examined. Then a follow up study is done with 4162 cases, 4810 controls and 21,955 SNPs. As a result there are two loci found very strongly associated with ovarian cancer.

Most widely observed ovarian cancer type is the epithelial ovarian cancer [198]. A study by P. Raska [199] worked on European American population. They used Principal Component analysis (PCA) collaborated by GWAS and identified three important variant regions.

There can be non-epithelial ovarian cancers which are most likely caused by egg cells and fallopian tubes [200]. One of the papers that worked on the non-epithelial ovarian cancer in the literature is the work of K. V. Kee [201] and friends. Because non-epithelial cases are so low they could only work on eight cases and seven controls. They found 804 genes with 443 of them are over expressed and the rest is under expressed in cancer.

Another study focused on the common genes in different types of cancer. With this aim H. Song [202] searched the common genes effecting the breast and ovarian cancers. Eleven SNPs related with breast cancer is tested over ovarian dataset of 2927 cases and 4143 controls. As a result three SNPs are found to be weekly related and one SNP is very strongly related with the diseases in non-Hispanic population.

Pancreatic cancer is another cancer type studied by GWAS. It forms the 2% of the total cancer in worldwide [203].

A work by D. Li [204] searched the known SNPs and tries to identify the biological pathways for pancreatic cancer disease. They gather data from different case control and cohort studies to work on a dataset with 3851 cases and 3934 controls. As a result of their work, they identified 23 pathways related with the disease.

Because pancreatic cancer has a high rank among most observed cancers in Japan S. K. Low and friends [205] conducted a research of pancreatic cancer on Japanese population. They worked on 3851 cases and 3934. As a result they identified three loci and their respective chromosomes related with pancreatic cancer.

C. Rizzato [206] made a genetic survival analysis on pancreatic cancer. They used a dataset of 690 cases and 1277 controls. They worked on 15 SNPs of this population and identify one strong SNP to be strongly related with the prognosis of the disease but they offer further replication studies to be done.

F.J. Couch [207] studied the common genetic variants of breast cancer and pancreatic cancer. They worked on Caucasian ethnicity dataset with 1143 cases and 1097 controls. They detected two SNPs which are strongly associated with the disease when the subjects are smokers or heavy smokers. They stated that there are many common SNPs associated with breast and pancreatic cancer diseases.

1.3 Data Mining Methodologies on GWAS and SNP Selection

In this section of the thesis single and hybrid data mining approaches of SNP selection in complex and oncological diseases will be given. According to our literature search most widely used methods are Artificial Neural Networks, Decision Trees and Support Vector Machines. In the following, recent and important works are given and grouped according to the methods used.

1.3.1 Artificial Neural Networks

Y. Tomita and friends used ANN structure for determining the SNPs for Childhood Allergic Asthma (CAA) [28]. They worked on a dataset consists of 344 Japanese subjects and focused on 17 genes that have a total number of 25 SNPs. As ANN structure, they used one hidden layer, 25 input nodes for 25 SNPs and one output node to identify the disease condition. The reason behind using ANN was the high accuracy rate of their ANN models from their previous studies [22]. They got good results with ANN especially when the number of SNPs is very small in number. To find SNPs related with CAA, they used Parameter Decreasing Method (PDM). In this method they excluded one of the inputs in each step and then construct ANN structure. Based on cross validation, minimum error rate and learning time they tried to find the best model and input combinations. For different SNP combinations chi-square test is used. As a result they found 21 SNPs related with disease according to their high P values. They interpreted the SNP with lowest P value as indicator of cases and controls with 54.4% of accuracy, 12.8% of sensitivity an 95.5% of specificity. The result of ANN is compared to Logistic Regression (LR) based on accuracy, sensitivity and specificity both in learning and evaluation phases. In both steps ANN performed better than LR.As a result of their study they were able to identify ten SNPs that are closely related with CAA.

In previous sections we mentioned that SNPs and GWAS are also used in personalized medicine. One of the application areas is the drug response in therapy. N-acetylation enzyme in liver is responsible for metabolism of some drugs. If this enzyme is slow then given drug stays in body for a long time and well absorbed. But if it is fast then the drug is thrown away from body. So this enzyme is very important when designing a drug based therapy.

Work of A. Sabbagh and P. Darlu [23] searched the SNPs do detect acetylation phenotype in different ethnic groups. They used a dataset of 258 Spanish, 137 Nicaraguans, 1000 Koreans, 101 Africans, 564 Germans, 284 Polish, 303 Turkish and 50 Mali. One of the learning models they used was the ANN. They constructed a 2 hidden layer ANN structure by using NNPERM tool. For SNP relations with acetylation condition, statistical tests such as T statistics, Unpaired T test and Permutation Test was used. As a result they were able to identify seven common SNPs related within all populations.

S. Tomida and friends used ANN for SNP selection to predict allergic diseases such as Atopic Dermatitis (AD), Allergic Conjunctivitis (AC), Allergic Rhinitis (AR) and Bronchial Asthma (BA) [24]. They used SNPs from six genetic regions of the Japanese population with 82 cases. They designed a special numeric coding for SNPs in order to use them as numeric inputs in ANN. Another data preprocessing was the elimination of some data tuples. They found that two subjects with the same SNP pattern had different diagnosis. Such data were eliminated. After preprocessing they constructed four different ANNs for each of the diseases. Each ANN had different number of hidden layers and ANN with the highest performance according to its accuracy rate is chosen as the model for the given disease.
Each selected ANN is then compared to Multiple Regression Analysis (MRA). As a result for the given diseases ANN outperformed the MRA.

There can be many ANN approaches used for GWAS and SNP selection can be found in literature [25-26-27]. A. Motsinger explains the reason behind choosing this method over others in three matters [208]:

- Handle high volume-high dimensional data
- High performance in non-linear problems
- Model free structure

Although ANN has the advantages mentioned above, its high performance is based on parameters such as its structure (the number of hidden layers and nodes to be used in different layers), convergence and error criteria, learning rate and number of training epochs. So in order to get best results from ANN one should optimize these parameters. This can be done by using grid search or trial and error approaches but they require high computational power and long time.

A new approach in literature for optimizing ANN is to use Genetic Evaluation and Genetic Programming techniques and to form a hybrid system. The basics of such approach are given in the different works of A. Motsinger [45-209], M. D. Ritchie [45-208], P. Yang and Z. Zhang [210].

The purposed approach, Genetic Evaluation Neural Network (GENN), works as follows:

1- ANN parameters form the population and initial solution pool
2- As in genetic algorithms
    a. Parameter combinations for ANN is selected
    b. ANN is constructed according to a
    c. Performance of ANN is tested and ranked
    By this way step 2 populates the solution pool.
3- Step 2 is repeated until GA finds the best performing ANN among the solution pool.

A very well-known application of GENN is the Athena, developed for detecting gene-gene interactions by Stephen D Turner, Scott M Dudek, and Marylyn D Ritchie [211].

1.3.2 Decision Trees

L. Fiaschi, J. M. Garibaldi and N. Krasnogor conducted a research in 2009 to compare the performance of different decision tree algorithms on SNP selection of pre-eclampsia disease [14]. Pre-eclampsia is the high blood pressure observed in pregnancy followed by high amounts of protein in urine [15].The dataset they used consists of 4529 subjects which are pre-eclampsia, eclampsia, other hypertensive patients and health ones. A total of 52 SNPs are researched. There was also some clinical information included. Most important clinical attribute was CBC value which indicates the corrected value of weight based on baby's sex, ethnicity and mother's weight, height and number of pregnancies. In pre-processing step they made data cleaning by removing redundant subjects. After that they eliminated the rows with high missing value rate. At last step they tried to balance the dataset by random selection strategy. They compared the performance of ADTree, ID3 and C4.5. ADTree and ID3 works with categorical variables so some continuous valued clinical attributes needed to be converted to categorical data. To do that, they used both clinical and statistical information to obtain the appropriate thresholds to divide the data in to category folds. According to statistical approach, Kappa Test, three significant thresholds are found for CBC. For this reason three datasets are formed for these thresholds. C4.5 and ADTree performed similar results where ID3 could not give any significant results. In the second phase of study a new dataset is formed according to the common attributes chosen by ADTree and C4.5. Then these two algorithms ran on the new dataset. Associations are found among the sex of the baby, birth week, CBC and particular SNPs. But they sated that there is no significant association found between the given diseases and SNPs.

Another performance comparison study of decision trees is made by S. Uhmn [16]. They compare the performance of different machine learning methods including C4.5 decision tree to detect SNP chronic hepatitis relation. The dataset they worked on consisted of 194 subjects with 28 SNPs. They stated that they've chosen decision tree approach for its known advantages such as easy construction, easy interpretation and relatively a few parameters to optimize. They used C4.5 tree and for SNP future selection different selection algorithms are used such as Incremental Information Gain, Forward Selection and Backtracking, Rule Based Future Selection. Accuracy, sensitivity and specificity of machine learning methods, SVM-KNN-C4.5, are compared and the best performance result are gathered from C4.5 with backward elimination.

Different decision tree techniques such as Decision Stamp, Alternating decision tree (ADT), Multiclass ADT and Logistic decision tree are compared in the work of Y. Jiao and friends [17]. As a control mechanism performance of SVM is also compared with decision trees. Performances of the given techniques are compared on Autism Spectrum Disorder (ASD).

The dataset consisted of 36 children of whom 18 of them have ASD, 13 of them with high functioning autism and five of them with Asperger syndrome. Eight genes with 25 SNPs over them are studied. In addition to genotype information thickness and volume information of brain cross sections gathered from MR are included. By this way decision performance of SNPs, volume information and thickness information could also be compared individually. Performances are compared by accuracy; sensitivity and specificity are given in the Table-4 below which is adopted from their study.

Table 4: Performance comparison of decision stamp, decision trees and SVM. (Adapted from [17])

	Decision Stamp	**Alternating Decision Tree**	**Multiclass Alternating Decision Tree**	**Logistic Decision Tree**	**SVM**
Accuracy	90	89	90	81	81
Sensitivity	0.95	0.94	0.95	0.93	0.96
Specificity	0.75	0.73	0.72	0.54	0.35

The table above shows that the decision tree performances are very good when detecting ASD based on SNPs. The authors state that the low performance of SVM could be related to type of the data variables which were categorical. All decision tree methods found significant relation between the disease and particular SNPs. Two important outcomes of this study were all trees found the same SNP for root node and they had a good on separating ASD and high functional autism.

Performance of four decision trees, C4.5-PART-ID3-PRISM, were compared by J. T. Horng and friends in 2004 [18]. They tried to find genetic factors that underlie in cervical cancer disease. The dataset used were consisted of Asian population with 224 healthy subjects, 238 cervical cancer (SCC) patients and 106 patients with low degree (LSIL) and 152 patients with high degree (HSIL of another cervical disease. From these patients they constructed three different datasets which were SNP dataset, micro satellite dataset and hybrid dataset. Among all datasets PART outperformed the other tree structures. This performance comparison is given in the Table-5 below, adopted from their work.

Table 5: Performance comparison of decision trees. Comparison is based on precision, recall and specificity where D1: Microsatellite. D2: SNP. D3: Hybrid datasets. (Adapted from [18])

Algorithm	**Precision (DS1-DS2-DS3)**	**Recall (DS1-DS2-DS3)**	**Specificity (DS1-DS2-DS3)**
PART	63.75 - 53.13 – 59.73	65.39 – 56.57 – 60.50	82.05 – 76.17 – 79.55
J48	60.14 – 56.76 – 58.40	60.88 – 60.27 – 59.66	80.20 – 77.66 – 78.53
ID3	52.80 – 55.84 – 53.34	53.24 – 55.43 – 53.19	74.01 – 73.35 – 73.76
PRISM	53.33 – 54.65 – 56.18	53.61 – 53.49 – 53.81	74.58 – 76.48 – 77.89

As a result they were able to identify five important SNP markers for cervical cancer disease.

Work of M. P. Rocha [212] used C4.5 trees to identify the relations between SNPs and breast cancer. They used a dataset of 94 cases and 164 controls with 32 SNPs. They used WEKA tool for C4.5 application. To find the best performing tree they used a set of different parameters such as confidence factor, pruning criteria and binary splitting in a tenfold cross validation. In each validation the average accuracy of the C4.5 model is calculated and the model with highest accuracy is used as a best C4.5 model. After the best tree ran on the dataset, they found significant relations between breast cancer and the given SNPs. Although the relations are found to be significant they stated that the research must be conducted on a larger dataset.

K. Miyaki and friends conducted a research on stroke risk and SNP relation by using exact tree method [19]. They worked on 68 cases and 189 controls with 14S SNPs from Japanese population. They made crosstab genotype analysis. Prevalence of alleles are calculated by Fisher's Exact Test. P values obtained are used for splitting criteria of decision trees. The number of trees to be used was determined by two step process. In step one best splitting point are found according to given P value threshold. In second step, talus plot of P values are drawn. In this approach the variables corresponding to sharp slope changes of the plot were chosen. By this method different candidate trees were constructed and the best performing one is chosen. As a result of this study, statistically significant differences are found in some alleles of cases and controls.

FlexTree is a different application of classification and regression trees (CART). Its performance is tested on Chinese women population with hypertension disease in the work of J. Huang et al [20]. Their dataset consisted of 307 cases and 206 controls. In addition to genotype data, some phenotype data such as insulin resistance and menopausal status are included. FlexTree used in the research finds its root node as in conventional tree structures. The difference comes from branching strategy. For each branch a linear combination of variables is used based on a scoring mechanism of regression. This approach is stated to be robust and could deal with overfitting. As a result FlexTree had 65% of sensitivity, 54% of specificity in diagnosing the hypertension among Chinese women. It was also able to identify six SNPs related with the disease.

Decision trees can be combined with other learning methods to form a hybrid structure in order to improve its performance. As in ANNs the most widely used method is the combining the trees with genetic programming and/or grammatical evaluation.

Genetic programming uses the ideas behind genetic algorithms. The population base consists of solutions; new generations of solutions are obtained by reproduction, recombination and mutations. Performance of each solution is measured by a fitness function and parents are selected according to their rankings on fitness function. This idea is used over decision trees by J.K. Estreda [213]. They designed a genetic programming decision tree induction algorithm for epistatic interactions. The basic approach explained above is adapted to decision trees where each solution tree is an individual in population pool. New trees are created by mutation and crossover operations, based on fitness function and tournament selection. They used classification error as fitness function and set the probabilities of reproduction and recombination to ten and ninety respectively. Five different epistasis sets were selected from literature and performance comparison is done between Genetic Programming Decision Tree (GPDT) and Genetic Programming Neural Network (GPNN). The results showed that GPDT had a better performance of prediction and classification error over GPNN on most of the problem sets.

Grammatical Evolution (GE), like genetic programming (GP), is a subtype of genetic algorithms. The main difference between GE and GP is that GP uses tree structure for evolution process where GE uses binary strings and pre-defined grammar rules.

In a biological point of view DNA (genotype) is represented by binary strings, codons are represented by specific sequence of eight bits and with transcription process binary strings are converted to integer strings (phenotypes) [46]. Then these phenotypes are translated into decision [213]. Fitness function that determines the parents of next generations is calculated by using specificity and sensitivity of decision tree build [213].

This idea, grammatical evolution decision tree (GEDT), is depicted in the Figure-28 below.

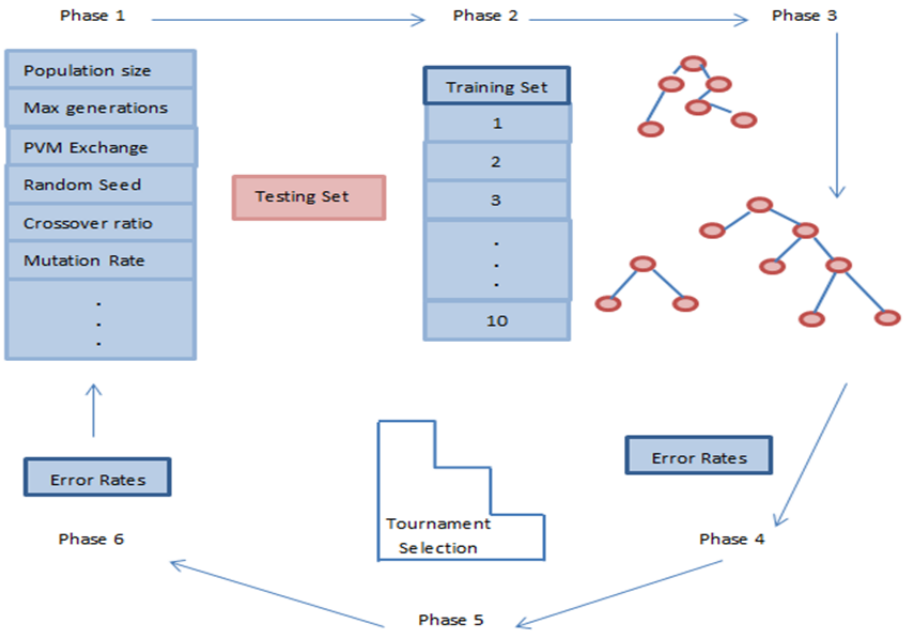


Figure 28: Process of forming a Grammatical Evolution Tree (Adapted from [213])

To evaluate GEDT, S. Deodhar and A. Motsinger generated synthetic data of 250 cases and 250 [213]. In this data set only two SNPs out of hundred were related. They generated hundred datasets with the specifications above. Each of these datasets is modeled according to XOR and ZZ problems from related literature. All models had three different heritability percentages. This made a total of 600 tests conducted. With different models used GEDT had the power ranging between 70% and 86%. Although this was a good performance, authors suggest further investigation of GEDT.

1.3.3 Support Vector Machines

An early study is conducted by M. P. S. Brown in 1997 [41]. They used SVM to identify the functional and nonfunctional genes. According to their work SVM could be chosen over other learning mechanisms because of it can handle vast amount of data and perform well on nonlinear problems. Also it has the advantage of choosing different types of distance functions when compared to other methods based on distance learning. They've used SVM for classification and then for outlier detection in the training dataset.

Gene datasets from different sources are used for training and testing SVMs with Radial Basis Function (RBF) and linear kernels. These SVM models were also compared to different learning methods such as Fisher's Linear Discriminant, Parzen Windows and r-two decision trees C4.5 and MOC1. The performance criterion was cost factor based on true positive, true negative, false positive and false negative rates. In all datasets used SVM outperformed other techniques. There was a small performance difference between RBF and linear kernel SVMs.

N.Zhou and L. Wang used SVM for population classification [32]. Their aim was to find ethnicity related SNPs and by this way to identify populations. They used four populations from HapMap database which were European, Japanese, Chinese and Yoruba. They included all twenty three chromosomes for individuals which make nearly four million SNPs to be searched. To find representative SNP subsets they have chosen top hundred SNPs of each chromosome. This ranking was based on statistical methods such as modified T Test and F-Statistics. A second type of ranking was also made to find the top SNPs among 2300. SVM, Matlab toolbox is used, was used to find the classification accuracy in there and four populations. The classification accuracy changed between 69%-94% in three population classification and 55%-74% in four population classification.

Z. Wei stated that the reason behind some poor performance of personalized disease prevention based on susceptibility loci are the insufficient number of loci analyzed [33]. They suggest using large amount of genotype data and difficulties of conventional analysis methods rising from dimensionality of data could be solved by machine learning techniques. They used Affymatrix platform of type-1 diabetes. They evaluated their model by using another Affymatrix dataset with 1529 cases and 1458 controls and an Illumina dataset of 1008 cases and 1000 controls. The number of SNPs in these datasets was reduced by using different thresholds of P values selected. The reason that they've chosen SVM is that based on their literature search SVM showed high performance on such studies. They compared linear and RBF kernels by using Affymatrix dataset and threshold for SNP selection is done according to $P<10^{-5}$. Area under ROC curve (AUC) was used for comparison and SVM with (RBF) outperformed the linear kernel SVM. The same results were gathered on other datasets too. Then SVM with RBF kernel was compared to Logistic Regression (LR). Main performance criteria they used was AUC. SVM outperformed the LR with AUC ranging between 0.86 – 0.89 and sensitivity ranging between 0.85 – 0.88.

H. Y. Ban and friends used SVM approach to select related SNPs with type-2 diabetes [34]. They used a dataset of 452 cases and 456 controls. Each individual in this set had 408 SNPs on 87 genes. Like Z. Wei's work described above, Ban have also chosen SVM for its good reputation in computational biology literature. They used SVM with RBF kernel. Their model had 65.3% of prediction performance and identified 14SNPs on 12 genes. They have conducted sub researches on datasets with p-value filtered SNPs and sex attribute separated. When the dataset is separated according to sex attribute, the performance of SVM model increased to 71%. In order to improve the performance the dataset is filtered by p-value and top SNPs above threshold were selected. Then SVM is applied to filtered dataset but the performance decreased to 57.6%. Authors state that it can be caused by insufficient number of SNPs selected by filtering. As a result of their work, they suggest to use larger sets with high number of SNPs especially when the p-value filtering to be used. They also believe the SVM performance could be increased by including environmental factors that affect genetics into datasets.

J. Listgarten and friends conducted a research for detecting SNPs that underlie breast cancer [35]. 98 SNPs on 45 genes examined in a dataset of 174 cases and 158 controls. To analyzes this dataset three different machine learning methods are used such as SVM, Decision Tree and Naïve Bayes. Each model ran on datasets with varying number of SNPs. The SNPs in actual dataset are filtered by using statistical methods. In addition to genotype information some clinical information such as images, histology and patient interviews were included in the study. As in the most studies explained in this subsection, different kernel methods were used for detecting best performance SVM model.

Models are applied by Matlab and SVMLight software tools. For more accurate result they used 20 fold cross validation. Results of their study are given in the Table-6 below:

Table 6: Performance comparison of Naive Bayes, Decision Tree and SVMs. Different kernels were used for SVM and comparison is based on accuracy, sensitivity , specificity and number of SNPs used. (Adapted from [35])

	Accuracy (%)	**Sensitivity (%)**	**Specificity (%)**	**# of SNPs**
Naïve Bayes	67	54	79	3
Decision Tree	68	64	70	2
Linear Kernel SVM	62	57	57	60
Quadric Kernel SVM	69	53	83	3
Cubic Kernel SVM	67	47	84	3

Authors stated that the best performing model was SVM with Quadric kernel. But Naïve Bayes also gave similar performance and because of this model is much simpler it can be preferred over SVM.

Another performance comparison of SVM with Naïve Bayes and Decision Tree (C4.5) was made by L. C. Huang, S. Y. Hsu and E. Lin on classification of Chronic Fatigue Syndrome (CFS) [36]. 109 subjects with 55 cases and 54 controls formed the dataset and each individual had 42 SNPs. In the first phase of their study 1% of missing values were replaced by class modes then performances of three models were compared. In the second phase a hybrid feature selection algorithm which combined Information Gain and chi-Square was used to reduce the number of SNPs. All methods were applied by WEKA tool.
For SVM linear, polynomial, sigmoid and Gaussian RBF were used. The parameters of these kernels were set to constant values. Performances were compared by means of AUC, sensitivity and specificity. This comparison is given in the Table-7 below:

Table 7: Performance comparison of different machine learning techniques. Based on sensitivity, specificity, number of SNPS used and area under ROC curve. (Adapted from [36])

Performances without feature selection				
Algorithm	AUC	Sensitivity	Specificity	# SNPs
Naïve Bayes	0.60	0.64	0.52	42
Linear SVM	0.55	0.55	0.56	42
Polynomial SVM	0.59	0.46	0.71	42
Sigmoid SVM	0.61	0.62	0.61	42
RBF SVM	0.62	0.60	0.64	42
C4.5	0.50	0.52	0.48	11
Performances with feature selection				
Algorithm	AUC	Sensitivity	Specificity	# SNPs
Naïve Bayes	0.70	0.64	0.63	8
Linear SVM	0.63	0.71	0.55	9
Polynomial SVM	0.63	0.43	0.82	12
Sigmoid SVM	0.64	0.59	0.70	6
RBF SVM	0.63	0.60	0.66	7
C4.5	0.59	0.65	0.55	6

Although authors stated that Naïve Bayes outperformed other learning methods based on AUC, there were no such big difference between SVM and Naïve Bayes and some SVM models were better than Naïve Bayes in terms of sensitivity and specificity based on the table.

T. Abeel and friends worked on identification of robust biomarkers for cancer detection [37]. They've worked on four different cancer datasets of which are Leukemia, Colon, Lymphoma and Prostate. Each dataset had the following number of genes, 50, 2000, 4026 and 400. Although the number of genes was high the number of subjects in these datasets was very low. They used a linear SVM for future selection process. They designed a back propagation elimination algorithm based on the weights of vector features. The algorithm starts with initial dataset and genetic features were eliminated according to their weight scores that were determined by SVM. This continued until a desired number of features were reached. One of the important features of SVM is C that determines how well the model will learn. In their research although different C values are tested by cross-validation a stable value of C is set to one. Authors stated that the performance of their model was significantly promising.

S. Uhmn and friends conducted a research on identification of chronic hepatitis [15]. In their work they compared performance of different machine learning methods such as decision trees and SVMs. The data they worked on consisted of 194 subjects with 28 SNPs. These SNPs were selected from the genes that are known to be related with chronic hepatitis. Although the number of SNPs was low in number they applied four feature selection algorithms. The reason behind this elimination is when the SNP set is too large selection of related SNPs becomes a NP-Hard problem [15]. Algorithms used were Incremental Information Gain (IIG), Forward Selection and Backtracking (FSBT), Backward Elimination with Backtracking (BEBT) and Rule Based Feature Selection (RBFS). They applied SVM by using SVMLight software tool. They've chosen linear kernel and constructed SVM model with leave one out cross validation.A numeric to categorical value conversion applied because SVMs work on numeric values. The performance comparison of SVM and decision tree, with given feature selection methods, is given in the Table-8 below.

Table 8: Performance comparison of SVM and decision tree. Different types of feature selection algorithms were used for each method. (Adapted from [15])

Method	Feature Selection	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM	IIG	67.53	74.23	60.82
	FSBT	67.53	74.23	60.82
	BEBT	65.98	65.98	65.98
	RBFS	48.97	35.05	62.89
Decision Tree	IIG	68.56	63.92	73.20
	FSBT	74.13	68.04	74.23
	BEBT	72.68	65.98	79.38
	RBFS	70.10	70.10	69.07

According to the table decision tree with different feature selection algorithms performed better than SVM with linear kernel.

SVM was applied by M. Waddell, D. Page and F. Zhan [38] for early detection of risk in multiple myeloma under age forty. They worked on a dataset with forty cases and the same number of controls and 3000 SNPs. In order to prevent bias, SNPs were not selected from previous studies as candidate susceptible SNPs. Rather they were chosen to represent the general human genome. SVM was chosen because of its high power in determining the relevancies and eliminating the redundancies. Before constructing the model, a numeric conversion of SNP data, similar to S. Uhmn's hepatitis [15] work, was done. Authors stated that SVM could handle large datasets but feature selection could improve their performance. For this reason an entropy based feature selection is done on the dataset. SNPs were ranked according to their Information Gain and then top 10% of them selected. By this way a subset of 300 SNPs was used in the analysis. SVM model was constructed by SVMLight and linear kernel function was preferred. Their model had a performance of 71% accuracy, 65% sensitivity and 77% specificity.

Another disease prevention study was conducted by L. Y. Chuang and friends [39]. They worked over oral cancer patients with 238 subjects consisted of cases and controls. The machine learning algorithm they used was SVM. As most of the SVM studies [15-35] given in this subsection, a numerical to categorical conversion of SNP values are done based on heterozygous and homozygous alleles. SVM model was constructed by using WEKA software tool. RBF kernel and default parameters of WEKA were used with two different cross validation techniques which were 10-fold and holdout. By this way they were able to make a performance comparison of these two cross validation methods. The results were evaluated by means of accuracy, sensitivity and specificity.SVM with holdout cross validation performed better than 10-fold cross validation. The best performance in holdout set had the performance indicators of 64.2% accuracy and 95% specificity but a lower value of sensitivity which were 23%. The average performance of the SVM was 55.4% of accuracy, 65.2% specificity and 14.2% sensitivity.

Cancer Somatic Variation (CSM), Mendelian Variant (SVD) and neutral polymorphism (SVP) were researched by using non synonymous SNPs and machine learning algorithms such as Random forests and SVM by the work of M. Wang and friends [40]. The dataset they used were consisted of 280 SNPs. They conducted two phase feature selection by using Random forest (RF) and SVM. In first phase RF was used to eliminate features and to construct a candidate feature set. In second phase SVM with fivefold cross validation was used. In SVM feature selection, candidate set from RF was used in initial step then in following steps each feature was excluded from dataset and performance of SVM was calculated. The candidate set with max performance was chosen as the final dataset to be used with SVM model. By this way 18 of 280 features were selected for final dataset. SVM model was applied by using a software tool LibSVM. Radial basis function (RBF) is used as kernel and a grid search was conducted to optimize RBF parameters. Four different SVM models were used to classify the given conditions. Three class SVM is used to classify all three classes CSM, SVD and SVP. But in fact this classification was converted to binary classification instead of multi classification. This was done by assigning CSM to positive class and other two remaining ones to negative class. For each of these classes, CSM-SVD-SVP, an individual SVM model was constructed. Their performances were based on accuracy, sensitivity and specificity. These indicators and values are given in the Table-9.

Table 9: Performance of SVM on different subsets. Based on sensitivity, accuracy and specificity. (Adapted from [40])

SVM	Sensitivity	Specificity	Accuracy
3 Class	-	-	84.97
CSM	98.44	96.63	96.93
SVD	86.65	87.64	86.98
SVP	76.60	90.19	88.24

CHAPTER 2

MATERIAL and METHODS

2.1 Data

Our datasets used in this work are downloaded from NCBI's dbGaP database. This database contains the results of GWAS and phenotype information of many studies. In order to access these data one should have NIH eRA Commons account or accepted as principal investigator by NIH and NCBI.

The datasets are stored in a single folder. This folder contains two subfolders which are genotype and phenotype folders. All data in these folders are encrypted by dbGaP. In order to open these files a decryption code was provided by dbGaP. This code is used in command prompt to open the files. In phenotype folder, there are files about subjects, samples, pedigrees and all phenotypes of given cases and controls. In genotype folder there are different types of compressed files. These file types and contents are listed in the Table-10 below.

Table 10: dbGaP file types and their contents

File Type	Content
indfmt	Genotype information of individual subjects stores as *.ind* extension
matixfmt	Stores the *.bed .bim .fam* files to be used in Plink analysis
idat	Raw data of individuals coded
info	Stores general information about the data and manifest files

2.1.1 Prostate Cancer Data

First data set chosen was "Multi Ethnic Genome Wide Scan of Prostate Cancer" with dbGap Study accession number phs000306.v2.p1. As a part of GENEVA study, genotyping of the data was done by Broad Institute of MIT and Harvard. This data was collected from cohort and nested case control studies conducted on different ethnicities such as African Americans, Latinos and Japanese that live in California and Hawaii. This data set consists of a total of 9457 subjects which 4650 cases and 4795 controls. Genotype, phenotype and pedigree distribution among subjects is given in the Figure-29.

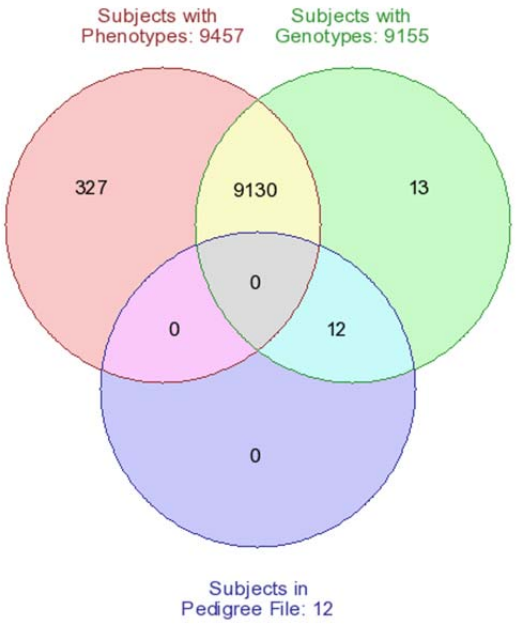


Figure 29: Prostate cancer data, distribution of phenotypes and genotypes among subjects. (Taken from http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000306.v2.p1&phv=124440&phd=3229&pha=&pht=1911&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1)

Each individual in the study has 600,000 SNPs represented by Rsid's. The phenotype information consists of twenty attributes that are listed in the Table-11 below:

Table 11: Phenotype variables of prostate cancer data

Attribute Name	Explanation
sex	Gender
Status	Case/Control status
age_cat	Age at entry into cohort
agedx_cat	Age at diagnosis for cancer cases
ageco_cat	Age at blood draw controls
bmi_cat	Body mass index
fh_prca	Family history of prostate cancer (brother or father)
pa_cat	Hours per day of moderate or vigorous physical activity
Packyrs_ca	Pack years of smoking cigarettes
ethanol_ca	Alcohol drinks per day
d_lyco_cat	Density of lycopene intake
p_fat_cat	Percentage of calories from fat
d_calc_cat	Density for calcium intake
currsmoke	Currently smoker?
eversmoke	Ever smoked?
severity	Aggressiveness of disease for cases

2.1.2 Melanoma Data

The second dataset used is the "High Density SNP Association Analysis of Melanoma: Case-Control and Outcomes Investigation" with dbGaP study accession number phs000187.v1.p1. The subjects of this study are gathered from UTMD Anderson Cancer Center through several years. As a part of GENEVA study, genotyping of the data was done by Johns Hopkins University Center for Inherited Disease Research (CIDR). This data was collected from case control studies conducted on European ancestry. This data set consists of a total of 3115 subjects which 2053 cases and 1062 controls. Genotype, phenotype and pedigree distribution among subjects is given in the Figure-30 below:

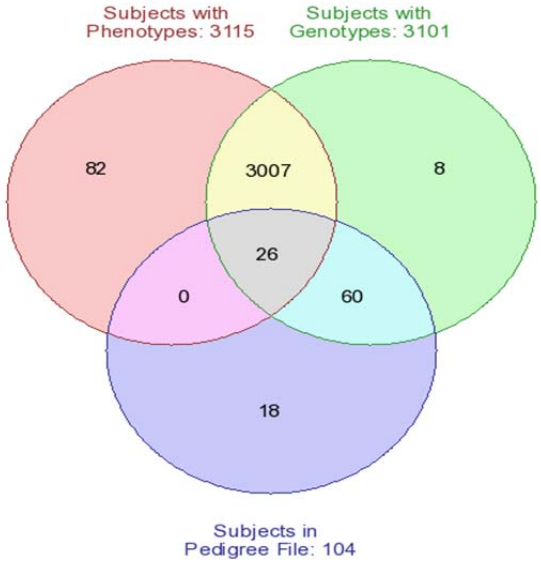


Figure 30: Melanoma data, distribution of phenotypes and genotypes among subjects. (Taken from http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000187.v1.p1&phv=73647&phd=2432&pha=&pht=814&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1)

Each individual in the study has 600,000 SNPs represented by Rsid's. The phenotype information consists of twenty four attributes that are listed in the Table-12 below:

Table 12: Phenotype variables of melanoma data

Attribute Name	Explanation
Geneva_ID	subject ıd
Gender	subject's gender
case/control	case-control status
Race	subject's race
moles	ıf moles are present
dysplastic_nevi	presence of dysplastic nevi
skin_color	Grading of skin color
eye_color	eye color
hair_color	hair color
sunburn	presence of severe blistering
freckles	freckling in sun
untanned	tan after exposure to sun for 30-40min

family_history	cancer history in 1st degree relatives
stage_cat	stage of disease
Age_at_DX	age at diagnosis melanoma
State	Patient's state of residence
Country	Patient's country
Control_Age_at_DX	patient's age as recorded
Breslow_tumor_thickness	Breslow thickness expressed in millimeters
Clark_Level	Clark level based on MDACC
Ulceration	presence of ulceration
Blood_Draw_Rate	date of patient's blood drow
Months_follow_up	the months of follow up
vital_status	Patient's vital status at time

Table-12 (Cont'd.)

2.2 Data Preprocessing

Our data preprocessing consists of three steps. In first step Plink analysis was conducted in order to find the statistical power of relations between genotype and given disease. At second step METU-SNP analysis was done to detect both statistically and biologically significant genotype attributes. In last step data matching, cleaning and transformation were done.

2.2.1 Plink Analysis

Plink is an open source GWAS tool developed by Shaun Purcell and the Broad Institute of Harvard & MIT. This tool has features such as Data Management, Summary Statistics, Population Stratification, Association Testing, Haplotype Testing and Meta-Analysis. We benefitted from its Association Testing feature. This feature calculates the allele frequencies between cases and controls by using statistical tests such as Fisher's Exact Test and T Test. In detail Association Testing feature includes Case/control association by using Standard allelic test, Fisher's exact test, Cochran-Armitage trend test, Mantel-Haenszel and Breslow-Day tests for stratified samples, Dominant/recessive and general models, Model comparison tests (e.g. general versus multiplicative). In addition to case-control test, different tests can be conducted such as: Family-based association (TDT, sibship tests), Quantitative traits (association and interaction), Association conditional on one or more SNPs, Asymptotic and empirical p-values, Flexible clustered permutation scheme, Analysis of genotype probability data and fractional allele counts (post-imputation).

In order to conduct Plink association tests, special file formats of the data are needed. These file formats are ***.bed***, ***.bim*** and ***.fam*** files. The ***.bed*** file is a binary data format and gives the genotype information to Plink. ***.bim*** file stores the information about Chromosome, Marker ID, Genetic distance, Physical position and Alleles. ***.fam*** file stores the information about Family ID, Sample ID, Paternal ID, Maternal ID, Sex and Affection status which determines and individual have the disease or not. All these files are found in the genotype folder of the dataset with a ***matrixfmt*** extension. In association analysis, affection status in the ***.fam*** file plays and important role. In our first analysis Plink retuned zero associated genotypes.
When we searched for the problem, we've seen that the ***.fam*** file in our dataset had all missing values for the affection status attribute.

In order to set the right values for this attribute, ***.fam*** file was matched by phenotype file according to Subject ID's and affection status are gathered from case-control attribute in phenotype file. After this process, Plink was successfully run and gave the desired results. The results are stored in a special file format which is ***.assoc.adjusted***. This file contains information about chromosome, SNP ID, code for alleles, the frequency of variant in cases and controls, Chi-squared statistics, asymptotic significance value and Odds Ratio. By this analysis, all genotype information is listed and sorted according the calculated P values.

2.2.2 METU SNP Analysis

METU-SNP [214] is a software tool that is designed for to conduct different types of analysis based on genotyping data. It offers configuration, preprocessing, GWAS, SNP prioritization, SNP Selection and Performance tests.

In our work we've used METU SNP's AHP (Analytical Hierarchical Process) feature. Basically AHP takes ***.assoc.adjusted*** file from Plink as input and creates and SNP prioritization list that is based on both biological and statistical significance. This is done by combining P value statistics, genomic locations, functional consequences, evolutionary conservation and gene disease relations.

After AHP score analysis, we determined a threshold for P values and eliminated the genotypes under this value. This is done for three reasons.

- To find the most related genome information with the given disease.
- To shorten SNP list and to work on a better representative dataset which is also done in similar works in literature [32-33-34-35-39-41] and such.
- To include both the statistically and biologically significant gene information

2.3 Data Matching-Cleaning-Transformation

As explained in Section 2.1 datasets downloaded from dbGaP consists of two main folders which are genotype and phenotype. In genotype folder ***.indfmt*** extension folders store the individual genotype information. In each ***.indfmt*** compressed dataset file there are a number of ***.indfmt*** files with unique ID. These files contain the whole genomic information of the individual subjects. The phenotype information of the subjects is stored in phenotype folder. Phenotype file contains information such as dbGaP ID, ID of the platform used in GWAS, case-control information and specific phenotype information of the given dataset such as age, sex, other demographic information, clinical results and lab test results.

In order to combine genotype and phenotype information of the individuals and merge all individual in a single data set, a Python script is written. This script matches the genotype information of individuals from ***.indfmt*** folder and phenotype information from phenotype folder based on subject IDs. In some cases subject IDs from genotype and phenotype folders do not directly match. In such cases a manifest data that is given by dbGaP is used to convert the IDs and make the appropriate data matching. After all data are matched and all phenotype and genotype information is combined, the script filters the genotype information by using the SNPS in METU-SNP's AHP Score list.

Python script outputs the combined, merged and filtered dataset in ***.csv*** format. This is because the ***.csv*** format is compatible for many decision support tools.

After matching process the output file is examined for missing values. These missing values were caused by ***.fam*** file problem explained in Section 2.2.1. Luckily the number of missing files based on this problem is very low in numbers. The second reason behind missing values is the phenotype files. Some phenotype attributes were missing. This is mainly caused by these attributes were collected by questionnaires. Three approaches were applied for these values. Only attribute missing in genotype file is the Case-Control attribute. The rate of missing values for this attribute is very low in number and this is the main attribute that our hybrid system is based on. So the subjects with missing case-control status were removed from dataset. For missing values in phenotype set, missing value rate for the attribute is considered. If the rate is low and class mean for that attribute can be calculated effectively, then class rate value is used instead of missing values. But if the missing value rate for an attribute is high, over 40%, than this attribute is removed from dataset.

Data transformation is needed to code the alleles because SVMs use numerical values instead of categorical ones. In literature allele combinations are coded by three numeric values. These three categories are based on the heterozygous and homozygous major alleles. In [38] {-1, 0, 1} is used. For example major allele CC was assigned to -1, CT was assigned to 0 and TT was assigned to 1. The same coding scheme is used in [39]. Similar coding scheme is preferred in [35]. In this work value 1 is used for heterozygous alleles, value 2 is used for homozygous alleles and value 3 for ambiguous alleles. The same values are used in [16] but values 1 and 3 are used for homozygous alleles where value 2 is used for heterozygous alleles. The work in [35] states the disadvantage of this coding scheme as "*alleles are not treated symmetrically*". In order to avoid this situation and as the parent of origin was not indicated in the datasets used, a different coding scheme was used. This coding scheme is given in the Table-13 below:

Table 13: Major allele coding scheme

Major Alleles	Coding Value
AA	1
AT / TA	2
AC / CA	3
AG / GA	4
TT	5
CT / TC	6
GT / TG	7
CC	8
GC/CG	9
GG	10

2.4 Proposed SVM-ID3 Hybrid System

According to our literature search the most widely used algorithms for detecting the relations between genotype information and the disease are ANN, SVM and Decision Trees. These methods are used either individually or combined with other methods to form hybrid structures. Generally these hybrid structures are formed with one of the methods given above and a type of Genetic Algorithm, like genetic programming or grammar evolution. Genetic Algorithms in these systems are used for optimizing the main method. It is either to find the optimum parameters such as learning rate and/or the number of epochs in ANN [45-208-209] or to find the best structure, the number of splits, branches and depth for Decision Trees [20-213].

In our model we've combined to different methods and for each of these methods an appropriate optimization was applied rather than combining a main method with an advanced optimization as stated above. By this way instead of benefitting from one strong method, we've combined the strengths of different methodologies.

In order to find the methods to be used, first we conducted a literature search to find out the best methods used in GWAS. Although some methods are most widely used, there is no general method that performs well in all conditions. For example in some works decision trees outperformed SVMs [16], in some works Bayesian approaches give better performance than Decision Trees and SVMs [34] and in some works SVMs perform better [37-38-39]. The performances of the methods are strongly related to the properties of the dataset used, the structure and self-parameters of the method and how they were applied. Optimization of features and data preprocessing also play an important role. In this situation selection of the methods is based on our aim.

That is the selection of the features which are strongly related with the disease rather than a medical diagnosis. By this way underlying genetic factors could be interpreted and further to be investigated by potential future studies. At this point interpretability and easy determination of the selected features comes forward with performance criteria.

The first method selected for our hybrid system is the Decision Trees. The reason behind selecting this method is its easy to understand structure. As the name indicates, Decision Trees form a tree structure with a root node, inner nodes, branches and leafs. This structure is given either graphically or hierarchical text presentation as in the Figure-31 below.

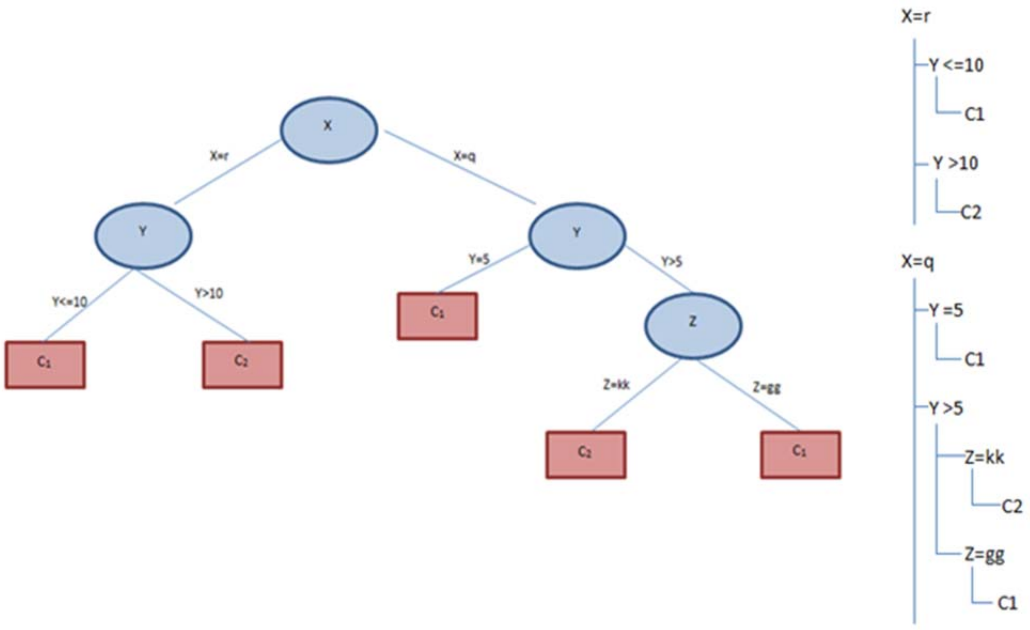


Figure 31: Decision tree structure and its hierarchical text representation

Either by looking the graphical or hierarchical structure, one can easily understands the decision mechanism that the method uses. The other reasons are its easy construction, easy application, handling numeric and discrete values, good performance on dealing with data with noise and missing values [47]. Because of such advantages, many woks in literature use decision trees in clinical diagnosis, GWAS and SNP diseases relations. One important point of decision trees is: although they are simple and easy, they have competitive performance among more complex methods.

For example in [34], Decision Tree gave better performance than Naïve Bayes, SVM with linear and cubic kernel in the means of accuracy on detecting breast cancer based on SNPs. The only method that was better than Decision Tree was SVM with quadratic kernel in that study. Behind these advantages decision trees can suffer from the sub tree problem. This problem occurs when the data features are not quality enough to separate the data space evenly, resulting many sub trees appearing more than once in the tree structure [47]. Such disadvantage can be ignored in our situation where the representation power is very important. Also we are working on related SNPs so it's not quite important that an SNP appears more than once in sub trees.

Among other decision trees such as C4.5 and CART, ID3 is selected because it is the best model appropriate for our data. Our data formed by two general parts. One of them is phenotype data with all features are categorical. Other is genotype data that consist of specific SNPs and their allele modifications. These modifications were converted to discreet structure as explained in Section 2.1. ID3 trees are especially designed for categorical data so it was chosen for our application. Application of tree, parameter adjustments and other details will be given in the Results Section.

ANNs and SVMs are two methods that are widely used in GWAS, SNP disease relations and in diagnosis of complex diseases. Our literature search given in Section 1.3 shows that both methods perform very well in selecting suspicious SNPs that cause complex diseases. The main reason of this high performance comes from their ability to deal with non-separable problems. To handle non-separable problems, ANNs use feed forward back propagation and weight adjustment methods while SVMs use kernel trick and different kernel types such as RBF and quadratic kernels. These two methods can be selected over each other but SVMs were selected for our application instead of ANNs.

The main reason behind this selection is the ANNs black box structure. That is the decision logic behind ANNs is not as clear as in decision trees. Some methods are being developed and used based on finding attribute weights and ranking these attributes by their weights as the highest weighted attribute is the most relevant feature for the given problem [76-118-119-120]. But such ranking approach is much easier in SVMs. In most of the SVM applications the output hyperplane is given by data features and their coefficients. These coefficients can be interpreted as the highest value most relevant and lowest value least relevant [37]. The other reasons that SVM chosen over ANN are, less overfitting, lower chance to stuck at local minima, flexible structure when separating the classes and preferred in many bioinformatics applications [49-50]. As stated in Section1, there is no agreed on standard for choosing the right kernel type. For our application SVM with RBF kernel is chosen. RBF kernel is preferred because of its faster learning speed, in some special conditions RBF could behave like linear function and another kernel which is sigmoid also can behave like RBF [215]. By this way only with adjusting the parameters, we had the chance of combining both linear and non-linear SVM with one kernel function. Application, parameter adjustments and other details will be given in the Results Section.

By the reasons explained above, our hybrid system formed by ID3 Decision Tree and SVM with RBF kernel. By this way advantages of entropy and hierarchical based methods by ID3 tree and advantages of regression methods by SVM combined [36].

First a data preprocessing step is conducted. By this way some SNPs with different representation are changed to Rsid format and allele modifications are coded. After that Plink analysis is made for selecting the statistically related SNPs in GWAS step.

Another filtering is carried on by using METU-SNP. The reason behind this filtering is to choose not only high statistically significant SNPs but also to consider the biological background. By this way both biologically and statistically significant SNPs could be chosen.

Downloaded datasets contain genotype and phenotype information in different folders. Phenotypes of all subjects are given in a single file but genotype information is given individually. That is each subject has its own folder about his/her genomic information. So a matching code was written to match phenotype and genotype information.

After this dataset is constructed it is given to hybrid system. In hybrid system SVM works first to select the related phenotype and genotype attributes. SVM weights are given to ID3 Tree and the features that it selected are the system's output. This whole process is given in the Figure-32.

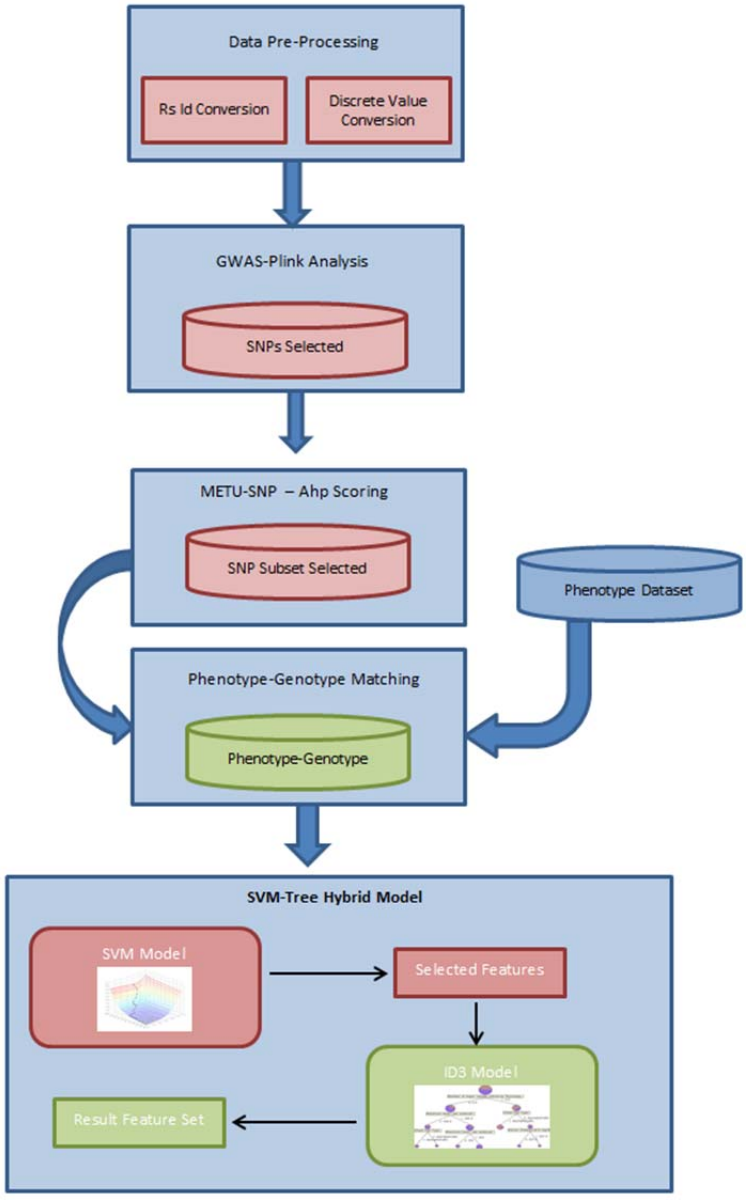


Figure 32: Work flow and the structure of the Hybrid System

CHAPTER 3

PROSTATE CANCER MODELING

3.1 Prostate Cancer Data Preprocessing

The hybrid system build is firstly tested on Multi Ethnic Genome Wide Scan of Prostate Cancer" with dbGap Study accession number phs000306.v2.p1. This data set consists of a total of 9457 subjects which 4650 cases and 4795 controls. Among these subjects 9130 of them have both phenotype and genotype information. Each individual in this dataset has approximately 600,000 SNPs and 20 phenotype attributes which are given in Table-11. Among these 20 phenotype attributes 12 of them were used for our analysis. The attributes, ***genava_id*** and ***study_id***, were excluded because they give only identification information. The attributes ***age_cat*** and ***ageco_cat*** excluded from our study because the rate of missing values were so high, missing value rate over 40%, and they could not be calculated by using class mean of the given attributes. Another attribute that was excluded is ***agedx_cat*** which is the age at cancer diagnosis. This feature is valid for only cases and not for controls. Aggressiveness of disease is given with the ***severity*** attribute. Like ***agedx_cat*** it is valid for only cases and gives misleading strong relation power with the disease. The remaining attributes, which were used in our study, are given with their value range in Table-14 below:

Table 14: Phenotype attributes, their explanations and value ranges of prostate cancer data

Attibute	Explanation	Range
bmi_cat	Body Mass index. Calculated by weight and height.	1:<= 22.5 2:<=24.9 3:<=29.9 4:>=30
fh_prca	Family history of prostate cancer, brother or father	0: No 1:Yes 9:Unknown
pa_cat	Hours per day of moderate or vigorous physical activity. Categories are based on gender-specific, cohort-wide quintiles with the following cut-points: 0.32, 0.71, 1.07, 2.04	1:Quintile 1 2:Quintile 2 3:Quintile 3 4:Quintile 4 5:Quintile 5
packyrs_ca	number of packs of cigarettes smoked per day times the number of years smoked	1:None 2: <5 years 3:<10 years 4:<20 years 5:<30 years 6:30+ years

ethanol_ca	**Alcohol drinks per day. grams per day of alcohol consumption (12 grams=1 drink)**	**1: None 2:<1 Drink 3:<2 Drink 4: 2+ Drink**
d_lyco_cat	Density for lycopene intake (micrograms per 1000 kilocalories per day. Categories are based on gender-specific, cohort-wide quintiles with the following cut-points: 752, 1077, 1437, 2023	1:Quintile 1 2:Quintile 2 3:Quintile 3 4:Quintile 4 5:Quintile 5
p_fat_cat	Percent of calories from fat (%) reported at baseline. Categories are based on gender-specific, cohort-wide quintiles with the following cut-points: 24.1, 28.6, 32.3, 36.3	1:Quintile 2 2:Quintile 3 3:Quintile 4 4:Quintile 5 5:Quintile 6
d_calc_cat	Density for calcium intake milligrams per 1000 kilocalories per day. Categories are based on gender-specific, cohort-wide quintiles with the following cut-points: 245, 304, 362, 439	1:Quintile 3 2:Quintile 4 3:Quintile 5 4:Quintile 6 5:Quintile 7
currsmoke	currently smoker?	0: No 1: Yes
eversmoke	eversmoke ?	0: No 1: Yes

Table-14 (Cont'd.)

The distribution of these attributes among cases and controls are given in the Appendix-A

3.2 Plink and METU SNP Analysis

Result of Plink analysis, ***.assoc.adjusted*** file is analyzed by using ENSEMBL database. This database allows searching important biological information. It offers searches for genes, variations and regulations. For our study we used variation search. In variations different types of species can be searched. For Homo Sapiens somatic structural variations, somatic variations, structural variations and variations can be listed. Homo Sapiens variations are used in our study. After the appropriate database is selected filters can be applied by using regions, variations and gene associations. In region, filters for chromosome, base-pairs and markers can be used. In variation, results are filtered according given SNPs list. Allele frequencies and significance values can also be included. In gene association, results are displayed according to given gene id list. In last step attributes to be displayed can be selected by sequence variations and gene association. In sequence variation, attributes such as name, chromosome name, position on chromosome, allele frequency, phenotype description, information about study, associated genes, p-values and such can be selected. With gene association, attributes such as specific alleles, gene ids, start and end of translations and transcription ids can be selected.

In summary by using ENSEMBL database, biological information of SNPs given in ***.assoc.adjusted*** file can be examined. This is important for pre examining the biological relevance of the dataset with given and similar diseases. The number of SNPs biologically related with given disease can be low. This causes the number of SNPs to be found by hybrid system also low in number. But we must emphasize that the SNPs found related with given diseases are based on the previous studies. This means SNPs found related with disease in Plink, METU-SNP and Hybrid system could not be found in ENSEMBL search because they are not yet identified by previous studies.

SNP list formed by Plink analysis is given to ENSEMBL and among 600,000 SNPs only 22,848 of them are found to be associated with specific phenotypes. The number of SNPs that are directly related with prostate cancer, related with other types of cancer such as ovarian, lung, pancreatic cancer and related with the phenotypic attributes of prostate cancer is given in the Table-15 below:

Table 15: The number of SNPs, after Plink analysis in prostate cancer. SNPs are either related with ed with prostate cancer or phenotypic attributes of prostate cancer orother types of cancer gathered from ENSEMBL database.

Relation	Number of SNPs
Prostate Cancer	340
Phenotypic attributes for Prostate Cancer	200
Other Cancer Types	441

After Plink analysis, gathered ***.assoc.adjusted*** file is given to METU-SNP in order to find both statistically and biologically relevant SNPs. For this aim, AHP Scoring Mechanism of METU-SNP is used. Threshold for p value was determined as 0.05. AHP Mechanism selected 1000 SNPs. When the performance of the hybrid system was tested by using this dataset, the observed performance of genotype set was low. In order to increase this performance we decided to increase the number of genetic features. To do that assoc.adjusted file is directly used and SNPs with p value<0.05 were chosen. With this analysis, the number of related SNPs was increased to 2710. ENSEMBL analysis found 87 of these SNPs related with a specific phenotype. The number of SNPs that are directly related with prostate cancer, related with other types of cancer such as ovarian, lung, pancreatic cancer and related with the phenotypic attributes of prostate cancer is given in the Table-16 below:

Table 16: ENSEMBL SNPs results for prostate cancer. The number of SNPs, after selecting the representative subset, related with prostate cancer, phenotypic attributes of prostate cancer and other types of cancer gathered from ENSEMBL database.

Relation	Number of SNPs
Prostate Cancer	2
Phenotypic attributes for Prostate Cancer	2
Other Cancer Types	1

We conducted a second search by using RegulomeDB. This database takes the SNP list in rsID format and gives information about the corresponding SNP's chromosome, location by coordinates and a score point which shows the importance of the relation. This scoring mechanism is given in the Table-17 below:

Table 17: RegulomeDB SNP scoring mechanism. And our interpretation of the scores

Score	Data Support	Our Score
1a	eQTL + TF binding + matched TF notif + matched Dnase Footprint+Dnase peak	High Score
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak	
1c	eQTL + TF binding + matched TF motif + DNase peak	
1d	eQTL + TF binding + any motif + DNase peak	
1e	eQTL + TF binding + matched TF motif	
1f	eQTL + TF binding / DNase peak	
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak	
2b	TF binding + any motif + DNase Footprint + DNase peak	
2c	TF binding + matched TF motif + DNase peak	
3a	TF binding + any motif + DNase peak	Medium Score
3b	TF binding + matched TF motif	
4	TF binding + DNase peak	
5	TF binding or DNase peak	Low Score
6	other	

Out of 2710 SNPs, which were gathered from Plink, 2697 of them were found in this database. And among these 81 of them had high score and 200 of them had medium score.

3.3 Support Vector Machine Analysis

The first machine learning method in our hybrid system is the support vector machine (SVM). The main reason for selecting this method is its good performance in bioinformatics studies [49-50]. Interpretation of the decision logic is easier when compared to ANNs. Attribute weights given in SVM can be interpreted as relevance to the given problem with higher value weighted attributes are the most relevant ones [37]. Less overfitting, lower chance to stuck at local minima, flexible structure when separating the classes are the other appealing features of SVMs. For our SVM application RBF kernel is chosen. This kernel is widely used in GWAS [33-36] and preferred in our study for its faster learning speed and its advantage of to be used as both linear kernel and sigmoid kernel in some special conditions [49-50-215].

SVM model is applied with RapidMiner 5.0 which is a free open source software tool for data mining applications. This software is preferred in various applications in the literature such as [216-217-218]. SVM has many parameters to adjust, the most important parameters for our application are:

- Kernel Type: Different types of kernels can be chosen such as linear, radial basis, polynomial, sigmoid and quadric kernels.
- Kernel Gamma: The gamma parameter of kernel with range [0.0, ∞]
- Kernel Sigma: The sigma parameter of kernel with range [0.0, ∞]
- C: Penalty constant with range [0.0, ∞]
- Epsilon: The epsilon parameter of kernel with range [0.0, ∞]

In order to get good performance results, the parameters given above must carefully be adjusted. This requires an optimization to be run in order to find the best performing SVM model. Finding the optimum parameters for SVMs examined by various works in literature such as [55-85-86-87] and grid search for optimization is recommended by [88-89-90]. The main idea behind grid search is to test the model by trying the given ranges of values of the parameters at the hand. The five SVM parameters above introduced a total of 10,648 combinations tested. These parameters are kernel type, gamma, sigma, C and epsilon. The kernel type chosen is the RBF kernel and the reason behind this is explained above and in Section 2.4. The epsilon parameter is used for Regression SVMs [220], which are used to predict continuous values. The parameters Gamma and Sigma are interchangeably used for RBF as given in the following equations [86-221-222]:

$$k(x,y) = \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right) = \exp(-\gamma\,|x-y|^2) \qquad \textbf{(EQUATION 43)}$$

This means in order to optimize our SVM model two parameters, gamma and C, must be tested. Although the number of parameters to be tested is low, the number of combinations for these parameters can be very high according to the value range to be used. To find an appropriate value range, purpose of these parameters in RBF and relations between them must be known clearly.

The ***C*** constant is used to adjust the margin of the hyperplane that separates the classes. If C is too high then the separating vectors closer to the data points, if C is low then the margin increases [222]. The data points which are very close to separating vector become margin error in the second case [222]. This situation is depicted in the Figure-33 below:

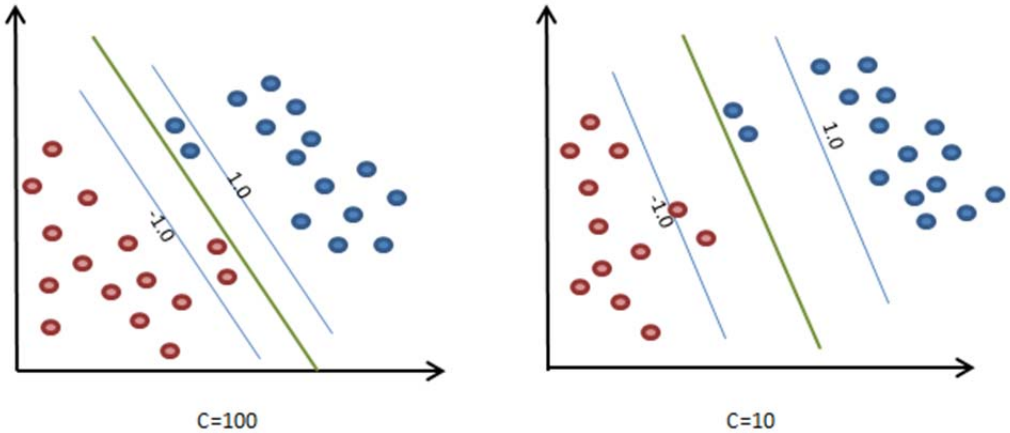


Figure 33: Margin of the hyperplane changes with respect to C (Adapted from[219]).

The gamma parameter gives its shape to decision boundary. In more formal description: *"Gamma determines the flexibility of SVM*[15]*"*. Although this flexibility allows a good separation between classes, if it is overestimated the problem of overfitting can occur. So the gamma parameter must be chosen carefully, with higher value have the risk of overfitting and smaller values form a linear like structure. This scheme is given in the Figure-34 below:

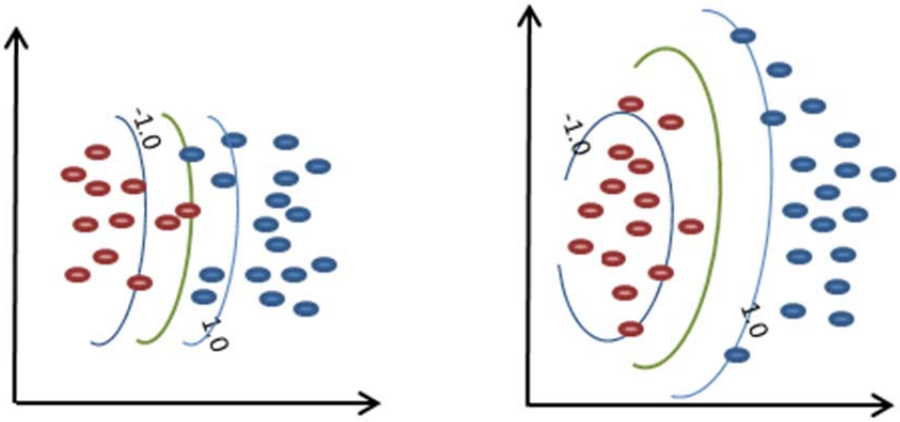


Figure 34: The effect of gamma parameter in decision boundary (Adapted from [219])

The relation between C and gamma explained in [222]. According to this work *"If gamma decreased, the curve of decision boundary decreases; at the same time if C is increased then curve increases to accommodate a larger margin*[14]*."* This phenomenon allows the same performance result to be gathered from SVM by using different C and gamma combinations.

In the light of the information given above the value ranges for C and gamma are searched through the literature and with our testing applications. For gamma the value range was determined by testing and suggestions of [221] and set in between [0.0001, 100] with powers of ten. The value range for C was determined by testing and suggestions in [87] and set in between [0-10] with five linear steps. These value ranges for two parameters revealed 42 combinations to be tested in grid optimization. After the optimum parameters were found, SVM was constructed by using these parameters and ten-fold cross validation. Validation sets were formed by stratified sampling. By this way each data element in validation set is chosen randomly and class distribution of the original dataset was kept.

In order to prove the advantage of combining phenotype and genotype data three datasets were used. One of the sets contain only the genotype information, the other set contains only phenotypes and the last set contains both phenotype and genotype data. In order to run SVM on these datasets a numerical conversion is needed because SVMs can not perform well on categorical data. The conversion is based on the value range of the selected attribute. If there are ***n*** categories for the chosen attribute these categories are represented by a total of ***n*** bits. All bits except the corresponding category set to zero.

[15] A User's Guide to Support Vector Machines. Ben-Hur A, Weston J. Methods Mol Biol. 2010;609:223-39

The corresponding bit of that category value is set to one. Then the binary value is converted to a decimal numeric value. After the conversion is done, Grid search was run and SVM model was applied on all datasets. Performance comparisons and the optimum values used for SVM parameters are given in the Table-18 below:

Table 18: Performance of SVM Model on prostate cancer datasets.Three sets were used as only genotype, only phenotype and combined set of phenotypes and genotypes.

	Only Genotype Dataset	**Only Phenotype Dataset**	**Genotype and Phenotype Dataset**
SVM Parameters (C, G)	$(2, 10^{-4})$	$(4, 10^{-2})$	$(10, 10^{-3})$
Accuracy	59.02	68.23	72.46
Precision	61.29	76.8	82.68
Recall	63.15	70.12	71.34
AUC	0.606	0.768	0.829

These results in the table above clearly shows that combining phenotypic information with genotype data slightly increases the decision performance in all aspects of accuracy, precision, recall and AUC.

When attribute weights of the kernel were examined and these attributes were searched through ENSEMBL, two SNPs (rs3792693 and rs8066529) were found to be related with prostate cancer; three SNPs (rs925946, rs3027409 and rs925946) were found to be related with phenotypic attributes of the prostate cancer dataset and one SNP (rs13016963) was found to be related with another cancer type which is melanoma. This result shows that our SVM model was able to find all related SNPs in the dataset which were given in Table-16. The number of related SNPs found can be low in number but again it is better had to state that ENSEMBL shows only the results of previous studies. So the SNPs found in this work are the candidates for further investigation.

3.4 Hybrid System Analysis

The hybrid system offered in this work is formed by two methods which are SVM and ID3 Decision Tree. The reason behind why these methods were selected is given in Section 2.4.

In literature there are various studies that combine SVMs and decision trees. This combination is generally used for multi-classification and multi-clustering problems, where there are more than two classes and/or clusters, such as in the works of [223-224-225-226-227]. Versions of the same approach are used to combine SVMs with decision trees for multi-classification tasks. The approach is to divide the multi-classification problem into *N*-binary classification tasks [226], then to use SVM for the decisions in inner nodes of the tree.

In this method SVM deals with binary classification in each node but combinations of these different classifications form the multi-classification task and by this way the whole structure has the SVM's high power for binary classification. To divide the problem into *N*-binary classification, the distance between each class is calculated and the furthest two classes is used for the initial binary problem and takes its place in the root of the tree [226].

This scheme continues for inner nodes. Euclidean Distance is generally preferred to calculate the distance between classes [224-226]. This approach is depicted in the Figure-35 below:

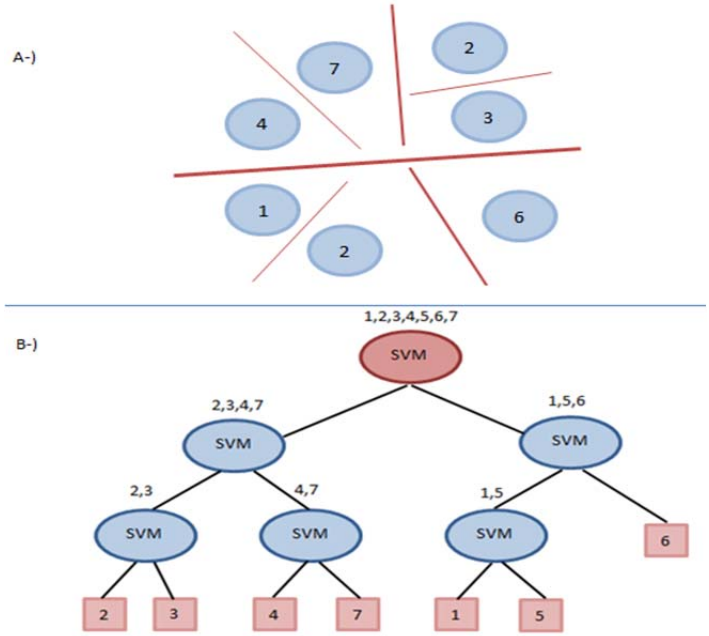


Figure 35: Solution of Multi-classification problem by SVM-Decision Tree. Problem is given in 45-a and solution is given in 45-b (adpted from [226])

Although this approach gives significant performance results, it is not appropriate for our purposes because we are concerning on binary classification rather than multi-classification.

SVM-Decision tree combination for binary classification problems are researched in the literature such as in the works of [227-228]. Both of these works first apply SVM in order to optimize the parameters and datasets to be used in decision tree. A cross validation scheme is applied to test and train the SVM, for each validation the prediction results are saved and correctly predicted points are stored in a new data set which to be used for constructing the decision tree by cross validation. [227-228].

In our application the main problem in combining the SVM and ID3 was how to use the output results of SVM in ID3. There were two choices. One of them was to use the entire result set from SVM and use attribute weights as a feature that will increase the information gain. The other approach was to rank the attributes according to their weights given by SVM and then select the top ranked SNPs.

In the second case there is a small risk of losing information of the data which are not in the top rank. Because of this risk factor, the first approach was preferred. By this way all data features will be used in ID3 tree and the importance factor that was given by SVM also will be taken into consideration.

Assigning weights to attributes in decision tree construction have been studied in various works such as [229-230-231-232]. The aim of the attribute weighting in these studies is to avoid tendency of decision trees to choose the attributes with higher range of values. [231-232].

In order to balance the dataset these studies use some weight coefficients to decrease the information gain of higher value ranged attributes. By this way the importance of attributes with low value range somewhat equalized with high value ranged attributes.

For example in [232] a correlation function for all attributes is calculated first by:

$$\mathrm{AF_c(A)} = \frac{\sum_{i=1}^{v} |\mathrm{A_{i1}} - \mathrm{A_{i2}}|}{\mathrm{V}} \qquad \textbf{(EQUATION 44)}$$

"*Where* $\boldsymbol{A}$ *is the attribute,* $\boldsymbol{V}$ *is the number of values for that attribute and* $\boldsymbol{A_{ij}}$ *is the number of tuples that* $\boldsymbol{i_{th}}$ *value of A belongs to* $\boldsymbol{j_{th}}$ *class*."

If an attribute has both a higher value range and high information gain then the weight is applied by:

$$\mathrm{W} = \frac{\mathrm{AF_C(A)}}{\sum_{i=1}^{n} \mathrm{AF_c(i)}} \qquad \textbf{(EQUATION 45)}$$

Where $\boldsymbol{n}$ is the number of attributes and $\boldsymbol{W_A}$ is the weight coefficient for attribute $\boldsymbol{A}$.

By this scheme if an attribute has high value range and high information gain its information gain is decreased by its weight which is dependent to the number of values of that attribute and the correlation value for the rest of the attributes.

These approaches are not sufficient for our application for two reasons. First, the attributes in our dataset has almost the same value range. For genotype attributes the value range is between one and ten, for phenotype attributes the value range is between one and six. Some phenotype attributes has a value rang between one and three. But this did not affect the tree performance Our ID3 prototype did not suffer from the high value ranged attributes. In tests we've seen that some low value ranged attributes, such as fh_prca, has taken part in higher levels of the tree where high value ranged attributes, such as d_lyco_cat, have settled in lower levels. The other reason is that if the weights from SVM are used for decreasing the information gain it will conflict with the idea of using SVM. SVM is used to detect the higher related attributes in our application.

So in our weighted tree approach the weights are taken from SVM and used to increase the information gain of the certain attributes. Kernel results of SVM can be extracted from RapidMiner. These results list the all attributes and their corresponding final kernel weights. But the given range of weights, which is between -289 and 328, is very high.
In order to narrow down this gap Min-Max normalization given by [48] was applied. The formulation for this normalization is given in the equation below:

$$Norm(A) = \frac{Value_A - MinValue_A}{MaxValue_A - MinValue_A}(NewMax - NewMin) + NewMin \qquad \textbf{(EQUATION 46)}$$

By Min-Max Normalization given above, value ranges of the weight attributes were matched to [0, 1] range. Then the normalized weights are used as coefficients for information gain ratio.

After the weighting strategy was determined, we moved on constructing ID3 Tree on RapidMiner. As in SVM, ID3 Tree has also parameters to set. These parameters are:

- Criterion: Defines the criteria to be used for selecting the splitting attribute. Gain Ratio, Information Gain, Gini Index and Accuracy are the offered criterions.
- Minimal Size of Split: The minimal size of a node for splitting
- Minimal Leaf Size: Minimum size for leaf nodes
- Minimal Gain: Minimum value for the gain

For criterion, gain ratio was used instead of information gain. That's because information gain tends to select higher value ranged attributes as explained above [61-231-232]. Gain ratio uses entropy measure to normalize the information gain. This scheme was given in the Equation 9.

Minimal size of split and minimal leaf size attributes left as default for two reasons. Firstly our prototype ID3 models did not show significant difference in performance when these values were changed. The second reason is that when these attributes were selected for grid search optimization a minimum of 1331 combinations had to be tested which exceeds our computational power.

The most important parameter that affects the performance of ID3 tree is the minimum value for the chosen information gain criteria. If this value is chosen too low, the tree will have too many levels and branches. This can cause overfitting. If this value is chosen too high then the depth of the tree can be too low which causes the loss of valuable information and so underfitting. In order to avoid such situations grid search is done for parameter optimization. In grid search maximum value for information gain ratio was set to ten and minimum value was set to 10^{-3}. This value range was searched by fifty logarithmic steps which resulted in fifty one combinations to be tested. Each test lasted approximately eight to eleven hours according to the dataset used.

In order to prove the advantage of combining phenotype and genotype data three datasets were used. One of the sets contain only the genotype information, the other set contains phenotype information and the last set contains both phenotype and genotype data. Grid search was run and our hybrid model was applied on all datasets.

Optimum values used for ID3 part of the hybrid system and performance comparisons of the hybrid system are given in the Table-19 below:

Table 19: Performance of Hybrid system on prostate cancer datasets. Three datasets were used as only genotype, only phenotype and combined genotype-phenotype

	Only Genotype Dataset	**Only Phenotype Dataset**	**Genotype and Phenotype Dataset**
Minimal Gain	7.43	3.56	4.1
Accuracy	71.67	84.23	93.81
Precision	72.69	86.20	96.55
Recall	68.96	83.78	90.92
AUC	0.674	0.857	0.91

Unfortunately the graphical tree structure is too big in width to give it here as a figure. Its hierarchical text structure is given in the Appendix-C. Maximum depth of the tree is sixteen and minimum depth is three. According to SVM ID3 hybrid system structure the most important attribute is the ethnicity phenotype information. Our system made a strict distinction on ethnicity attribute which means for the prostate cancer data used, the disease or healthy condition firstly depends on the subject's ethnicity. The second important attribute, which is found on second level of the tree, is the body mass index, ***bmi_cat***, attribute. For each ethnicity, this attribute is searched first when diagnosing the prostate cancer condition. Other phenotypic attributes that are found to be important are ***packyrs_ca*** which is related with smoking behavior of the subjects and ***ethanol_ca*** which is related with alcohol consumption of the subjects.

When genotypic attributes were examined, hybrid system found twenty eight SNPs related with the given disease in African American population. When these SNPs were searched by using RegulomeDB, sixteen of them are found to be related. For Japanese population twenty two SNPs were found. According to RegulomeDB scoring sixteen of them are found to be related. For the Latino population the number of SNPs found by hybrid system is sixty five and thirty of them are predicted to effect biological function according to RegulomeDB.

CHAPTER 4

MELANOMA MODELING

4.1 Melanoma Data Preprocessing

The second dataset used for testing the hybrid system was "High Density SNP Association Analysis of Melanoma: Case-Control and Outcomes Investigation" with dbGaP study accession number phs000187.v1.p1. This data set consists of a total of 3115 subjects which 2053 cases and 1062 controls. Among these subjects 995 of them has value ***unknown*** for all phenotype attributes. These subjects were excluded from dataset. The remaining subjects have both phenotype and genotype information. Each individual in this dataset has approximately 600,000 SNPs and 24 phenotype attributes which are given in Table-12. Among these 24 phenotype attributes 11 of them were used for our analysis. The attributes ***country, state***, ***Geneva_id*** and ***blood_draw_rate***, which give the date of blood draw, are irrelevant with disease so they were excluded from the study. The following attributes, ***Breslow_tumor_thickness***, ***Clark_level***, ***vital_status***, ***months_follow_up***, ***aged_DX***, ***stage_cat***, and ***ulceration***, were also excluded because they are only related with the case subjects and values for them are missing for control subjects. The attribute ***race*** was excluded because all subjects were white European ancestry and the attribute ***control_Age_at_DX*** also excluded because of its high missing value rate. The remaining attributes, which were used in our study, are given with their value range in Table-20 below:

Table 20: Phenotype attributes their explanations and value ranges of melanoma data

Attribute	Explanation	Range
gender	gender of subjects	M:Male F:Female
moles	indicator of moles presence of moles	1:Yes 2:No 1000: Unknown
dysplastic_nevi	indicator of moles presence of dysplastic nevi	1:Yes 2:No 1000: Unknown
skin_color	color of skin	Value Rang [1-10] 1:Lightest 10:Highest
eye_color	cloro of eye	1: Blue 2: Green 3: Grey 4: Brown 5:Hazel 9:missing 1000:Unknown

hair_color	**color of hair**	**1: Blonde 2: Red 3: Brown 4: Black 5:Other 9:missing 1000:Unknown**
sunburn	Indicator of the presence of severe blistering before age 16	1:Yes 2:No 9: Missing 1000: Unknown
freckle	Indicator of freckling	1:Yes 2:No 9: Missing 1000: Unknown
untanned	presence of a tan after exposure to the sun	1: Always 2: Usually 3:Moderate 4: Minimal 5:Never 6:Don't know 9:missing 1000:Unknown
family_history	cancer history in the 1st degree relatives	1:Yes 2:No 1000: Unknown

Table-20 (Cont'd.)

When melanoma data was examined we noticed that 995 of the cases have "unknown" label for the phenotype attributes to be used. These cases were removed from data set because they form 48% of the cases and missing value replacement strategies can lead undesired results with such high rates. With this removal our melanoma dataset consisted of 1095 cases and 1062 controls. The distribution of the phenotype attributes among cases and controls are given in Appendix-B.

4.2 Plink and METU SNP Analysis

ENSEMBL database, explained in detail in Section 3.2, used for examining the biological information of SNPs given in ***.assoc.adjusted*** file gathered from Plink analysis. This is important for pre examining the biological relevance of the dataset with given and similar diseases. The number of SNPs biologically related with given disease can be low. This causes the number of SNPs to be found by hybrid system also low in number. But we must emphasize that the SNPs found related with given diseases are based on the previous studies. This means SNPs found related with disease in Plink, METU-SNP and Hybrid system could not be found in ENSEMBL search because they are not yet identified by previous studies.

SNP list formed by Plink analysis is given to ENSEMBL and among 600,000 SNPs only 28,472 of them are found to be associated with specific phenotypes. The number of SNPs that are directly related with melanoma, related with other types of cancer such as ovarian, lung, breast, prostate, pancreatic cancer and related with the phenotypic attributes of melanoma is given in the Table-21.

Table 21: The number of SNPs, after Plink analysis in melanoma. SNPs are either related with ed with melanoma or phenotypic attributes of melanoma or other types of cancer gathered from ENSEMBL database.

Relation	Number of SNPs
Melanoma	33
Phenotypic attributes for Melanoma	217
Other Cancer Types	643

After Plink analysis, gathered ***.assoc.adjusted*** file is given to METU-SNP in order to find both statistically and biologically relevant SNPs. For this aim, AHP Scoring Mechanism of METU-SNP is used. Threshold for p value was determined as 0.05. With this analysis, the number of related SNPs is decreased to 2783. Although these SNPs are found to be highly related with prostate cancer in both statistical and biological point of view, ENSEMBL analysis found 207 of them related with a specific phenotype. The number of SNPs that are directly related with melanoma and other types of cancer such as ovarian, lung, breast, prostate, pancreatic cancer and related with the phenotypic attributes of melanoma is given in the Table-22 below:

Table 22: ENSEMBL SNPs results for melanoma. The number of SNPs, after selecting the representative subset, related with melanoma, phenotypic attributes of melanoma and other types of cancer gathered from ENSEMBL database

Relation	Number of SNPs
Melanoma	2
Phenotypic attributes for Melanoma	31
Other Cancer Types	1

When SNP list of METU-SNP was searched with RegulomeDB, the number of SNPs found that are related with binding was 1605. Among them 301 SNPs are ranked with higher scores.

4.3 Support Vector Machine Analysis

RapidMiner tool is used to construct SVM model. In order to find optimum parameters to use, a grid search was conducted on parameters C and gamma. The value ranges used for gamma was $[10^{-4}, 10^{2}]$ with powers of ten and for C [0, 10] with five linear steps. The reason behind choosing these value ranges is explained in Section 3.3. These value ranges for two parameters revealed 42 combinations to be tested in grid optimization. After the optimum parameters were found, SVM was constructed by using these parameters and ten-fold cross validation.

In order to prove the advantage of combining phenotype and genotype data three datasets were used. One of the sets contain only the genotype information, other set contains only phenotype information and the last set contains both phenotype and genotype data. In order to run SVM on these datasets a numerical conversion, which is explained in Section 3.3, was made.

The performance results of SVM model that was run on only genotype and both genotype and phenotype datasets of melanoma and the values of optimum parameters used are given in the Table-23 below.

Table 23: Performance of SVM Model on melanoma datasets.Three sets were used as only genotype, only phenotype and combined set of phenotypes and genotypes.

	Only Genotype Dataset	**Only Phenotype Dataset**	**Genotype and Phenotype Dataset**
SVM Parameters (C, G)	$(2, 10^{-4})$	$(6, 10^{-2})$	$(0, 10^{-4})$
Accuracy	78.41	70.37	78.6
Precision	75.52	64.32	76.45
Recall	76.38	74.64	75.07
AUC	0.84	0.756	0.846

As in the prostate cancer case, for melanoma combining genotype and phenotype information performed better when compared to only genotype and only phenotype datasets. This again supports our hypothesis on combining genotype and phenotype information will give better performance.

According to SVM result set, 68 SNPs are found to be associated with a specific gene in ENSEMBL database. 40 of them found to be related with phenotypic attributes of melanoma.

4.4 Hybrid System Analysis

The hybrid system constructed in this work combines ID3 decision tree and SVM methods. The reason behind choosing these methods was explained in previous sections. In order to combine these methods, kernel weights that were extracted from RapidMiner were used as a weight coefficient for ID3's information gain ratio. These kernel weights were normalized to map the values into [0, 1] range.

ID3 decision tree part of the hybrid system was constructed in RapidMiner. In order to find the optimum parameters, a grid search was conducted on information gain ratio parameter. In grid search maximum value for information gain ratio was set to ten and minimum value was set to 10^{-3}. This value range was searched by fifty logarithmic steps which resulted in fifty one combinations to be tested. Each test lasted approximately ten to twelve hours according to the dataset used.

In order to prove the advantage of combining phenotype and genotype data three datasets were used. One of the sets contain only the genotype information, other set contains only phenotype information and the last set contains both phenotype and genotype data. Optimum values used for ID3 part of the hybrid system and performance comparisons of the hybrid system are given in the Table-24.

Table 24: Performance of Hybrid system on melanoma datasets. Three datasets were used as only genotype, only phenotype and combined genotype-phenotype

	Only Genotype Dataset	Only Phenotype Dataset	Genotype and Phenotype Dataset
Minimal Gain	10^{-3}	0.334	0.6
Accuracy	57.12	75.48	86.35
Precision	57.77	79.54	82.27
Recall	53.47	68.9	79.07
AUC	0.567	0.799	0.81

Unfortunately the graphical tree structure is too big in width to give it here as a figure. Its hierarchical text structure is given in the Appendix-D. Maximum depth of the tree is nine and minimum depth is four. According to SVM ID3 hybrid system structure the most important attribute is the gender phenotype information. It was an expected result because attributes such as gender successfully divides the datasets into to even parts. Other important attributes, which are found on second level of the tree, are the ***moles*** and ***dysplastic_nevi*** attributes. All ten attributes in dataset were found in different levels of tree.

When genotypic attributes were examined, hybrid system found 53 SNPs related with the given disease. Among them the SNPs, rs2246095, rs2768343, rs10406787, rs239695, rs1467414, rs10028824 either matched to specific genes or have an important role in the genes regulation and binding according to ENSEMBL and Regulomedb databases.

CHAPTER 5

DISCUSSION

The aim of this thesis is to combine number phenotypes including clinical information, demographic information and life style habits. By this way the underlying reasons behind complex diseases could be better explained and using both genotyping and phenotypes can improve diagnostic performance and a system build on genotype-phenoype information could be used as an alternative preventive or early detection system. To prove that an SVM-ID3 Hybrid model is constructed and tested on prostate cancer and melanoma datasets downloaded from dbGaP.

In this section of the thesis, the result of the analysis, given in the Section 3 and Section 4, will be discussed.

5.1 Discussion of Prostate Cancer Results

The prostate dataset used in this study consists of 9457 subjects which 4650 cases and 4795 controls. After data preprocessing step the number of phenotypic attributes were reduced to 12 and number of genotypic attributes represented with Rsid were reduced to 2710.

In order to use this dataset in SVM Model, a polynomial to numeric data conversion was conducted. This first part of the hybrid system gave 59.2% of accuracy on the dataset that only used genotype information, 68.23% of accuracy for the dataset with only phenotypes and gave 72.46% of accuracy with the dataset that combines genotype and phenotype information.

Attribute weights of the SVM's kernel were normalized and used as coefficients for information gain ratio in ID3 decision tree part of the hybrid system. When this system was run on only genotype and only phenotype datasets, the accuracy performance was 71.67% and 84.23% respectively. When the Hybrid System ran on the dataset that has genotype and phenotype information, the accuracy performance increased to 93.81%.

These performance measures can be compared to Prostate Specific Antigen (PSA) test in clinical point of view. PSA is an antigen and its levels can be detected by blood draw. The levels of PSA are used for early detection of prostate cancer condition before biopsy.

PSA levels those are smaller than 4ng/ml is known as normal levels, levels between 4ng/ml – 10ng/ml are known as suspicious and levels higher than 10ng/ml known as high [237]. The problem with this test is the cutting points. Although cutting point is generally taken as 4ng/ml, 4ng/ml – 10ng/ml range creates a grey area for decision and even some subjects below this value could be diagnosed as cancer and subjects above this value could be diagnosed as healthy [238]. Table-25 below shows the associated prostate cancer level risk according to given PSA levels.

Table 25: The risk of prostate cancer at given PSA levels(Taken from [239])

PSA ng/ml	Risk of Prostate Cancer	Risk of Aggressive Prostate Cancer
<0.5	7%	1%
0,6-1.0	10%	1%
1.1-2.0	17%	2%
2.1-3.0	24%	5%
3.1-4.0	27%	7%

The cutoff value of PSA test also changes with respect to the subject's age [240]. That is why there are various studies in the literature that offer different cutoff values for PSA test [240-241]. These also result in a second problem that is the changing values of sensitivity and specificity of the PSA test. This inconsistency of performance results stated in the work of Oesterling as: *"it is unlikely that PSA by itself will become an effective screening tool for the early diagnosis of prostate cancer*"[16]. Although there is an inconsistency, our literature search so far states if the cutoff point is taken as 4ng/ml, the highest performance rates found as: 86% of sensitivity, and 0.67 of AUC [240]. Some works in the literature calls attention to performance of the PSA test must be developed in terms of sensitivity and specificity [242-243].

According to the information given above there is a clear need for a test that will be used for early detection of the prostate cancer before biopsy. Our hybrid system with performance criteria of: %90.92 sensitivity and 0.91 AUC is a good candidate for early detection of prostate cancer.

Resulting feature sets of our hybrid system was examined and phenotypic attribute ethnicity was found to be the most related attribute with the prostate cancer. This result was not surprising because several works in the literature show that there is a relation with ethnic features and prostate cancer disease. Kleinmann's work shows that the ethnic background of the patients plays an important role in the prostate cancer related quality of life [244]. According to Hoffman, the etiology of the prostate cancer is highly depended on ethnicity and African American's has the highest risk for having prostate cancer [245]. As a supporting result, our hybrid system strictly divides the prostate dataset according to ethnicity and for each ethnicity different paths were observed.

[16] Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate. Oesterling JE. The Journal of Urology 1991, 145(5):907-23

If the subject's ethnicity is African American and its body mass index (BMI) is in first category, which is bmi<22.5, then just by looking to SNP with rsid 11729739 our hybrid system can decide whether the subject is a case or control. If the major allele of this SNP is TT then the subject is diagnosed as case and if the major allele is CT than the subject is diagnosed as control. By looking to the tree structure given in the Appendix-C, many rules can be extracted like the one above. For African American ethnicity our system found 29 important SNPs related with the disease. Among them 17 of SNPs are found to be important by RegulomeDB. These SNPs, their related chromosome information and Regulome score is given in the Table-26 below:

Table 26: SNPs found by hybrid system for African American Population. The list is gathered from RegulomeDB.

rsid	chromosome	Coordinate	score
rs1433369	chr12	76424152	2b
rs17701543	chr10	50426612	3a
rs17375010	chr1	49227960	4
rs10788555	chr10	89368766	4
rs10745253	chr10	50439214	5
rs2296370	chr19	55224784	5
rs2120806	chr3	122715811	5
rs17001078	chr19	11113650	5
rs918285	chr7	78537750	5
rs11729739	chr4	27220863	6
rs964130	chr4	147633044	6
rs766045	chr8	3980977	6
rs4908656	chr1	7616592	6
rs12980509	chr19	22952945	6
rs7843255	chr8	4531834	6
rs7067548	chr10	9110787	6

When the results of hybrid system for Japanese population are examined, the following relations are noticed first. If the subjects are in fourth category of BMI, which is >=30, then these subject are classified as controls. If the subjects are in first category of BMI, which is <22.5, then the decision is made based on the SNP rs2442602. The major allele AA for this SNP indicates the subjects are cases. A total of 22 SNPs found to be related with the disease for Japanese population. Regulome results for these SNPs are given in the Table-27 below:

Table 27: SNPs found by hybrid system for Japanese Population. The list is gathered from RegulomeDB

rsid	#chromosome	Coordinate	score
rs12644498	chr4	57372629	3a
rs6887293	chr5	67889233	4
rs3812906	chr15	97325117	5
rs2666205	chr2	9356116	5
rs12247568	chr10	74025366	5
rs504207	chr11	94060401	5
rs6708126	chr2	67852991	5

rs2853668	chr5	1300024	5
rs2442602	chr8	6390750	6
rs3093679	chr17	26664214	6
rs7010457	chr8	120955011	6
rs10854395	chr21	40349876	6
rs7843255	chr8	4531834	6
rs524534	chr11	102851924	6
rs10854395	chr21	40349876	6
rs7010457	chr8	120955011	6

Table 27 (Cont'd.)

The tree structure shows that the decision path for Latino population is more complex than the Japanese and African American populations. Unfortunately there is no significant evidence for this complexity. But overall decision path remains the same. The branches of tree first constructed by the BMI attribute. Some remarkable rules for this population can be stated as follows: If the subjects are in first category of BMI, which is <22.5, then the SNP rs17799219 with major allele AG states these subjects are controls. If the subjects are in third category of BMI, which is <29.9, then a second phenotypic attribute, family history must be examined. If these subjects have first degree relatives with prostate cancer, then SNP rs6475584 is examined. AA major allele of this SNP states that the subjects are cases. Our hybrid system found 65 SNPs related with prostate cancer. Regulome results for these SNPs are given in the Table-28 below:

Table 28: SNPs found by hybrid system for Latino Population. The list is gathered from RegulomeDB

rsid	#chromosome	Coordinate	score
rs11790106	chr9	38047709	2b
rs6774902	chr3	10515718	2b
rs12644498	chr4	57372629	3a
rs744346	chr12	106424743	4
rs4562278	chr8	128664589	4
rs17799219	chr7	51137855	5
rs197265	chr2	161902483	5
rs17363393	chr2	180006051	5
rs280986	chr18	4449083	5
rs6475584	chr9	21811026	5
rs2115101	chr19	52957923	5
rs4793790	chr17	53446278	5
rs517036	chr1	76680524	5
rs10106027	chr8	35721526	5
rs3760903	chr19	4011264	5
rs7584223	chr2	54818894	5
rs2826802	chr21	22705793	6
rs11126869	chr2	82727487	6
rs9401290	chr6	120866631	6
rs6779266	chr3	157876322	6
rs2948268	chr7	96438317	6
rs6676372	chr1	71297209	6

rs2711134	chr7	24632110	6
rs7843255	chr8	4531834	6
rs17595858	chr5	154650099	6
rs501700	chr1	100520310	6
rs6747704	chr2	77468722	6
rs17152800	chr5	124813666	6
rs10068915	chr5	84386903	6
rs7152946	chr14	54051075	6

Table 28 (Cont'd.)

When phenotypic attributes were examined eight of them are found to be related with the prostate cancer. For African American population only three attributes were used which are body mass index, alcohol consumption, and smoking behaviors. For this population, all of BMI categories and all of categories for smoking behavior indicated by packets of cigarettes smoked per day times the number of years smoked, were evaluated in tree. The alcohol consumption was evaluated with only using two categories which is non-drinkers and subjects that consume more than 24 grams of alcohol for African American population. And these consumers are found to be more likely to develop prostate cancer. Surprisingly only one phenotypic attribute, which is BMI, is used in Japanese populations. Seven phenotypic attributes used for detecting the prostate cancer in Latino population. These attributes are body mass index, family history, physical activity, smoking behavior that is indicated by current smoke and ever smoke, alcohol consumption and lycopene intake. All category ranges of these attributes were used for detecting prostate cancer disease in Latino population.

From these findings we can state that the most important phenotypic attribute for prostate cancer is the body mass index. This indicator also studied for its relations with different types of cancer. A general analysis is made by Renehan and friends, BMI found related with different cancer types in this study [246]. The relation of specific cancers with BMI, such as breast cancer [247] and esophagus [248] are also studied in the literature. BMI and prostate cancer relation is examined by the works such as [249-250-251-252]. A strong relation is found between prostate cancer and BMI in [249]. This research also states that the relation between BMI, age and family history, which are also a selected attributes by our hybrid system, is also very important when detecting the prostate cancer. In another work [250] a reverse relation was found between body mass index and prostate cancer. This works also explains why some of low BMI subjects are classified as cases in our hybrid system. The works [251-252] also shows the relation between BMI and prostate cancer. These works in the literature supports our findings that the most important phenotypic attribute in prostate cancer diagnosis is the BMI.

Another related phenotypic attribute found by our hybrid system is the family history. This attribute was used for determining the prostate cancer condition especially in the Latin population. This attribute was found in upper levels, level three, of the tree. When the specific SNP has the major allele AC then subjects that have prostate cancer in their first degree relatives has a chance to develop the disease. Relation between prostate cancer and family history is studied in the literature. Positive relations were found with the subjects who had prostate cancer in their relatives. In these works generally prostate cancer condition of the first degree relatives such as brother and father is searched and statistically significant relations were found [253]. If a subject's both father and brother had prostate cancer then the risk increases [254]. In another work a strong relation is found, in terms of family history, in the Latino population [255].

Smoking behavior is represented by 3 attributes in prostate cancer dataset. These are eversmoke, currentsmoke and packyrs_ca. Our hybrid system found the last attribute that gives packs of cigarettes smoked per day times the number of years smoked, to be related with the disease condition in African American population. In this population level three of the tree shows that three categories of packyrs_ca attribute are related with three different SNPs and together they are used for diagnosis. These relations are given below:

- If the subject's packyrs attribute is less than ten then rs10745253 is checked. If the major allele of this SNP is AA then subjects are classified as healthy. If the major allele is AG then the subjects are classified as having the disease.
- If the subject's packyrs attribute is less than thirty then rs7843255 is checked. If the major allele of this SNP is AA or GG then subjects are classified as having the disease. If the major allele is AT then the subjects are classified as healthy.
- If the subject's packyrs attribute is greater than thirty then rs2194505 is checked. If the major allele of this SNP is TT then subjects are classified as healthy. If the major allele is CT then the subjects are classified as having the disease.

Effect of smoking behavior on prostate cancer is studied in various works. For example in [256] incremental dose of smoking and prostate cancer relation was studied. As a result a relation was found especially packet years of smoking is greater than forty. Coughlin's work [257] found a relation between deaths from prostate cancer and smoking behavior in black races and also a dose relation was found in this study too. The effect of smoking behavior combined by other effects such as BMI, alcohol consumption and physical activity on prostate cancer is studied in the works such as [258-259]. In these works the same positive relations were stated.

Overall structure of our hybrid system shows that a total of 108 SNPs selected out of 2710. The list of these SNPs is given in the Appendix-E. When these SNPs are searched in ENSEMBL, one of them is found to be associated with a specific gene which is crr9,tert. This gene plays an important role in the regulation of telomerase activity and "*Overexpression of telomerase is key component of the transformation process in many malignant cancer cells*"[17].

We could state that our hybrid system can identify the functional SNPs that map to a gene. The none-coding SNPs are then searched by RegulomeDB in order to show their effect on regulation. As a result 57 SNPs were listed. Among them ten of them got high score therefore these SNPs are important in binding. The list of these SNPs is given in the Table-29 below:

Table 29: SNPs affect binding in prostate cancer dataset. SNPs are found by hybrid system.

#chromosome	rsid	score
chr12	rs1433369	2b
chr9	rs11790106	2b
chr3	rs6774902	2b
chr10	rs17701543	3a
chr4	rs12644498	3a
chr1	rs17375010	4
chr10	rs10788555	4
chr5	rs6887293	4

[17] http://www.genecards.org/cgi-bin/carddisp.pl?gene=TERT accessed on 15.12.2012

chr12	rs744346	4
chr8	rs4562278	4

Table-29 (Cont'd.)

The SNP with rs11790106 affects the regulation of ATP2B2 gene which is important for energy production and calcium transportation of the cells. rs12644498 affects regulation of ARL9 gene and rs6887293 affects the regulation of AGBL4 which are also important for ATP/GTP cycle in cells. These genes are closely related to IGF1 gene which plays an important role in insulin like growth. It seems all these genes are related with growth and energy processes which in fact could be related with BMI, the most important phenotypic attribute found by our hybrid system.

The SNPs found by hybrid model are also searched through SNPnexus and 107 unique rsIDs matched with 62 unique Entrez GeneID and 42 of them were previously found to be associated with a condition listed in Genetic Association of Complex Diseases and Disorders (GAD) database. A representative set of genes- phenotypes and disease classes is given in the Table-30 and the whole list can be found in Appendix-F material.

Table 30: SNPnexus results of Prostate Cancer .

Gene	Entrez gene	Phenotype	Disease Class	Pubmed
MCPH1	79648	Adenocarcinoma|Pancreatic Neoplasms	CANCER	19690177
MCPH1	79648	breast cancer	CANCER	20508983
SMARCA4	6597	breast cancer	CANCER	19183483
CSMD1	64478	Chromosomal Instability|Cystadenocarcinoma , Serous|Ovarian Neoplasms	CANCER	19383911
CSMD1	64478	Chromosomal Instability|Cystadenocarcinoma , Serous|Ovarian Neoplasms	CANCER	19383911
MTAP	4507	Melanoma|Nevus|Precancerous Conditions|Skin Neoplasms	CANCER	19578365
MTAP	4507	melanoma|Nevus|Skin Neoplasms	CANCER	20574843
MTAP	4507	melanoma|Nevus|Skin Neoplasms|Sunburn	CANCER	20647408
MTAP	4507	Precursor Cell Lymphoblastic Leukemia-Lymphoma	CANCER	19665068
ST6GALNAC3	256435	Alcoholism	CHEMDEPENDENCY	20421487
ANGPT2	285	BMI- Edema rosiglitazone or pioglitazone	PHARMACOGENOMIC	18996102
KLF7	8609	Body Weight|Diabetes Mellitus, Type 2|Obesity|Overweight	METABOLIC	19147600
MTAP	4507	diabetes, type 2	METABOLIC	11985785
PACRG	135138	male infertility	REPRODUCTION	19268936
SEMA5B	54437	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
CAMTA1	23261	Type 2 diabetes	METABOLIC	18210030

SEMA5B	54437	Type 2 Diabetes| edema | rosiglitazone	PHARMACO GENOMIC	20628086

Table 30 (Cont'd.)

These findings show that our SVM – ID3 Tree Hybrid system was able to identify the functional and regulatory SNPs related with prostate cancer. The number of functional and regulatory related SNPs that were found is highly depended to the dataset used, Plink Analysis and METU-SNP AHP scoring mechanism. In addition, these SNP relations were checked by using the databases such as RegulomeDB and ENSEMBL which store the information according to previously proven studies. This means that the SNPs found by our hybrid system are also candidates for further biological studies for their relation with prostate cancer.

5.2 Discussion of Melanoma Results

The melanoma dataset used in this study consists of 3115 subjects which 2053 cases and 1062 controls. There were 958 subjects with all phenotype attributes were missing so these subjects were excluded from dataset. After data preprocessing step the number of phenotypic attributes were reduced from 24 to 11 and number of genotypic attributes represented with Rsid were reduced to 2783.

In order to use this dataset in SVM Model, a polynomial to numeric data conversion was conducted. SVM part of the hybrid system gave 78.41% of accuracy on the dataset that only used genotype information and gave 70.37% of accuracy with the dataset of only phenotypes and accuracy increased to 78.6% when genotyping and phenotype information are combined.

Attribute weights of the SVM's kernel were normalized and used as coefficients for information gain ratio in ID3 decision tree part of the hybrid system. When this system was run on only genotype and only phenotype datasets, the accuracy performance was 57.12% and 75.48% respectively. When genotyping and phenotype information are combined, the accuracy performance increased to 86.35%.

In clinical point of view the gold standard for diagnosing melanoma is of course biopsy. Noninvasive methods are dermascopy and the ABCD test. Dermoscopy is the examining of skin with a device called dermoscope. And in ABCD test each letter corresponds to specific control condition as follows [260]:

- A: Asymmetry
- B: Border irregularity
- C: Color Variability
- D: Diameter

Different values for performance indicators of these tests are found in the literature. But according to out literature search so far maximum sensitivity for ABCD test is given as 91% [261-262-263]. For dermoscopy the maximum performance indicator values we came through are between 0.75-0.96 for sensitivity [264].

It is clear that these tests give better performance when compared to output of melanoma of our hybrid system with 79.07% of sensitivity. In our opinion the reason behind these performance results is the phenotypic attributes. Although our dataset contains common phenotypic attributes for melanoma such has moles, dysplastic nevi, family history etc. important clinical phenotypic attributes were excluded from dataset. These attributes, such as Clark's level and Breslow's measure, are important clinical findings that are used to diagnose melanoma. These attributes were excluded because they were given only for cases. So it was impossible to make a distinction between cases and controls based on such attributes although they are the most important ones in clinical point of view. The other important attributes that were researched in the literature are moles, dysplastic nevi, freckles and coloring features of eye, hair and skin. These attributes were included in our dataset. But moles, dysplastic nevi and freckles are binary attributes. In literature the density of these attributes were examined. Unfortunately in our dataset we had only the information about if these attributes present or not. No density or amount information was given. We believe that performance of our hybrid system could be increased by using a proper dataset on melanoma.

Resulting feature sets of our hybrid system was examined and phenotypic attribute gender was found in the root of the tree. ID3 algorithm likely to choose such attributes, divides the dataset into two even parts, as o root node. So in this case attributes below root must be examined for their relation with the given disease. The attributes moles and dysplatic_nevi were found as most important attributes, in level two and level tree of tree, in our hybrid system.

The relation of moles is well known and studied in the literature. Linda T.'s work shows a strong relation between moles and melanoma and states that the risk of having melanoma increases with the number of moles in the subjects [265]. Another work [266] also examined this incremental relation. The relation between melanoma and the structure of moles in terms of shape and size was also studied and positive relations were found [267].

Dysplastic nevus is in fact a type of mole. It differs from common observed moles with its structure. Generally this type of mole has irregular shape, their color varies and easily can be diagnosed as melanoma; but not all dysplastic nevi are melanoma although they are benign [268]. Several studies show that subjects with dysplastic nevi are likely to develop melanoma [268-269-270].

Other attributes found in tree are skin color, eye color, hair color, sunburn, tanned, freckle and family history. Although these attributes are in lower levels of the tree each of these attributes were studied according to their relations with melanoma. A study by B. Langholz found hair color and skin color as important indicators for melanoma [271]. Eye color is found to be related with eye melanoma. Work of C. M. Vajdic states that subjects with grey, hazel and blue eye color has a bigger chance to have eye melanoma than subjects with brown eye color [272]. The work of V. Beral examined all factors that are given above, among them hair color, especially red, and skin color are the factors that came forward in their relations with melanoma [273]. In melanoma studies instead of examining one particular attribute, the attributes given above are examined together.

A survey like study [274] that examined the publications on melanoma between years 2002 and 2005 has found the following relations:

- Family history that indicates the first degree relatives that have melanoma is highly related with the disease
- Subjects with high density of freckles has a higher risk to develop melanoma
- Sun exposure condition which is determined by tanning and sunburn was examined. Subjects with no sunburn and/or easy tan are in lower risk groups for developing melanoma
- Subjects with light eye color have a greater chance to have melanoma than subjects with darker colors. The same relation was found for hair color.

Some important relations that were extracted from our hybrid system can be summarized as follows:

For female subjects the most descriptive phenotype is having moles at the second level of the tree structure. If a subject has moles then it is directly classified as a case. Deeper branching of the tree structure is formed by the subjects that have no moles. At the third level subjects that have dysplastic nevi attribute are classified as a case. The female subjects that do not have moles and dysplastic nevi are classified according to particular set of SNPs and other phenotypic attributes such as skin color, eye color, hair color and freckles.

First three levels of the tree structure for male subjects are almost identical with the tree structure of female subjects. Again having moles and dysplastic nevi attributes at the second and third levels show a distinction between cases and healthy subjects, where subjects having moles or dysplastic nevi are classified as cases. In deeper levels the decision is based on particular SNPs and phenotypic attributes. Besides skin, eye and hair color and freckles for male subjects, having sunburn or being untanned and family history of melanoma are found to be nodes at the decision tree.

Our SVM-ID3 Hybrid model identified total of 53 SNPs out of 2783 SNPs selected by METU SNP tool. Among all SNPs in the model 17 SNPs found to be descriptive for females and 36 SNPS for males. The list of these SNPs is given in the Appendix-G. We have investigated the SNPs mapping to genes in the SNPnexus database and the non-coding SNPs in the RegulomeDB in order to see if they have been associated with melanoma or any other condition before. SNPnexus results show that a total of 16 unique SNPs (6 for female and 10 for male) match to a gene which are associated with a condition listed in Genetic Association of Complex Diseases and Disorders (GAD) database. The SNPs that are associated with cancer diseases are listed in Table-31 and the whole list of SNPs that are associated with a specific condition is given in the Appendix-H.

Table 31: SNPnexus results of Melanoma:

SNP	GAD Id	Entrez gene	Phenotype	Disease Class	Pubmed
rs12207699	587660	401237	Cell Transformation, Neoplastic|Neuroblastoma	CANCER	18463370
rs12207699	597635	401237	neuroblastoma	CANCER	18463370
rs17747388	147169	2066	lung cancer	CANCER	17487277
rs17747388	574563	2066	Cell Transformation, Neoplastic|Melanoma| Skin Neoplasms	CANCER	19718025

rs17747388	685880	2066	lung cancer	CANCER	20881644
rs17747388	685883	2066	Brain Neoplasms|Glioma	CANCER	20446891
rs17747388	685878	2066	lung cancer	CANCER	20975381
rs17747388	683848	2066	colorectal cancer	CANCER	18094435
rs2392695	599543	6262	Acute lymphoblastic leukemia (childhood)	CANCER	19684603
rs10051060	691268	2890	Drug Hypersensitivity|Precursor T-Cell Lymphoblastic Leukemia-Lymphoma	CANCER	20592726
rs1043848	566239	10611	prostate cancer	CANCER	19767753
rs1043848	675861	10611	prostate cancer	CANCER	20564319
rs1043848	675860	10611	prostate cancer	CANCER	20878950
rs2768343	676985	57118	prostate cancer	CANCER	20080650
rs2768343	676982	57118	breast cancer	CANCER	20418484

Table 31 (Cont'd.)

Besides the SNPs that map to a specific gene, the non-coding SNPs are also investigated by using RegulomeDB, the top scoring SNPs found here are listed in Table-32.

Table 32: High score SNPs from RegulomeDB for melanoma

rsid	hits	score
rs491322	Single_Nucleotides|C9orf52|eQTL, Chromatin_Structure|DNase-seq	1f
rs10171924	Single_Nucleotides|KLF11|eQTL, Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq	1f
rs2246095	Single_Nucleotides|ADCK4|eQTL, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|IKZF1, Protein_Binding|ChIP-seq|POLR2A	1f
rs2288704	Motifs|PWM|c-Ets-2, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|HNF4A	3a
rs10097728	Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|EBF1	4
rs12466022	Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|GATA1, Protein_Binding|ChIP-seq|SPI1	4
rs10211242	Chromatin_Structure|FAIRE, Chromatin_Structure|DNase-seq, Protein_Binding|ChIP-seq|SMARCA4, Protein_Binding|ChIP-seq|BATF, Protein_Binding|ChIP-seq|IRF4	4

These findings show that our SVM – ID3 Tree Hybrid system was able to identify the functional and regulatory SNPs related with prostate cancer. The number of functional and regulatory related SNPs that were found is highly depended to the dataset used, Plink Analysis and METU-SNP AHP scoring mechanism. In addition, these SNP relations were checked by using the databases such as RegulomeDB and SNPnexus which store the information according to previously proven studies. This means that the SNPs found by our hybrid system are also candidates for further biological studies for their relation with melanoma.

CHAPTER 6

CONCLUSION

Through GWAS, SNP profiles related with complex diseases can be discovered. The GWAS outputs are big in amount and high in dimension, also relations between SNPs, phenotypes and diseases are most likely to be non-linear. In order to handle high volume-high dimensional data and to be able to find the nonlinear relations we have utilized data mining approaches and designed a hybrid feature selection model combining support vector machine and decision tree. The designed model is tested on prostate cancer and melanoma data and for the first time combined genotype and phenotype information is used to increase the classification performance.

The overall results in this work show that combining genotypic and phenotypic information gives better performance than using only genotypes and only phenotypes. Table-33 below summarizes the results and the comparison of performances of SVM and the Hybrid Model.

Table 33: Performance comparison of SVM and Hybrid Model on Melanoma and Prostate cancer datasets. Combined dataset outperformed the only genotype and only phenotype datasets in both cases and models

		SVM			**Hybrid Model**		
		Only Gen	Only Phen	Combined	Only Gen	Only Phen	Combined
Melanoma	Accuracy	78.41	70.37	78.6	57.12	75.48	86.35
	Precision	75.52	64.32	76.45	57.77	79.54	82.27
	Recall	76.38	74.64	75.07	53.47	68.9	79.07
	AUC	0.84	0.756	0.846	0.567	0.799	0.81
Prostate CA	Accuracy	59.02	68.23	72.46	71.67	84.23	93.81
	Precision	61.29	76.8	82.68	72.69	86.20	96.55
	Recall	63.15	70.12	71.34	68.96	83.78	90.92
	AUC	0.606	0.768	0.829	0.674	0.857	0.91

The results on both prostate cancer and melanoma show that the performance increases when phenotypes and genotypes are combined we expected. Prostate Specific Antigen (PSA), which is a gold standard in prostate cancer diagnosis, is known to have a sensitivity of %86 [240]. In our study the proposed hybrid system outperformed the performance of PSA for classification of prostate cancer cases with a sensitivity of 90.92%. Although developing a diagnostic test was not the major goal of this study, with this performance level, the prostate cancer model build by the hybrid system can be utilized as an alternative tool for diagnosis and a new measure to assess the prostate cancer risk at early stages.

When the features selected in the prostate cancer model is examined, ethnicity comes forward as the main the phenotypic attribute. SVM-ID3 hybrid model separates individuals based on subject's ethnicity in the first step of the classificiation. This result is not surprising as the effect of ethnicity in prostate cancer cases is well known in the literature [244-245]. Our structure constructed different decision paths for African American, Latino and Japanese populations. Although paths differ, the most important phenotypic attribute in the model next to the ethnicity is found to be BMI. This attribute also have been reported for its relation with prostate cancer in previous studies [246-247-248-249-250-251-252].

The proposed hybrid method selected 107 unique SNPs for the diagnostic model out of 2710 highly associated SNPs determined after GWAS. When these 107 SNPs are searched in databases such as ENSEMBL, SNPnexus and RegulomeDB, some of them are found to be related with specific genes and others affect regulation and bindings. For example, rs2853668 is found to be associated with *CRR9, TERT* which plays an important role in the regulation of telomerase activity. The rs11790106 affects the regulation of *ATP2B2* gene which is important for energy production and calcium transportation of the cells. rs12644498 affects regulation of *ARL9* gene and rs6887293 affects the regulation of *AGBL4* which are also important for ATP/GTP cycle in cells. These genes are closely related to *IGF1* gene which plays an important role in insulin metabolism. Additionally many of the genes, which the 107 SNPs in the disease model map, are related with growth and energy processes, which in fact related to BMI, the second most important phenotypic attribute found by our hybrid model. These results suggest that SVM-ID3 Hybrid model was able to identify functional and regulatory SNPs that can explain molecular mechanisms involved in prostate cancer.

Although the results on melanoma show that the performance increases when phenotypes and genotypes are integrated, the performance of the hybrid system was not as good as clinical standards such as ABCD test and dermoscopy. According to our literature search so far, maximum sensitivity for ABCD test is given as 91% [261-262-263], the maximum sensitivity reported for dermoscopy between was 0.75-0.96 [264]. This is most likely due to lack of some important attributes used for melanoma diagnosis in the dataset. For example Clark's level and Breslow's measure are two important attributes that are used in diagnosing melanoma. Unfortunately these attributes were only valid for cases so no distinction could be made between cases and controls by using these attributes. Also only the presence of other attributes like moles, freckles and dysplastic nevi were provided as binary values, where as their amount and density is important features in melanoma diagnosis.

The melanoma model build with the hybrid approach is not suitable to use in clinical diagnosis but should be utilized to further understand the related genotypic and phenotypic features with the disease. Phenotypic features that are selected by our hybrid system were also studied in previous works. This shows that the hybrid system for melanoma could be used to determine risk of individuals based on their genotype and phenotype as a predictive tool rather than a diagnostic tool. The selected phenotypic and genotyping features could be used as markers for risk groups and the subjects in the risk group could further appointed to diagnostic tests such as ABCD or dermoscopy.

Resulting melanoma feature sets of our hybrid system was examined and phenotypic attribute gender was found at the root of the tree. ID3 algorithm likely to choose such attributes, divides the dataset into two even parts, as a root node. When the attributes of the melanoma model below the root were examined; moles and dysplatic_nevi were found as the most important attributes, at level two and level three of the tree. The relation of these attributes with melanoma was confirmed in the literature [265-266-267-268-269-270]. When the general structure was examined the same decision pattern was observed for both males and females.

The proposed melanoma model selected 53 unique SNPs out of 2783 highly associated SNPs determined after GWAS. When these SNPs are searched in SNPnexus and RegulomeDB six of them are found to be related with specific genes and seven of them affect regulation and binding. For example; rs2246095 is found to affect the regulation of ADCK4 which has an important role as a protein kinase, and its relation has been previously observed with melanoma [275] and also other types of cancer such as colon cancer and breast cancer [276,277]. rs2768343 affects *CAMK1D* which has an important role in Ca2+ cycle and reported for its role in various cancer studies as well as melanoma [278]. Another SNP rs239695 identified in the proposed melanoma model affects binding of *RYR2* gene, which also has a role in calcium transport. *RYR2* is also studied previously for its relation with melanoma [279,280].

The aim of this thesis was to combine genotyping data with number phenotypes including clinical information, demographic information and life style habits, downloaded from dbGaP. Both SVM-ID3 Hybrid models constructed and tested on prostate cancer and melanoma presented good classification performances and the best performance gathered by using integrated dataset of phenotypes and genotypes as in our hypothesis. Here, the details of the hybrid model build is described and comparisons between only SVM and the Hybrid Model is presented along with the comparisons between only genotype, only phenotype and integrated genotype phenotype data. Additionally, we have found important descriptive phenotypic attributes that are also studied by previous works and identified SNPs that map to specific genes. Overall, we have shown that using both genotyping data and phenotypes we can improve diagnostic performance and help build new hypothesis to further investigate the underlying reasons behind complex diseases. Decision support systems build on genotype-phenotype information as described in this study could be used as an alternative preventive or early detection system. Further studies on the proposed hybrid SVM-ID3 method and other data mining approaches for the integrative analysis of the GWAS results and phenotypic information would aid in development of other successful disease models, which would excel the translation of variant-disease association finding into the clinical setting for the development of new personalized medicine approaches.

REFERENCES

[1]- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (April 2005). "Complement Factor H Polymorphism in Age-Related Macular Degeneration". *Science* 308 (5720): 385–9. doi:10.1126/science.1109557

[2]- Genome Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. Plos Genetics February 2011 | Volume 7 | Issue 2 | e1001300. Guillaume Lettre, Cameron D. Palmer, Taylor Young, Kenechi G. Ejebe, Hooman Allayee, Emelia J. Benjamin, Franklyn Bennett, Donald W. Bowden, Aravinda Chakravarti, Al Dreisbach, Deborah N. Farlow, Aaron R. Folsom, Myriam Fornage, Terrence Forrester, Ervin Fox, Christopher A. Haiman, Jaana Hartiala, Tamara B. Harris, Stanley L. Hazen, Susan R. Heckbert, Brian E. Henderson, Joel N. Hirschhorn, Brendan J. Keating, Stephen B. Kritchevsky, Emma Larkin, Mingyao Li, Megan E. Rudock, Colin A. McKenzie, James B. Meigs, Yang A. Meng, Tom H. Mosley Jr., Anne B. Newman, Christopher H. Newton-Cheh, Dina N. Paltoo, George J. Papanicolaou, Nick Patterson, Wendy S. Post, Bruce M. Psaty, Atif N. Qasim, Liming Qu, Daniel J. Rader, Susan Redline, Muredach P. Reilly, Alexander P. Reiner, Stephen S. Rich, Jerome I. Rotter, Yongmei Liu, Peter Shrader, David S. Siscovick, W. H. Wilson Tang, Herman A. Taylor Jr., Russell P. Tracy, Ramachandran S. Vasan, Kevin M. Waters, Rainford Wilks, James G. Wilson, Richard R. Fabsitz, Stacey B. Gabriel, Sekar Kathiresan, Eric Boerwinkle

[3]- Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population. MV Prasad Linga Reddy, H Wang, S Liu, B Bode, J C Reed, R D Steed, S W Anderson, L Steed, D Hopkins and J-X She. *Genes and Immunity* 12, 208-212 (April 2011) | doi:10.1038/gene.2010.7

[4]- Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Eli A Stahl, Soumya Raychaudhuri, Elaine F Remmers, Gang Xie,Stephen Eyre, Brian P Thomson,Yonghong Li, Fina A S Kurreeman, Alexandra Zhernakova, Anne Hinks, Candace Guiducci, Robert Chen,Lars Alfredsson, Christopher I Amos, Kristin G Ardlie, BIRAC Consortium, Anne Barton, John BowesElisabeth Brouwer, Noel P Burtt, Joseph J Catanese, Jonathan Coblyn,Marieke J H Coenen, Karen H Costenbader, Lindsey A Criswell. Nature Genetics 42,508–514(2010)doi:10.1038/ng.582

[5]- Genome-wide association studies and Crohn's disease. James C. Lee, Miles Parkes. Oxford Journals Life Sciences Briefings in Functional Genomics Volume 10, Issue 2 Pp. 71-76

[6]- Genome-Wide Association and Meta-Analysis of Bipolar Disorder in Individuals of European Ancestry. Laura J. Scott, Pierandrea Muglia, Xiangyang Q. Kong, Weihua Guan, Matthew Flickinger, Ruchi Upmanyu, Federica Tozzi, Jun Z. Li, Margit Burmeister, Devin Absher, Robert C. Thompson, Clyde Francks, Fan Meng,Athos Antoniades, Audrey M. Southwick, Alan F. Schatzberg, William E. Bunney, Jack D. Barchas, Edward G. Jones, Richard Day, Keith Matthews, Peter McGuffin, John S. Strauss, James L. Kennedy, Lefkos Middleton, Allen D. Roses, Stanley J. Watson, John B. Vincent, Richard M. Myers, Ann E. Farmer, Huda Akil, Daniel K. Burns, and Michael Boehnke. PNAS May 5, 2009 vol. 106 no. 18: 7501–7506

[7]- A Genome-Wide Association Study of Hypertension and Blood Pressure in African Americans. Adebowale Adeyemo, Norman Gerry, Guanjie Chen, Alan Herbert, Ayo Doumatey, Hanxia Huang, Jie Zhou, Kerrie Lashley, Yuanxiu Chen, Michael Christman, Charles Rotimi. Plos Genetics. July 2009 | Volume 5 | Issue 7 | e1000564

[8]- Genome-wide Association Study in a High-Risk Isolate for Multiple Sclerosis Reveals Associated Variants in STAT3 Gene. Eveliina Jakkula, Virpi Leppa, Anna-Maija Sulonen, Teppo Varilo, Suvi Kallio, Anu Kemppinen, Shaun Purcell, Keijo Koivisto, Pentti Tienari, Marja-Liisa Sumelahti, Irina Elovaara, Tuula Pirttila, Mauri Reunanen, Arpo Aromaa, Annette Bang Oturai, Helle Bach Søndergaard, Hanne F. Harbo, Inger-Lise Mero, Stacey B. Gabriel, Daniel B. Mirel,Stephen L. Hauser, Ludwig Kappos, Chris Polman, Philip L. De Jager, David A. Hafler, Mark J. Daly, Aarno Palotie, Janna Saarela and Leena Peltonen. The American Journal of Human Genetics 86, 285–291, February 12, 2010

[9]- Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Meredith Yeager, Nick Orr, Richard B Hayes, Kevin B Jacobs, Peter Kraft, Sholom Wacholder, Mark J Minichiello, Paul Fearnhead, Kai Yu, Nilanjan Chatterjee, Zhaoming Wang, Robert Welch, Brian J Staats, Eugenia E Calle, Heather Spencer Feigelson, Michael J Thun, Carmen Rodriguez, Demetrius Albanes, Jarmo Virtamo, Stephanie Weinstein, Fredrick R Schumacher, Edward Giovannucci, Walter C Willett, Geraldine Cancel-Tassin, Olivier Cussenot, Antoine Valeri, Gerald L Andriole, Edward P Gelmann, Margaret Tucker, Daniela S Gerhard, Joseph F Fraumeni, Jr, Robert Hoover, David J Hunter, Stephen J Chanock, Gilles Thomas. Nature Genetics 39, 645 - 649 (2007). doi:10.1038/ng2022)

[10]- Genome-wide association studies in cancer. Douglas F. Easton, Rosalind A. Eeles. Oxford Journals Life Sciences and Medicine Human Molecular Genetics Volume 17, Issue R2 Pp. R109-R115

[11]- Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis. Meg R. Gerstenblith, Jianxin Shi, Maria Teresa Landi. Pigment Cell & Melanoma Research. Volume 23, Issue 5, pages 587–606, October 2010. DOI: 10.1111/j.1755-148X.2010.00730.x

[12]- Detection of Gene - Gene Interactions in Genome-Wide Association Studies of Human Population Data. Solomon K. Musani, Daniel Shriner, Nianjun Liu, Rui Feng, Christopher S. Coffey, Nengjun Yi, Hemant K. Tiwari, David B. Allison. Hum Hered 2007;63:67–84 DOI: 10.1159/000099179

[13]- V. Aguiar, Seoane, J. A., Freire, A., & Guo, L. , "GA-Based Data Mining Applied to Genetic Data for the Diagnosis of Complex Diseases.," *Soft Computing Methods for Practical Environment Solutions: Techniques and Studies.,* vol. In Gestal Pose, M., & Rivero Cebrián, D. (Eds.), pp. 219-239, 2010

[14]- A Framework for the Application of Decision Trees to the Analysis of SNPs Data. Linda Fiaschi, Jonathan M Garibaldi and Natalio Krasnogor. Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB '09. IEEE Symposium. P 106 – 113

[15]- Pre-eclampsia: more than pregnancy-induced hypertension. J.M Roberts, C.W.G Redman. The Lancet, Volume 341, Issue 8858, Pages 1447 - 1451, 5 June 1993.

[16]- A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis Saangyong Uhmn, Dong-Hoi Kim, Young-Woong Ko, Sungwon Cho, Jaeyoun Cheong and Jin Kim. Expert Systems, February 2009, Vol. 26, No. 1

[17]- Predictive models for subtypes of autism spectrum disorder based on single-nucleotide polymorphisms and magnetic resonance imaging. Jiao Y, Chen R, Ke X, Cheng L, Chu K, Lu Z, Herskovits EH. Advances in Medical Sciences · Vol. 56 · 2011 · pp 334-342

[18]- Identifying the Combination of Genetic Factors that Determine Susceptibility to Cervical Cancer. Jorng-Tzong Horng, K. C. Hu, Li-Cheng Wu, Hsien-Da Huang, Feng-Mao Lin, S. L. Huang, H. C. Lai, and T. Y. Chu. IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 8, NO. 1, MARCH 2004.

[19]- High throughput multiple combination extraction from large scale polymorphism data by exact tree method. Koichi Miyaki, Kazuyuki Omae, Mitsuru Murata, Norio Tanahashi, Ikuo Saito, Kiyoaki Watanabe. J Hum Genet (2004) 49:455–462

[20]- Tree-structured supervised learning and the genetics of hypertension. Jing Huang, Alfred Lin, Balasubramanian Narasimhan, Thomas Quertermous, C. Agnes Hsiung, Low-Tone Ho, John S. Grove, Michael Olivier h, Koustubh Ranade, Neil J. Risch and Richard A. Olshen. PNAS. July 20, 2004. vol. 101 no. 29. 10529–10534

[21]- A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups. Orlando Anunciação, Bruno C. Gomes, Susana Vinga, Jorge Gaspar, Arlindo L. Oliveira, José Rueff. Advances in Bioinformatics. Advances in Intelligent and Soft Computing Volume 74, 2010, pp 43-51

[22]- Tomida S, Hanai T, Koma N, Suzuki Y, Kobayashi T, Honda H: Artificial neural network predictive model for allergic disease using single nucleotide polymorphisms data. J Biosci Bioeng 2002, 93:470-478)

[23]- SNP selection at the NAT2 locus for an accurate prediction of the acetylation phenotype. Audrey Sabbagh, P.Darlu. Genet Med 2008:8(2):76–85

[24]- Artificial Neural Network Predictive Model for Allergic Disease Using Single Nucleotide Polymorphisms Data. S. Tomida, T. Hanai, N. Koma, Y. Suzuki, T. Kobayashi, H. Honda. Journal of Biosience and Bioengineering. Vol 93. No 5. P 470-478. 2002

[25]- Lucek PR, Hanke J, Reich J, Solla SA, Ott J. 1998. Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. Hum Hered 48:275–284

[26]- Lucek PR, Ott J. 1997. Neural network analysis of complex traits. Genet Epidemiol 14:1101–1106

[27]- Marinov M, Weeks D. 2001. The complexity of linkage analysis with neural networks. Hum Hered 51:169–176

[28]- Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. Yasuyuki Tomita, Shuta Tomida, Yuko Hasegawa, Yoichi Suzuki, Taro Shirakawa, Takeshi Kobayashi and Hiroyuki Honda. BMC Bioinformatics 2004, 5:120

[29]- A bayesian network approach to model local dependencies among SNPs. Raphael Mourad, Christine Sinoquet, Philippe Leray. MODGRAPH 2009 Probabilistic graphical models for integration of complex data and discovery of causal models in biology, satellite meeting of JOBIM 2009, Nantes : France (2009)

[30]- A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. Raphaël Mourad, Christine Sinoquet, Philippe Leray. *BMC Bioinformatics* 2011, 12:16 doi:10.1186/1471-2105-12-16

[31]- Identfying Genetic Interactions in Genome Wide Data Using Bayesian Networks. Xia Jiang, M. Michael Barmada, Shyam Visweswaran. Genet. Epidemiol., 34: 575–581. doi: 10.1002/gepi.20514

[32]- Effective selection of informative SNPs and classification on the HapMap genotype data. Nina Zhou and Lipo Wang. BMC Bioinformatics 2007, 8:484 doi:10.1186/1471-2105-8-484

[33]-From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. Zhi Wei, Kai Wang, Hui-Qi Qu, Haitao Zhang, Jonathan Bradfield, Cecilia Kim, Edward Frackleton, Cuiping Hou, Joseph T. Glessner, Rosetta Chiavacci, Charles Stanley, Dimitri Monos, Struan F. A. Grant, Constantin Polychronakos, Hakon Hakonarson. Plosone. October 2009. Volume 5. Issue 10. e1000678

[34]- Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. Hyo-Jeong Ban, Jee Yeon Heo, Kyung-Soo Oh and Keun-Joon Park. BMC Genetics 2010, 11:26.

[35]- Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Jennifer Listgarten, Sambasivarao Damaraju, Brett Poulin, Lillian Cook, Jennifer Dufour,Adrian Driga, John Mackey, David Wishart, Russ Greiner and Brent Zanke. Clinical cancer reseach. Vol. 10, 2725–2737, April 15, 2004

[36]- A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data. Lung-Cheng Huang, Sen-Yen Hsu and Eugene Lin. Journal of Translational Medicine 2009, 7:81 doi:10.1186/1479-5876-7-81

[37]- Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont and Yvan Saeys. Advanced Access Publication. Bioinformatics. *Vol. 26 no. 3 2010, pages 392–398 doi:10.1093/bioinformatics/btp630*

[38]- Predicting Cancer Susceptibility from SingleNucleotide Polymorphism Data: A Case Study in Multiple Myeloma. M. Waddell, D. Page and F. Zhan. KDD conference. Proceedings of the 5th international workshop on Bioinformatics. 2005

[39]- Support Vector Machine-based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes. Li-Yeh Chuang, Kuo-Chuan Wu, Hsueh-Wei Chang, and Cheng-Hong Yang. Proceedings of International Multiconference of Engineers and Computer Scientists. IMECS 2011. Vol-1

[40]- Predicting Functional Impact of Single Amino Acid Polymorphisms by Integrating Sequence and Structural Features. Mingjun Wang, Hong-Bin Shen, Tatsuya Akutsu, Jiangning Song. 2011 IEEE International Conference on Systems Biology (ISB) 978-1-4577-1666-9/11

[41]- Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D. Proceedings of the National Academy of Science. 97(1):262-267

[42]- Sensible Initialization of a Computational Evolution System Using Expert Knowledge for Epistasis Analysis in Human Genetics**.** Joshua L. Payne, Casey S. Greene, Douglas P. Hill, Jason H. Moore**.** Exploitation of Linkage Learning in Evolutionary Algorithms Evolutionary Learning and Optimization Volume 3, 2010, pp 215-226

[43]- Exploiting Expert Knowledge of Protein-Protein Interactions in a Computational Evolution System for Detecting Epistasis**.** Kristine A. Pattin, Joshua L. Payne, Douglas P. Hill, Thomas Caldwell, Jonathan M. Fisher, Jason H. Moore**.** Genetic Programming Theory and Practice VIII. Genetic and Evolutionary Computation Volume 8, 2011, pp 195-210

[44]- The GA and the GWAS: Using Genetic Algorithms to Search for Multi-locus Associations. Mooney M, Wilmot B, Bipolar Genome Study T, McWeeney S. IEEE/ACM Trans Comput Biol Bioinform. 2011 Oct 19. PMID: 22025762

[45]- Understanding the Evolutionary Process of Grammatical Evolution Neural Networks for Feature Selection in Genetic Epidemiology Alison A. Motsinger, David M. Reif, Scott M. Dudek, and Marylyn D. Ritchie. Proc IEEE Symp Comput Intell Bioinforma Comput Biol. 2006 Sep 28;2006:1-8

[46]- Grammatical Evolution Decision Trees for Detecting Gene-Gene Interactions. Sushamna Deodhar and Alison Motsinger-Reif. *BioData Mining* 2010, 3:8 doi:10.1186/1756-0381-3-8

[47]- *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*. Chapter 9 DECISION TREES. Lior Rokach,Oded Maimon.2005 Spinger. ISBN-13: 978-0-387-24435-8.

[48]- Data Mining: Concepts and Techniques. Jiawei Han, M. Kamber. 2001 Academic Press. ISBN 1-55860-489-8

[49]- Scholkopf, B., Tsuda, K., and Vert, J-P.editors (2004) Kernel Methods in Computational Biology. MIT Press series on Computational Molecular Biology

[50]- Auria, Laura; Moro, Rouslan A. (2008): Support Vector Machines (SVM) as a technique for solvency analysis, Discussion papers // German Institute for Economic Research, No. 811, http://hdl.handle.net/10419/27334

[51]- Spectral Feature Selection for Supervised and Unsupervised Learning. Zheng Zhao, Huan Liu. Department of Computer Science and Engineering, Arizona State University. Appearing in *Proceedings of the 24 th International Confer-ence on Machine Learning*, Corvallis, OR, 2007

[52]- Constraint Classification for Multiclass Classification and Ranking Sariel Har-Peled Dan Roth Dav Zimak Department of Computer Science. University of Illinois Urbana, IL 61801

[53]- Simulation metamodeling through artificial neural networks D.J. Fonseca, D.O. Navaresse, G.P. Moynihan Department of Industrial Engineering, The University of Alabama. Engineering Applications of Artificial Intelligence 16 (2003) 177–183. Received 10 June 2002; accepted 12 April 2003

[54]- On Strategies for Imbalanced Text Classi¯cation Using SVM: A Comparative Study Aixin Sun, Ee-Peng Lim, Ying Liu. Preprint submitted to Decision Support Systems. July 7, 2009

[55]- Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70,* 320-328).

[56]- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B), 36,* 111-147

[57]- k-Nearest Neighbour Classifiers. Padraig Cunningham and Sarah Jane Delany. University College Dublin, Dublin Institute of Technology. Technical Report UCD-CSI-2007-4. March 27, 2007

[58]- Comprehensive Decision Tree Models in Bioinformatics. Gregor Stiglic, Simon Kocbek, Igor Pernek, Peter Kokol. PLoS ONE 7(3): e33812. doi:10.1371/ journal.pone.0033812. Received October 31, 2011; Accepted February 17, 2012; Published March 30, 2012

[59]- Survey of Evolutionary Algorithms for Decision-Tree Induction. Rodrigo Coelho Barros, M´arcio Porto Basgalupp, Andr´e C. P. L. F. de Carvalho, and Alex A. Freitas. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 42, NO. 3, MAY 2012

[60]- Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the micro electronic age.* Edinburgh University Press

[61]- Information Gain Versus Gain Ratio: A Study of Split Method Biases. Earl Harris Jr. The MITRE Corporation. Technical Report. October 2001

[62]- Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos. Morgan Kaufmann Publishers. ISBN:1-55860-238-0. 1993

[63]- Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8

[64]- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29 (2), 119-127

[65]- Loh W.Y. and Shih X. Split selection methods for classification trees. Statistica Sinica, 7: 815-840, 1997.

[66]- A complete fuzzy decision tree technique. Cristina Olaru, Louis Wehenkel. Fuzzy Sets and Systems. Volume 138 Issue 2, September 1, 2003. P 221-254.

[67]- Random Forests. LEO BREIMAN *Statistics Department, University of California.* Machine Learning, 45, 5–32, 2001.Kluwer Academic Publishers. Manufactured in The Netherlands.

[68]- Bayesian Networks: A model of Self Activated Memory for Evidential Reasoning. Judea Pearl. Computer Science Deparment, University of Califormia. April 1985. Report no CSD-850017. Submitted to seventh annual conference of the Cognitive Sicence Society. 15-17 August 1985.

[69]- Learning Bayesian Belief Network Classifiers: Algorithms and System. Jie Cheng, Russell Greiner. Proceedings of 14 th Biennial conference. 2001. P 141-151

[70]- Heckerman, D. (1995). A tutorial on learning Bayesian networks. Communications of the ACM. *Tech Report MSR-TR-95-06*, Microsoft Research. Vol.38

[71]- Cooper, G.F. and Herskovits, E. (1992). A Bayesian Method for the induction of probabilistic networks from data. *Machine Learning*, 9. pp. 309-347

[72]- Cheng, J., Bell, D.A. and Liu, W. (1997). An algorithm for Bayesian belief network construction from data. In *Proceedings of AI & STAT'97* (pp.83-90)

[73]- Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction, and Search*. MIT Press; 2nd Revised edition edition (16 Mar 2001) ISBN-10: 0262194406

[74]- Chow, C.K. and Liu, C.N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans.on Information Theory*, 14. pp. 462-467

[75]- Introduction to Artificial Neural Network (ANN) Methods: What They Are and How to Use Them. Jure Zupan, Department of Chemistry, University Rovira.I. Virgili, Tarragona, Spain. Acta Chimica Slovenica 41/3/1994, pp. 327-352

[76]- An introduction to artificial neural networks in bioinformatic application to complex microarray and mass spectrometry datasets in cancer studies. Lee J. Lancashire, Christophe Lemetre and Graham R. Ball. BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 3. 315-329. Advance Access publication March 23, 2009

[77]- Comparison of Approaches for Machine-Learning Optimization of Neural Networks for Detecting Gene-Gene Interactions in Genetic Epidemiology Alison A. Motsinger-Reif, Scott M. Dudek, Lance W. Hahn and Marylyn D. Ritchie. Genet Epidemiol. 2008 May;32(4):325-40.

[78]- E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. 5th Annual ACM Workshop on COLT, pages 144 152. 1992. ACM Press.

[79]- Kernel-based Methods and Function Approximation. BAUDAT *G . , ANOUAR F. MEI,* International Joint Conference on Neural Networks, pp. 1244 - 1249. 2001

[80]- IEEE Transactions on Neural Networks, Vol. 12 No. 2 March 2005. An Introduction to Kernel-Based Learning Algorthms. Klaus Robert Müller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, Bernhard Schölkopf.

[81]- Lin, H. and C. Lin: 2003, 'A Study on Sigmoid Kernels for SVM and the Training of non- PSD Kernels by SMO-type Methods'. Technical report, Dept. of Computer Science and Information Engineering, National Taiwan University

[82]- On the Kernel Widths in Radial-Basis Function Networks. NABIL BENOUDJIT and MICHEL VERLEYSEN Universite´ Catholique de Louvain, Microelectronics Laboratoy. Neural ProcessingLetters 18: 139–154, 2003. 2003 Kluwer Academic Publishers. Printed in the Netherlands.

[83]- Park, J. and Sandberg, I. W.: Universal approximation using radial-basis-function networks. Neural Comput. 3 (1991), 246–257

[84]- Kernel Methods in Machine Learning by Thomas Hofmann, Bernhard Scholkopf, Alexander J. Smola. Darmstadt University of Technology, Max Planck Institute for biological Cybernetics and National ICT Australia. The Annals of statistics 2008, Vol 36 no 3, 1171-1220

[85]- Introduction to Support Vector Machines Dustin Boswell August 6, 2002

[86]- A practical guide to support vector classification. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. Department of Computer Science. National Taiwan University. 2010. Advances in Pattern Recognition Support Vector Machines for Pattern Classification Shiqeo Abe 2nd edition. Springler. ISSN 1617-7916. P 58-59

[87]- Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Kuo-PingWu, Sheng-DeWang. *Elsevier* Pattern Recognition. Volume 42, Issue 5, May 2009, Pages 710–717

[88]- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B), 36,* 111-147.

[89]- http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29 accessed on 29.09.2012

[90]- Parameter selection for support vector machines Carl Staelin, *Senior Member IEEE.* HP Laboratories Israel, HPL-2002-354 (R.1) November 10th , 2003. Technion City, Haifa, 32000, Israel ã Copyright Hewlett-Packard Company 2002

[91]- Evolutionary tuning of multiple SVM parameters. Frauke Friedrichs, Christian Igel. Institut fur Neuroinformatik, Ruhr-Universitat Bochum, 44780 Bochum, Germany. Elsevier Neurocomputing 64 (2005) 107–117 Available online 7 January 2005.

[92]- S.A. Rojas, D. Fernandez-Reyes, Adapting multiple kernel parameters for support vector machines using genetic algorithms, The 2005 IEEE Congress on Evolutionary Computation, vol. 1, 02–05 September, 2005, pp. 626–631

[93]- Genetic Algorithms for Multi Criterion Classification and Clustering in Data Mining. Satchidanada Dehuri, Ashish Ghosh, Rajib Mall. International Journal of Computing & Information Sciences. Vol. 4, No. 3, December 2006.

[94]- Baker, J.E. (1985). Adaptive selection methods for genetic algorithms. Proceedings of an International conference on Genetic Algorithms and Their Applications, 110-111.

[95]- A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. David E. Goldberg , Kalyanmoy Deb. Foundations of Genetic Algorithms Conference. 1991. Pages 69-93

[96]- W. M. Spears and V. Anand, "A study of crossover operators in genetic programming," in *Proc. 6th Int. Symp. Methodologies for Intelligent Systems (ISMIS'91)*, Z. W. Ras and M. Zemankova, Eds. Berlin, Germany: Springer-Verlag, 1991, pp. 409–418.

[97]- Prediction For Traffic Accident Severity: Comparing The Artificial Neural Network, Genetic Algorithm, Combined Genetic Algorithm And Pattern Search Methods: Mehmet Metin Kunt, Iman Aghayan, Nima Noii. Taylor & Francis Group. TRANSPORT ISSN 1648-4142 print 2011 Volume 26(4): 353–366

[98]- An Application of Genetic Algorithm Search Techniques to the Future Total Exergy Input/Output Estimation. *Energy Sources, Part A*, 28:715–725, 2006.

[99]- Simulated annealing based parallel genetic algorithm for facility layout problem. Meei-Yuh Kuab, Michael H. Hub and Ming-Jaan Wang. International Journal of Production Research Vol. 49, No. 6, 15 March 2011, 1801–1812

[100]- Evaluation of Edge Configuration in Medical Echo Images Using Genetic Algorithms. Ching-Fen Jiang. World Academy of Science, Engineering and Technology. 43. 2010

[101]- Forward–Backward Time-Stepping Method Combined with Genetic Algorithm Applied to Breast Cancer Detection. Toshifumi Moriyama, Zhiqi Meng, and Takashi Takenaka. MICROWAVE AND OPTICAL TECHNOLOGY LETTERS / Vol. 53, No. 2, February 2011

[102]- A Hybrid Automatic System for the Diagnosis of Lung Cancer Based on Genetic Algorithm and Fuzzy Extreme Learning Machines. Daliri, Mohammad. Journal of Medical Systems; Apr2012, Vol. 36 Issue 2, p1001-1005

[103]- Robot-Assisted Stroke Rehabilitation: Joint Torque/Force Conversion From Emg Using Ga Process. S. Parasuraman, Arif Wicaksono Oyong. Journal of Mechanics in Medicine and Biology Vol. 11, No. 4 (2011) 827–843

[104]- Introduction to partitioning based clustering methods with a robust example. Sami Ayramo Tommi Karkkainen. Reports of the Department of Mathematical Information Technology Series C. Software and Computational Engineering. No. C. 1/2006. ISBN 951-39-2467-X, ISSN 14564-378.

[105]- Least squares quantization in PCM. Sttuart P. Lloyd. IEEE Transactions On Information Theory,Vol. It-28, No. 2, March 1982.

[106]- Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. T. Velmurugan and T. Santhanam. Journal of Computer Science 6 (3): 363-368, 2010

[107]- CLUSTER ANALYSIS. Steven M. Holland. Jan 2006. Department of Geology, University of Georgia, Athens, GA 30602-2501

[108]- Density Based Clustering. Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek. Wiley Interdisciplinary reviews: Data Mining and Knowledge Discovery 1 (3): 231–240. June 2011

[109]- OPTICS: Ordering Points to Identify the Clustering Structure. Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander. Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

[110]- An Efficient Approach to Clustering in Large Multimedia Databases with Noise. Alexander Hinneburg , Er Hinneburg , Daniel A. Keim. Proceedings of the 4th International Conference on Knowledge Discovery and Datamining, 1998

[111]- A Survey of Grid Based Clustering Algorithms. MR ILANGO, Dr V MOHAN. International Journal of Engineering Science and Technology Vol. 2(8), 2010, 3441-3446

[112]- A Grid-based Clustering Algorithm using Adaptive Mesh Refinement. Wei-keng Liao, Ying Liu, Alok Choudhary. Appears in the 7th Workshop on Mining Scientific and Engineering Datasets 2004

[113]- A comparison between conceptual clustering and conventional clustering**.** Srivastava, Anurag and Murty, MN (1990) Pattern Recognition, 23 (9). pp. 975-981

[114]- MKNN: Modified K-Nearest Neighbor. Hamid Parvin, Hosein Alizadeh and Behrouz Minaei-Bidgoli. Proceedings of the World Congress on Engineering and Computer Science 2008.

[115]- Data Mining with Decision Trees. Theory and Applications. Lior Rokach, Oded Maimon. Series in machine perception artificial intelligence. Volume 69. Published by world scientific publishing. ISBN: 13 978-981-277-171-1

[116]- BAYESIAN BELIEF NETWORKS. Scott Wooldridge. Australian Institute of Marine Science. Prepared for CSIRO Centre for Complex Systems Science CSIRO 2003.

[117]- Scaling Up the Accuracy of Naive Bayes Classifiers: A Decision Tree Hybrid. Ron Kohavi. PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. Conference. 1996. P 202-207

[118]- Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. EcolModell 2004;178: 389–97

[119]- Silva A, Cortez P, Santos MF, Gomes L, Neves J. Rating organ failure via adverse events using data mining in the intensive care unit. Artif IntellMed 2008;43:179–93

[120]- Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. EcolModell 2003;160:249–64

[121]- Comparison of Genetic Algorithm and Quantum Genetic Algorithm. Zakaria Laboudi and Salim Chikhi. The International Arab Journal of Information Technology, Vol. 9, No. 3, May 2012

[122]- LEARNING SVM WITH COMPLEX MULTIPLE KERNELS EVOLVED BY GENETIC PROGRAMMING. LAURA DIOSAN, ALEXANDRINA ROGOZAN and JEAN-PIERRE PECUCHET. International Journal on Artificial Intelligence Tools Vol. 19, No. 5 (2010) 647–677)

[123]- Neural Network and Genetic Algorithm Hybrid Model for Modeling Exchange Rate: The Dollar – Kuwaiti Dinar Case. Mustapha Djennas, Mohamed BENBOUZIANE, Meriem Djennas. International Research Journal of Finance and Economics ISSN 1450-2887 Issue 92. 2012

[124]- Cascade of genetic algorithm and decision tree for cancer classification on gene expression data. Jinn-Yi Yeh, and Tai-Hsi Wu.Expert systems. The Journal of Knowledge Engineering. July 2010, Vol. 27, No. 3.

[125]- Understanding the Human Genome. Douglas L. Brutlag, Scientific American: Introduction to Molecular Medicine. 1994 (pp. 153- 168).

[126]- Coghill, Anne M.; Garson, Lorrin R., ed. (2006). The ACS style guide: effective communication of scientific information (3rd ed.). Washington, D.C.: American ChemicalSociety. p. 244. ISBN 978-0-8412-3999-9

[127]- Jankowski JAZ, Polak JM (1996). Clinical gene analysis and manipulation: Tools, techniques and troubleshooting. Cambridge University Press. p. 14. ISBN 0-521-47896-0. OCLC 33838261

[128]- Yakovchuk P, Protozanova E, Frank-Kamenetskii MD (2006). "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix". Nucleic Acids Res. 34 (2): 564–74. doi:10.1093/nar/gkj454

[129]- Gilbert W (February 1978). "Why genes in pieces?". Nature 271 (5645): 501. doi:10.1038/271501a0

[130]- Rearick D, Prakash A, McSweeny A, Shepard SS, Fedorova L, Fedorov A (March 2011). "Critical association of ncRNA with introns". Nucleic Acids Res. 39 (6): 2357–66. doi:10.1093/nar/gkq1080

[131]- The Human Genome Project: Lessons from Large-Scale Biology Francis S. Collins, Michael Morgan, Aristides Patrinos. Science 300, 286 (2003); DOI: 10.1126

[132]-www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml accessed on 10.10.2012

[133]- http://www.genome.gov/25019925 accessed on 10.10.2012

[134]- Genomes, 2nd edition**.** Chapter 5. Terence A Brown. Oxford: Wiley-Liss; 2002. ISBN-10: 0-471-25046-5

[135]- Bioinformatics: A Primer. P. Narayanan. New Age Internationla (P) Limited Publishers. ISBN: 81-224-1610-1. P 28.

[136]- A review on SNP and other types of molecular markers and their use in animal genetics. Alain VIGNAL, Denis MILAN, Magali SANCRISTOBAL, André EGGEN. Genet. Sel. Evol. 34 (2002) 275-305

[137]- Finding the missing heritability of complex diseases. Teri A. Manolio, Francis S. Collins. Nature 461, 747-753 (8 October 2009)

[138]- Genomewide Association Studies and Human Disease. John Hardy and Andrew Singleton. N Engl J Med 2009; 360:1759-1768

[139]- Genotypic analysis of gene expression in the dissection of the aetiology of complex neurological and psychiatric diseases. Mina Ryten, Danyah Trabzuni and John Hardy. Oxford Journals Life Sciences Briefings in Functional Genomics Volume 8, Issue 3 Pp. 194-198

[140]- Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics. William E. Evans, Mary V. Relling. Science 15 October 1999; Vol. 286 No. 5439 pp. 487-491

[141]- How to Interpret a Genome-wide Association Study. Thomas A. Pearson; Teri A. Manolio. JAMA. 2008;299(11):1335-134

[142]- Concato J, Feinstein AR, Holford TR: The risk of determining risk with multivariable models. Ann Intern Med 1993; 118: 201– 210

[143]- Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49: 1373–1379

[144]- Nelson MR, Kardia SL, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 2001; 11: 458–470

[145]- Ritchie MD, Hahn LW, and Moore JH: Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003; 24: 150–157

[146]- International HapMap Project web site.
http://hapmap.ncbi.nlm.nih.gov/thehapmap.html.en accessed on 20.10.2012

[147]- The International HapMap Project. The International HapMap Consortium. NATURE. VOL 426-18/25 DECEMBER 2003

[148]- A second generation human haplotype map of over 3.1 million SNPs. The International HapMap Consortium. Nature Vol 449-18. October 2007

[149]- A haplotype map of the human genome. The International HapMap Consortium. Nature. Vol 437-27. October 2005

[150]- dbSNP web site.
http://www.ncbi.nlm.nih.gov/projects/SNP/get_html.cgi?whichHtml=overview aceessed on 20.10.2012

[151]- Regulome web site. http://regulome.stanford.edu/about accessed on 20.10.2012

[152]- The NFI-Regulome Database: A tool for annotation and analysis of control regions of genes regulated by Nuclear Factor I transcription factors. Richard M Gronostajski, Joseph Guaneri, Dong Hyun Lee, Steven M Gallo. Journal of Clinical Bioinformatics 2011, 1:4

[153]- The NCBI dbGaP database of genotypes and phenotypes. Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura. Nature Genetics 39, 1181 - 1186 (2007)

[154]- A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. Mar Matarín, W Mark Brown. Lancet Neurol 2007; 6: 414–20

[155]- A genome-wide association study of serum uric acid in African Americans. Bashira A Charles, Daniel Shriner. BMC Medical Genomics 2011, 4:17

[156]- Causal Graph-Based Analysis of Genome-Wide Association Data in Rheumatoid. Arthritis. Alexander V. Alekseyenko, Nikita I Lytkin, Jizhuo Ai. *Biology Direct* 2011, 6:25

[157]- Genome-wide genotyping in Parkinson's disease and neurologically normal controls: fi rst stage analysis and public release of data. Hon-Chung Fung, Sonja Scholz, Mar Matarin, Javier Simón-Sánchez. Lancet Neurol 2006; 5: 911–16

[158]- Cell Adhesion Molecules Contribute to Alzheimer's disease: Multiple Pathway Analysis of two GWAS. Guiyou Liu, Youngshuai Jiang, Ping Wang. JOURNAL OF NEUROCHEMISTRY 2012. 120. 190–198

[159]- Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behçet's disease. E. F. Remmers e.t al. Nature Genetics. Vol 42. Number 8. 2010

[160]- Genome-wide association studies in type 1 diabetes, inflammatory bowel disease and other immune-mediated disorders. Hakon Hakonarson, Struan F.A. Grant. Seminars in Immunology 21 (2009) 355–362

[161]- A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population.Nature Genetics volume 41. Number 12. December 2009. UK IBD Genetics Consortium, Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H, Nimmo ER, Massey D, Blaszczyk K, Elliott T, Cotterill L, Dallal H, Lobo AJ, Mowat C, Sanderson JD, Jewell DP, Newman WG, Edwards C, Ahmad T, Mansfield JC, Satsangi J, Parkes M, Mathew CG; Wellcome Trust Case Control Consortium 2, Donnelly P, Peltonen L, Blackwell JM, Bramon E, Brown MA, Casas JP, Corvin A, Craddock N, Deloukas P, Duncanson A, Jankowski J, Markus HS, Mathew CG, McCarthy MI, Palmer CN, Plomin R, Rautanen A, Sawcer SJ, Samani N, Trembath RC, Viswanathan AC, Wood N, Spencer CC, Barrett JC, Bellenguez C, Davison D, Freeman C, Strange A, Donnelly P, Langford C, Hunt SE, Edkins S, Gwilliam R, Blackburn H, Bumpstead SJ, Dronov S, Gillman M, Gray E, Hammond N, Jayakumar A, McCann OT, Liddle J, Perez ML, Potter SC, Ravindrarajah R, Ricketts M, Waller M, Weston P, Widaa S, Whittaker P, Deloukas P, Peltonen L, Mathew CG, Blackwell JM, Brown MA, Corvin A, McCarthy MI, Spencer CC, Attwood AP, Stephens J, Sambrook J, Ouwehand WH, McArdle WL, Ring SM, Strachan DP

[162]- Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. Mousheng Xu, Kelan G Tantisira. BMC Medical Genetics 2011, 12:90

[163]- Genome-wide association study of skin complex diseases. Xuejun Zhang. Journal of Dermatological Science 66 (2012) 89–97

[164]- HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* 314, 989–992 (2006). Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, Tam PO, Chan WM, Lam DS, Snyder M, Barnstable C, Pang CP, Hoh J

[165]- Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (*LIPC*). *Proc. Natl. Acad. Sci. USA* 107, 7395–7400 (2010). Neale BM, Fagerness J, Reynolds R, Sobrin L, Parker M, Raychaudhuri S, Tan PL, Oh EC, Merriam JE, Souied E, Bernstein PS, Li B, Frederick JM, Zhang K, Brantley MA Jr, Lee AY, Zack DJ, Campochiaro B, Campochiaro P, Ripke S, Smith RT, Barile GR, Katsanis N, Allikmets R, Daly MJ, Seddon JM

[166]- Genome-wide association studiesin multiple sclerosis: lessons and future prospects. Anu Kemppinen, Stephen Sawcer and Alastair Compston. BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 10. NO 2. 61-70

[167]- Genome-Wide Association Studies in Myocardial Infarction and Coronary Artery Disease. Pier Mannuccio Mannucci, Luca A Lotta, Flora Peyvandi. The Journal of Tehran University Heart Center. (2010) 116-121

[168]- Network-Assisted Investigation of Combined Causal Signals from Genome-Wide Association Studies in Schizophrenia. PLoS Comput Biol 8(7): e1002587. doi:10.1371/journal.pcbi.1002587. 2012. Jia P, Wang L, Fanous AH, Pato CN, Edwards TL; International Schizophrenia Consortium, Zhao Z

[169]- Genome-wide association studies and type 2 diabetes. Eleanor Wheeler and Ines Barroso. BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 10. NO 2. 52- 60

[170]- A genome-wide search for loci interacting with known prostate cancer risk-associated genetic variants. Tao S, Wang Z, Feng J, Hsu FC, Jin G, Kim ST, Zhang Z, Gronberg H, Zheng LS, Isaacs WB, Xu J, Sun J. . Advanced Access publication. Carcinogenesis vol.33 no.3 pp.598–603, 2012

[171]- Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. Nature Genetics. VOLUME 43. NUMBER 6. JUNE 2011. Christopher A. Haiman, Gary K. Chen, William J. Blot, Sara S. Strom, Sonja I. Berndt, Rick A. Kittles, Benjamin A. Rybicki, William Isaacs, Sue A. Ingles, Janet L. Stanford, W. Ryan Diver, John S. Witte, Ann W. Hsing, Barbara Nemesure, Timothy R. Rebbeck, Kathleen A. Cooney, Jianfeng Xu, Adam S. Kibel, Jennifer J. Hu, Esther M. John, Serigne M. Gueye, Stephen Watya, Lisa B. Signorello, Richard B. Hayes, Zhaoming Wang, Edward Yeboah, Yao Tettey, Qiuyin Cai, Suzanne Kolb, Elaine A. Ostrander, Charnita Zeigler-Johnson, Yuko Yamamura, Christine Neslund-Dudas, Jennifer Haslag-Minoff, William Wu, Venetta Thomas, Glenn O. Allen, Adam Murphy, Bao-Li Chang, S. Lilly Zheng, M. Cristina Leske, Suh-Yuh Wu, Anna M. Ray, Anselm JM Hennis, Michael J. Thun, John Carpten, Graham Casey, Erin N. Carter, Edder R. Duarte, Lucy Y. Xia, Xin Sheng, Peggy Wan, Loreall C. Pooler, Iona Cheng, Kristine R. Monroe, Fredrick Schumacher, Loic Le Marchand, Laurence N. Kolonel, Stephen J. Chanock, David Van Den Berg, Daniel O. Stram and Brian E. Henderson

[172]- Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population.Takata R, Akamatsu S, Kubo M, Takahashi A, Hosono N, Kawaguchi T, Tsunoda T, Inazawa J, Kamatani N, Ogawa O, Fujioka T, Nakamura Y, Nakagawa H.. Nature Genetics VOLUME 42. NUMBER 9. SEPTEMBER 2010

[173]- Genome-wide Association Study Identifies a Genetic Variant Associated with Risk for More Aggressive Prostate Cancer. Liesel M. FitzGerald, Erika M. Kwon, Matthew P. Conomos, Suzanne Kolb, Sarah K. Holt, David Levine, Ziding Feng, Elaine A. Ostrander, and Janet L. Stanford. American Association for Cancer Research. Cancer Epidemiol Biomarkers Prev; 20(6) June 2011

[174]- Genome-wide Association Study of Prostate Cancer Mortality. Penney KL, Pyne S, Schumacher FR, Sinnott JA, Mucci LA, Kraft PL, Ma J, Oh WK, Kurth T, Kantoff PW, Giovannucci EL, Stampfer MJ, Hunter DJ, Freedman ML. . Cancer Epidemiol Biomarkers Prev 2010;19:2869-2876..

[175]- Large-scale Exploration of Gene–Gene Interactions in Prostate Cancer Using a Multistage Genome-wide Association Study. Ciampa J, Yeager M, Amundadottir L, Jacobs K, Kraft P, Chung C, Wacholder S, Yu K, Wheeler W, Thun MJ, Divers WR, Gapstur S, Albanes D, Virtamo J, Weinstein S, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Hunter D, Hoover R, Thomas G, Chanock S, Chatterjee N. American Association for Cancer Research. Cancer Res 2011;71:3287-3295.

[176]- Drugs in Clinical Development for Melanoma. Pharm Med 26 (3): 171-183. 2012.

[177]- Jerant AF, Johnson JT, Sheridan CD, Caffrey TJ (July 2000). Early detection and treatment of skin cancer. Am Fam Physician 62 (2): 357–68, 375–6, 381–2. PMID 10929700

[178]- Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S; NISC Comparative Sequencing Program, Stemke-Hale K, Davies MA, Gershenwald JE, Robinson W, Robinson S, Rosenberg SA, Samuels Y. Nature Genetics. VOLUME 43. NUMBER 5. may 2011

[179]- Genome-wide screen of promoter methylation identifies novel markers in melanoma. Yasuo Koga, Mattia Pelizzola, Elaine Cheng, Michael Krauthammer, Mario Sznol, Stephan Ariyan, Deepak Narayan, Annette M. Molinaro, Ruth Halaban and Sherman M. Weissman. Genome Res. 2009 19: 1462-1470)

[180]- Genome-wide association study identifies a new melanoma susceptibility locus at 1q21.3. Nature Genetics. VOLUME 43. NUMBER 11. NOVEMBER 2011. Macgregor S, Montgomery GW, Liu JZ, Zhao ZZ, Henders AK, Stark M, Schmid H, Holland EA, Duffy DL, Zhang M, Painter JN, Nyholt DR, Maskiell JA, Jetann J, Ferguson M, Cust AE, Jenkins MA, Whiteman DC, Olsson H, Puig S, Bianchi-Scarrà G, Hansson J, Demenais F, Landi MT, Dębniak T, Mackie R, Azizi E, Bressac-de Paillerets B, Goldstein AM, Kanetsky PA, Gruis NA, Elder DE, Newton-Bishop JA, Bishop DT, Iles MM, Helsing P, Amos CI, Wei Q, Wang LE, Lee JE, Qureshi AA, Kefford RF, Giles GG, Armstrong BK, Aitken JF, Han J, Hopper JL, Trent JM, Brown KM, Martin NG, Mann GJ, Hayward NK

[181]- Genome-wide association study identifies nidogen 1 (NID1) as a susceptibility locus to cutaneous nevi and melanoma risk. Nan H, Xu M, Zhang J, Zhang M, Kraft P, Qureshi AA, Chen C, Guo Q, Hu FB, Rimm EB, Curhan G, Song Y, Amos CI, Wang LE, Lee JE, Wei Q, Hunter DJ, Han J. Human Molecular Genetics, 2011, Vol. 20, No. 13 2673–2679.

[182]- Genome-wide association study identifies three new melanoma susceptibility loci. Nature Genetics. VOLUME 43. NUMBER 11. NOVEMBER 2011. Barrett JH, Iles MM, Harland M, Taylor JC, Aitken JF, Andresen PA, Akslen LA, Armstrong BK, Avril MF, Azizi E, Bakker B, Bergman W, Bianchi-Scarrà G, Bressac-de Paillerets B, Calista D, Cannon-Albright LA, Corda E, Cust AE, Dębniak T, Duffy D, Dunning AM, Easton DF, Friedman E, Galan P, Ghiorzo P, Giles GG, Hansson J, Hocevar M, Höiom V, Hopper JL, Ingvar C, Janssen B, Jenkins MA, Jönsson G, Kefford RF, Landi G, Landi MT, Lang J, Lubiński J, Mackie R, Malvehy J, Martin NG, Molven A, Montgomery GW, van Nieuwpoort FA, Novakovic S, Olsson H, Pastorino L, Puig S, Puig-Butille JA, Randerson-Moor J, Snowden H, Tuominen R, Van Belle P, van der Stoep N, Whiteman DC, Zelenika D, Han J, Fang S, Lee JE, Wei Q, Lathrop GM, Gillanders EM, Brown KM, Goldstein AM,

Kanetsky PA, Mann GJ, Macgregor S, Elder DE, Amos CI, Hayward NK, Gruis NA, Demenais F, Bishop JA, Bishop DT; GenoMEL Consortium

[183]- Pathway-Based Analysis of a Melanoma Genome-Wide Association Study: Analysis of Genes Related to Tumour- Immunosuppression. Nils Schoof, Mark M. Iles, D. Timothy Bishop, Julia A. Newton-Bishop, Jennifer H. Barrett, GenoMEL consortium. Plosone. December 2011 | Volume 6 | Issue 12 | e29451

[184]- A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Rothman et al. Nature Genetics. VOLUME 42. NUMBER 11. NOVEMBER 2010

[185]- Large-Scale Pathway-Based Analysis of Bladder Cancer Genome-Wide Association Data from Five Studies of European Background. Plosone. January 2012 | Volume 7 | Issue 1 | e29396. Menashe I, Figueroa JD, Garcia-Closas M, Chatterjee N, Malats N, Picornell A, Maeder D, Yang Q, Prokunina-Olsson L, Wang Z, Real FX, Jacobs KB, Baris D, Thun M, Albanes D, Purdue MP, Kogevinas M, Hutchinson A, Fu YP, Tang W, Burdette L, Tardón A, Serra C, Carrato A, García-Closas R, Lloreta J, Johnson A, Schwenn M, Schned A, Andriole G Jr, Black A, Jacobs EJ, Diver RW, Gapstur SM, Weinstein SJ, Virtamo J, Caporaso NE, Landi MT, Fraumeni JF Jr, Chanock SJ, Silverman DT, Rothman N

[186]- Cancer Peptide Vaccine Therapy Developed from Oncoantigens Identified through Genome-wide Expression Profile Analysis for Bladder Cancer. Obara W, Ohsawa R, Kanehira M, Takata R, Tsunoda T, Yoshida K, Takeda K, Katagiri T, Nakamura Y, Fujioka T. Jpn J Clin Oncol 2012;42(7)591–600

[187]- Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41: 324–328.2009. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL, Gu K, Fair AM, Cai Q, Lu W, Shu XO

[188]- Ancestry-shift refinement mapping of the C6orf97-ESR1 breast cancer susceptibility locus. PLoS Genet 6: e1001029. 2010. Stacey SN, Sulem P, Zanon C, Gudjonsson SA, Thorleifsson G, Helgason A, Jonasdottir A, Besenbacher S, Kostic JP, Fackenthal JD, Huo D, Adebamowo C, Ogundiran T, Olson JE, Fredericksen ZS, Wang X, Look MP, Sieuwerts AM, Martens JW, Pajares I, Garcia-Prats MD, Ramon-Cajal JM, de Juan A, Panadero A, Ortega E, Aben KK, Vermeulen SH, Asadzadeh F, van Engelenburg KC, Margolin S, Shen CY, Wu PE, Försti A, Lenner P, Henriksson R, Johansson R, Enquist K, Hallmans G, Jonsson T, Sigurdsson H, Alexiusdottir K, Gudmundsson J, Sigurdsson A, Frigge ML, Gudmundsson L, Kristjansson K, Halldorsson BV, Styrkarsdottir U, Gulcher JR, Hemminki K, Lindblom A, Kiemeney LA, Mayordomo JI, Foekens JA, Couch FJ, Olopade OI, Gudbjartsson DF, Thorsteinsdottir U, Rafnar T, Johannsson OT, Stefansson K

[189]- Comparison of 6q25 Breast Cancer Hits from Asian and European Genome Wide Association Studies in the Breast Cancer Association Consortium (BCAC). Plosone. August 2012. Volume 7. Issue 8. e42380. Hein R, Maranian M, Hopper JL, Kapuscinski MK, Southey MC, Park DJ, Schmidt MK, Broeks A, Hogervorst FB, Bueno-de-Mesquit HB, Muir KR, Lophatananon A, Rattanamongkongul S, Puttawibul P, Fasching PA, Hein A, Ekici AB, Beckmann MW, Fletcher O, Johnson N, dos Santos Silva I, Peto J, Sawyer E, Tomlinson I, Kerin M, Miller N, Marmee F, Schneeweiss A, Sohn C, Burwinkel B, Guénel P, Cordina-Duverger E, Menegaux F, Truong T, Bojesen SE, Nordestgaard BG, Flyger H, Milne RL, Perez JI, Zamora MP, Benítez J, Anton-Culver H, Ziogas A, Bernstein L, Clarke CA, Brenner H, Müller H, Arndt V, Stegmaier C, Rahman N, Seal S, Turnbull C, Renwick A, Meindl A, Schott S, Bartram CR, Schmutzler RK, Brauch H, Hamann U, Ko YD; GENICA Network, Wang-Gohrke S, Dörk T, Schürmann P, Karstens JH, Hillemanns P, Nevanlinna H, Heikkinen T, Aittomäki K, Blomqvist C, Bogdanova NV, Zalutsky IV, Antonenkova NN, Bermisheva M, Prokovieva D, Farahtdinova A, Khusnutdinova E, Lindblom A, Margolin S, Mannermaa A, Kataja V, Kosma VM, Hartikainen J, Chen X, Beesley J; Kconfab Investigators; AOCS Group, Lambrechts D, Zhao H, Neven P, Wildiers H, Nickels S, Flesch-Janys D, Radice P, Peterlongo P, Manoukian S, Barile M, Couch FJ, Olson JE, Wang X, Fredericksen Z, Giles GG, Baglietto L, McLean CA, Severi G, Offit K, Robson M, Gaudet MM, Vijai J, Alnæs GG, Kristensen V, Børresen-Dale AL, John EM, Miron A, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Grip M, Andrulis IL, Knight JA, Glendon G, Mulligan AM, Figueroa JD, García-Closas M, Lissowska J, Sherman ME, Hooning M, Martens JW, Seynaeve C, Collée M, Hall P, Humpreys K, Czene K, Liu J, Cox A, Brock IW, Cross SS, Reed MW, Ahmed S, Ghoussaini M, Pharoah PD, Kang D, Yoo KY, Noh DY, Jakubowska A, Jaworska K, Durda K, Złowocka E, Sangrajrang S, Gaborieau V, Brennan P, McKay J, Shen CY, Yu JC, Hsu HM, Hou MF, Orr N, Schoemaker M, Ashworth A, Swerdlow A, Trentham-Dietz A, Newcomb PA, Titus L, Egan KM, Chenevix-Trench G, Antoniou AC, Humphreys MK, Morrison J, Chang-Claude J, Easton DF, Dunning AM

[190]- Caution in generalizing known genetic risk markers for breast cancer across all ethnic/racial populations. Fang Chen, Daniel O Stram, Loic Le Marchand, Kristine R Monroe, Laurence N Kolonel, Brian E Henderson and Christopher A Haiman. European Journal of Human Genetics (2011) 19, 243–245

[191]- Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM.GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet] Lyon, France: International Agency for Research on Cancer, 2010. Accessed on 22.10.2012

[192]- Genome-Wide Significant Association Between a Sequence Variant at 15q15 Genome-Wide Significant Association Between a Sequence Variant at 15q15.2 and Lung Cancer Risk. American Association for Cancer Research. Cancer Res; 71(4) February 15, 2011. Rafnar T, Sulem P, Besenbacher S, Gudbjartsson DF, Zanon C, Gudmundsson J, Stacey SN, Kostic JP, Thorgeirsson TE, Thorleifsson G, Bjarnason H, Skuladottir H, Gudbjartsson T, Isaksson HJ, Isla D, Murillo L, García-Prats MD, Panadero A, Aben KK, Vermeulen SH, van der Heijden HF, Feser WJ, Miller YE, Bunn PA, Kong A, Wolf HJ, Franklin WA, Mayordomo JI, Kiemeney LA, Jonsson S, Thorsteinsdottir U, Stefansson K

[193]- Replication of results of genome-wide association studies on lung cancer susceptibility loci in a Korean population. Asian Pacific Society of Respirology (2012) 17, 699–706. Bae EY, Lee SY, Kang BK, Lee EJ, Choi YY, Kang HG, Choi JE, Jeon HS, Lee WK, Kam S, Shin KM, Jin G, Yoo SS, Lee J, Cha SI, Kim CH, Jung TH, Park JY.

[194]- A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2. Nature Genetics. VOLUME 43. NUMBER 8. AUGUST 2011. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, Yang L, Dai J, Hu L, Tan W, Li Z, Deng Q, Wang J, Wu W, Jin G, Jiang Y, Yu D, Zhou G, Chen H, Guan P, Chen Y, Shu Y, Xu L, Liu X, Liu L, Xu P, Han B, Bai C, Zhao Y, Zhang H, Yan Y, Ma H, Chen J, Chu M, Lu F, Zhang Z, Chen F, Wang X, Jin L, Lu J, Zhou B, Lu D, Wu T, Lin D, Shen H.

[195]- Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. Thorax 2011;66:413 doi:10.1136/thx.2010.14558. Yafei Li, Chau-Chyun Sheu, Yuanqing Ye, Prof Mariza de Andrade, Liang Wang, Shen-Chih Chang, Marie C Aubry, Jeremiah A Aakre, Mark S Allen, Feng Chen, Julie M Cunningham, Claude Deschamps, Ruoxiang Jiang, Jie Ling, Randolph S Marks, V Shane Pankratz, Li Su, Yan Li, Zhifu Sun, Hui Tang, George Vasmatzis, Curtis C Harris, Prof Margaret R Spitz, Jin Jen, Renyi Wang, Zuo-Feng Zhang, Prof David C Christiani, Xifeng Wu, Prof Ping Yang

[196]-GLOBACAN 2008 Fact Sheet.
http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900 . Accessed on 21.10.2012).

[197]- A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nature Genetics VOLUME 42. NUMBER 10. OCTOBER 2010

[198]- Inherited Variants in Mitochondrial Biogenesis Genes May Influence Epithelial Ovarian Cancer Risk. Cancer Epidemiology Biomarkers & Prevention.2011. Permuth-Wey J, Chen YA, Tsai YY, Chen Z, Qu X, Lancaster JM, Stockwell H, Dagne G, Iversen E, Risch H, Barnholtz-Sloan J, Cunningham JM, Vierkant RA, Fridley BL, Sutphen R, McLaughlin J, Narod SA, Goode EL, Schildkraut JM, Fenstermacher D, Phelan CM, Sellers TA.

[199]- European American Stratification in Ovarian Cancer Case Control Data: The Utility of Genome-Wide Data for Inferring Ancestry. Plosone. May 2012. Volume 7. Issue 5. e35235. Paola Raska, Edwin Iversen, Ann Chen, Zhihua Chen, Brooke L. Fridley, Jennifer Permuth-Wey, Ya-Yu Tsai, Robert A. Vierkant, Ellen L. Goode, Harvey Risch, Joellen M. Schildkraut, Thomas A. Sellers, Jill Barnholtz-Sloan

[200]- Piek JM, van Diest PJ, Verheijen RH (2008). "Ovarian carcinogenesis: an alternative hypothesis". Adv. Exp. Med. Biol. Advances in Experimental Medicine and Biology 622: 79–87

[201]- Molecular markers associated with nonepithelial ovarian cancer in formalin-fixed, paraffin-embedded specimens by genome wide expression profiling. Koon Vui-Kee, Ahmad Zailani Hatta Mohd Dali, Isa Mohamed Rose, Razmin Ghazali, Rahman Jamal, Norfilza Mohd Mokhtar. Kaohsiung Journal of Medical Sciences (2012) 28, 243-250

[202]- Association between invasive ovarian cancer susceptibility and 11 best candidate SNPs from breast cancer genome-wide association study. Human Molecular Genetics, 2009, Vol. 18, No. 12. Song H, Ramus SJ, Kjaer SK, DiCioccio RA, Chenevix-Trench G, Pearce CL, Hogdall E, Whittemore AS, McGuire V, Hogdall C, Blaakaer J, Wu AH, Van Den Berg DJ, Stram DO, Menon U, Gentry-Maharaj A, Jacobs IJ, Webb PM, Beesley J, Chen X; Australian Cancer (Ovarian) Study; Australian Ovarian Cancer Study Group, Rossing MA, Doherty JA, Chang-Claude J, Wang-Gohrke S, Goodman MT, Lurie G, Thompson PJ, Carney ME, Ness RB, Moysich K, Goode EL, Vierkant RA, Cunningham JM, Anderson S, Schildkraut JM, Berchuck A, Iversen ES, Moorman PG, Garcia-Closas M, Chanock S, Lissowska J, Brinton L, Anton-Culver H, Ziogas A, Brewster WR, Ponder BA, Easton DF, Gayther SA, Pharoah PD; Ovarian Cancer Association Consortium (OCAC)

[203]- GLOBACAN. 2008 database version 1.2

[204]- Pathway analysis of genome-wide association study data highlights pancreatic development genes as susceptibility factors for pancreatic cancer. Advance Acccess publication. Carcinogenesis vol. 33 no.7 pp. 1384-1390, 2012. Li D, Duell EJ, Yu K, Risch HA, Olson SH, Kooperberg C, Wolpin BM, Jiao L, Dong X, Wheeler B, Arslan AA, Bueno-de-Mesquita HB, Fuchs CS, Gallinger S, Gross M, Hartge P, Hoover RN, Holly EA, Jacobs EJ, Klein AP, LaCroix A, Mandelson MT, Petersen G, Zheng W, Agalliu I, Albanes D, Boutron-Ruault MC, Bracci PM, Buring JE, Canzian F, Chang K, Chanock SJ, Cotterchio M, Gaziano JM, Giovannucci EL, Goggins M, Hallmans G, Hankinson SE, Hoffman Bolton JA, Hunter DJ, Hutchinson A, Jacobs KB, Jenab M, Khaw KT, Kraft P, Krogh V, Kurtz RC, McWilliams RR, Mendelsohn JB, Patel AV, Rabe KG, Riboli E, Shu XO, Tjønneland A, Tobias GS, Trichopoulos D, Virtamo J, Visvanathan K, Watters J, Yu H, Zeleniuch-Jacquotte A, Amundadottir L, Stolzenberg-Solomon RZ.

[205]- Genome-Wide Association Study of Pancreatic Cancer in Japanese Population. Plosone. July 2010. Volume 5. Issue 7. e11824. Siew-Kee Low, Aya Kuchiba, Hitoshi Zembutsu, Akira Saito, Atsushi Takahashi, Michiaki Kubo, Yataro Daigo, Naoyuki Kamatani, Suenori Chiku, Hirohiko Totsuka, Sumiko Ohnami, Hiroshi Hirose, Kazuaki Shimada, Takuji Okusaka, Teruhiko Yoshida, Yusuke Nakamura, Hiromi Sakamoto

[206]- Pancreatic Cancer Susceptibility Loci and Their Role in Survival. Plosone. November 2011. Volume 6. Issue 11. e27921. Rizzato C, Campa D, Giese N, Werner J, Rachakonda PS, Kumar R, Schanné M, Greenhalf W, Costello E, Khaw KT, Key TJ, Siddiq A, Lorenzo-Bermejo J, Burwinkel B, Neoptolemos JP, Büchler MW, Hoheisel JD, Bauer A, Canzian F.

[207]- Association of Breast Cancer Susceptibility Variants with Risk of Pancreatic Cancer. Fergus J. Couch, Xianshu Wang, Robert R. McWilliams, William R. Bamlet, Mariza de Andrade and Gloria M. Petersen. *Cancer Epidemiol Biomarkers Prev* 2009;18:3044-3048

[208]- Neural networks for genetic epidemiology: past, present, and future. Alison A Motsinger-Reif and Marylyn D Ritchie. BioData Mining 2008, 1:3

[209]- Genetic programming neural networks: A powerful bioinformatics tool for human genetics. Marylyn D. Ritchie, Alison A. Motsinger, William S. Bush, Christopher S. Coffey, Jason H. Moore. Applied Soft Computing 7 (2007) 471–479

[210]- A Hybrid Approach to Selecting Susceptible Single Nucleotide Polymorphisms for Complex Disease Analysis. Pengyi Yang and Zili Zhang. 2008 International Conference on BioMedical Engineering and Informatics

[211]- ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. Stephen D Turner, Scott M Dudek, Marylyn D Ritchie. BioData Mining 2010, 3:5

[212]- A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups Orlando Anunciacao, Bruno C. Gomes, Susana Vinga, Jorge Gaspar, Arlindo L. Oliveira, and Jos´e Rueff. IWPACBB 2010, AISC 74, pp. 43–51

[213]- GPDTI: A Genetic Programming Decision Tree Induction method to find epistatic effects in common complex diseases. Vol. 23 ISMB/ECCB 2007, pages i167–i174. Jesús K. Estrada-Gil, Juan C. Fernández-López , Enrique Hernández-Lemus, Irma Silva-Zolezzi, Alfredo Hidalgo-Miranda, Gerardo Jiménez-Sánchez and Edgar E. Vallejo-Clemente

[214]- METU-SNP: an integrated software system for SNP-complex disease association analysis. Ustünkar G, Aydın Son Y. J Integr Bioinform. 2011 Dec 12;8(1):187

[215]- Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. S. Sathiya Keerthi, Chih-Jen Lin. Neural Computation, Volume 15, Number 7, p.1667-1689 (2003)

[216]- Ebrahimi M, Lakizadeh A, Agha-Golzadeh P, Ebrahimie E, Ebrahimi M (2011) Prediction of Thermostability from Amino Acid Attributes by Combination of Clustering with Attribute Weighting: A New Vista in Engineering Enzymes. PLoS ONE 6(8): e23146. doi:10.1371/journal.pone.0023146

[217]- Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA. Magdalena Graczyk, Tadeusz Lasota, Bogdan Trawiński. Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. Lecture Notes in Computer Science Volume 5796, 2009, pp 800-812

[218]- Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner. Jianchao Han, Rodriguez, J.C.; Behest,, M. Future Generation Communication and Networking, 2008

[219]- Evolutionary learning with kernels: a generic solution for large margin problems. Ingo Mierswa. GECCO '06 Proceedings of the 8th annual conference on Genetic and evolutionary computation. Pages 1553-1560

[220]- SVMTorch: support vector machines for large-scale regression problems. The Journal of Machine Learning Research. Volume 1, 9/1/2001. Pages 143-160

[221]- The Genetic Kernel Support Vector Machine: Description and Evaluation. Tom Howley. Michale G. Madden. Artificial Intelligence Review. Volume 24 Issue 3-4, November 2005 Pages 379-395

[222]- A User's Guide to Support Vector Machines. Ben-Hur A, Weston J. Methods Mol Biol. 2010;609:223-39

[223]- A Multi-class SVM Classifier Utilizing Binary Decision Tree. Gjorgji Madzarov, Dejan Gjorgjevikj and Ivan Chorbev. Informatica 33 (2009) 233-241

[224]- An Improved Algorithm for Decision-Tree-Based SVM. Xiaodan Wang, Zhaohui Shi, Chongming Wu and Wei Wang. Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006

[225]- Combining Support Vector Machines With a Pairwise Decision Tree. Jin Chen, Cheng Wang, *Member, IEEE*, and Runsheng Wang. IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 5, NO. 3, JULY 2008

[226]- Ensembles of Binary SVM Decision Trees. Gjorgji Madjarov, Dejan Gjorgjevikj and Tomche Delev. ICT Innovations 2010 Web Proceedings ISSN 1857-7288

[227]- Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree. Jieyue He, Hae-Jin Hu, Robert Harrison, Phang C. Tai, and Yi Pan. IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 5, NO. 1, MARCH 2006

[228]- A Hybrid SVM based Decision Tree M. ArunKumar, M.Gopal. Pattern Recognition 43 (2010) 3977–3987

[229]- A New Pruning Heuristic Based on Variance Analysis of Sensitivity Information. Andries P. Engelbrecht IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 12, NO. 6, NOVEMBER 2001

[230]- Input Feature Selection for Classification Problems. Nojun Kwak and Chong-Ho Choi. IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 1, JANUARY 2002

[231]- An Improved ID3 Decision Tree Algorithm. Chen Jin, Luo De-lin, Mu Fen-xiang. Proceedings of 2009 4th International Conference on Computer Science & Education

[232]- An Improved ID3 Based on Weighted Modified Information Gain. Chun Guan, Xiaoqin Zeng. 2011 Seventh International Conference on Computational Intelligence and Security. P 1283-1285

[233]- Rapini, Ronald P.; Bolognia, Jean L.; Jorizzo, Joseph L. (2007). Dermatology: 2-Volume Set. St. Louis: Mosby. pp. 1732.

[234]- Melanocortin 1 Receptor (*MC1R*) Gene Variants are Associated with an Increased Risk for Cutaneous Melanoma Which is Largely Independent of Skin Type and Hair Color. Cornelis Kennedy, Jeanet ter Huurne, Marjo Berkhout, Nelleke Gruis, Maarten Bastiaens, W Bergman, R Willemze and Jan Nico Bouwes Bavinck. *Journal of Investigative Dermatology* (2001) 117, 294–300

[235]- Skin colour and skin cancer -MC1R, the genetic link. Sturm, R. A. Melanoma Research: September 2002 - Volume 12 - Issue 5 - pp 405-416

[236]- The epidemiology of UV induced skin cancer. Bruce K Armstrong, Anne Kricker. Journal of Photochemistry and Photobiology B: Biology. Volume 63, Issues 1–3, October 2001, Pages 8–18

[237]- Prostate-Specific Antigen and the Early Diagnosis of Prostate Cancer Aaron Caplan, Alexander Kratz. Am J Clin Pathol 2002;117(Suppl 1):S104-S108

[238]- PSA Density and PSA Transition Zone Density in the Diagnosis of Prostate Cancer in PSA Gray Zone Cases. Yilmaz Aksoy, Aytekin Oral, Hulya Aksoy, Azam Demirel, Fatih Akcay. Annals of Clinical & Laboratory Science, vol. 33, no. 3, 200

[239]- PSA and Prostate Cancer. New England Journal of Medicine 2004;350:2239-46.

[240]- Prostate-specific antigen testing accuracy in community practice. Richard M Hoffman, Frank D Gilliland, Meg Adams-Cameron, William C Hunt, Charles R Key. BMC Fam Pract. 2002; 3: 19.

[241]- Serum prostate specific antigen as pre-screening test for prostate cancer. Labrie F, Dupont A, Suburu R, Cusan L, Tremblay M, Gomez JL, Emond J. The Journal of Urology [1992, 147(3 Pt 2):846-51

[242]- The use of prostate specific antigen density to enhance the predictive value of intermediate levels ofserum prostate specific antigen. Benson MC, Whang IS, Olsson CA, McMahon DJ, Cooner WH. J Urol 1992;147:817-821.

[243]- The use of prostate specific antigen density to improve the sensitivity of prostate specific antigen in detecting prostate carcinoma.Bretton PR, Evans WB, Borden JD, Castellanos RD. Cancer 1994;74:2991-2995

[244]- The effect of ethnicity and sexual preference on prostate-cancer-related quality of life. Nir Kleinmann, Nicholas G. Zaorsky, Timothy N. Showalter, Leonard G. Gomella, Costas D. Lallas & Edouard J. Trabulsi. Nature Reviews Urology 9, 258-265 (May 2012)

[245]- Racial and Ethnic Differences in Advanced-Stage Prostate Cancer: the Prostate Cancer Outcomes Study**.** Richard M. Hoffman, Frank D.Gilliland, J. William Eley, Linda C. Harlan, Robert A. Stephenson, Janet L. Stanford, Peter C. Albertson, Ann S. Hamilton, W. Curtis Hunt, ArnoldL. Potosky. JNCI J Natl Cancer Inst *(2001) 93 (5): 388-395.*

[246]- Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. Andrew G Renehan, Margaret Tyson, Matthias Egger , Richard F Heller, Marcel Zwahlen. The Lancet. Volume 371, Issue 9612, 16–22 February 2008, Pages 569–578.

[247]- Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women. Journal of the National Cancer Institute .2003, 95(16):1218-1226 . Key TJ, Appleby PN, Reeves GK, Roddam A, Dorgan JF, Longcope C, Stanczyk FZ, Stephenson HE Jr, Falk RT, Miller R, Schatzkin A, Allen DS, Fentiman IS, Key TJ, Wang DY, Dowsett M, Thomas HV, Hankinson SE, Toniolo P, Akhmedkhanov A, Koenig K, Shore RE, Zeleniuch-Jacquotte A, Berrino F, Muti P, Micheli A, Krogh V, Sieri S, Pala V, Venturelli E, Secreto G, Barrett-Connor E, Laughlin GA, Kabuto M, Akiba S, Stevens RG, Neriishi K, Land CE, Cauley JA, Kuller LH, Cummings SR, Helzlsouer KJ, Alberg AJ, Bush TL, Comstock GW, Gordon GB, Miller SR, Longcope C, Endogenous Hormones Breast Cancer Collaborative Group

[248]- Body Mass Index and Risk of Adenocarcinomas of the Esophagus and Gastric Cardia. Wong-Ho ChowJoseph F. Fraumeni Jr., William J. Blot, Thomas L. VaughanJanet L. StanfordDiana C. Farrow, Harvey A. RischRobert DubrowSusan T. Mayne, Marilie D. GammonHabibul Ahsan, Janet B. Schoenberg, A. Brian West, Heidi Rotterdam and Shelley Niwa. Oxford Journals. Medicine. JNCI J Natl Cancer Inst. Volume 90, Issue 2 Pp. 150-155.

[249]- Body Mass Index and Risk of Prostate Cancer in U.S. Health Professionals. Edward Giovannucci, Eric B. Rimm, Yan Liu, Michael Leitzmann, Kana Wu, Meir J. Stampfer and Walter C. Willett. Oxford Journals. Medicine. JNCI J Natl Cancer Inst Volume 95, Issue 16 Pp. 1240-1244.

[250]- Body Mass Index, Weight Change, and Risk of Prostate Cancer in the Cancer Prevention Study II Nutrition Cohort. Carmen Rodriguez, Stephen J. Freedland, Anusila Deka, Eric J. Jacobs, Marjorie L. McCullough, Alpa V. Patel, Michael J. Thun and Eugenia E. Calle. Cancer Epidemiol Biomarkers Prev. 2007 Jan;16(1):63-9. Epub 2006 Dec 19.

[251]- Body mass index, prostate cancer-specific mortality, and biochemical recurrence: a systematic review and meta-analysis. Cao Y, Ma J. Cancer Prev Res (Phila). 2011 Apr;4(4):486-501

[252]- Body mass index, height, and prostate cancer mortality in two large cohorts of adult men in the United States. Rodriguez C, Patel AV, Calle EE, Jacobs EJ, Chao A, Thun MJ. Cancer Epidemiol Biomarkers Prev. 2001 Apr;10(4):345-53.

[253]- Family history and the risk of prostate cancer. Gary D. Steinberg Bob S. Carter, Terri H. Beaty, Barton Childs, Dr. Patrick C. Walsh. The Prostate Volume 17, Issue 4, pages 337–347. JUL 2006

[254]- Family History and Prostate Cancer Risk. Samuel M. Lesko, Lynn Rosenberg and Samuel Shapiro. Oxford Journals Medicine American Journal of Epidemiology Volume 144, Issue 11 Pp. 1041-1047

[255]- Family history, Hispanic ethnicity, and prostate cancer risk. Stone SN, Hoffman RM, Tollestrup K, Stidley CA, Witter JL, Gilliland FD. Ethn Dis. 2003 Spring;13(2):233-9.

[256]- Cigarette Smoking and Risk of Prostate Cancer in Middle-Aged Men. Lora A. Plaskon, David F. Penson, Thomas L. Vaughan. Cancer Epidemiol Biomarkers Prev 2003;12:604-609

[257]- Cigarette Smoking as a Predictor of Death from Prostate Cancer in 348,874 Men Screened for the Multiple Risk Factor Intervention Trial. Steven S. Coughlin James D. Neatonand Anjana Sengupta. Oxford Journals Medicine American Journal of Epidemiology Volume 143, Issue 10 Pp. 1002-1006

[258]- Association of smoking, body mass, and physical activity with risk of prostate cancer in the Iowa 65+ Rural Health Study (United States). James R. Cerhan, James C. Torner, Charles F. Lynch, Linda M. Rubenstein, Jon H. Lemke, Michael B. Cohen, David M. Lubaroff, Robert B. Wallace . Cancer Causes & Control. 1997, Volume 8, Issue 2, pp 229-238

[259]- Alcohol consumption, smoking, and other risk factors and prostate cancer in a large health plan cohort in California (United States). Dr Robert A. Hiatt, Ms Mary Anne Armstrong, Mr Arthur L. Klatsky, Stephen Sidney . Cancer Causes & Control. January 1994, Volume 5, Issue 1, pp 66-72

[260]- The ABCD System of Melanoma Detection. A Spectrophotometric Analysis of the Asymmetry, Border, Color, and Dimension. Aldo Bono, Stefano Tomatis, Cesare Bartoli, Gabrina Tragni, Giovanni Radaelli, Andrea Maurichi, Renato Marchesini. Cancer. Volume 85, Issue 1, Article first published online: 19 NOV 2000

[261]- The seven features for melanoma: a new dermoscopic algorithm for the diagnosis of malignant melanma. V. Dal Pozzo, C. Benelli, E. Roscetti. European Journal of Dermatology. Volume 9, number 4, 303-8, June 1999

[262]- Stolz W, Braun-Falco O, Bilek P, Landthaler M, Cognetta AB. *Color Atlas of Dermatoscopy.* Blackwell Science Ltd, 1994.

[263]Nachbar F, Stolz W, Merkle T. The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *J Am Acad Dermatol* 1994; 30: 551-9.

[264]- Is Dermoscopy (Epiluminescence Microscopy) Useful for the Diagnosis of Melanoma?Results of a Meta-analysis Using Techniques Adapted to the Evaluation of Diagnostic Tests. Marie-Lise Bafounta, Alain Beauchet, Philippe Aegerter, Philippe Saiag. Evidence-Based Dermatology. October 2001, Vol 137, No. 10

[265]- Pigmentary characteristics and moles in relation to melanoma risk. Linda Titus-Ernstoff, Ann E. Perry, Steven K. Spencer, Jennifer J. Gibson, Bernard F. Cole, Marc S. Ernstoff. International Journal of Cancer. Volume 116, Issue 1, pages 144–149, 10 August 2005

[266]- Malignant melanoma in relation to moles, pigmentation, and exposure to fluorescent and other lighting sources. J. M. Elwood, C. Williamson, and P. J. Stapleton. Br J Cancer. 1986. January; 53(1): 65–74.

[267]- Moles and melanoma. Research Update " Le point sur la recherche" Canadian Dermatology Association. CAN MED ASSOC J JULY. 1, 1997; 157 (1)

[268]- Demographic Study of Clinically Atypical (Dysplastic) Nevi in Patients with Melanoma and Comparison Subjects. James J. Nordlund, John Kirkwood, Bernadette M. Forget, Adrianna Scheibner,Daniel M. Albert, Emmanuel Lemer, and G. W. Milton. Cancer Res April 1985 45; 1855

[269]- Independence of dysplastic nevi from total nevi in determining risk for nonfamilial melanoma George C. Roush, James J. Nordlund, Bernadette Forget, Stephen B. Gruber, John M. Kirkwood. Preventive Medicine Volume 17, Issue 3, May 1988, Pages 273–279

[270]- Dysplastic Nevi in Association with Multiple Primary Melanoma. Linda Titus-Ernstoff, Paul H. Duray, Marc S. Ernstoff, Raymond L. Barnhill, Pamela L. Horn, John M. Kirkwood. Cancer Res February 15, 1988 48; 1016

[271]- Skin characteristics and risk of superficial spreading and nodular melanoma (United States) Bryan Langholz, Jean Richardson, Edward Rappaport, Jerry Waisman, Myles Cockburn, Thomas Mack. Cancer Causes and Control 11: 741-750, 2000.

[272]- Eye color and cutaneous nevi predict risk of ocular melanoma in Australia. Claire M. Vajdic, Anne Kricker, Michael Giblin, John McKenzie, Joanne Aitken, Graham G. Giles, Bruce K. Armstrong. International Journal of Cancer Volume 92, Issue 6, pages 906–912, 15 June 2001

[273]- Cutaneous factors related to the risk of malignant melanoma. VALERIE BERAL, SUSAN EVANS, HELEN SHAW, G. MILTON. British Journal of Dermatology. Volume 109, Issue 2, pages 165–172, August 1983

[274]-Meta-analysis of risk factors for cutaneous melanoma: III. Family history, actinic damage and phenotypic factorsSara Gandini, Francesco Sera, Maria Sofia Cattaruzza, Paolo Pasquini, Roberto Zanetti, Cinzia Masini, Peter Boyle, Carmelo Francesco Melchi. European Journal of Cancer. Volume 41, Issue 14, September 2005, Pages 2040–2059.

[275]- Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, Harshman K, Guipponi M, Bukach O, Zoete, et al.. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet. 2011; 44(2):133-139.

[276]- Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, Chaudhuri S, Guan Y, Janakiraman V, Jaiswal BS.. Reccurent R-spondin functions in colon cancer. Nature. 2012; 488(7413):660-664

[277]- Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, et al.. The landscape of cancer genes and mutational processes in breast cancer. Nature. 201; 486(7403):400-404.

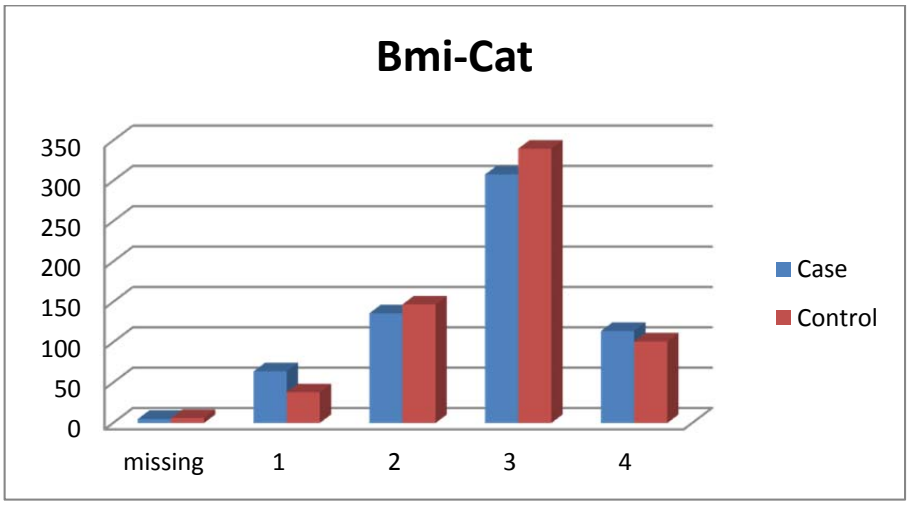
[278]- Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012 May 9;485(7399):502-506

[279]- Deli T, Varga N, Adám A, Kenessey I, Rásó E, Puskás LG, Tóvári J, Fodor J, Fehér M, Szigeti GP, et al.. Functional genomics calcium channles in human melanoma cells. Int J Cancer. 2007; 121(1):55-65.

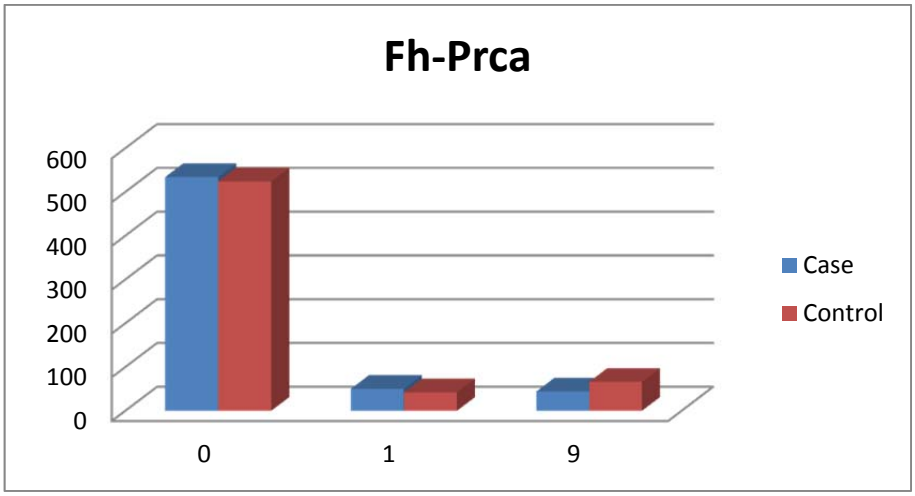
[280]- Dutton-Regester K, Aoude LG, Nancarrow DJ, Stark MS, O'Connor L, Lanagan C, Pupo GM, Tembe V, Carter CD, O'Rourke M. Et al.. Identification of TFG (TRK-fused gene) as a putative metastatic melanoma tumor suppressor gene. Genes Chromosomes Cancer. 2012; 51(5):452-461

APPENDICIES

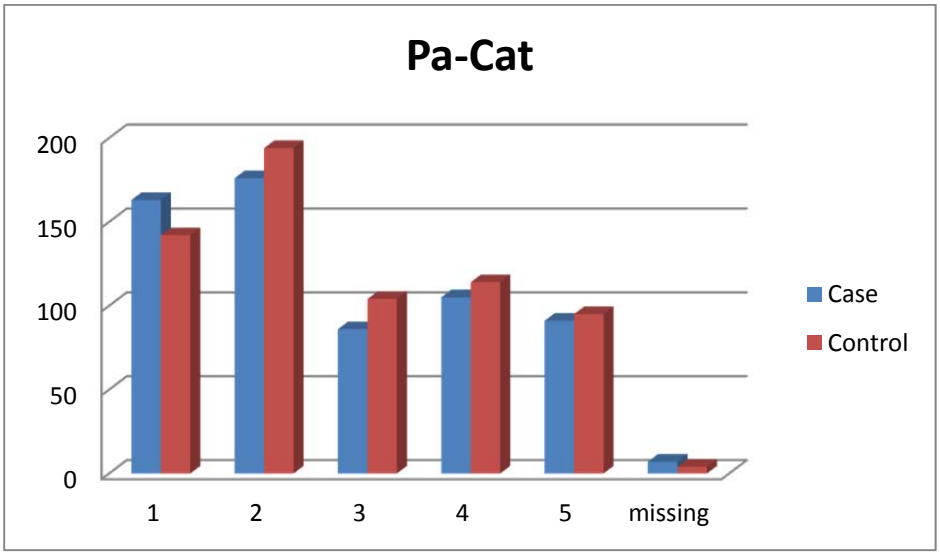
APPENDIX A-) Distribution of Phenotypic Attributes among Cases and Controls in Prostate Cancer Dataset



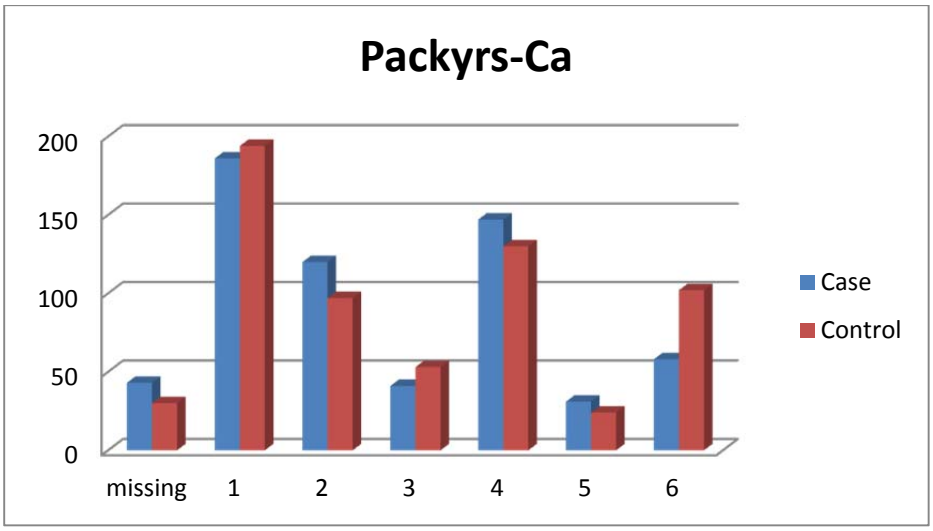
Distribution of bmi_cat among cases and controls



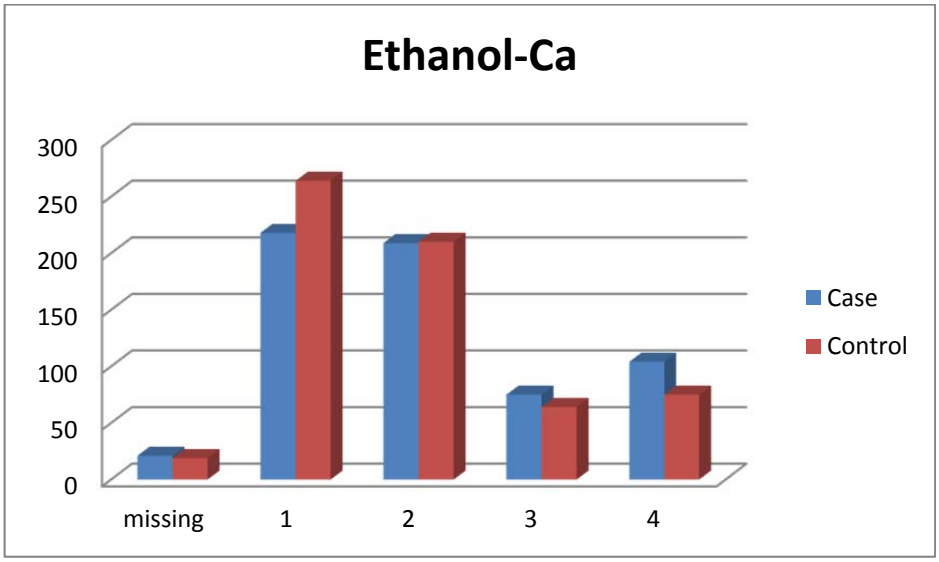
Distribution of fh_prca among cases and controls



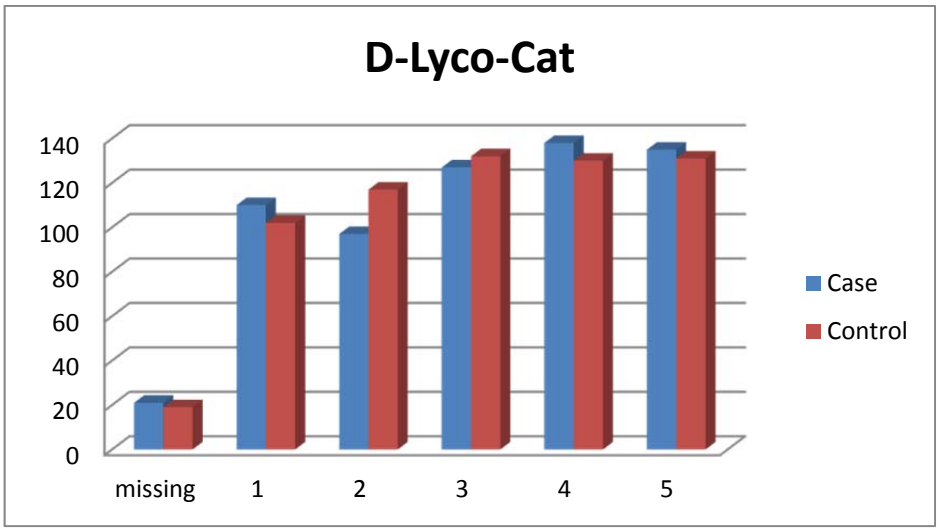
Distribution of pa_cat among cases and controls



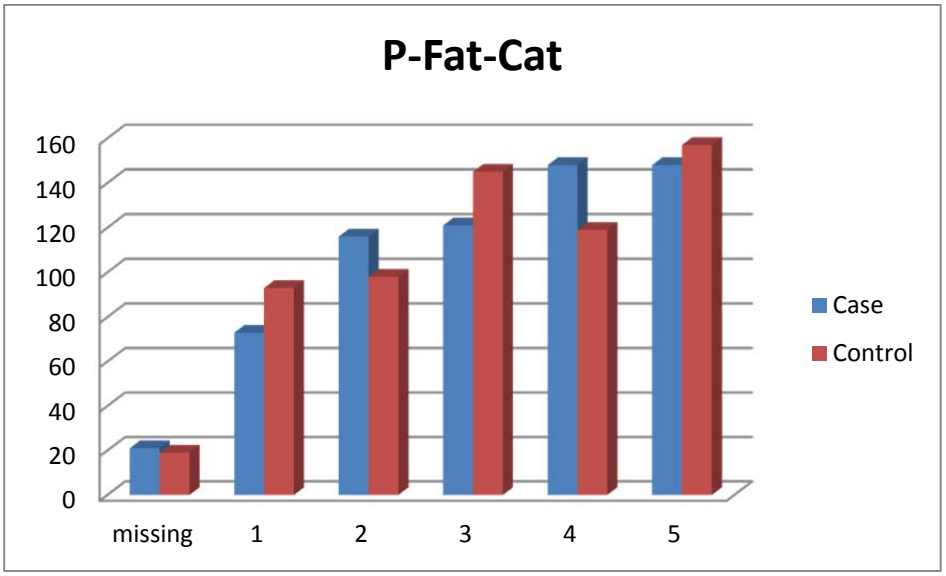
Distribution of packyrs_ca among cases and controls



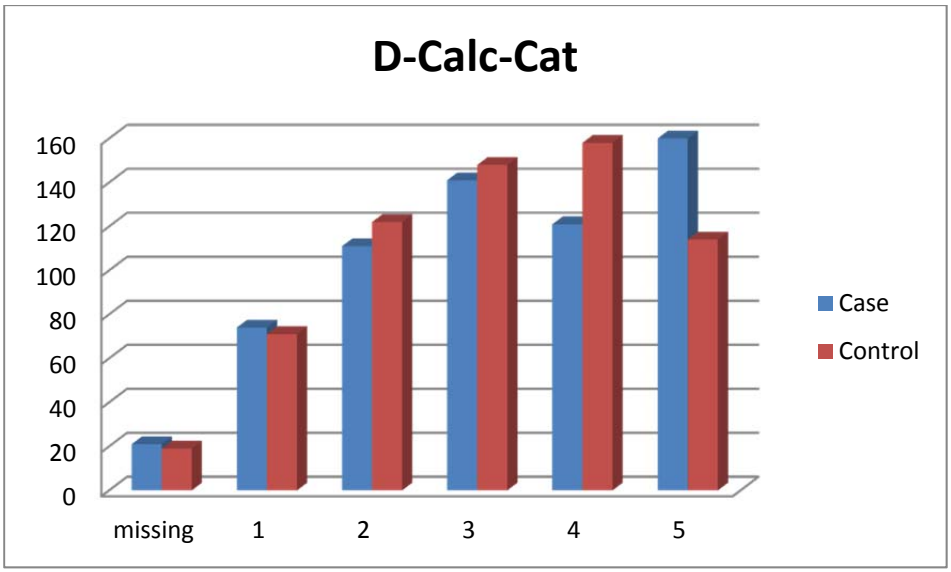
Distribution of ethanol_ca among cases and controls



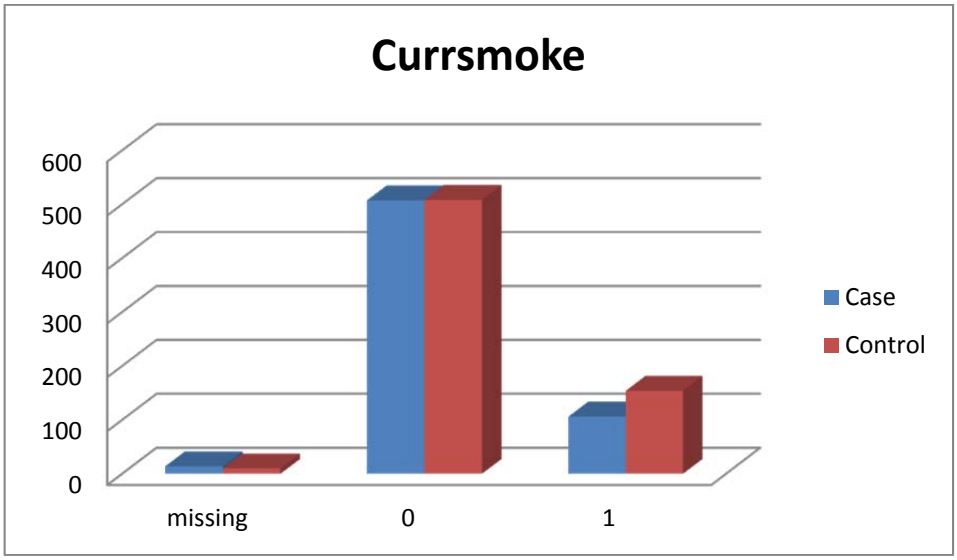
Distribution of d_lyco_cat among cases and controls



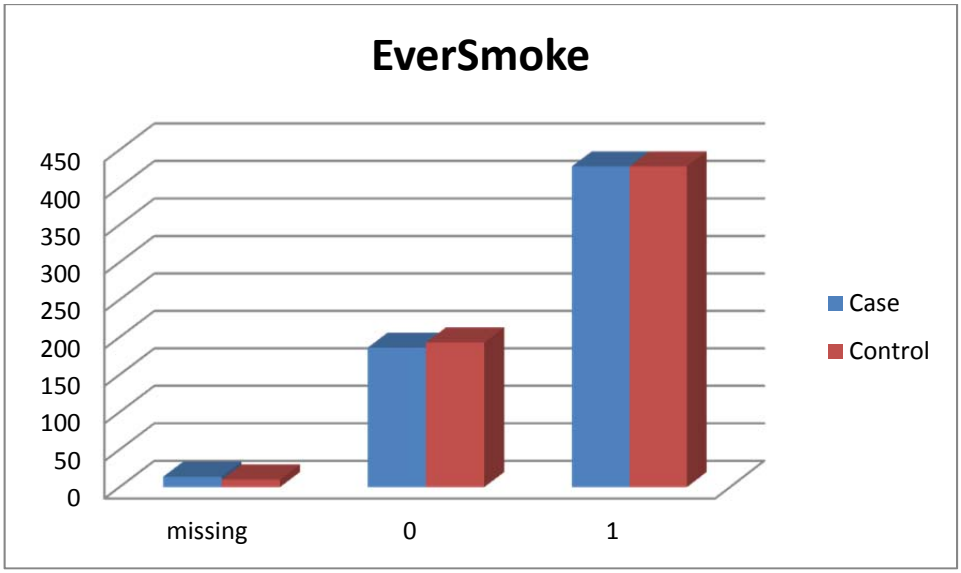
Distribution of p_fat_cat among cases and controls



Distribution of d_calc_cat among cases and controls

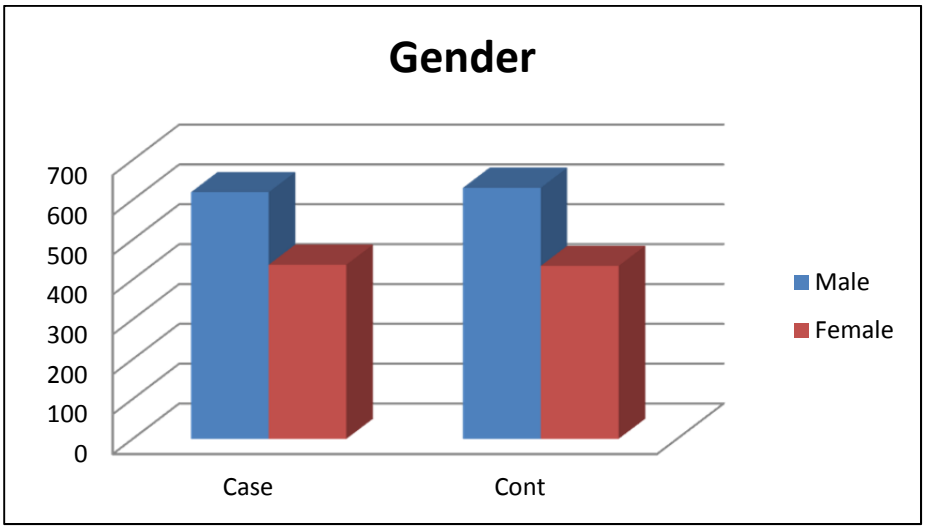


Distribution of currsmoke among cases and controls

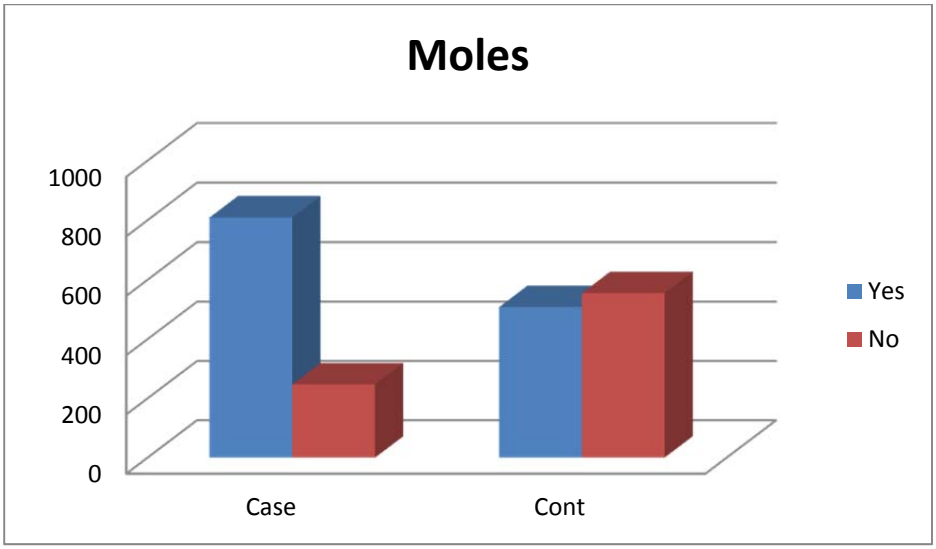


Distribution of eversmoke among cases and controls

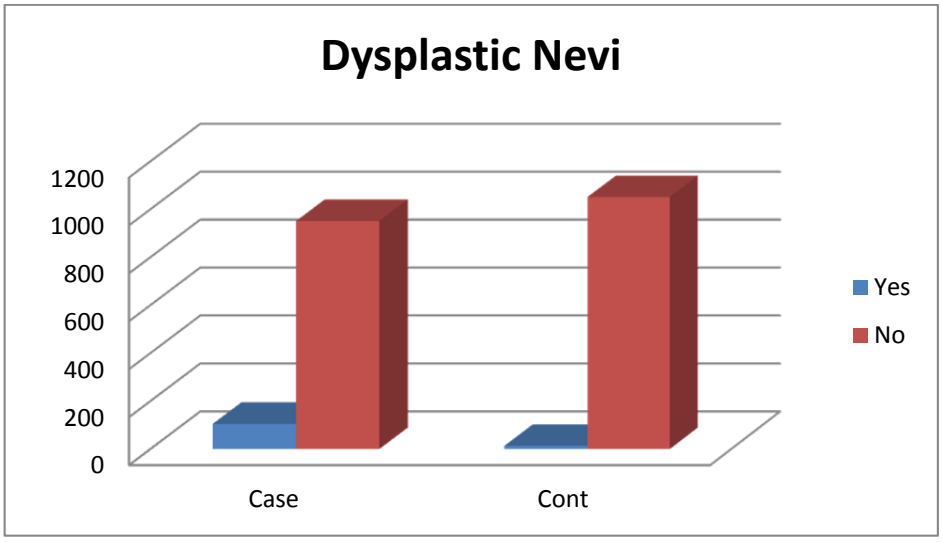
APPENDIX B-) Distribution of Phenotypic Attributes among Cases and Controls in Melanoma Dataset



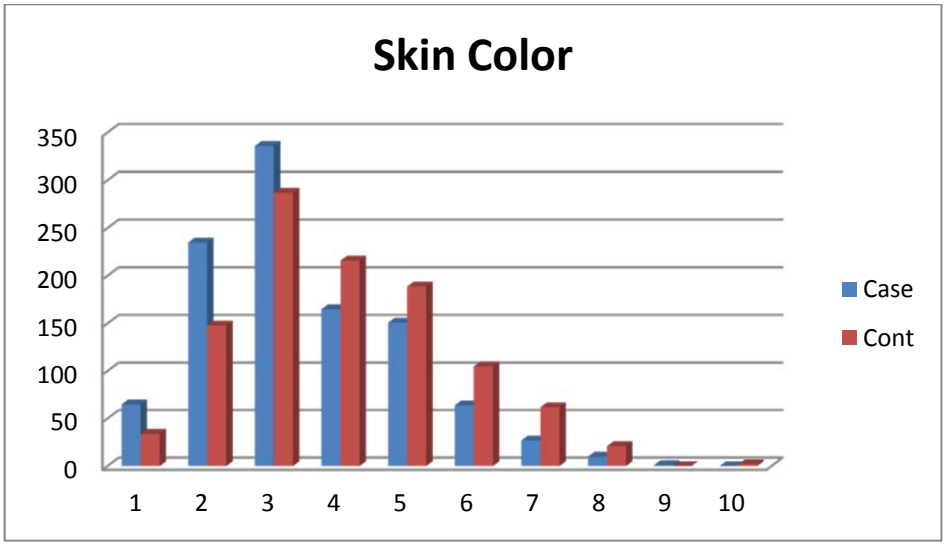
Distribution of attribute gender among cases and controls



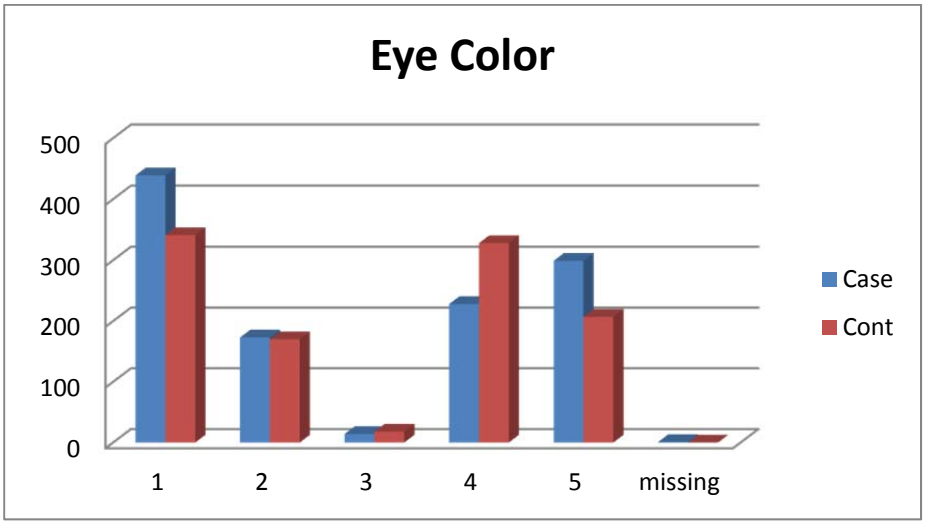
Distribution of attribute moles among cases and controls



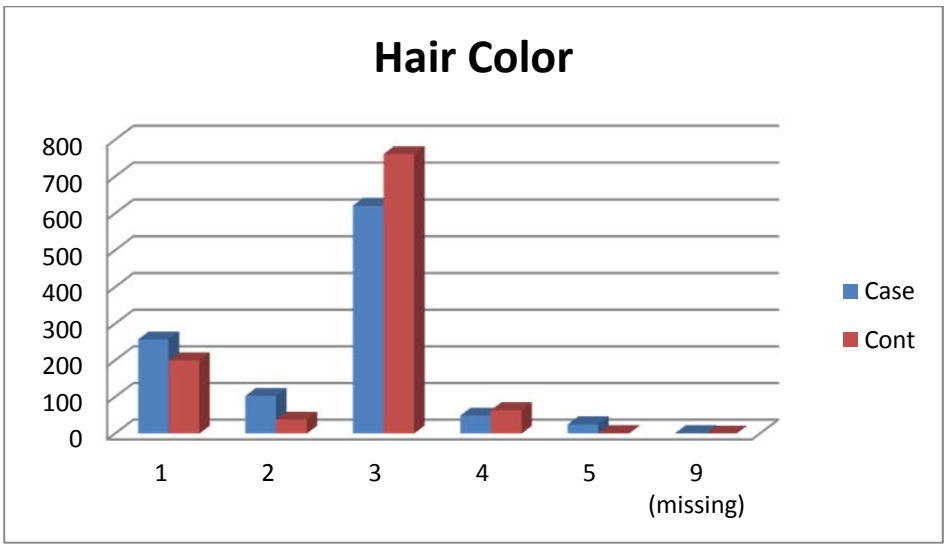
Distribution of attribute dysplastic_nevi among cases and controls



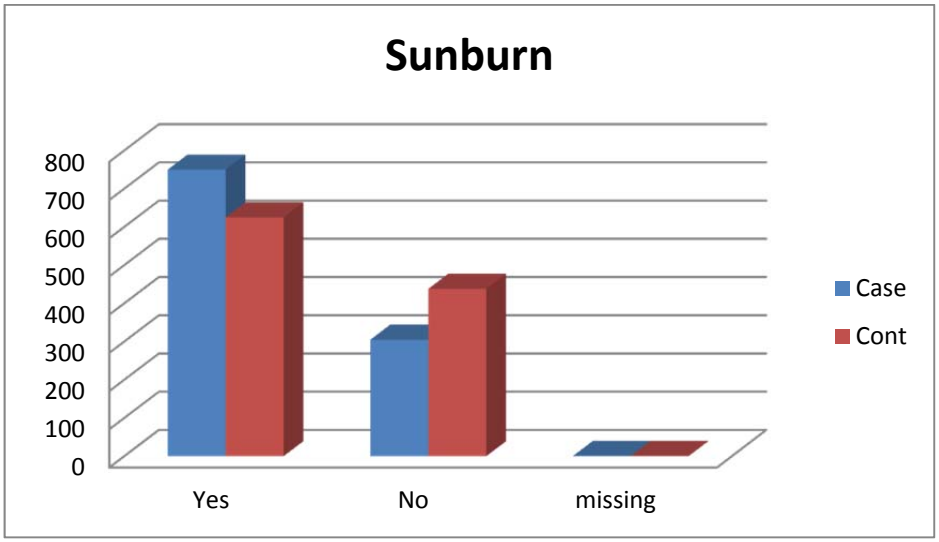
Distribution of attribute skin_color among cases and controls



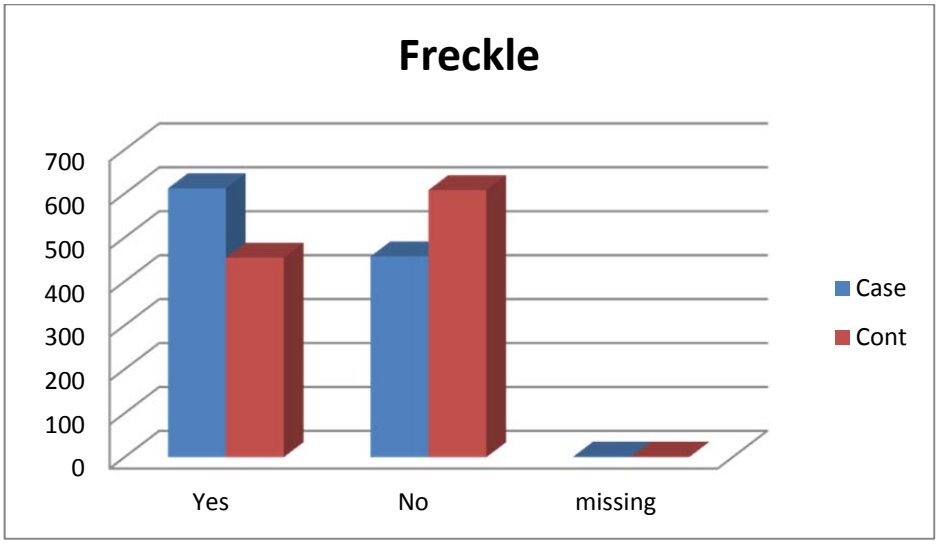
Distribution of attribute eye_color among cases and controls



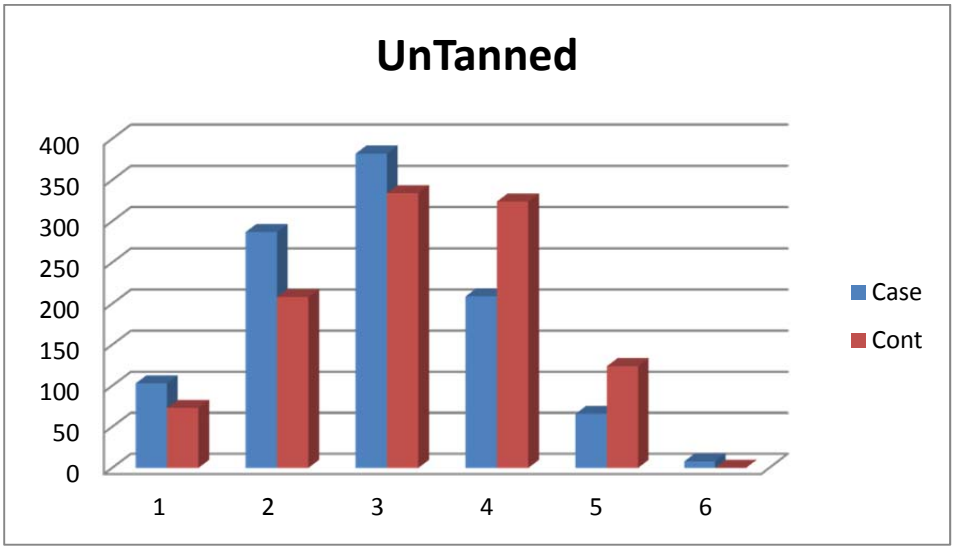
Distribution of attribute hair_color among cases and controls



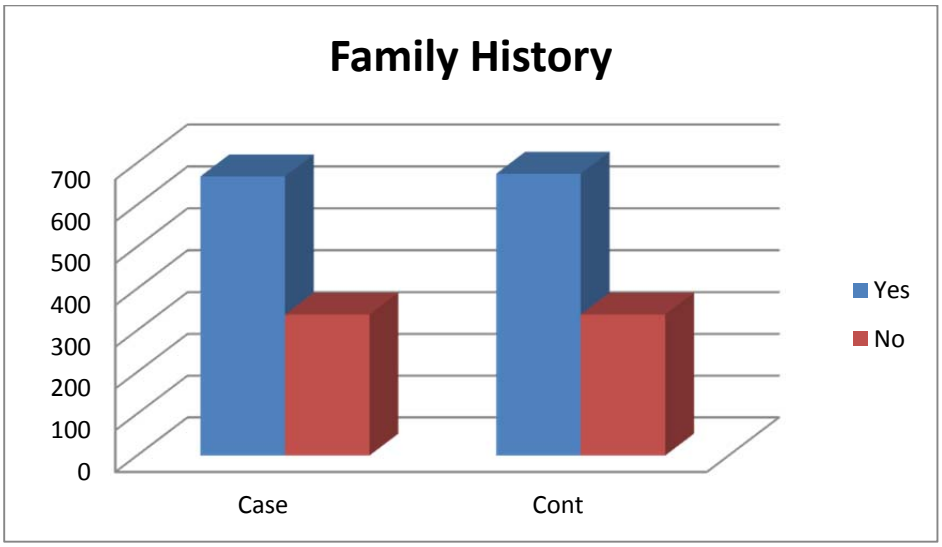
Distribution of attribute sunburn among cases and controls



Distribution of attribute freckle among cases and controls



Distribution of attribute un-tanned among cases and controls



Distribution of attribute family_history among cases and controls

APPENDIX C-) Decision Tree Structure of Hybrid System on Prostate Cancer

```
ethni = B
|   bmi_cat = aa
|   |   rs11729739 = ee: 1 {2=0, 1=12}
|   |   rs11729739 = ff: 2 {2=5, 1=0}
|   bmi_cat = bb
|   |   rs17701543 = dd: 1 {2=0, 1=10}
|   |   rs17701543 = kk
|   |   |   rs9848588 = dd: 1 {2=0, 1=5}
|   |   |   rs9848588 = kk
|   |   |   |   rs964130 = aa
|   |   |   |   |   rs10195113 = dd: 1 {2=0, 1=3}
|   |   |   |   |   rs10195113 = kk
|   |   |   |   |   |   rs1433369 = ff: 1 {2=0, 1=2}
|   |   |   |   |   |   rs1433369 = hh
|   |   |   |   |   |   |   rs12733054 = ff: 1 {2=0, 1=1}
|   |   |   |   |   |   |   rs12733054 = hh
|   |   |   |   |   |   |   |   rs17375010 = ff: 1 {2=0, 1=1}
|   |   |   |   |   |   |   |   rs17375010 = hh
|   |   |   |   |   |   |   |   |   rs766045 = aa: 2 {2=38, 1=0}
|   |   |   |   |   |   |   |   |   rs766045 = dd
|   |   |   |   |   |   |   |   |   |   ethanol_ca = aa: 2 {2=2, 1=0}
|   |   |   |   |   |   |   |   |   |   ethanol_ca = cc: 1 {2=0, 1=3}
|   |   |   |   rs964130 = dd: 1 {2=0, 1=3}
|   bmi_cat = cc
|   |   packyrs_ca = aa
|   |   |   rs12201462 = ff: 1 {2=0, 1=7}
|   |   |   rs12201462 = hh
|   |   |   |   rs4908656 = aa
|   |   |   |   |   rs9462806 = aa
|   |   |   |   |   |   rs1974562 = dd: 1 {2=0, 1=1}
|   |   |   |   |   |   rs1974562 = kk: 2 {2=31, 1=0}
|   |   |   |   |   rs9462806 = dd: 1 {2=0, 1=1}
|   |   |   |   rs4908656 = cc: 1 {2=0, 1=2}
```

```
|   |   packyrs_ca = bb
|   |   |   rs10954845 = aa
|   |   |   |   rs6997228 = aa: 1 {2=0, 1=16}
|   |   |   |   rs6997228 = dd: 2 {2=2, 1=0}
|   |   |   rs10954845 = dd: 2 {2=5, 1=0}
|   |   packyrs_ca = cc
|   |   |   rs10745253 = aa: 2 {2=8, 1=0}
|   |   |   rs10745253 = dd: 1 {2=0, 1=5}
|   |   packyrs_ca = dd
|   |   |   rs12980509 = ff: 1 {2=0, 1=14}
|   |   |   rs12980509 = hh
|   |   |   |   rs2296370 = aa: 1 {2=0, 1=1}
|   |   |   |   rs2296370 = dd: 1 {2=0, 1=6}
|   |   |   |   rs2296370 = kk: 2 {2=16, 1=0}
|   |   packyrs_ca = ee
|   |   |   rs7843255 = aa: 1 {2=0, 1=1}
|   |   |   rs7843255 = dd: 2 {2=8, 1=0}
|   |   |   rs7843255 = kk: 1 {2=0, 1=7}
|   |   packyrs_ca = ff
|   |   |   rs2194505 = ee: 2 {2=13, 1=0}
|   |   |   rs2194505 = ff: 1 {2=0, 1=4}
|   |   packyrs_ca = xx
|   |   |   rs2120806 = ff: 1 {2=0, 1=9}
|   |   |   rs2120806 = hh: 2 {2=2, 1=0}
|   bmi_cat = dd
|   |   rs10517581 = aa
|   |   |   rs2103869 = dd: 2 {2=4, 1=0}
|   |   |   rs2103869 = kk
|   |   |   |   rs10788555 = gg: 2 {2=2, 1=0}
|   |   |   |   rs10788555 = kk
|   |   |   |   |   rs7067548 = aa
|   |   |   |   |   |   rs17001078 = ee
|   |   |   |   |   |   |   rs918285 = ee: 1 {2=0, 1=41}
|   |   |   |   |   |   |   rs918285 = xx: 2 {2=1, 1=0}
|   |   |   |   |   |   rs17001078 = ff: 2 {2=2, 1=0}
|   |   |   |   |   rs7067548 = cc: 2 {2=2, 1=0}
|   |   rs10517581 = dd
|   |   |   rs9462806 = aa: 2 {2=13, 1=0}
|   |   |   rs9462806 = dd: 1 {2=0, 1=1}
|   bmi_cat = xx
|   |   rs7034430 = ff: 2 {2=5, 1=0}
|   |   rs7034430 = hh: 1 {2=0, 1=4}
ethni = J
|   bmi_cat = aa
|   |   rs2442602 = aa: 1 {2=0, 1=15}
|   |   rs2442602 = dd
|   |   |   rs3093679 = dd: 2 {2=1, 1=0}
|   |   |   rs3093679 = kk: 1 {2=0, 1=9}
|   |   rs2442602 = kk: 2 {2=9, 1=0}
|   bmi_cat = bb
|   |   rs9347691 = aa
|   |   |   rs6851444 = aa
|   |   |   |   rs11086671 = aa
```

```
|  |  |  |  |  rs6887293 = dd: 2 {2=3, 1=0}
|  |  |  |  |  rs6887293 = kk
|  |  |  |  |  |  rs3812906 = dd: 2 {2=2, 1=0}
|  |  |  |  |  |  rs3812906 = kk
|  |  |  |  |  |  |  rs6549458 = aa
|  |  |  |  |  |  |  |  rs11944117 = aa: 1 {2=0, 1=27}
|  |  |  |  |  |  |  |  rs11944117 = dd: 2 {2=1, 1=0}
|  |  |  |  |  |  |  rs6549458 = cc: 2 {2=1, 1=0}
|  |  |  |  rs11086671 = dd: 2 {2=3, 1=0}
|  |  |  rs6851444 = dd: 2 {2=5, 1=0}
|  |  rs9347691 = dd
|  |  |  rs2666205 = aa: 2 {2=21, 1=0}
|  |  |  rs2666205 = cc: 1 {2=0, 1=1}
|  bmi_cat = cc
|  |  rs7010457 = aa
|  |  |  rs10854395 = aa
|  |  |  |  rs12644498 = aa: 1 {2=0, 1=2}
|  |  |  |  rs12644498 = dd: 2 {2=17, 1=0}
|  |  |  |  rs12644498 = kk
|  |  |  |  |  rs12247568 = aa
|  |  |  |  |  |  rs6686571 = aa
|  |  |  |  |  |  |  rs7843255 = dd: 1 {2=0, 1=2}
|  |  |  |  |  |  |  rs7843255 = kk
|  |  |  |  |  |  |  |  rs504207 = aa: 2 {2=36, 1=0}
|  |  |  |  |  |  |  |  rs504207 = dd
|  |  |  |  |  |  |  |  |  rs12119983 = aa: 2 {2=9, 1=0}
|  |  |  |  |  |  |  |  |  rs12119983 = cc: 1 {2=0, 1=2}
|  |  |  |  |  |  |  |  rs504207 = kk: 1 {2=0, 1=1}
|  |  |  |  |  |  rs6686571 = cc
|  |  |  |  |  |  |  rs6708126 = aa: 1 {2=0, 1=6}
|  |  |  |  |  |  |  rs6708126 = dd: 2 {2=3, 1=0}
|  |  |  |  |  rs12247568 = dd: 1 {2=0, 1=2}
|  |  |  rs10854395 = dd
|  |  |  |  rs2853668 = cc: 2 {2=4, 1=0}
|  |  |  |  rs2853668 = hh
|  |  |  |  |  rs524534 = aa: 1 {2=0, 1=13}
|  |  |  |  |  rs524534 = dd: 2 {2=1, 1=0}
|  |  |  |  |  rs524534 = kk: 2 {2=1, 1=0}
|  |  |  rs10854395 = kk: 2 {2=1, 1=0}
|  |  rs7010457 = dd: 1 {2=0, 1=6}
|  bmi_cat = dd: 2 {2=7, 1=0}
ethni = L
|  bmi_cat = aa
|  |  rs17799219 = dd: 2 {2=6, 1=0}
|  |  rs17799219 = kk
|  |  |  rs7183502 = aa
|  |  |  |  rs13011951 = aa
|  |  |  |  |  rs12266639 = dd: 2 {2=2, 1=0}
|  |  |  |  |  rs12266639 = kk
|  |  |  |  |  |  rs2826802 = dd: 2 {2=1, 1=0}
|  |  |  |  |  |  rs2826802 = kk
|  |  |  |  |  |  |  rs7024840 = aa: 1 {2=0, 1=23}
|  |  |  |  |  |  |  rs7024840 = hh: 2 {2=1, 1=0}
```

```
|  |  |  |  rs13011951 = dd: 2 {2=4, 1=0}
|  |  |  rs7183502 = dd: 2 {2=4, 1=0}
|  bmi_cat = bb
|  |  rs197265 = aa
|  |  |  rs17363393 = dd: 1 {2=0, 1=10}
|  |  |  rs17363393 = kk
|  |  |  |  rs280986 = dd: 2 {2=13, 1=0}
|  |  |  |  rs280986 = kk
|  |  |  |  |  rs11790106 = dd: 2 {2=8, 1=0}
|  |  |  |  |  rs11790106 = kk
|  |  |  |  |  |  rs11126869 = dd: 2 {2=5, 1=0}
|  |  |  |  |  |  rs11126869 = kk
|  |  |  |  |  |  |  fh_prca = aa
|  |  |  |  |  |  |  |  pa_cat = cc: 2 {2=1, 1=0}
|  |  |  |  |  |  |  |  pa_cat = dd: 1 {2=0, 1=1}
|  |  |  |  |  |  |  fh_prca = jj: 2 {2=6, 1=0}
|  |  |  |  |  |  |  fh_prca = ww
|  |  |  |  |  |  |  |  rs7775829 = cc: 2 {2=3, 1=0}
|  |  |  |  |  |  |  |  rs7775829 = hh
|  |  |  |  |  |  |  |  |  rs9401290 = aa
|  |  |  |  |  |  |  |  |  |  rs17284653 = dd: 2 {2=2, 1=0}
|  |  |  |  |  |  |  |  |  |  rs17284653 = kk
|  |  |  |  |  |  |  |  |  |  |  MitoG752A = kk
|  |  |  |  |  |  |  |  |  |  |  |  rs1379015 = dd: 2 {2=1, 1=0}
|  |  |  |  |  |  |  |  |  |  |  |  rs1379015 = kk
|  |  |  |  |  |  |  |  |  |  |  |  |  rs1965340 = aa: 1 {2=0, 1=34}
|  |  |  |  |  |  |  |  |  |  |  |  |  rs1965340 = dd: 2 {2=1, 1=0}
|  |  |  |  |  |  |  |  |  |  |  MitoG752A = xx: 2 {2=1, 1=0}
|  |  |  |  |  |  |  |  |  rs9401290 = dd
|  |  |  |  |  |  |  |  |  |  rs6704731 = aa: 2 {2=8, 1=0}
|  |  |  |  |  |  |  |  |  |  rs6704731 = dd: 1 {2=0, 1=3}
|  |  |  |  |  |  |  |  |  rs9401290 = kk: 1 {2=0, 1=1}
|  |  rs197265 = dd: 1 {2=0, 1=15}
|  |  rs197265 = kk: 2 {2=2, 1=0}
|  |  rs197265 = xx: 1 {2=0, 1=1}
|  bmi_cat = cc
|  |  fh_prca = aa
|  |  |  rs6475584 = aa: 1 {2=0, 1=12}
|  |  |  rs6475584 = dd: 2 {2=4, 1=0}
|  |  |  rs6475584 = kk: 1 {2=0, 1=1}
|  |  fh_prca = jj
|  |  |  rs7876199 = aa: 1 {2=0, 1=7}
|  |  |  rs7876199 = kk
|  |  |  |  rs17673975 = aa
|  |  |  |  |  rs6779266 = cc: 1 {2=0, 1=1}
|  |  |  |  |  rs6779266 = hh
|  |  |  |  |  |  rs7024840 = aa
|  |  |  |  |  |  |  rs16863955 = aa: 2 {2=20, 1=0}
|  |  |  |  |  |  |  rs16863955 = dd: 1 {2=0, 1=1}
|  |  |  |  |  |  rs7024840 = cc: 1 {2=0, 1=1}
|  |  |  |  rs17673975 = dd: 1 {2=0, 1=4}
|  |  fh_prca = ww
|  |  |  pa_cat = aa
```

```
|  |  |  |  rs9963110 = cc: 2 {2=6, 1=0}
|  |  |  |  rs9963110 = hh
|  |  |  |  |  rs960278 = aa
|  |  |  |  |  |  rs2115101 = aa
|  |  |  |  |  |  |  rs2602296 = aa
|  |  |  |  |  |  |  |  rs17400029 = dd: 2 {2=1, 1=0}
|  |  |  |  |  |  |  |  rs17400029 = kk
|  |  |  |  |  |  |  |  |  rs2948268 = aa
|  |  |  |  |  |  |  |  |  |  rs11685549 = dd: 2 {2=1, 1=0}
|  |  |  |  |  |  |  |  |  |  rs11685549 = kk
|  |  |  |  |  |  |  |  |  |  |  rs6676372 = dd
|  |  |  |  |  |  |  |  |  |  |  |  rs4793790 = aa: 1 {2=0, 1=2}
|  |  |  |  |  |  |  |  |  |  |  |  rs4793790 = dd: 2 {2=3, 1=0}
|  |  |  |  |  |  |  |  |  |  |  |  rs6676372 = kk: 1 {2=0, 1=28}
|  |  |  |  |  |  |  |  |  |  rs2948268 = dd: 2 {2=1, 1=0}
|  |  |  |  |  |  |  rs2602296 = dd: 2 {2=2, 1=0}
|  |  |  |  |  |  rs2115101 = dd
|  |  |  |  |  |  |  eversmoke = aa: 2 {2=5, 1=0}
|  |  |  |  |  |  |  eversmoke = ww: 1 {2=0, 1=1}
|  |  |  |  |  rs960278 = dd: 2 {2=6, 1=0}
|  |  |  pa_cat = bb
|  |  |  |  rs2711134 = aa
|  |  |  |  |  rs4517938 = dd: 1 {2=0, 1=12}
|  |  |  |  |  rs4517938 = kk
|  |  |  |  |  |  rs517036 = aa
|  |  |  |  |  |  |  rs7843255 = dd: 1 {2=0, 1=6}
|  |  |  |  |  |  |  rs7843255 = kk
|  |  |  |  |  |  |  |  rs7562894 = dd: 1 {2=0, 1=3}
|  |  |  |  |  |  |  |  rs7562894 = kk
|  |  |  |  |  |  |  |  |  rs17595858 = dd: 1 {2=0, 1=2}
|  |  |  |  |  |  |  |  |  rs17595858 = kk
|  |  |  |  |  |  |  |  |  |  rs5972169 = aa: 1 {2=0, 1=2}
|  |  |  |  |  |  |  |  |  |  rs5972169 = hh
|  |  |  |  |  |  |  |  |  |  |  rs4782945 = cc: 1 {2=0, 1=2}
|  |  |  |  |  |  |  |  |  |  |  rs4782945 = hh
|  |  |  |  |  |  |  |  |  |  |  |  rs12243805 = dd: 1 {2=0, 1=2}
|  |  |  |  |  |  |  |  |  |  |  |  rs12243805 = kk
|  |  |  |  |  |  |  |  |  |  |  |  |  rs1454186 = aa: 1 {2=0, 1=1}
|  |  |  |  |  |  |  |  |  |  |  |  |  rs1454186 = kk
|  |  |  |  |  |  |  |  |  |  |  |  |  |  rs4827384 = aa
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs11221701 = aa
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs501700 = aa: 2 {2=1, 1=0}
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs501700 = dd: 1 {2=0, 1=1}
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs501700 = kk: 2 {2=30, 1=0}
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  rs11221701 = dd: 1 {2=0, 1=1}
|  |  |  |  |  |  |  |  |  |  |  |  |  |  rs4827384 = kk: 1 {2=0, 1=1}
|  |  |  |  |  |  rs517036 = dd: 1 {2=0, 1=11}
|  |  |  |  |  |  rs517036 = kk: 1 {2=0, 1=2}
|  |  |  |  rs2711134 = dd: 2 {2=8, 1=0}
|  |  |  pa_cat = cc
|  |  |  |  currsmoke = aa: 2 {2=8, 1=0}
|  |  |  |  currsmoke = ww
|  |  |  |  |  rs6686571 = aa
```

```
|  |  |  |  |  |  rs17432165 = aa: 1 {2=0, 1=18}
|  |  |  |  |  |  rs17432165 = dd: 2 {2=1, 1=0}
|  |  |  |  |  rs6686571 = cc: 2 {2=3, 1=0}
|  |  |  |  |  rs6686571 = hh: 2 {2=1, 1=0}
|  |  |  pa_cat = dd
|  |  |  |  rs1470494 = aa
|  |  |  |  |  rs744346 = aa
|  |  |  |  |  |  rs12644498 = dd: 1 {2=0, 1=2}
|  |  |  |  |  |  rs12644498 = kk
|  |  |  |  |  |  |  rs6774902 = dd: 1 {2=0, 1=1}
|  |  |  |  |  |  |  rs6774902 = kk
|  |  |  |  |  |  |  |  rs6747704 = dd: 1 {2=0, 1=1}
|  |  |  |  |  |  |  |  rs6747704 = kk
|  |  |  |  |  |  |  |  |  rs17152800 = dd: 1 {2=0, 1=1}
|  |  |  |  |  |  |  |  |  rs17152800 = kk: 2 {2=26, 1=0}
|  |  |  |  |  rs744346 = cc: 1 {2=0, 1=3}
|  |  |  |  rs1470494 = cc
|  |  |  |  |  rs6549458 = aa: 1 {2=0, 1=10}
|  |  |  |  |  rs6549458 = hh: 2 {2=1, 1=0}
|  |  |  pa_cat = ee
|  |  |  |  rs10068915 = aa: 2 {2=2, 1=0}
|  |  |  |  rs10068915 = dd: 1 {2=0, 1=11}
|  |  |  |  rs10068915 = kk
|  |  |  |  |  rs1020235 = aa
|  |  |  |  |  |  rs10106027 = dd: 1 {2=0, 1=4}
|  |  |  |  |  |  rs10106027 = kk
|  |  |  |  |  |  |  rs17111584 = aa
|  |  |  |  |  |  |  |  rs17178580 = aa: 2 {2=21, 1=0}
|  |  |  |  |  |  |  |  rs17178580 = cc: 1 {2=0, 1=1}
|  |  |  |  |  |  |  rs17111584 = dd: 1 {2=0, 1=2}
|  |  |  |  |  rs1020235 = dd: 1 {2=0, 1=7}
|  |  |  pa_cat = xx
|  |  |  |  ethanol_ca = aa: 2 {2=3, 1=0}
|  |  |  |  ethanol_ca = bb: 1 {2=0, 1=5}
|  bmi_cat = dd
|  |  d_lyco_cat = aa
|  |  |  eversmoke = aa: 1 {2=0, 1=9}
|  |  |  eversmoke = ww: 2 {2=1, 1=0}
|  |  d_lyco_cat = bb
|  |  |  rs3760903 = dd
|  |  |  |  rs12266639 = dd: 2 {2=3, 1=0}
|  |  |  |  rs12266639 = kk
|  |  |  |  |  rs11584032 = dd: 1 {2=0, 1=19}
|  |  |  |  |  rs11584032 = xx: 2 {2=1, 1=0}
|  |  |  rs3760903 = kk: 2 {2=4, 1=0}
|  |  d_lyco_cat = cc
|  |  |  rs4562278 = aa: 1 {2=0, 1=7}
|  |  |  rs4562278 = cc
|  |  |  |  fh_prca = aa: 1 {2=0, 1=1}
|  |  |  |  fh_prca = ww: 2 {2=8, 1=0}
|  |  d_lyco_cat = dd
|  |  |  rs17363393 = dd: 1 {2=0, 1=3}
|  |  |  rs17363393 = kk
```

```
|  |  |  |  rs11885120 = dd: 1 {2=0, 1=2}
|  |  |  |  rs11885120 = kk
|  |  |  |  |  rs6779266 = cc: 1 {2=0, 1=1}
|  |  |  |  |  rs6779266 = hh: 2 {2=26, 1=0}
|  |  d_lyco_cat = ee
|  |  |  rs7584223 = aa
|  |  |  |  rs7152946 = aa
|  |  |  |  |  rs340542 = aa: 2 {2=14, 1=0}
|  |  |  |  |  rs340542 = dd
|  |  |  |  |  |  pa_cat = bb: 1 {2=0, 1=3}
|  |  |  |  |  |  pa_cat = cc: 2 {2=1, 1=0}
|  |  |  |  rs7152946 = cc: 1 {2=0, 1=7}
|  |  |  rs7584223 = dd: 1 {2=0, 1=10}
|  |  d_lyco_cat = xx: 2 {2=3, 1=0}
```

APPENDIX D-) Decision Tree Structure of Hybrid System on Melanoma

```
Gender_ = F
|  moles = aa: aa {aa=146, bb=0}
|  moles = bb
|  |  dysplastic_nevi = aa: aa {aa=2, bb=0}
|  |  dysplastic_nevi = bb
|  |  |  skin_color = aa
|  |  |  |  rs3790623 = dd: aa {aa=4, bb=0}
|  |  |  |  rs3790623 = kk: bb {aa=0, bb=5}
|  |  |  skin_color = bb
|  |  |  |  eye_color = aa
|  |  |  |  |  rs1157492 = ee: aa {aa=3, bb=0}
|  |  |  |  |  rs1157492 = ff: aa {aa=2, bb=0}
|  |  |  |  |  rs1157492 = hh: bb {aa=0, bb=6}
|  |  |  |  eye_color = bb
|  |  |  |  |  rs1481003 = ee: bb {aa=0, bb=4}
|  |  |  |  |  rs1481003 = ff: aa {aa=4, bb=0}
|  |  |  |  eye_color = cc: bb {aa=0, bb=1}
|  |  |  |  eye_color = dd
|  |  |  |  |  freckle = aa: aa {aa=2, bb=0}
|  |  |  |  |  freckle = bb: bb {aa=0, bb=1}
|  |  |  |  eye_color = ee
|  |  |  |  |  rs491322 = aa: aa {aa=3, bb=0}
|  |  |  |  |  rs491322 = dd: bb {aa=0, bb=8}
|  |  |  skin_color = cc
|  |  |  |  eye_color = aa
|  |  |  |  |  hair_color = aa
|  |  |  |  |  |  rs10046376 = ee: bb {aa=0, bb=6}
|  |  |  |  |  |  rs10046376 = ff: aa {aa=2, bb=0}
|  |  |  |  |  hair_color = bb
|  |  |  |  |  |  rs10179662 = ff: bb {aa=0, bb=3}
|  |  |  |  |  |  rs10179662 = hh: aa {aa=2, bb=0}
|  |  |  |  |  hair_color = cc
|  |  |  |  |  |  rs12102923 = dd: aa {aa=3, bb=0}
|  |  |  |  |  |  rs12102923 = kk: bb {aa=0, bb=6}
|  |  |  |  eye_color = bb
|  |  |  |  |  rs10898422 = ff: aa {aa=1, bb=0}
|  |  |  |  |  rs10898422 = hh: bb {aa=0, bb=13}
|  |  |  |  eye_color = dd
|  |  |  |  |  rs1467414 = ff: aa {aa=2, bb=0}
|  |  |  |  |  rs1467414 = hh: bb {aa=0, bb=9}
|  |  |  |  eye_color = ee: bb {aa=0, bb=13}
|  |  |  skin_color = dd
|  |  |  |  eye_color = aa
|  |  |  |  |  rs10804551 = ee: bb {aa=0, bb=10}
|  |  |  |  |  rs10804551 = ff: aa {aa=1, bb=0}
```

```
|  |  |  |  eye_color = bb
|  |  |  |  |  rs10097728 = dd: aa {aa=1, bb=0}
|  |  |  |  |  rs10097728 = kk: bb {aa=0, bb=4}
|  |  |  |  eye_color = dd: bb {aa=0, bb=15}
|  |  |  |  eye_color = ee
|  |  |  |  |  rs10028824 = ee: bb {aa=0, bb=9}
|  |  |  |  |  rs10028824 = gg: aa {aa=2, bb=0}
|  |  |  skin_color = ee
|  |  |  |  eye_color = aa
|  |  |  |  |  rs12207699 = aa: aa {aa=3, bb=0}
|  |  |  |  |  rs12207699 = dd: bb {aa=0, bb=6}
|  |  |  |  eye_color = bb: bb {aa=0, bb=5}
|  |  |  |  eye_color = dd
|  |  |  |  |  rs17747388 = ee
|  |  |  |  |  |  rs11675159 = ee: bb {aa=0, bb=12}
|  |  |  |  |  |  rs11675159 = ff: aa {aa=1, bb=0}
|  |  |  |  |  rs17747388 = ff: aa {aa=3, bb=0}
|  |  |  |  eye_color = ee
|  |  |  |  |  rs10039077 = dd: bb {aa=0, bb=4}
|  |  |  |  |  rs10039077 = kk: aa {aa=1, bb=0}
|  |  |  skin_color = ff
|  |  |  |  rs10435832 = dd: aa {aa=1, bb=0}
|  |  |  |  rs10435832 = kk: bb {aa=0, bb=12}
|  |  |  skin_color = gg: bb {aa=0, bb=4}
|  |  |  skin_color = hh: bb {aa=0, bb=3}
|  moles = qq: bb {aa=0, bb=1}
Gender_ = M
|  moles = aa: aa {aa=163, bb=0}
|  moles = bb
|  |  dysplastic_nevi = aa: aa {aa=4, bb=0}
|  |  dysplastic_nevi = bb
|  |  |  skin_color = aa
|  |  |  |  rs2392695 = aa
|  |  |  |  |  rs10033683 = aa: aa {aa=4, bb=0}
|  |  |  |  |  rs10033683 = kk: bb {aa=0, bb=1}
|  |  |  |  rs2392695 = dd: bb {aa=0, bb=8}
|  |  |  skin_color = bb
|  |  |  |  eye_color = aa
|  |  |  |  |  hair_color = aa
|  |  |  |  |  |  rs11143592 = ff: bb {aa=0, bb=3}
|  |  |  |  |  |  rs11143592 = hh: aa {aa=8, bb=0}
|  |  |  |  |  hair_color = bb
|  |  |  |  |  |  rs10174761 = dd: bb {aa=0, bb=3}
|  |  |  |  |  |  rs10174761 = kk: aa {aa=4, bb=0}
|  |  |  |  |  hair_color = cc
|  |  |  |  |  |  rs2455854 = ee: bb {aa=0, bb=8}
|  |  |  |  |  |  rs2455854 = ff: aa {aa=3, bb=0}
|  |  |  |  eye_color = bb
|  |  |  |  |  rs2010344 = ff
|  |  |  |  |  |  rs10051060 = aa: aa {aa=4, bb=0}
|  |  |  |  |  |  rs10051060 = dd: bb {aa=0, bb=1}
|  |  |  |  |  rs2010344 = hh: bb {aa=0, bb=8}
|  |  |  |  eye_color = cc: bb {aa=0, bb=1}
```

```
|  |  |  |  eye_color = dd
|  |  |  |  |  rs1043848 = ee: aa {aa=4, bb=0}
|  |  |  |  |  rs1043848 = ff: bb {aa=0, bb=4}
|  |  |  |  eye_color = ee
|  |  |  |  |  rs10171924 = dd: aa {aa=2, bb=0}
|  |  |  |  |  rs10171924 = kk: bb {aa=0, bb=4}
|  |  |  skin_color = cc
|  |  |  |  eye_color = aa
|  |  |  |  |  hair_color = aa
|  |  |  |  |  |  rs10181522 = cc: bb {aa=0, bb=7}
|  |  |  |  |  |  rs10181522 = hh
|  |  |  |  |  |  |  rs10051060 = aa: aa {aa=7, bb=0}
|  |  |  |  |  |  |  rs10051060 = dd: bb {aa=0, bb=1}
|  |  |  |  |  hair_color = bb: aa {aa=2, bb=0}
|  |  |  |  |  hair_color = cc
|  |  |  |  |  |  sunburn = aa
|  |  |  |  |  |  |  freckle = aa
|  |  |  |  |  |  |  |  rs17055114 = aa: bb {aa=0, bb=7}
|  |  |  |  |  |  |  |  rs17055114 = dd
|  |  |  |  |  |  |  |  |  rs10000494 = aa: aa {aa=6, bb=0}
|  |  |  |  |  |  |  |  |  rs10000494 = dd: bb {aa=0, bb=1}
|  |  |  |  |  |  |  freckle = bb
|  |  |  |  |  |  |  |  rs10417461 = dd: aa {aa=2, bb=0}
|  |  |  |  |  |  |  |  rs10417461 = kk: bb {aa=0, bb=5}
|  |  |  |  |  |  sunburn = bb
|  |  |  |  |  |  |  rs10186968 = ee: bb {aa=0, bb=7}
|  |  |  |  |  |  |  rs10186968 = ff: aa {aa=2, bb=0}
|  |  |  |  eye_color = bb
|  |  |  |  |  rs2288704 = aa: aa {aa=4, bb=0}
|  |  |  |  |  rs2288704 = cc: bb {aa=0, bb=7}
|  |  |  |  |  rs2288704 = hh: aa {aa=1, bb=0}
|  |  |  |  eye_color = cc: bb {aa=0, bb=2}
|  |  |  |  eye_color = dd
|  |  |  |  |  rs11675159 = ee: bb {aa=0, bb=21}
|  |  |  |  |  rs11675159 = ff: aa {aa=1, bb=0}
|  |  |  |  eye_color = ee
|  |  |  |  |  rs12466022 = aa: aa {aa=3, bb=0}
|  |  |  |  |  rs12466022 = cc
|  |  |  |  |  |  rs10149366 = aa: aa {aa=2, bb=0}
|  |  |  |  |  |  rs10149366 = dd: bb {aa=0, bb=3}
|  |  |  |  |  rs12466022 = hh: bb {aa=0, bb=7}
|  |  |  skin_color = dd
|  |  |  |  eye_color = aa
|  |  |  |  |  hair_color = aa
|  |  |  |  |  |  rs1032002 = ee: aa {aa=1, bb=0}
|  |  |  |  |  |  rs1032002 = hh: bb {aa=0, bb=7}
|  |  |  |  |  hair_color = cc
|  |  |  |  |  |  rs12200877 = ee: bb {aa=0, bb=1}
|  |  |  |  |  |  rs12200877 = ff: aa {aa=3, bb=0}
|  |  |  |  |  |  rs12200877 = hh: bb {aa=0, bb=13}
|  |  |  |  eye_color = bb
|  |  |  |  |  rs2768343 = dd
|  |  |  |  |  |  untanned = cc: aa {aa=3, bb=0}
```

| | | | | | untanned = dd: bb {aa=0, bb=1}
| | | | | rs2768343 = kk: bb {aa=0, bb=9}
| | | | eye_color = cc: bb {aa=0, bb=2}
| | | | eye_color = dd
| | | | | rs4663447 = ee: bb {aa=0, bb=15}
| | | | | rs4663447 = ff
| | | | | | family_history = aa: aa {aa=3, bb=0}
| | | | | | family_history = bb: bb {aa=0, bb=2}
| | | | eye_color = ee
| | | | | rs10186968 = ee: bb {aa=0, bb=11}
| | | | | rs10186968 = ff: aa {aa=1, bb=0}
| | | skin_color = ee
| | | | eye_color = aa
| | | | | rs7690372 = ee
| | | | | | rs11896268 = ff: aa {aa=1, bb=0}
| | | | | | rs11896268 = hh: bb {aa=0, bb=10}
| | | | | rs7690372 = ff: bb {aa=0, bb=4}
| | | | | rs7690372 = hh: aa {aa=3, bb=0}
| | | | eye_color = bb
| | | | | rs10051060 = aa: bb {aa=0, bb=5}
| | | | | rs10051060 = dd: aa {aa=1, bb=0}
| | | | eye_color = cc
| | | | | rs10011621 = aa: aa {aa=2, bb=0}
| | | | | rs10011621 = cc: bb {aa=0, bb=1}
| | | | eye_color = dd
| | | | | hair_color = cc
| | | | | | rs10211242 = ff: aa {aa=4, bb=0}
| | | | | | rs10211242 = hh: bb {aa=0, bb=17}
| | | | | hair_color = dd
| | | | | | rs10008985 = cc: aa {aa=1, bb=0}
| | | | | | rs10008985 = hh: bb {aa=0, bb=6}
| | | | eye_color = ee
| | | | | rs2246095 = ee: aa {aa=2, bb=0}
| | | | | rs2246095 = ff
| | | | | | hair_color = cc: bb {aa=0, bb=13}
| | | | | | hair_color = dd: aa {aa=1, bb=0}
| | | | | rs2246095 = hh: bb {aa=0, bb=2}
| | | | | rs2246095 = xx: aa {aa=1, bb=0}
| | | skin_color = ff
| | | | eye_color = aa: bb {aa=0, bb=10}
| | | | eye_color = bb
| | | | | rs11108542 = dd: aa {aa=1, bb=0}
| | | | | rs11108542 = kk: bb {aa=0, bb=7}
| | | | eye_color = cc: bb {aa=0, bb=1}
| | | | eye_color = dd
| | | | | rs11988061 = ee: bb {aa=0, bb=14}
| | | | | rs11988061 = ff: aa {aa=1, bb=0}
| | | | eye_color = ee
| | | | | rs10406787 = aa: bb {aa=0, bb=5}
| | | | | rs10406787 = dd: aa {aa=2, bb=0}
| | | skin_color = gg
| | | | rs7265926 = ee: bb {aa=0, bb=19}
| | | | rs7265926 = ff

| | | | | rs10238965 = ff: bb {aa=0, bb=2}
| | | | | rs10238965 = hh: aa {aa=4, bb=0}
| | | skin_color = hh: bb {aa=0, bb=5}
| moles = tt: aa {aa=1, bb=0}
| moles = ww: aa {aa=1, bb=0}

APPENDIX E-) SNP List of Prostate Cancer Found by Hybrid System

RS-Ids			
rs11729739	rs2442602	rs17363393	rs7562894
rs17701543	rs3093679	rs280986	rs17595858
rs9848588	rs9347691	rs11790106	rs5972169
rs964130	rs6851444	rs11126869	rs4782945
rs10195113	rs11086671	rs7775829	rs12243805
rs1433369	rs6887293	rs9401290	rs1454186
rs12733054	rs3812906	rs17284653	rs4827384
rs17375010	rs6549458	rs1379015	rs11221701
rs766045	rs2666205	rs1965340	rs501700
rs12201462	rs7010457	rs6704731	rs17432165
rs4908656	rs10854395	rs6475584	rs1470494
rs9462806	rs12644498	rs7876199	rs744346
rs1974562	rs12247568	rs17673975	rs6774902
rs10954845	rs6686571	rs6779266	rs6747704
rs10745253	rs504207	rs16863955	rs17152800
rs12980509	rs12119983	rs9963110	rs10068915
rs2296370	rs6708126	rs960278	rs1020235
rs7843255	rs2853668	rs2115101	rs10106027
rs2194505	rs524534	rs2602296	rs17111584
rs2120806	rs17799219	rs17400029	rs17178580
rs10517581	rs7183502	rs2948268	rs3760903
rs2103869	rs13011951	rs11685549	rs11584032
rs10788555	rs12266639	rs6676372	rs4562278
rs7067548	rs2826802	rs4793790	rs11885120
rs17001078	rs7024840	rs2711134	rs7584223
rs918285	rs197265	rs4517938	rs7152946
rs7034430		rs517036	rs340542

APPENDIX F-) SNPnexus Results of Prostate Cancer

Gene	Entrez gene	Phenotype	Disease Class	Pubmed
GPR109A	338442	schizophrenia | bipolar disorder	PSYCH	19502010
DDEF2	8853	multiple sclerosis	IMMUNE	19010793
SEMA5B	54437	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
SEMA5B	54437	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
MTAP	4507	diabetes, type 2	METABOLIC	11985785
MTAP	4507	Melanoma|Nevus|Precancerous Conditions|Skin Neoplasms	CANCER	19578365
MTAP	4507	Brain Ischemia|Diabetes Mellitus|Hyperlipidemias|Hypertension|Intracranial Embolism	CARDIOVASCULAR	19427650
MTAP	4507	Myocardial Infarction	CARDIOVASCULAR	19272367
MTAP	4507		UNKNOWN	19887491
MTAP	4507	Precursor Cell Lymphoblastic Leukemia-Lymphoma	CANCER	19665068
MTAP	4507	Cutaneous nevi	OTHER	19578365
MTAP	4507	melanoma|Nevus|Skin Neoplasms|Sunburn	CANCER	20647408
MTAP	4507	melanoma|Nevus|Skin Neoplasms	CANCER	20574843
CSMD1	64478	Tobacco Use Disorder	CHEMDEPENDENCY	18519826
CSMD1	64478	Chromosomal Instability|Cystadenocarcinoma, Serous|Ovarian Neoplasms	CANCER	19383911
CSMD1	64478	Celiac Disease|	IMMUNE	19240061
CSMD1	64478	Mucocutaneous Lymph Node Syndrome	IMMUNE	19132087
CSMD1	64478	hypertension	CARDIOVASCULAR	19960030
CSMD1	64478	multiple sclerosis	IMMUNE	19010793

CSMD1	**64478**	**smoking cessation**	**CHEMDEPENDENCY**	**20235792**
CSMD1	64478	smoking cessation	CHEMDEPENDENCY	20235792
CSMD1	64478	Psoriasis	IMMUNE	20953187
CSMD1	64478	Peripheral Vascular Diseases	CARDIOVASCULAR	20610895
CSMD1	64478	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
TNFAIP1	7126	anorexia nervosa	PSYCH	11702059
TNFAIP1	7126	Malaria infection	INFECTION	11929592
TNFAIP1	7126	arthritis, rheumatoid	IMMUNE	11791643
TNFAIP1	7126	systemic lupus erythematosus	IMMUNE	11704801
TNFAIP1	7126	nephropathy, IgA	RENAL	11849463
TNFAIP1	7126	sarcoidosis	IMMUNE	12039524
TNFAIP1	7126	Asthma|Obesity	METABOLIC	19196817
TNFAIP1	7126	Alzheimer's disease	NEUROLOGICAL	19141999
ATP2B2	491	schizophrenia	PSYCH	19850283
ATP2B2	491	serum metabolites	METABOLIC	19043545
ATP2B2	491	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
ATP2B2	491	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
FCAMR	83953	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
DPT	1805	Cardiovascular Diseases	CARDIOVASCULAR	17903295
DPT	1805	Hypertension	CARDIOVASCULAR	19536175
DPT	1805	Osteoporosis	METABOLIC	19064610
DPT	1805	morbidity-free survival	AGING	17903295
USP24	23358	Parkinson's disease	NEUROLOGICAL	16917932
USP24	23358	Parkinson's disease	NEUROLOGICAL	20302855
NCAM2	4685	several psychiatric disorders	PSYCH	19086053
NCAM2	4685	Alzheimer Disease|Alzheimer's Disease	NEUROLOGICAL	20932310

NCAM2	4685	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
ST6GALNAC3	256435	Tobacco Use Disorder	CHEMDEPENDENCY	18519826
ST6GALNAC3	256435	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
ST6GALNAC3	256435	Alcoholism	CHEMDEPENDENCY	20421487
IL1RAPL1	11141	schizophrenia | autism	DEVELOPMENTAL	19736351
IL1RAPL1	11141	cognitive ability	NEUROLOGICAL	18467032
IL1RAPL1	11141	Chorioamnionitis|Fetal Membranes, Premature Rupture|Infection of amniotic sac and membranes|Obstetric Labor, Premature|Pre-Eclampsia|Premature Birth	REPRODUCTION	20452482
IL1RAPL1	11141	Chorioamnionitis|Fetal Membranes, Premature Rupture|Infection of amniotic sac and membranes	REPRODUCTION	20673868
IL1RAPL1	11141	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
ECOP		Apoplexy|Cerebral Hemorrhage|Cerebral Hemorrhages|Intracranial Hemorrhages|Stroke|Subarachnoid Hemorrhage	CARDIOVASCULAR	20198315
SHB	6461	Alzheimer's disease	NEUROLOGICAL	19141999
SHB	6461	Brain structure	NEUROLOGICAL	20171287
IMMP2L	83943	Autism	PSYCH	19401682
IMMP2L	83943	Celiac Disease|	IMMUNE	19240061
IMMP2L	83943	Autism	PSYCH	19058789
IMMP2L	83943	ADHD | attention-deficit hyperactivity disorder	PSYCH	19546859
IMMP2L	83943	Cognitive performance	NEUROLOGICAL	19734545
IMMP2L	83943	Cognitive performance	NEUROLOGICAL	19734545
IMMP2L	83943	Acquired Immunodeficiency Syndrome|Disease Progression	INFECTION	20877624
IMMP2L	83943	Psychiatric Disorders	PSYCH	20398908
IMMP2L	83943	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
COBL	2324	Celiac Disease|	IMMUNE	192400

	2			61
COBL	23242	Autism	PSYCH	19058789
COBL	23242	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
SPTBN1	6711	Fractures, Bone	METABOLIC	19801982
SPTBN1	6711	Dengue Hemorrhagic Fever	INFECTION	20588308
SPTBN1	6711	Osteoporosis	METABOLIC	20554715
SPTBN1	6711	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
SPTBN1	6711	HIV Infections|[X]Human immunodeficiency virus disease	INFECTION	21083371
LRRTM4	80059	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
GPR109A	338442	schizophrenia | bipolar disorder	PSYCH	19502010
CPM	1368	bronchodilator response	IMMUNE	18617639
SMARCA4	6597	Cardiovascular Diseases	CARDIOVASCULAR	19913121
SMARCA4	6597	breast cancer	CANCER	19183483
SMARCA4	6597	plasma HDL cholesterol (HDL-C) levels	METABOLIC	18660489
SMARCA4	6597	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
SMARCA4	6597	Coronary Artery Disease	CARDIOVASCULAR	20810930
ANGPT2	285	pregnancy loss, recurrent	OTHER	14556828
ANGPT2	285	fetal loss, late	REPRODUCTION	15694966
ANGPT2	285	uterine leiomyomas	OTHER	16009172
ANGPT2	285	idiopathic recurrent miscarriage	OTHER	14556828
ANGPT2	285	retinopathy of prematurity	VISION	16877277
MCPH1	79648	brain size	NEUROLOGICAL	17566767
MCPH1	79648	cognitive function head circumference social intelligence	PSYCH	17251122
MCPH1	79648	Mental Retardation|Microcephaly	DEVELOPMENTAL	19267414
MCPH1	79648	Coronary Artery Disease|	CARDIOVASCULAR	18651322
MCPH1	79648	Adenocarcinoma|Pancreatic Neoplasms	CANCER	19690177

MCPH1	79648	Multiple System Atrophy	UNKNOWN	19475667
ANGPT2	285	Lymphedema	HEMATOLOGICAL	18564921
ANGPT2	285	Birth Weight|Retinopathy of Prematurity	METABOLIC	19018553
ANGPT2	285	BMI- Edema rosiglitazone or pioglitazone	PHARMACOGENOMIC	18996102
ANGPT2	285	Retinopathy of Prematurity	VISION	18568888
ANGPT2	285	Stroke	CARDIOVASCULAR	19341361
ANGPT2	285	Respiratory Distress Syndrome, Adult	UNKNOWN	19271210
MCPH1	79648	atherosclerosis	CARDIOVASCULAR	20485444
MCPH1	79648	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
MCPH1	79648	Micrencephaly |Microcephaly	DEVELOPMENTAL	20080800
MCPH1	79648	Hypercholesterolemia|LDLC levels	METABOLIC	20602615
MCPH1	79648	breast cancer	CANCER	20508983
MCPH1	79648	Microcephaly	DEVELOPMENTAL	18204051
ANGPT2	285	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
ANGPT2	285	Chorioamnionitis|Fetal Membranes, Premature Rupture|Infection of amniotic sac and membranes	REPRODUCTION	20673868
ANGPT2	285	Apoplexy|Brain Infarction|Recurrence|Stroke	CARDIOVASCULAR	20599737
ANGPT2	285	Chorioamnionitis|Fetal Membranes, Premature Rupture|Infection of amniotic sac and membranes|Obstetric Labor, Premature|Pre-Eclampsia|Premature Birth	REPRODUCTION	20452482
NRG1	3084	Schizophrenia	PSYCH	19575259
NRG1	3084	Infant, Newborn, Diseases	UNKNOWN	20472376
NRG1	3084	Schizophrenia|bipolar disorder	PSYCH	20435087
NRG1	3084	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
NRG1	3084	Alzheimer's disease	NEUROLOGICAL	20182055
NRG1	3084	Schizophrenia	PSYCH	20638435

NRG1	3084	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
NRG1	3084	Schizophrenia	PSYCH	20526724
NRG1	3084	Marijuana Abuse|Psychoses, Substance-Induced	CHEMDEPENDENCY	21041608
NRG1	3084	Schizophrenia	PSYCH	20921115
NRG1	3084	longevity	AGING	20800603
NRG1	3084	Schizophrenia	PSYCH	18282690
NRG1	3084	Schizophrenia	PSYCH	18198266
NRG1	3084	Schizophrenia	PSYCH	18032396
NRG1	3084	prepulse inhibition	UNKNOWN	17631867
NRG1	3084	Hippocampal Atrophy	UNKNOWN	18291420
NRG1	3084	Schizophrenia	PSYCH	18286587
PHACTR2	9749	Parkinson's disease	NEUROLOGICAL	19429005
PHACTR2	9749	Alzheimer's disease	NEUROLOGICAL	19141999
PHACTR2	9749	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
PHACTR2	9749	Multiple Sclerosis	IMMUNE	20546594
PHACTR2	9749	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
GPR109A	338442	schizophrenia | bipolar disorder	PSYCH	19502010
CSMD1	64478	Tobacco Use Disorder	CHEMDEPENDENCY	18519826
CSMD1	64478	Chromosomal Instability|Cystadenocarcinoma, Serous|Ovarian Neoplasms	CANCER	19383911
CSMD1	64478	Celiac Disease|	IMMUNE	19240061
CSMD1	64478	Mucocutaneous Lymph Node Syndrome	IMMUNE	19132087
CSMD1	64478	hypertension	CARDIOVASCULAR	19960030
CSMD1	64478	multiple sclerosis	IMMUNE	19010793
CSMD1	64478	smoking cessation	CHEMDEPENDENCY	20235792
CSMD1	64478	smoking cessation	CHEMDEPENDENCY	20235792
CSMD1	6447	Psoriasis	IMMUNE	209531

	8			87
CSMD1	64478	Peripheral Vascular Diseases	CARDIOVASCULAR	20610895
CSMD1	64478	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
CAMTA1	23261	memory disturbance	PSYCH	17470457
CAMTA1	23261	hypertension	CARDIOVASCULAR	19851296
CAMTA1	23261	Coronary Disease	CARDIOVASCULAR	19336475
CAMTA1	23261	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
CAMTA1	23261	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
CAMTA1	23261	Apoplexy|Cerebral Hemorrhage|Cerebral Hemorrhages|Intracranial Hemorrhages|Stroke|Subarachnoid Hemorrhage	CARDIOVASCULAR	20198315
CAMTA1	23261	Type 2 diabetes	METABOLIC	18210030
AGBL4	84871	Celiac Disease|	IMMUNE	19240061
AGBL4	84871	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
PACRG	135138	leprosy	INFECTION	14737177
PACRG	135138	leprosy	INFECTION	16391553
PACRG	135138	Parkinson's disease	NEUROLOGICAL	15925106
PACRG	135138	leprosy	INFECTION	14737177
PACRG	135138	Parkinson's disease	NEUROLOGICAL	19196541
PACRG	135138	Tuberculosis	INFECTION	19723394
PACRG	135138	male infertility	REPRODUCTION	19268936
PACRG	135138	Acquired Immunodeficiency Syndrome|Disease Progression	INFECTION	20877624
PACRG	135138	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
CPNE4	131034	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
TTC29	83894	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
GPR109A	338442	schizophrenia | bipolar disorder	PSYCH	19502010
DEPDC6	64798	Tobacco Use Disorder	CHEMDEPENDENCY	20379614

CSMD1	64478	Tobacco Use Disorder	CHEMDEPENDENCY	18519826
CSMD1	64478	Chromosomal Instability|Cystadenocarcinoma, Serous|Ovarian Neoplasms	CANCER	19383911
CSMD1	64478	Celiac Disease|	IMMUNE	19240061
CSMD1	64478	Mucocutaneous Lymph Node Syndrome	IMMUNE	19132087
CSMD1	64478	hypertension	CARDIOVASCULAR	19960030
CSMD1	64478	multiple sclerosis	IMMUNE	19010793
CSMD1	64478	smoking cessation	CHEMDEPENDENCY	20235792
CSMD1	64478	smoking cessation	CHEMDEPENDENCY	20235792
CSMD1	64478	Psoriasis	IMMUNE	20953187
CSMD1	64478	Peripheral Vascular Diseases	CARDIOVASCULAR	20610895
CSMD1	64478	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
SESTD1	91404	Celiac Disease|	IMMUNE	19240061
SESTD1	91404	HIV Infections|[X]Human immunodeficiency virus disease	INFECTION	21083371
SESTD1	91404	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
LRP1B	53353	cognitive ability	NEUROLOGICAL	19367585
LRP1B	53353	Aging	AGING	19367585
LRP1B	53353	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
KLF7	8609	diabetes, type 2	METABOLIC	15937668
KLF7	8609	Body Weight|Diabetes Mellitus, Type 2|Obesity|Overweight	METABOLIC	19147600
MAGI2	9863	several psychiatric disorders	PSYCH	19086053
MAGI2	9863	inflammatory bowel disease	IMMUNE	18720471
MAGI2	9863	Celiac Disease|	IMMUNE	19240061
MAGI2	9863	hippocampal atrophy	NEUROLOGICAL	19668339
MAGI2	9863	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
MAGI2	9863	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
MAGI2	9863	Celiac Disease|Down Syndrome	IMMUNE	200967

				42
MAGI2	9863	Celiac Disease|Colitis, Ulcerative	IMMUNE	17989107
BARX2	8538	Cleft Lip|Cleft Palate	DEVELOPMENTAL	20634891
ATP2C2	9914	Migraine without Aura	NEUROLOGICAL	18676988
ATP2C2	9914	ADHD | attention-deficit hyperactivity disorder	PSYCH	18839057
ATP2C2	9914	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
CD247	919	Osteoporosis	METABOLIC	19064610
CD247	919	Arthritis, Rheumatoid|	IMMUNE	19898481
CD247	919	Hypertension	CARDIOVASCULAR	19536175
CD247	919	Lupus Erythematosus, Systemic	IMMUNE	19422667
CD247	919	Celiac disease	IMMUNE	20190752
CD247	919	Scleroderma, Systemic|Systemic Scleroderma	IMMUNE	20383147
CD247	919	Lupus Erythematosus, Systemic|Systemic lupus erythematosus	IMMUNE	18178846
CD247	919	Lupus Erythematosus, Systemic|Systemic lupus erythematosus	IMMUNE	18174230
SLC8A1	6546	Hyperparathyroidism, Secondary	METABOLIC	20424473
SLC8A1	6546	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
SLC8A1	6546	Cardiovascular Diseases	CARDIOVASCULAR	20109173
RSRC1	51319	schizophrenia	PSYCH	19065146
RSRC1	51319	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
RSRC1	51319	Tobacco Use Disorder	CHEMDEPENDENCY	20379614

APPENDIX G-) SNP List of Melanoma

Female	Male	
rs3790623	rs2392695	rs12466022
rs1157492	rs10033683	rs10149366
rs1481003	rs11143592	rs1032002
rs491322	rs10174761	rs12200877
rs10046376	rs2455854	rs2768343
rs10179662	rs2010344	rs4663447
rs12102923	rs10051060	rs10186968
rs10898422	rs1043848	rs11896268
rs1467414	rs10171924	rs7690372
rs10804551	rs10181522	rs10051060
rs10097728	rs10051060	rs10011621
rs10028824	rs17055114	rs10211242
rs12207699	rs10000494	rs10008985
rs17747388	rs10417461	rs2246095
rs11675159	rs10186968	rs11108542
rs10039077	rs2288704	rs11988061
rs10435832	rs11675159	rs10406787
	rs10238965	rs7265926

APPENDIX H-) SNPnexus Results of Melanoma

Gene	Entrez gene	Phenotype	Disease Class	Pubmed
FLJ22536	401237	Cell Transformation, Neoplastic|Neuroblastoma	CANCER	18463370
FLJ22536	401237	neuroblastoma	CANCER	18463370
ERBB4	2066	lung cancer	CANCER	17487277
ERBB4	2066	Cell Transformation, Neoplastic|Melanoma|Skin Neoplasms	CANCER	19718025
ERBB4	2066	lung cancer	CANCER	20881644
ERBB4	2066	Brain Neoplasms|Glioma	CANCER	20446891
ERBB4	2066	lung cancer	CANCER	20975381
ERBB4	2066	colorectal cancer	CANCER	18094435
RYR2	6262	Acute lymphoblastic leukemia (childhood)	CANCER	19684603
GRIA1	2890	Drug Hypersensitivity|Precursor T-Cell Lymphoblastic Leukemia-Lymphoma	CANCER	20592726
PDLIM5	10611	prostate cancer	CANCER	19767753
PDLIM5	10611	prostate cancer	CANCER	20564319
PDLIM5	10611	prostate cancer	CANCER	20878950
CAMK1D	57118	prostate cancer	CANCER	20080650
CAMK1D	57118	breast cancer	CANCER	20418484
KCNK2	3776	atherosclerosis	CARDIOVASCULAR	19948975
RYR2	6262	long QT syndrome	CARDIOVASCULAR	16188589
RYR2	6262	cardiomyopathy	CARDIOVASCULAR	16769042
RYR2	6262	hypertension	CARDIOVASCULAR	17554300

RYR2	**6262**	**Arrhythmias, Cardiac|Death, Sudden, Cardiac|Heart Failure|Sudden Cardiac Death**	**CARDIOVASCULAR**	**20408814**
KCNK2	3776	Tobacco Use Disorder	CHEMDEPENDENCY	18519826
KCNK2	3776	smoking cessation	CHEMDEPENDENCY	20235792
KCNK2	3776	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
DNAH3	55567	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
ATP9A	10079	Tobacco Use Disorder	CHEMDEPENDENCY	18519826
ATP9A	10079	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
GRID2	2895	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
ERBB4	2066	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
RYR2	6262	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
GRIA1	2890	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
NRXN1	9378	nicotine dependence	CHEMDEPENDENCY	17158188
NRXN1	9378	Alcoholism	CHEMDEPENDENCY	20201926
NRXN1	9378	smoking	CHEMDEPENDENCY	20414139
NRXN1	9378	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
NRXN1	9378	Tobacco Use Disorder	CHEMDEPENDENCY	18270208

C4ORF27		Tobacco Use Disorder	CHEMDEPENDENCY	20379614
GMDS	2762	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
CAMK1D	57118	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
NAV3	89795	Tobacco Use Disorder	CHEMDEPENDENCY	20379614
ERBB4	2066	schizophrenia | autism	DEVELOPMENTAL	19736351
NRXN1	9378	schizophrenia | autism	DEVELOPMENTAL	19736351
ERBB4	2066	Celiac Disease|	IMMUNE	19240061
RYR2	6262	Multiple Sclerosis	IMMUNE	19626040
ADCK4	79934	Acquired Immunodeficiency Syndrome|Disease Progression	INFECTION	20877624
RYR2	6262	Hyperparathyroidism, Secondary	METABOLIC	20424473
GRIA1	2890	Body Weight	METABOLIC	19260139
GRIA1	2890	Weight Gain	METABOLIC	19156168
CAMK1D	57118	Posttransplantation diabetes mellitus (PTDM)	METABOLIC	
CAMK1D	57118	diabetes, type 2	METABOLIC	19020323
CAMK1D	57118	diabetes, type 2	METABOLIC	19324937
CAMK1D	57118	Diabetes Mellitus|Diabetes Mellitus, Type 2|	METABOLIC	19139842
CAMK1D	57118	diabetes, type 2	METABOLIC	19020324
CAMK1D	57118	diabetes, type 2	METABOLIC	19670153
CAMK1D	57118	diabetes, type 2	METABOLIC	19602701
CAMK1D	57118	diabetes, type 2	METABOLIC	19455301

CAMK1D	57118	diabetes, type 2	METABOLIC	19794065
CAMK1D	57118	diabetes, type 2	METABOLIC	19789630
CAMK1D	57118	diabetes, type 2	METABOLIC	19741467
CAMK1D	57118	diabetes, type 2	METABOLIC	19720844
CAMK1D	57118	diabetes, type 2	METABOLIC	19933996
CAMK1D	57118	diabetes, type 2	METABOLIC	19833888
CAMK1D	57118	type 2 diabetes	METABOLIC	18372903
CAMK1D	57118	Type 2 diabetes	METABOLIC	20161779
CAMK1D	57118	Type 2 diabetes	METABOLIC	20571754
CAMK1D	57118	obesity	METABOLIC	20712903
CAMK1D	57118	Type 2 diabetes	METABOLIC	21084393
CAMK1D	57118	Type 2 diabetes	METABOLIC	20927120
KCNK2	3776	Migraine without Aura	NEUROLOGICAL	18676988
GPR156	165829	Narcolepsy	NEUROLOGICAL	20677014
GRIA1	2890	Migraine Disorders	NEUROLOGICAL	20579352
NRXN1	9378	cognitive ability	NEUROLOGICAL	19734545
CAMK1D	57118	Alzheimer's disease	NEUROLOGICAL	16385451
CAMK1D	57118	Narcolepsy	NEUROLOGICAL	20677014
RYR2	6262	exercise treadmill test traits	OTHER	17903301
GRIA1	2890	Anthropometric traits	OTHER	19260139
KCNK2	3776	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
KCNK2	3776	antidepressant response	PHARMACOGENOMIC	18288090
ERBB4	2066	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086

RYR2	6262	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
SPEG	10290	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
NAV3	89795	Type 2 Diabetes| edema | rosiglitazone	PHARMACOGENOMIC	20628086
KCNK2	3776	major depressive disorder	PSYCH	19741570
KCNK2	3776	depression	PSYCH	19621370
ERBB4	2066	schizophrenia	PSYCH	16249994
ERBB4	2066	schizophrenia	PSYCH	16891421
ERBB4	2066	schizophrenia	PSYCH	17598910
ERBB4	2066	schizophrenia	PSYCH	16402353
ERBB4	2066	schizophrenia	PSYCH	19367581
ERBB4	2066	Schizophrenia	PSYCH	20600594
ERBB4	2066	Schizophrenia	PSYCH	20921115
GRIA1	2890	schizophrenia	PSYCH	16526023
GRIA1	2890	several psychiatric disorders	PSYCH	19086053
GRIA1	2890	Bipolar Disorder	PSYCH	18484081
GRIA1	2890	Bipolar Disorder	PSYCH	18444252
GRIA1	2890	Bulimia	PSYCH	20468064
GRIA1	2890	Psychiatric Disorders	PSYCH	20398908
PDLIM5	10611	schizophrenia	PSYCH	16213469
PDLIM5	10611	schizophrenia; bipolar disorder	PSYCH	16044170
PDLIM5	10611	schizophrenia	PSYCH	17287082
PDLIM5	10611	Bipolar Disorder	PSYCH	18496208
PDLIM5	10611	Bipolar Disorder	PSYCH	18456508
PDLIM5	10611	Bipolar Disorder	PSYCH	19448850
PDLIM5	10611	Bipolar Disorder	PSYCH	19328558
PDLIM5	10611	Bipolar Disorder	PSYCH	18496210
PDLIM	10611	depression	PSYCH	18197271

5				
NRXN1	9378	schizophrenia	PSYCH	18940311
NRXN1	9378	Autism	PSYCH	18490107
NRXN1	9378	Autism	PSYCH	19557195
NRXN1	9378	schizophrenia	PSYCH	19197363
NRXN1	9378	several psychiatric disorders	PSYCH	19086053
NRXN1	9378	schizophrenia	PSYCH	18945720
NRXN1	9378	schizophrenia	PSYCH	19880096
NRXN1	9378	schizophrenia	PSYCH	19658047
NRXN1	9378	Schizophrenia	PSYCH	20967226
GPR109A	338442	schizophrenia | bipolar disorder	PSYCH	19502010
RYR2	6262	Chronic renal failure|Kidney Failure, Chronic	RENAL	21085059
PDLIM5	10611	Chronic renal failure|Kidney Failure, Chronic	RENAL	21085059
GRIA1	2890	Citalopram/adverse effects*	REPRODUCTION	19295509
ERBB4	2066		UNKNOWN	18668031
RYR2	6262	Death, Sudden, Cardiac|Long QT Syndrome|Syncope|Tachycardia, Ventricular	UNKNOWN	19926015
NRXN1	9378	Language Development Disorders|Mental Retardation	UNKNOWN	20468056
CAMK1D	57118	Diabetes mellitus|HIV Infections|[X]Human immunodeficiency virus disease	UNKNOWN	20879858

CURRICULUM VITAE

Personal Information

Surname, Name: Yücebaş, Sait Can
Nationality: Turkish
Date and Place of Birth: 05.03.1979, Ankara
Marital Status: Married
E-mail: can_yucebas@yahoo.com

Education

Degree	Institution	Year of Graduation
MS	Başkent University	2006
BS	Başkent University	2003

Work Experience

Year	Place	Enrollment
2006 - …	METU, Department of Health Informatics	Research Assistant
2005 - 2006	Gazi University, Department of Computer Engineering	Research Assistant
2003 - 2005	Başkent University, Department of Computer Engineering	Research Assistant

Project Experience

Consultant: Mobile Personal Health Platform (0638.TSGD.2011), Medica Medikal Bilişim ve Teknolojik Hizmetler, Teknogirişim Sermayesi Desteği Programı. 2011-2012

Consultant: Ekonomik Öngörü Sistemi, Adasoft Danışmanlık Ve Yazılım Hizmetleri Tic. Ltd. Şti., Sanayi Arge Destekleme Programı, 2009-2010

Participant: Preparatory Assistance Study for "Telemedicine: Quality Health Service for Poor and Remote Populations", Birleşmiş Milletler Kalkınma Programı, 2008-2009

Refreed Journals

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" (submitted in July 2013 and currently under revision for PLOS ONE).

C.Yücebaş and Y. Aydın Son. "A Novel SVM-ID3 Hybrid Feature Selection Method to Build Disease Model for Melanoma using both genotyping and phenotype data from dbGaP" (submitted in August 2013 and currently under revision for Turkish Journal of Medical Sciences).

Proceedings in Meetings

An Assessment of the Implementation of Mobile Solutions at a State Hospital in Turkey. Mehrdad A. Mizani, Can Sait Yucebas, Nazife Baykal. American Medical Informatics Association Annual Symposium 2009. Biomedical and Health Informatics: From Foundations to Applications to Policy. San Francisco, California, USA. 14-18 November 2009. Volume 1 of 2

The Effect of Data Set Characteristics on the Choice of Clustering Validity Index Type Taşkaya Temizel, T., Mizani M. A., İnkaya T., & Yücebaş S. C. (2007). . 22nd International Symposium on Computer and Information Sciences (ISCIS'07).

Hibrid Tıbbi Karar Destek Sistemi. H. Sever, G. Köse, E. Çorapçıoğlu, S. C. Yücebaş. 4. Ulusal Tıp Bilişimi Kongresi. 2007.

Sağlık Kuruluşlarındaki Bilgi Güvenliği Yönetim Sistemlerinin ISO/IEC 27001 Standardına Göre Değerlendirilmesi. Mehrdad Alizadeh MIZANI , Sait Can YÜCEBAŞ , Melahat Oya ÇINAR. 4. Ulusal Tıp Bilişimi Kongresi. 2007.

Project Papers

A Report on Preparatory Assistance Study for "Telemedicine: Quality Health Service for Poor and Remote Populations" Project. Can Yucebas, Sevgi Ozkan, Nazife Baykal Middle East Technical University Informatics Institute Department of Health Informatics. 2008. Prepared for United Nations Development Program.

Foreign Languages

Advanced English

Hobbies
Motorsports, Watersports, Fitness, Photography, Travel