2014

Finding Informative Genes in Subtypes of Breast Cancer

Forough Firoozbakht
University of Windsor

FINDING INFORMATIVE GENES IN SUBTYPES OF BREAST CANCER

by
Forough Firoozbakht

A Thesis
Submitted to the Faculty of Graduate Studies
through School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada
2015

FINDING INFORMATIVE GENES IN SUBTYPES OF BREAST CANCER

by

Forough Firoozbakht

APPROVED BY:

K. Tepe
Department of Electrical and Computer Engineering

A. Ngom
School of Computer Science

L. Rueda, Advisor
School of Computer Science

L. Porter, Co-Advisor
Department of Biological Sciences

December 17, 2014

Author's Declaration of Originality

I. Co-Authorship Declaration

I hereby declare that this thesis incorporates material that is result of joint research, as follows:

This thesis incorporates the outcome of a joint research undertaken in collaboration with Dr. Iman Rezaeian under supervision of Dr. Lisa Porter and Dr. Luis Rueda. The collaboration is covered in Chapter 4 and 5 of the thesis. The key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contribution of co-authors was primarily through guiding and supervising the author.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

Thesis Chapter	Publication title/full citation	Publication status
Chapters 4,5	Firoozbakht, Forough, et al. ”Breast cancer subtype identification using machine learning techniques.” Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference on. IEEE, 2014.	published

II. Declaration of Previous Publication

This thesis includes one original paper that has been previously published/submitted for publication in peer reviewed journals, as follows:

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyones copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis. I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

World wide, one in nine women is diagnosed with breast cancer in her lifetime and breast cancer is the second leading cause of death among women. Accurate diagnosis of the specific subtypes of this disease is vital to ensure that the patients will have the best possible response to therapy. In this thesis, we use different machine learning techniques to select the most informative biomarkers for the recently proposed ten subtypes of breast cancer. Unlike existing gene selection approaches, we use a hierarchical based classification approach that selects genes and builds the classifier concurrently in a top-down fashion. We also propose a new bottom-up hierarchical approach to obtain the most informative genes for different subtypes, while we identify the similarity level between these subtypes. Our results support that this modified approach to gene selection yields a small subset of genes that can predict each of these ten subtypes with very high accuracy. The bottom-up approach, on the other hand, provides an insightful structure for further analysis of these subtypes.

Dedication

To my beloved husband, for his support and encouragement.

Acknowledgements

The author wishes to express her gratefulness to her supervisor, Dr. Luis Rueda, who was abundantly helpful and supportive and offered priceless assistance and guidance. Special thanks to Dr. Lisa Porter for her help and guidance toward this thesis. Deepest appreciation also to the members of the thesis committee, Dr. Alioune Ngom and Dr. Kemal Tepe, for all their support and advice.

The author wishes to express her love and gratitude to her family; her loving parents, sister and brother, and also her husband for their understanding and endless love through the duration of her studies. The author would also like to convey thanks to all her friends and labmates for their support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human body consists of trillions of living cells [42]. Normal body cells grow, divide into new cells, and eventually die. The dividing rate of normal cells varies at different stages of each person's life. For example, normal cells divide faster in early years to allow the person to grow appropriately. In an adult person, most of the cells divide only to replace the existing cells that are about to die or to repair wounds and injuries. When cells in a specific part of body start to grow out of control, they become cancerous. Although there are many types of cancer, they all start because of out-of-control growth of abnormal cells. Cell growth in cancer cells is different from normal cells. Instead of dying, cancer cells continue to grow and form new cancer cells. They can also grow into (invade) other tissues, which is not a typical behavior in normal cells. Cancer cells may contain a damaged copy of DNA. DNA is in every cell and directs all cell actions. When DNA becomes damaged, the cell tries to repair itself and if it does not succeed, the cell dies. However, the damaged DNA in a cancer cell is not repaired, and the cell does not die as it is suppose to. Instead, that

cell keeps creating cells that contain the similar damaged DNA. While people can inherit damaged DNA, most of the DNA damage can be caused by mistakes that happen while the normal cell is reproducing or by the effects of our environment, though there is still no clear cause for most types of the cancer.

In most of the cases, the cancer cells form a tumor. Some cancers such as leukemia rarely form tumors. Instead, these cancer cells involve the blood and circulate through other tissues. Cancer cells often travel to other parts of the body, where they begin to grow and form new tumors that replace normal tissue; this process is called metastasis. Different types of cancer can behave very differently. For example, brain cancer and breast cancer are very different diseases that grow at different rates and respond to different treatments. For this reason, people need specific treatment that is aimed at their particular kind of cancer.

On the other hand, machine learning is the subfield of artificial intelligence that studies problems that generalize from past experience. This thesis looks at classification, where an algorithm tries to predict the label of an unknown sample. A sample is a single set of features (here, gene expression levels for a cancer sample) plus a label, which is the category (for example, basal or luminal) the sample falls in. The machine learning algorithm takes many of these samples, called the training set, and builds an internal model. Using that model, the system can then predict the labels of new samples, called the testing set. Feature selection, classification and clustering are among the most important applications of machine learning in bioinformatics, which we also use in this thesis to address the underlying problem.

As one of the examples of classification methods, support vector machines (SVM) are

machine learning methods that have been proposed based on statistical learning theory [69], and have been used extensively in a wide range of applications in bioinformatics. The SVM follows a data-driven approach towards solving classification problems. High accuracy, generalization, and capacity to handle high-dimensional data such as gene expression are among the advantages of using SVM for transcriptome analysis [58].

1.1 Motivation

Breast cancer is one of the leading causes of cancer related deaths among women in North America [1]. Breast cancer is not just one disease, but rather a combination of different diseases that occur in the same site. There are different types of breast cancer that can be distinguished based on gene expression characteristics. Appropriate diagnosis of the specific subtypes of breast cancer is very important to ensure that the patient response to the therapy in the best approach.

A study using a collection of over 2,000 breast tumors has revealed that breast cancer can be categorized into at least ten subtypes, based on Dunn's index [2]. In that study, using a combination of copy number variations (CNVs), copy number aberrations (CNAs) and single nucleotide polymorphisms (SNPs) along with gene expression data led the research group to obtain ten distinct subtypes of breast cancer. In this thesis, the focus is on identifying a small subset of genes that can reliably predict each of these ten breast cancer subtypes using gene expression information.

1.2 Problem

The problem that we have addressed in this work is to find a small subset of genes that can predict the newly discovered ten breast cancer subtypes with the highest accuracy. Since there are two main criteria (level of accuracy and number of genes), we face a multi-objective problem in optimization [37]. Moreover, we aim to find the similarity between these ten subtypes using a bottom-up clustering approach.

1.3 Contributions

The main contributions of this thesis are:

- Using a hierarchical top-down approach for developing a prediction model for newly discovered ten subtypes of breast cancer.
- Proposing a bottom-up approach based on different similarity measures for clustering subtypes, producing a tree-based model with a semi-balanced topology.
- Using different classification methods with in-depth comparison of their performances on the METABRIC dataset.
- Evaluating the proposed models using different performance measures and comparing their performances for various cases.

1.4 Thesis Organization

This thesis comprises six chapters. Chapter 2 discusses breast cancer and its subtypes. In Chapter 3, we provide an introduction to the machine learning techniques used in this thesis. Chapter 4 includes the methods and materials that we have used to address the underlying problem. Chapter 5 includes the results of the experiments that we conducted, as well as the relevant comparisons and discussions. Finally, in Chapter 6, a conclusion on this thesis is made and future works are discussed.

Chapter 2

Breast Cancer

Breast cancer consists of a group of cancer cells that can grow into surrounding tissues or spread to distant areas of the body. Breast cancer is a heterogeneous disease [50], which means that some tumors are very aggressive and respond poorly to treatment, while others respond well and lead to better patient survival.

Breast cancer is increasingly considered to be not just one disease, but a group of diseases distinguished by different molecular subtypes, risk factors, clinical behaviors, and responses to treatment. Distinct molecular subtypes of breast cancer have been identified using gene expression profiles, a process that is both complex and costly [50]. More convenient approximations of molecular subtypes have been identified by using biological markers, including the presence or absence of estrogen receptors (ER+/ER-), progesterone receptors (PR+/PR-), and human epidermal growth factor receptor 2 (HER2+/HER2-) [52]. Molecular subtypes are increasingly being used for research purposes; however, questions remain about their usefulness to further tailor breast cancer treatments and predict breast

cancer prognosis [26, 52].

2.1 Stages of Breast Cancer

The prognosis of breast cancer is strongly influenced by the stage of the disease. There are two main systems for describing the stage of cancer. The TNM classification system uses information on tumor size and the extend that the tumor spreads within the breast (T), the extend of spread to the nearby lymph nodes (N), and the absence or presence tumor invasion to distant organs (M) [22]. Once T, N, and M are determined, a stage of 0, 1, 2, 3, or 4 is assigned, with stage 0 being in situ, stage 1 being early stage of cancer, and stage 4 being the most advanced level of the disease. The TNM staging system is commonly used in clinical settings. The Surveillance, Epidemiology, and End Results (SEER) Summary Stage system is more simplified than the TNM system. The SEER system is commonly used for public health research and planning as well as reporting cancer registry data [77] . Based on the SEER system, the local stage refers to cancers that are confined to the breast, which corresponds to stage 1 and some stage 2 cancers in the TNM system. Regional stage refers to tumors that have spread to surrounding tissue or nearby lymph nodes (generally corresponding to stage 2 or 3 cancers, depending on the size of the tumor and lymph node involvement level). Distant stage refers to cancers tumors that have metastasized to other distant organs or lymph nodes above the collarbone (corresponding to stages 3c and 4 in the TNM system) [65].

2.2 Breast Cancer Subtypes

There are at least five major molecular breast cancer subtypes according to the gene expression profiles of tumor samples in previous studies [50]. They are basal-like, HER2, luminal A, luminal B and normal-like. These subtypes are characterized by different clinical outcomes and respond to different treatments [30, 50, 60, 61].

About 40% of breast cancers are luminal A, making it the most common breast cancer subtype [49]. These tumors tend to be ER+ and/or PR+ and HER2-, slow-growing, and less aggressive than other subtypes. Luminal A tumors are associated with the most favorable short-term prognosis, in part, because the expression of hormone receptors is predictive of a favorable response to hormonal therapy; however, long-term survival is similar to or even lower than some other subtypes [9].

About 10% to 20% of breast cancers are luminal B [49, 72]. Like luminal A tumors, most luminal B tumors are ER+ and/or PR+, but they are distinguished by either expression of HER2 or high proliferation rates (high numbers of cancer cells actively dividing) [17].

About 10% to 20% of breast cancers are basal-like, and the majority of basal-like breast cancers are referred to as triple negative because they are ER-, PR-, and HER2- [14, 72]. Basal-like tumors are more common in African American women and premenopausal women [32]. Women diagnosed with basal-like breast cancer have a poorer short-term prognosis than those diagnosed with other breast cancer subtypes because there are no targeted therapies for these tumors.

About 10% of breast cancers produce excessive HER2 (a growth-promoting protein) and do not express hormone receptors (ER- and PR-) [49]. Similar to the basal-like sub-

type, these cancers tend to grow and spread more aggressively than other breast cancers and are associated with poorer short-term prognosis compared to ER+ breast cancers [9]. However, the use of targeted therapies for HER2+ cancers has reversed much of the adverse prognostic impact of HER2 overexpression.

A recent study using a collection of over 2,000 breast tumors has revealed that breast cancer can be categorized into at least ten subtypes, based on Dunn's index [18]. Accurate prediction of cancer subtypes can aid in directing patient's therapies. Techniques such as breast MRI, mammography, and CT scan, examine the phenotypic mammary change, but do not provide information to direct therapy. Genomic techniques provide high-throughput tools useful for diagnosis and treatment of breast cancer. The huge amount of genomic data and resources that have been generated, have allowed researchers to use unsupervised machine learning techniques to establish distinct tumor subtypes [50, 66].

Chapter 3

Machine Learning

Machine learning is a subfield of artificial intelligence that studies problems that generalize from past experience. In this thesis, we use different machine learning methods such as classification, feature selection and clustering.

In general, a learning problem considers a set of n samples of data and then tries to predict properties of unknown data. If each sample is more complex than a single number, it is said to have several attributes or features. We can separate learning problems into two main categories: supervised learning and unsupervised learning.

In some cases, we have samples that belong to two or more classes and we want to learn, from already-labeled data, how to predict the class of unlabeled data. These types of problems are categorized as supervised learning problems. An example of a supervised learning problem could be cancer subtype identification, for example, in which the aim is to assign a patient (or sample) to one of the subtypes of a specific cancer, such as breast cancer.

Unsupervised learning, in which the training data consists of a set of input vectors X without any corresponding target values. The goal in such problems can be to discover groups of similar examples within the data, in which case it is called clustering, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space onto a two or three dimensional space for the purpose of visualization.

3.1 Classification

Classification is one of the main supervised learning applications. The aim of classification is to assign objects to one of the given classes. For a specific classification problem, the aim is to develop an application that can classify objects correctly with a reasonably high accuracy.

The classification problem can be stated as follows: A sample is a pair, (x_i, z_i), where x_i is a p-dimensional feature vector. Usually, $x_i \in R^p$, and z_i is the label of the sample, which indicates the class that the sample belongs to. The input to the classifier is a set of measurement vectors along with their known classes. This set, called the training set, is used to build the classifier. Once the classifier is built, given a new test sample, x, its class can be predicted by the classifier.

The high-dimensional nature of many classification tasks, i.e., the very large number of available features, may impose a problem for classifiers, especially when there are relatively few samples. Therefore, in many cases, a small subset of the available features need to be selected; this task is called feature selection. We discuss about feature selection in more

detail in the next sections.

The main goal of designing a classifier is to build an algorithm that is as accurate as possible when it comes to classifying new test samples. This challenge is not a straightforward task due to several reasons. The first problem is that we are often given a relatively small training set, which is usually the case when it comes to analyzing biomedical data. Thus, the classifier has to infer a general behavior from relatively few samples. The next problem is that we assume that the training set faithfully represents the test set, or the *real world*. However, specifically when the sample size is small, it is less likely to accurately represent the real world, and more likely to be biased. Another problem is the complexity of the model and its relationship to performance and generalization capabilities. If the classifier is too simple, it may fail to capture the underlying structure of the data. On the other hand, if the classifier is too complex and there are too many free parameters, it may lead to overfitting, where the learned model highly fits the training set, but performs poorly on test samples. Thus, achieving optimal performance on the training set is not the main goal and not even a requirement for a classifier. It may be possible that a classifier can achieve 100% classification accuracy on the training set but the expected performance on a test data is poorer than what could be achieved by different methods that have less performance on their training dataset.

Another problem is the definition of *optimality*. There are several ways of measuring the performance of the classifier. For binary classification problems the most common one is the error rate. But even this is not a simple task, as the error rate needs to be estimated and usually can not be directly calculated.

There are many methods for estimating the generalization error, e.g., the test-set method, cross validation, bootstrap, jackknife, and others. [76]. The focus of this work is on cross validation techniques, in particular, the ten-fold cross-validation method.

Cross-validation calculates the error by repeatedly partitioning the given training set into two disjoint subsets: the training subset and the test subset. When a sample belongs to the test subset, its label is hidden from the classifier that has been built based on the training subset only, and the prediction of its class can be compared to its true class. The process is repeated with several partitions and gives an estimate of the performance of the classifier. The k-fold cross-validation partitions the given training set into k subsets (preferably of equal size). Then, training is done on k-1 subsets and testing is done on the remaining subset. This process is repeated as each subset is taken to be a test set, in turn.

Generally speaking, there are two approaches for developing a classifier: parametric and nonparametric. In parametric approaches [25], there is *a priori* global knowledge of data distributions. The nonparametric approach, on the other hand, does not have this *a priori* knowledge about the data distributions. There are different types of nonparametric classification approaches such as neural networks, fuzzy systems, decision trees, random forests and SVMs. In the next section, we discuss some of these classifiers in more detail.

3.1.1 Support Vector Machines

SVMs [67, 68, 71] are very popular linear discrimination methods built based on a simple, yet powerful idea. The idea is to use the support vectors for separating classes. A separating hyperplane H is the best if it has the largest margin. The margin can be defined as the

largest distance between two hyperplanes parallel to H on both sides that do not contain sample points between them. For a better demonstration, Figure 3.1 shows an example of such hyperplane. We can see that there are different separating hyperplanes that can be used based on the given training set. The separating hyperplane that leaves the closest points from different classes at maximum distance from it is preferred, as the two groups of samples are separated from each other by the largest margin. This provides the least sensitive model to minor errors.

As mentioned earlier, SVMs are supervised learning methods, based on the statistical learning theory, which are designed for classification and pattern recognition. The SVM works by estimating a function called linear discriminant function that models the problem [6, 70, 71]. The SVM could also be modeled as a non-linear classifier with the use of different kernel functions. We examine both linear SVM and non-linear SVM in our approach. We have tried various kernel functions such as polynomial of degrees 2 and 3, radial basis function (RBF), and the sigmoid function as kernels.

In essence, the SVM maps the input samples onto a higher dimension feature space, and tries to find a hyperplane that separates the classes with the largest margin possible in the new space. In case that the problem is not linearly separable, based on the idea of the soft-margin, the SVM chooses a hyperplane that separates the samples as clearly as possible. In this thesis, we used Weka data mining tool with LibSVM for our implementation [27].

Normally, SVMs are formulated for two-class problems. However, there are some types of SVM that can handle multi-class problem, such as pairwise SVMs and one-against-all SVMs. Pairwise SVMs convert an n-class problem into an $n(n-1)/2$ pairwise problem

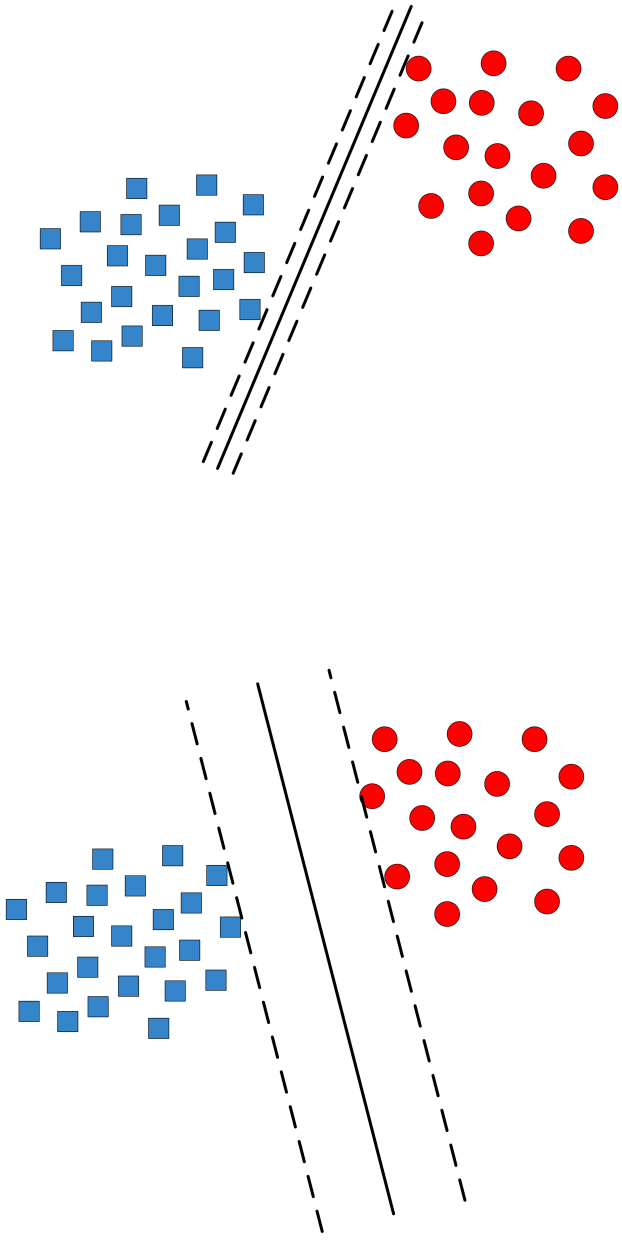

Figure 3.1: Separating hyperplanes and margin. Two different possible separating hyperplanes are shown (solid lines). A separating hyperplane with small distance from closest points (top) and a separating hyperplane that leaves the closest points at maximum distance from it (bottom). The dash lines identify the margin.

[64]. This covers all the pairs of the classes. In a one-against-all SVM, an n-class problem is converted into n two-class problems and for the j^{th} two-class problem, class j is separated from the remaining classes [70]. Both of the aforementioned formulations may produce unclassifiable regions if a discrete decision function is used. As discussed in the next chapter, we use a tree-based SVM scheme to overcome this issue.

3.1.2 Decision Trees

A decision tree is a tree structure that allows us to show dependencies among features and the underlying classes. It can also be used to identify the corresponding class for an unknown sample [54]. The tree consists of the root, internal nodes and leaves. All internal nodes and the root node represent features, while leaves indicate the classes depending on the value of the attributes on the path from the root to a particular leaf. An illustration of a decision tree for playing tennis is shown in Figure 3.2. As shown in the figure, the decision tree provides a decision based on three attributes: outlook, humidity and wind. For exmaple, if the outlook is sunny and humidity is normal, the decision for playing tennis is yes, while if the outlook is rainy and the wind is strong, the decision is no.

Each decision tree is constructed recursively, using a set of learning samples. In every recursive step, samples from previous steps are divided into smaller subsets depending on the value of the best attribute, which maximizes the criteria over the current subset of samples [55].

When the tree is constructed, it can be then used for classification of new unknown samples. Given the simple structure of the decision tree, finding the corresponding class

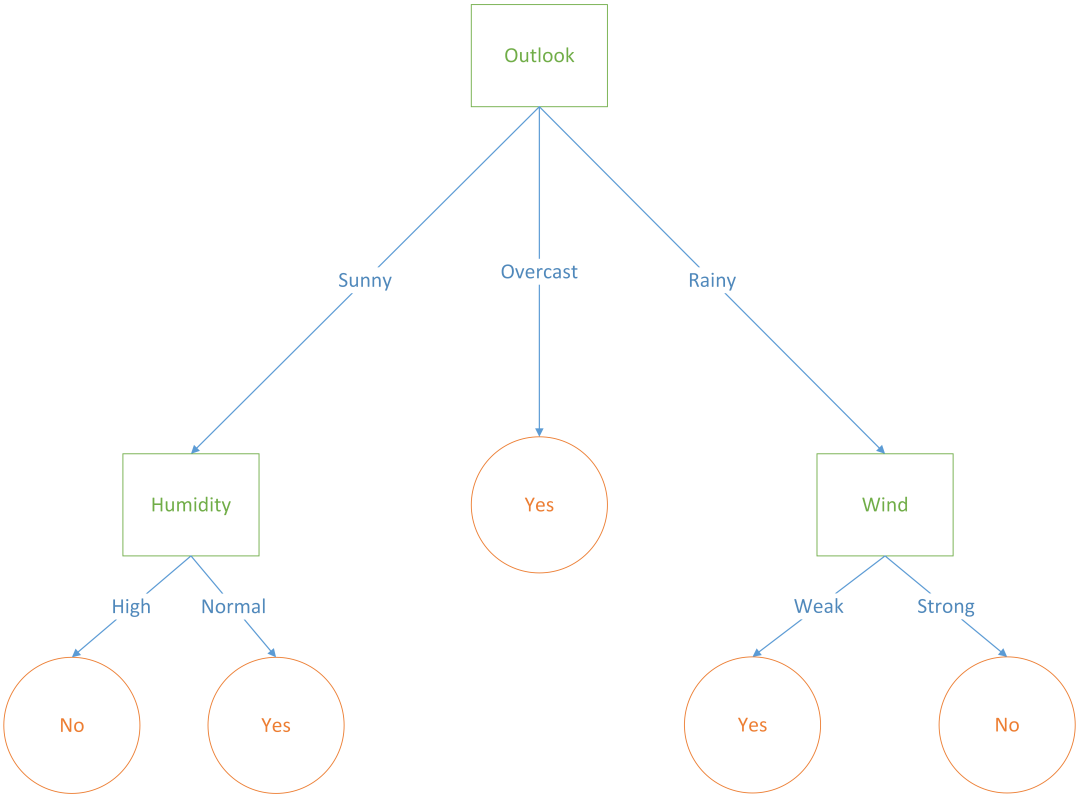

Figure 3.2: A decision tree for playing tennis based on three attributes: outlook, humidity and wind.

is straightforward. To acheive this, only a few comparisons on attribute values need to be performed starting from the root of the tree. At each internal node, the case's outcome is determined and the comparison shifts to the one of the sub-trees, based on the outcome of the previous step. When a leaf is reached, the class is determined.

3.1.3 Random forest

A random forest consists of a set of decision trees T, in which each of them depends on the value of a random vector that has been sampled independently and with the same distribution for all trees in the forest [47]. Figure 3.3 shows an illustration of a random forest. Each

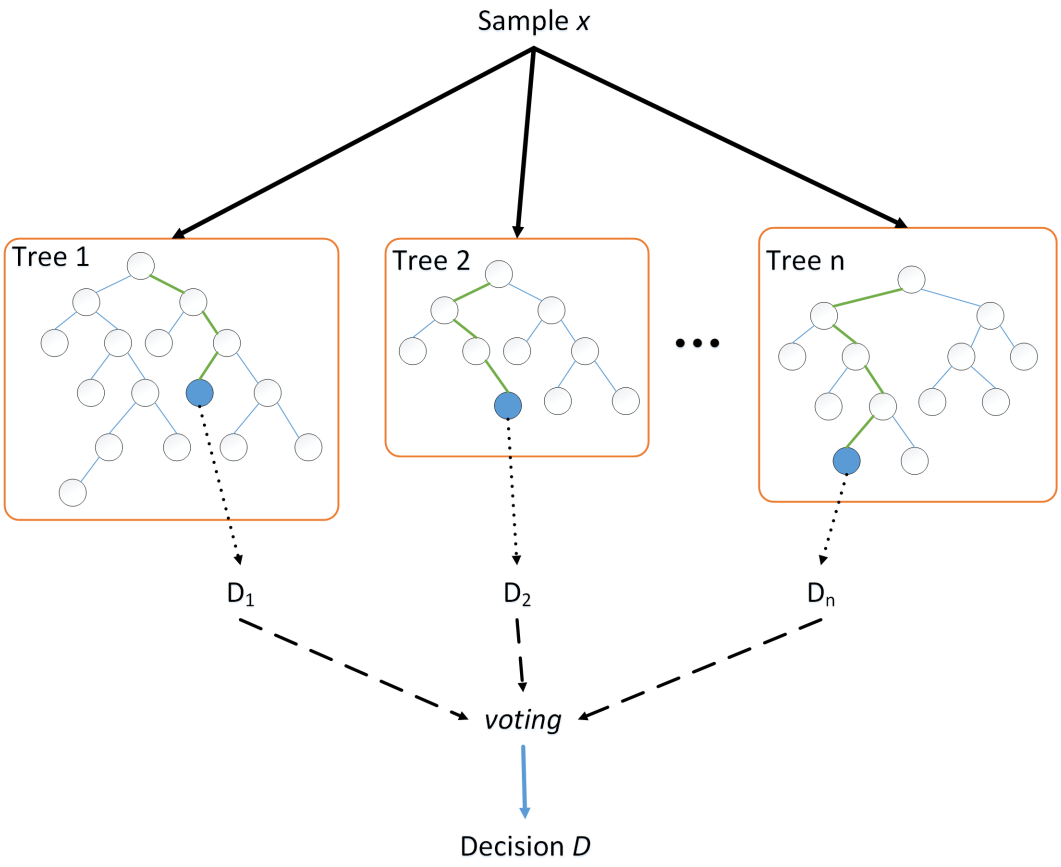

Figure 3.3: An example of a random forest containing n trees. For an unknown sample x, each tree votes for one of the classes, and the final class is determined by voting.

tree in the forest is grown as follows [11]:

- Let N be the number of samples in training set. We select N_r samples randomly with replacement to use them as training set for growing each tree.

- If F is the number of features in dataset, a number $f << F$ is specified in such a way that at each node, m variables are selected at random out of all the features. The value of m is fixed during the forest growing phase. There is no pruning to the trees and each tree can be grown to the largest extent possible.

Decision making in the random forest is based on voting. When a new sample comes, each tree in the forest classifies the sample and votes to one of the classes. At the end of the procedure, the class with the highest number of votes is declared as the class corresponding to the new sample.

The performance of the random forest depends on two factors [11]: The correlation between any two trees in the forest and the strength of each individual tree. Increasing the correlation between the trees reduces the performance of the forest, while increasing the performance of each individual tree improves the performance of the forest. Using a small number of features decreases the correlation between the trees, while at the same time, it reduces the performance of the individual trees as well. Increasing the number of features used in each tree, on the other hand, increases the performance of the tree, while, at the same time, it increases the correlation between the trees in the forest. Somewhere in between is an *optimal* range for the number of features used in each tree, which is usually quite wide [11].

3.2 Clustering

Unlike classification, which is usually supervised, clustering techniques are unsupervised thechniques used when the labels of a given dataset are unknown. The goal of clustering is to partition the instances into groups or clusters that are more similar to each other rather than to instances from other clusters [56].

Generally speaking, clustering techniques can be placed into one of two categories, partition-based and hierarchical [33]. Partition-based approaches divide the data into a

pre-set number of clusters. Hierarchical clustering algorithms, on the other hand, create a tree-like structure, wherein subgroups are merged until become one group.

3.2.1 Partition-based Clustering

One of the best-known partition-based clustering approaches is the *k*-means algorithm [43], where k is the desired number of clusters. The *k*-means algorithm begins by selecting K instances and defining an initial centroid for each of the k clusters as the location of the instance itself. The k centroids are then repeatedly updated by alternating between assigning instances to the closest centroid, in terms of a specific distance function. In the next step, the new centroid is calculated again based on the arithmetic or geometric mean of the features of the instances that have been assigned to the cluster. There are different distance metrics that can be used such as Euclidian distance, city block and Pearson correlation [19]. There are different methods for initialization of *k*-means, for which a comprehensive analysis of the common techniques can be found in [15].

Expectation-maximization (EM) [20] is another clustering technique that tries to cluster data with the assumption that the samples have been generated from n probability distributions, and attempts to estimate the parameters of these distributions. Cluster memberships are determined based on the probability distribution that most likely generates each sample. One of the common distribution used in EM is the mixture of Gaussians model, which assumes that the probability distributions are multivariate Gaussians [8]. The goal of EM is to compute the maximum likelihood estimation of the parameters for Gaussian distributions (means and the variances) based on the data.

Non-parametric clustering approaches have a different assumption about the definition of the clusters. For example, in density-based clustering techniques, the clusters are defined as contiguous regions of high density separated by regions of low density [35]. Since these methods do not make any assumption about the underlying distribution of the data, they can define clusters with arbitrary shapes. One of the most well-known non-parametric clustering methods is spectral clustering [73].

3.2.2 Hierarchical Clustering

One of the other categories of clustering methods consist of hierarchical clustering techniques. Hierarchical clustering is one of the most straightforward methods among different ways of forming clusters. The main idea behind this technique is that the objects being more related to nearby objects than to objects farther away. Hierarchical clustering can be either agglomerative, which considers each sample as a separate cluster and then merges the clusters; or divisive, which starts by considering all samples as a single cluster and then divides the cluster into sub-clusters, and so on. [34]. Agglomerative hierarchical clustering considers each sample as a cluster. Any two samples that have the smallest distance from each other are joined into a single cluster. At each step, one of these three scenarios can happens: individual samples are added to existing clusters, two individual samples are combined, or two existing clusters are combined.

The distance between two clusters with more than one sample in the cluster can be defined in various ways. For example, the average distance between all pairs of samples is formed by taking one member from each of the two clusters and calculating the distance

between them. The largest or smallest distance between two sample from different clusters can be taken. There are different methods for measuring the distance between clusters. However, different methods may produce different solutions. Linkage criteria determine the proximity of the objects to each other using distance information. Once the proximity between the objects in the data set has been computed, the objects are paired into binary clusters and the process continues until a hierarchical tree is formed. The arrangement of the clusters produced by the hierarchical clustering can be illustrated with a dendrogram, which is a useful approach to illustrate the clustering of genes or samples in bioinformatics and computational biology.

Most of the agglomerative hierarchal clustering algorithms can be considered as one of the variants of three standard algorithms: single linkage, complete-linkage, and minimum variance [33]. These algorithms initialize every instance as a separate cluster and combine the clusters based on the minimum, maximum or average distance between the clusters. The algorithm terminates once all the instances have been merged into a single cluster. The mergers are usually presented visually as a dendrogram, which shows every merger of the algorithm starting at the beginning with each instance as a single cluster.

In single linkage, the distance between two clusters is equal to the distance between the two nearest neighbors in such a way that two neighbors belong to different clusters. Figure 3.4 shows an example of single linkage distance evaluation between two clusters.

Complete linkage, on the other hand, evaluates the distance between two clusters based on the distance between the furthest neighbors in such a way that each neighbor belongs to one of the clusters. Figure 3.5 shows an example of complete linkage distance evaluation

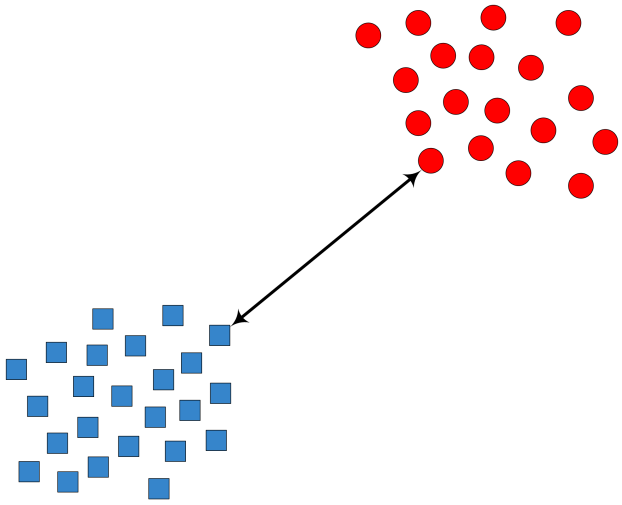

Figure 3.4: Single linkage distance evaluation between two clusters.

between two clusters.

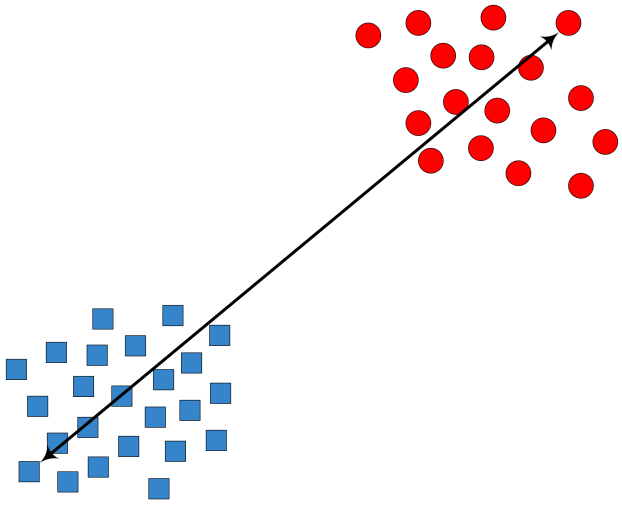

Figure 3.5: Complete linkage distance evaluation between two clusters.

Instead of relying on a pair of samples for determining the distance between two clusters, average linkage takes the distances between all pairs of samples into account and calculates the average of all possible distances. In other words, the distance between two clusters using the average linkage method can be computed as follows:

$$AverageDist = \frac{1}{|A|\,|B|}\sum_{x\in A}\sum_{y\in B} d(x,y) \tag{3.1}$$

where A and B are two clusters.

Ward's method is one of the other approaches that uses analysis of variance to evaluate the distances between clusters [34]. Ward's minimum variance method is a special case of the objective function approach originally presented in [75]. Ward's method works as follows:

- Using analysis of variance to evaluate the distances between clusters.
- Minimizing the sum of squares of any two (hypothetical) clusters that can be formed at each step, as follows:

$$d_{ij} = \frac{N_i \times N_j}{N_i + N_j}\sqrt{||c_i - c_j||^2} \tag{3.2}$$

  where N_i and N_j are the numbers of samples in cluster i and j respectively and c_i and c_j denote the centers of the clusters; $||.||$ is the Euclidean norm.

- The mean and cardinality of the new merged cluster, k, is computed as follows:

$$c_k = \frac{1}{N_i + N_j} N_i c_i + N_j c_j, \tag{3.3}$$

$$N_k = N_i + N_j \tag{3.4}$$

This process continues until all the clusters are merged into a single cluster.

3.3 Feature Selection

One of the main problems associated with machine learning and pattern recognition is the so-called *curse of dimensionality* [63]. If the samples have many features (which is usually the case in most of the biological data), it makes the data analysis time consuming. The task of feature selection is to reduce the dimension of the data (removing redundant features) as much as possible while still retaining as much relevant information as possible [76]. There are many benefits of using feature selection, especially when dealing with a very large number of features. Feature selection can remove redundant or irrelevant information and obtain a better classification performance. Moreover, reducing the number of features used by the classifier can increase the generalization capability of that model [63]. Thus, one of the goals in this study is to obtain a small subset of features (genes) that can effectively identify the subtypes of breast cancer without losing the generalization ability of the model.

Feature selection can be done in two ways: One way is to rank the features based on some criterion and select the top k features. This approach is called *filter method*, since the optimality of a feature is evaluated independently. Another way is to select a subset of features with minimum learning performance deterioration. This approach is called *wrapper method*. In this method, the subset selection algorithms can automatically determine the number of selected features, while filter methods need to rely on some given threshold to select relevant features. Filter methods, on the other hand, are usually much faster than the wrapper methods, since they evaluate each feature independently, while in wrapper methods, all the relationships between features also need to be considered. In the next section, we discus about two feature selection methods from both approaches.

3.3.1 Chi-squared Method

Chi-squared is an efficient feature selection method for numeric data that *automatically* and *adaptively* discretizes numeric features and selects features as well [38]. Chi-squared is a univariate filter based on the χ^2 statistic [39]. This method evaluates the worth of each feature by computing the value of the chi-squared statistic with respect to the class. The higher the value of chi-squared, the more relevant the feature will be, with respect to the class. Since this method evaluates the relevance of each feature independently, the method is usually fast in terms of the running time. This is very important when there are many features in a given dataset (such as biological datasets).

3.3.2 minimum Redundancy Maximum Relevance (mRMR)

The mRMR method [48] selects those features that have the highest relevance with the target class and the lowest redundancy. In other words, mRMR selects features that are maximally dissimilar to each other. Both optimization criteria (Maximum- Relevance and Minimum-Redundancy) are based on mutual information. Peng et al. [48] proposed the minimum redundancy maximum relevance (mRMR) method, where the criterion is given by the following formula:

$$mRMR = \max_{j \in Q-S} \left[I(f_j, y) - \frac{1}{|S|} \sum_{s \in S} I(f_j, f_s) \right] \tag{3.5}$$

where f_j is the j^{th} variable in the initial F-dimensional feature space, f_s is a variable that has already been selected in the feature subset S, s is an individual feature and Q contains all the features in the initial feature space. Moreover, S contains the selected features and

$Q - S$ contains those features that are not selected. Also, $|S|$ denotes the cardinality of the selected subset. In practice, the mRMR filter approach is highly successful in many applications [10, 13, 36, 45], thereby justifying the intuitive concept that selecting features based on the compromise between relevance and redundancy may be more appropriate than relying solely on the naïve idea of selecting features only on the basis of relevance.

3.4 Performance Evaluation

Assessing the performance of a classifier is very important and critical toward obtaining a good and robust model. There are different types of performance metrics that we briefly review here. Suppose that we deal with two classes: *positive* and *negative*. Table 3.1 shows the confusion matrix of these two classes.

Table 3.1: Confusion matrix corresponding to the original and classified samples for two classes.

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

- TP represents the number of true positives or the number of samples belong to class positive that have been classified as positive.

- TN represents the number of true negatives or the number of samples belong to class

negative that have been classified as negative.

- FP represents the number of false positives or the number of samples belong to class negative that have been classified *incorrectly* as positive.

- FN represents the number of false negatives or the number of samples belong to class positive that have been classified *incorrectly* as negative.

Using the above confusion matrix, different performance metrics can be calculated. For example, *Accuracy* can be computed as follows:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}, \tag{3.6}$$

F-measure uses both precision and recall measures to compute the score as follows:

$$F\text{-}measure = 2 \times \frac{Percision \times Recall}{Percision + Recall}, \tag{3.7}$$

where

$$Percision = \frac{TP}{TP+FP}, \tag{3.8}$$

$$Recall = \frac{TP}{TP+FN}, \tag{3.9}$$

Another measure is the area under the ROC curve (AUC). The receiving operating characteristics (ROC) curve shows the trade off between Specificity and Sensitivity (Recall).

where

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN}, \tag{3.10}$$

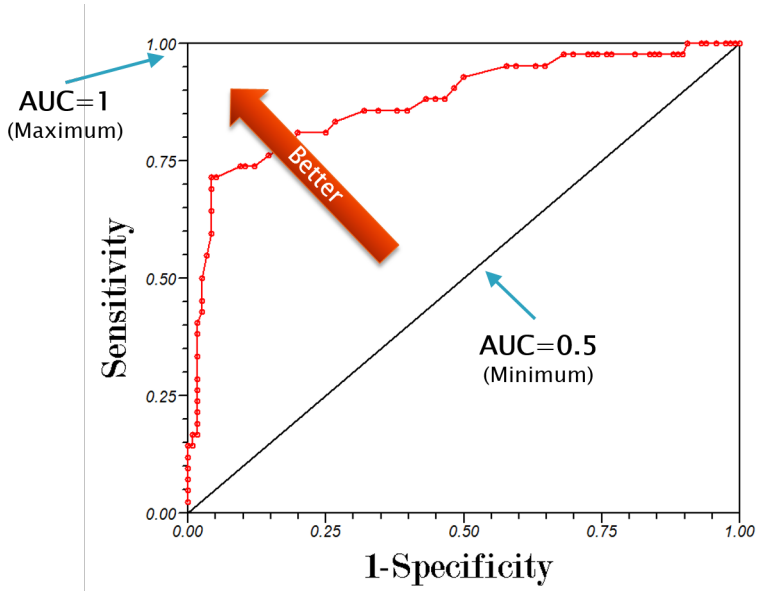

Figure 3.6: An example of the ROC curve and the relation between AUC, Specificity and Sensitivity.

$$Specifity = \frac{TN}{TN+FP}, \tag{3.11}$$

Figure 3.6 shows an example of an ROC curve. The closer the area under the ROC curve is to one, the better the performance of the classifier is.

3.5 Previous Work on Gene Selection for Breast Cancer Analysis

Modeling today's complex biological systems requires efficient computational models to extract the most valuable information from existing data. Machine learning approaches can help researchers to extract these information from biological and clinical data. for example using a classification model, we can predict the subtypes of breast cancer. Some of the

relevant approaches are briefly discussed here. Liu and colleagues proposed an entropy-based method for classifying cancer subtypes [41]. In that method, the genes related to the different cancer subtypes are selected by reducing the redundancy between genes. Another method uses a classification model containing a two-layer structure named as mixture of rough set (MRS) SVM [78]. In that model, the rough set classifier acts as the first layer to determine some singular samples in the data, while the SVM classifier acts as the second layer to classify the remaining samples. Recursive feature addition (RFA) is another method that combines supervised learning and statistical similarity measures to select genes that are relevant to the cancer subtype [40]. Mohamad et al. proposed a gene selection method based on binary particle swarm optimization (BPSO)[46]. In that model, a small subset of the most relevant genes is selected for cancer classification. Al-alaak et al. used a random forest classification model to predict survival in breast cancer [2]. Menden et al. [44] developed a machine learning model to predict the response of cancer cell lines to drug treatment based on both the genomic features of the cell lines and the chemical properties of the considered drugs. In another effort, Eshlaghy et al. [23] used decision trees, SVMs and artificial neural networks to predict breast cancer recurrence. Rezaeian et al. used a hierarchical classification model to predict the subtypes of breast cancer [53].

Chapter 4

Methods

In this chapter, we discuss the proposed methodology of our approach to the problem of finding relevant genes corresponding to each subtype. To solve the problem, we used two different strategies. In the first model, we used a top-down approach to obtain a hierarchical tree in such a way that each leaf consists of one of the ten subtypes. This model is similar to the one used in a previous study by Rezaeian et al. [53]. In that study, a smaller dataset was used and also, the classification model was based on the older five breast cancer subtypes. In the second approach, we used a bottom-up model to find the most similar subtypes in each stage and merge them together in an agglomerative clustering fashion. We describe each model in detail in the next sections.

4.1 Datasets

In this study, we have used the METABRIC dataset [18]. The METABRIC dataset (accession number EGAS00000000083) contains the copy numbers and gene expressions of 2000 primary breast tumors with long-term clinical follow-up. In that study, the copy number aberrations and copy number variations generated using Affymetrix SNP 6.0 arrays and gene expression data were obtained using Illumina HT 12 technology. The dataset contains two sets of data, *validation* set and *discovery* set. Due to the lack of class labels in the validation set, in this thesis we only used discovery set, which contains 997 samples from ten subtypes of breast cancer. Each sample contains expression information of 48,803 probe IDs. The numbers of samples corresponding to each subtype are listed in Table 4.1.

Table 4.1: The number of samples correspond to each of ten subtypes.

Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10
76	45	156	167	94	44	109	143	67	96

Table 4.2 shows a snapshot of the METABRIC dataset. The rows of the table correspond to the samples, while the columns of the table correspond to the expression level with respect to probe IDs. Moreover, each sample has been labeled based on both ten subtype and five subtype models.

Also, Table 4.3 shows the distribution of the samples across different subtypes based on both five and ten subtypes categorization. As shown in the table, most of the samples corresponding to *normal* subtype, are located in the Subtype 4. There is a similar relationship between *basal* and Subtype 10, as well as *Her2* and Subtype 5. All the other subtypes (subtypes 1, 2, 3, 6, 7, 8 and 9) are occupied mostly by *LumA* and *LumB* samples.

48803 Features (prob_IDs)

997 Samples (patients)

ILMN_1802380	ILMN_1893287	...	ILMN_1792389	10 Subtypes	5 Subtypes
5.3168	5.1416	...	5.3168	5	2
5.4664	5.5084	...	5.4616	9	4
5.1931	5.4199	...	5.2271	2	4
5.2991	5.4895	...	5.0578	9	5
5.5556	5.5483	...	4.992	5	1
5.2411	5.3613	...	5.494	1	5
5.2206	5.295	...	5.0625	10	1
5.3252	5.3375	...	5.5398	8	5
...	...	...	...	...	...
5.1908	5.2738	...	5.19	4	4
5.2952	5.5439	...	5.2425	8	4
5.1522	5.3534	...	5.4796	8	4

Table 4.2: Snapshot of the discovery set in the METABRIC dataset containing 997 samples and 48,803 features.

	Basal	Her2	Normal	LumA	LumB	*Total*
Subtype1	4	7	3	7	55	76
Subtype2	0	3	1	19	22	45
Subtype3	1	2	7	124	22	156
Subtype4	19	17	33	82	16	167
Subtype5	5	45	5	13	26	94
Subtype6	0	1	2	17	24	44
Subtype7	1	1	2	80	25	109
Subtype8	1	1	2	106	33	143
Subtype9	9	5	2	17	34	67
Subtype10	78	5	1	1	11	96
Total	118	87	58	466	268	997

Table 4.3: Distribution of the 997 samples across different subtypes based on both 5 and 10 subtypes categorization.

For a better illustration, Figures 4.1 and 4.2 show the distribution of the samples across both five and ten subtypes. The *y*-axis represents the percentage of samples, while the *x*-axis represents the subtypes.

In addition, we used Hu's dataset [31] for comparison proposes. Hu's dataset (CEO

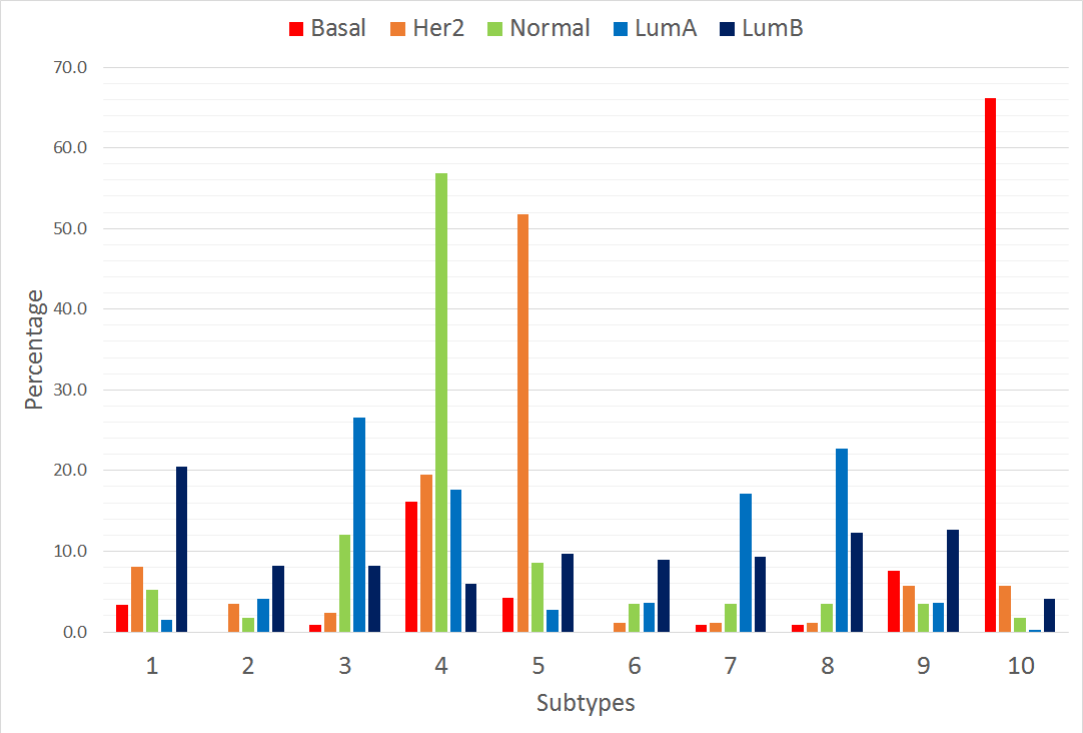

Figure 4.1: Distribution of samples across ten subtypes.

accession number GSE1992) was generated by using different Agilent platforms. Each platform contains 22,575 probe sets, and there are 14,460 common probe sets among these three platforms. The dataset contains 158 samples from five subtypes of breast cancer (13 Normal, 39 Basal, 22 Her2, 53 LumA and 31 LumB) and 13,582 genes with unique unigene IDs.

4.2 Data Pre-processing

The pre-processing of the data has been conducted based on the steps described in [18]. Using an R script [62], each BeadChip has been processed after its scan using the *beadarray*

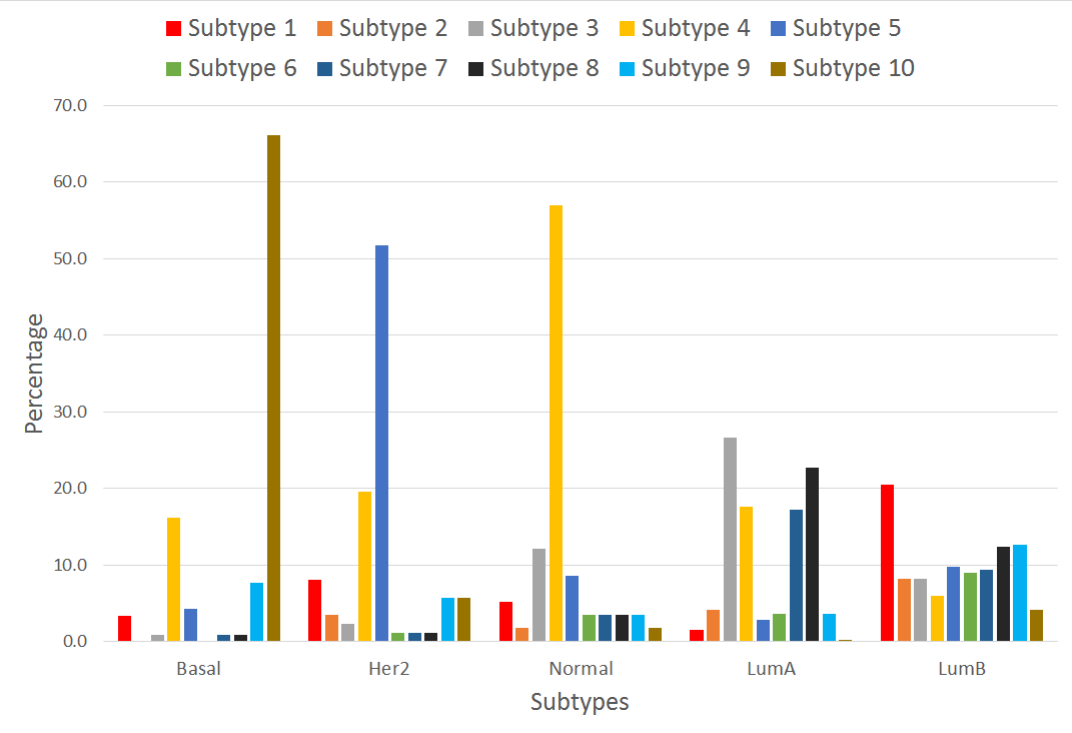

Figure 4.2: Distribution of samples across five subtypes.

package [21]. After processing the chip, the bead-level data are summarized and a series of $48,803 \times 12$ matrices of log2 of intensities are produced along with standard errors, and the number of observations. In the next step, potential outlier arrays are removed based on the bead-level QA scores obtained using the control probes on each array. A multivariate outlier testing procedure [4] from the *arrayMvout* Bioconductor package was used to identify arrays with poor quality based on the bead-level QA scores of all arrays. The arrays that remain after this three-step procedure are used for subsequent analysis. In addition, the comprehensive re-annotation of the Illumina HT-12 v3 platform [5] was used by generating a list of suitable probes based on the following criteria: having a perfect genomic match, not targeting the sex chromosomes, not containing a SNP, not containing

a polyG tail at the end of the probe sequence, and having GC content between 38% and 64%, while the probe does not target genes from the PAM50 list. ER-positive and ER-negative samples have been quantile normalized separately and averaged for obtaining the target distribution. Each of the arrays was then normalized by quantile normalizing probes belonging to the target distribution. Using the *limma* Bioconductor package [59], a linear model was fit to remove any array mis-positioning with respect to the BeadChip. In the next step, differential expression analysis was performed on a subset of the log2 of normalized intensity matrix.

4.3 Top-Down Approach

In this approach, we build the model starting from the top of tree (root). This model is similar to the one used in a previous study by Rezaeian et al. [53]. The difference between two models is that in this work we target a recently proposed ten breast cancer subtypes as our base model. Moreover, the dataset used for creating this model is based on METABRIC dataset, which is 20 times larger than the dataset used in the aforementioned study. We divide the process into two different steps, the training phase and the prediction phase. In the training phase for gene selection and breast cancer subtyping, the complete gene profile of each breast cancer subtype is compared against the others. The subtypes are then organized by two main criteria. The first criterion is the level of accuracy with which the selected genes identify the given subtype. The second criterion is the number of genes identified. Clearly, applying two or more gene selection criteria is a multi-objective optimization problem [37]. In this study, we select a small subset of genes that yields very high

accuracy. The subtype that is ranked highest is removed and the procedure is repeated for the remaining subtypes comparing each gene profile against the others. By repeating this step, a hierarchical tree is created, where each leaf of the tree becomes a unique subtype.

The flowchart of the training phase is illustrated in Figure 4.3. Also, an example of such a tree is illustrated in Figure 4.4. Suppose there are ten different subtypes, namely $\{C_1, C_2, \ldots, C_{10}\}$. The Training data is an $f \times n$ matrix $D = D_1, D_2, \ldots, D_{10}$ corresponding to the ten subtypes. D_i, with the size of $f \times n_i$ is the training data for subtype C_i, while f and N_i are the number of features (probe IDs) and the number of samples in subtype C_i, respectively. The total number of samples for all ten subtypes n is $n = \sum_{i=1}^{10} n_i$.

Note that the terms *genes* and *features* will be used indistinctly. In the first step, feature selection and classification are applied for each class against the other classes, along with a 10-fold cross-validation scheme. Assume that subtype C_6 obtains the highest rank based on both performance measure and number of genes contributing to that performance. Then, a leaf for subtype C_6 is created and the list of particular features is recorded to create an internal node for connecting C_6 and the rest of the tree. In the next step, the samples corresponding to subtype C_6 are removed from the dataset and the process continues in the same fashion for the remaining subtypes; i.e. $C_1, C_2, C_3, C_4, C_5, C_7, C_8, C_9, C_{10}$. In the next round, C_3 yields the highest rank among the remaining subtypes, and hence the feature list is retained for creating the internal node, while a leaf is created for C_3. This procedure is repeated until there is no subtype to classify.

In the prediction phase, we use the previously identified genes to predict breast cancer subtypes. Given the gene expression profile of a new patient, a sequence of classification

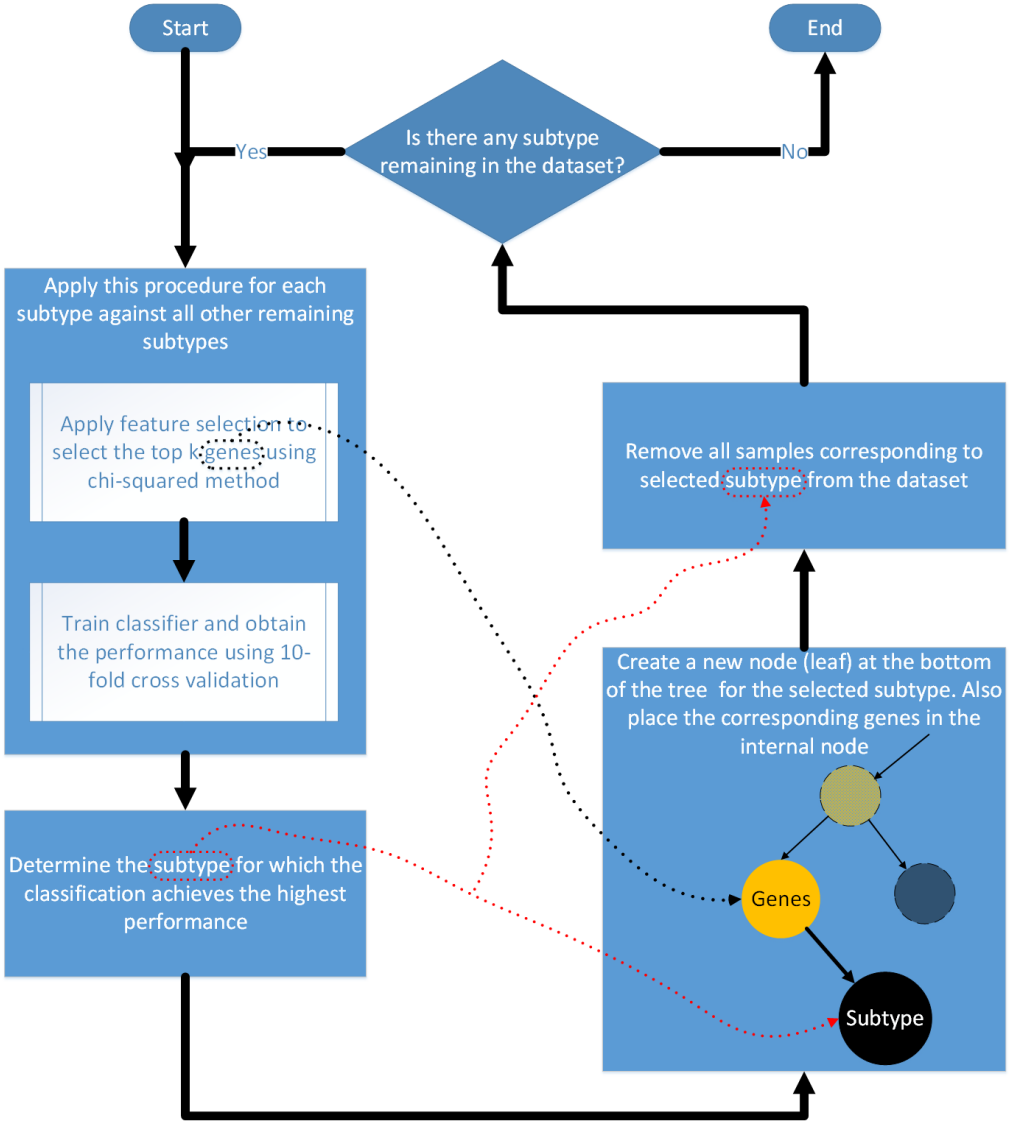

Figure 4.3: Flowchart of the training phase.

steps are performed by traversing the tree from the root toward a leaf. At each node in the path, only the selected genes for that node in the training phase are tested. If the patient's gene profile is classified as the corresponding subtype, then, the prediction phase terminates. Otherwise, the sequence of classification tests continues until a leaf is reached, in which case, the prediction outcome is the subtype associated with that leaf.

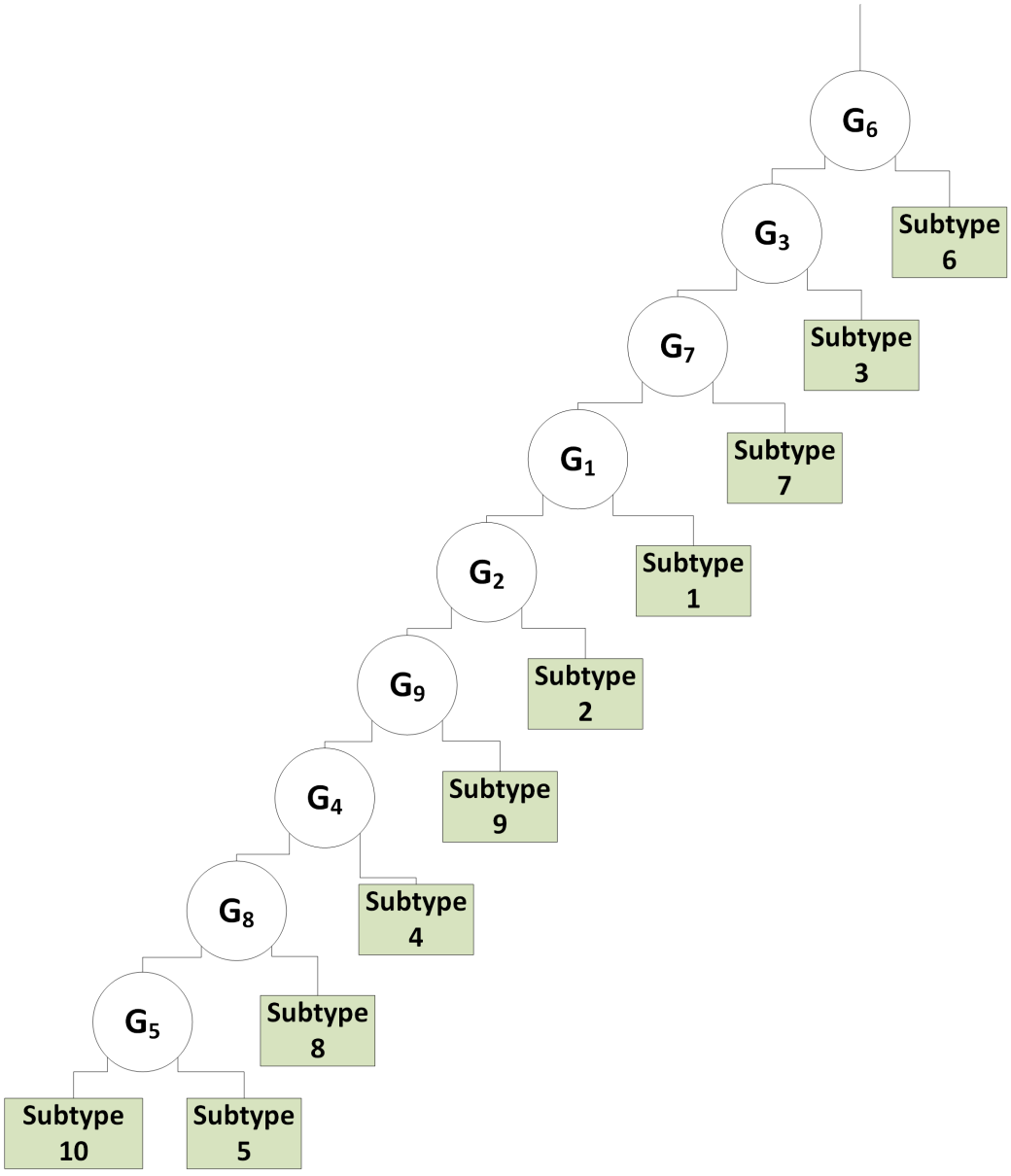

Figure 4.4: Determining genes related to each breast cancer subtype using the topdown approach. G_i is the subset of genes selected for subtype i.

To obtain a list of genes corresponding to each Illumina probe, we used the bioDBnet online tool (http://biodbnet.abcc.ncifcrf.gov/). BioDBnet is a network of the major biological databases with an easy to use Web resources. It contains a vast amount of infomation available in various formats and in various scattered resources. Some of the advantages of this tool are the simplicity of the model, the number and types of databases integrated, support for batch conversions and the integration process itself.

Also, to evaluate the performance of the model, we use accuracy, F-measure and AUC measures. We also use *Chi2* [38] as the feature selection method. After selecting relevant genes, the samples are classified using SVM [71]. In this study we used the Soft-margin SVM [1].

For the implementation, the Weka machine learning suite is used [27]. A gene selection method based on the *Chi2* feature evaluation algorithm is first used to find a subset of genes. For classification, LibSVM [16] in Weka is employed. The *Radial basis function* (RBF) kernel is used within the LibSVM classifier without normalizing the samples and with default parameters.

4.4 Bottom-Up Approach

Using the top-down approach has its own advantages and disadvantages, depending on the type of question we want to answer. One of the advantages of the top-down model is that it conducts gene selection and builds the classifier simultaneously. On the other hand, the samples corresponding to upper-level subtypes are removed during the classification and feature selection of the downstream subtypes. This makes it hard to conclude the connection

between genes corresponding to downstream nodes of the tree and their subtypes, since those genes are obtained under specific circumstances, by removing the samples of the upper-level subtypes from the dataset. To overcome this issue and achieve a more clear interpretation between the subset of genes in each node and their corresponding subtypes, we follow a bottom-up approach. In this approach, instead of building the tree from the root, we build the tree starting from the leaves. The flowchart of the training phase is illustrated in Figure 4.5

In the first step, the similarity between each pair of subtypes is computed and the ones with the highest similarity (lowest distance from each other) will be merged into a single subtype. Table 4.4 illustrates an example of pairwise comparisons between subtypes to find the most similar ones at each stage. $d_{i,j}$ represents the distance between subtype i and subtype j.

The selected pair of subtypes falls into one of these three scenarios: an individual subtype is merged with an existing combined subtypes, two individual subtypes are combined, or two existing subtypes are merged into a single cluster. Since at the beginning of the procedure, there is no combined subtype yet, two individual subtypes with lowest distance (highest similarity) are merged. Let us assume that these two individual subtypes are S_3 and S_8 because $d_{3,8}$ is the smallest distance in the table. These two subtypes are merged and form a new subtype which is the combination of these two. For the sake of simplicity, we call it subtype S_{38}. In the next step, these nine subtypes will be compared again in a pairwise fashion. Table 4.5 illustrates the pairwise comparison between the newly generated subtype S_{38} and the remaining subtypes. This process continues until only one subtype

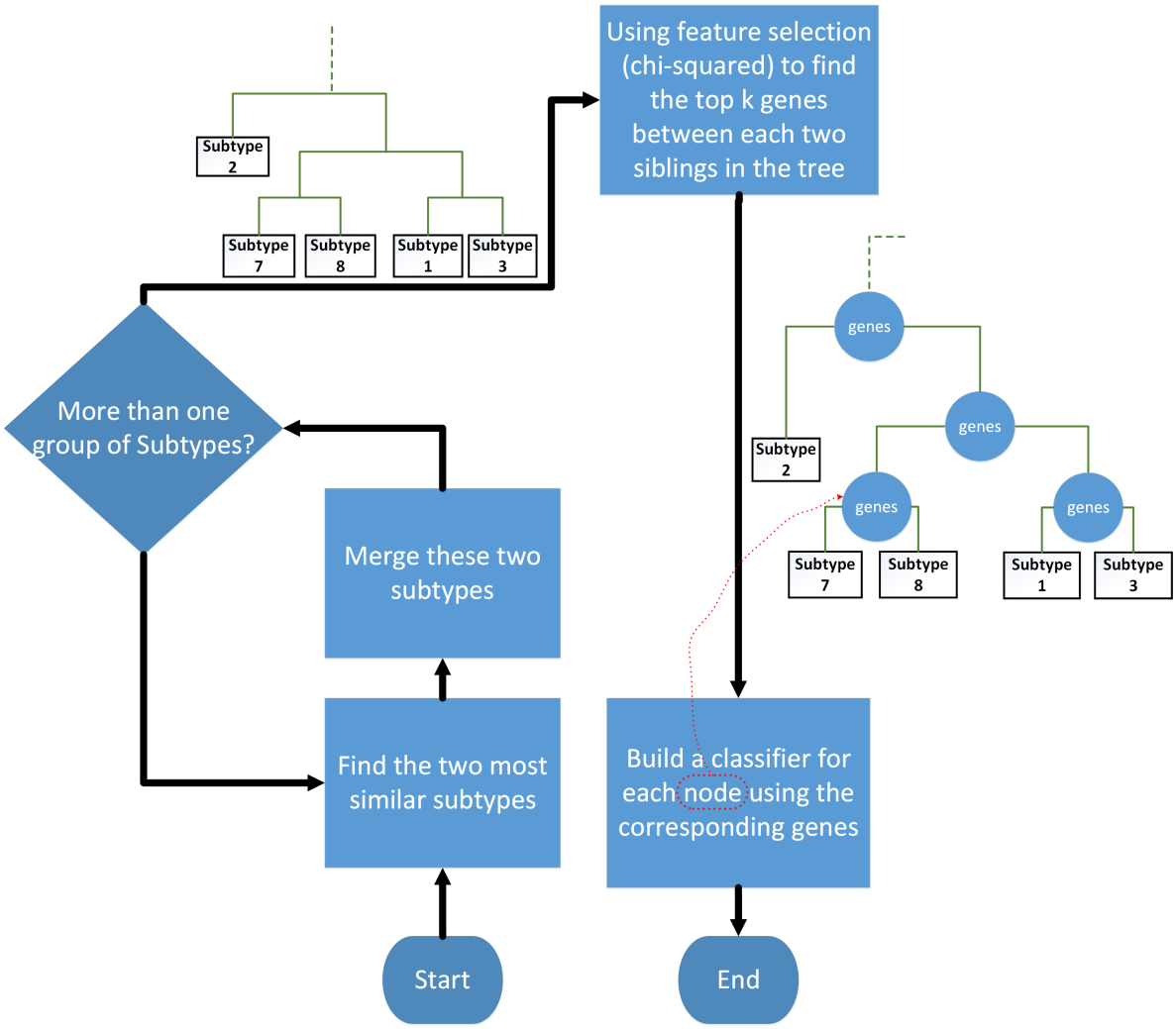

Figure 4.5: Flowchart of the training phase.

remains.

The distance metric used in this experiment is the Euclidean distance. Moreover, four different methods have been used to compute the distance between subtypes. These methods are: single linkage, complete linkage, average linkage and centroid-based linkage. Figure 4.6 shows an example of generated tree using the bottom-up approach and complete linkage as distance function.

Table 4.4: Computing the distance between each pair of subtypes; $d_{i,j}$ is the distance between subtypes i and j.

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
S_1		$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{1,5}$	$d_{1,6}$	$d_{1,7}$	$d_{1,8}$	$d_{1,9}$	$d_{1,10}$
S_2			$d_{2,3}$	$d_{2,4}$	$d_{2,5}$	$d_{2,6}$	$d_{2,7}$	$d_{2,8}$	$d_{2,9}$	$d_{2,10}$
S_3				$d_{3,4}$	$d_{3,5}$	$d_{3,6}$	$d_{3,7}$	$d_{3,8}$	$d_{3,9}$	$d_{3,10}$
S_4					$d_{4,5}$	$d_{4,6}$	$d_{4,7}$	$d_{4,8}$	$d_{4,9}$	$d_{4,10}$
S_5						$d_{5,6}$	$d_{5,7}$	$d_{5,8}$	$d_{5,9}$	$d_{5,10}$
S_6							$d_{6,7}$	$d_{6,8}$	$d_{6,9}$	$d_{6,10}$
S_7								$d_{7,8}$	$d_{7,9}$	$d_{7,10}$
S_8									$d_{8,9}$	$d_{8,10}$
S_9										$d_{9,10}$

Table 4.5: Computing the distance between each pair of subtypes after merging S_3 and S_8.

	S_1	S_2	S_4	S_5	S_6	S_7	S_9	S_{10}	S_{38}
S_1		$d_{1,2}$	$d_{1,4}$	$d_{1,5}$	$d_{1,6}$	$d_{1,7}$	$d_{1,9}$	$d_{1,10}$	$d_{1,38}$
S_2			$d_{2,4}$	$d_{2,5}$	$d_{2,6}$	$d_{2,7}$	$d_{2,9}$	$d_{2,10}$	$d_{2,38}$
S_4				$d_{4,5}$	$d_{4,6}$	$d_{4,7}$	$d_{4,9}$	$d_{4,10}$	$d_{4,38}$
S_5					$d_{5,6}$	$d_{5,7}$	$d_{5,9}$	$d_{5,10}$	$d_{5,38}$
S_6						$d_{6,7}$	$d_{6,9}$	$d_{6,10}$	$d_{6,38}$
S_7							$d_{7,9}$	$d_{7,10}$	$d_{7,38}$
S_9								$d_{9,10}$	$d_{9,38}$
S_{10}									$d_{10,38}$

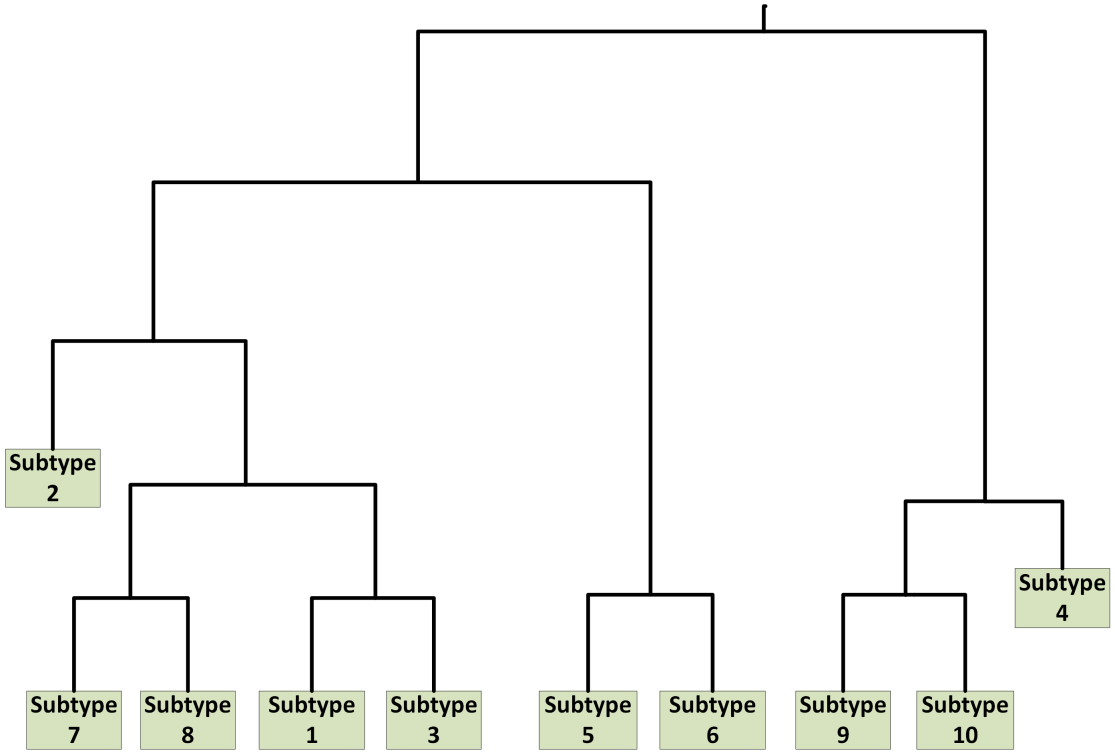

Figure 4.6: Tree generated using the bottom-up approach and complete linkage as distance function.

In the next phase, we use classification and feature selection to determine which genes are the most discriminative ones in terms of separating the subtypes in each branch of the tree. For this, we used Chi2 [38] along with different classifiers, such as decision tree, random forest and LibSVM [16], within the WEKA machine learning suite [27] to determine the top k genes corresponding to each node.

Chapter 5

Results and Discussion

5.1 Experimental Results

In this chapter, we present the results obtained after applying the top-down and the bottom-up approaches on the METABRIC dataset. Since two different approaches are presented in this thesis, the results and analysis of each approach are discussed in separate sections. The two approaches have some differences in terms of modeling the tree and the main reason behind creating such a hierarchical structure. In the top-down approach, the main goal is to obtain a tree with ability of predicting the subtype of an unknown sample with very high accuracy. Although in the bottom-up approach we still aim for a model that can classify an unknown sample with high accuracy, the main goal of the approach is to identify the similarities among the ten subtypes and those genes that are most responsible for differentiating these similar subtypes. On the other hand, there are some similarities between the two models. For example, to evaluate the performance of the model, we use

10-fold cross-validation for both approaches. Also, we consider accuracy, F-measure and area under the curve (AUC) as performance measures for both the top-down and bottom-up approaches.

5.1.1 Top-Down Approach

Table 5.1 shows the comparison between the SVM classifier with different numbers of genes versus the presented hierarchical classification model (HCL). As shown in the Table, using all genes decreases the overall accuracy of the model, since many of the genes are irrelevant or redundant. For example, using all 48,803 probes, accuracy, F-measure and AUC are just 59.27%, 0.554 and 0.765 respectively, while using a ranking algorithm and selecting the top 89 genes (corresponding to the top 100 probes) for prediction increases the accuracy, F-measure and AUC up to 72.31%, 0.718 and 0.842 respectively. Table 5.2 shows the top genes corresponding to the probes ranked by the Chi-Squared attribute evaluation algorithm used to classify samples as one of the ten subtypes. We repeated the same experiment using only the top 20 probes, which correspond to 18 genes. The corresponding accuracy, F-measure and AUC are 57.97%, 0.558 and 0.759 respectively. Table 5.3 shows the genes corresponding to the probes ranked by the Chi-Squared attribute evaluation algorithm to classify samples as one of the ten subtypes.

Using the hierarchical decision-tree-based model makes the prediction procedure more accurate. Using accuracy and F-measure provides the same hierarchical model and subset of genes, while using AUC as the performance measure changes the order of the nodes and the genes. Figure 5.1 shows the tree learned in the training phase and the set of genes

Table 5.1: Comparison between hierarchical and non-hierarchical classification setups using LibSVM.

Classification Method	Gene Selection Method	# of Genes	Accuracy	F-measure	AUC
LibSVM	—	all genes	59.27%	0.554	0.765
LibSVM	Chi-Squared	89	72.31%	0.718	0.842
LibSVM	Chi-Squared	18	57.97%	0.558	0.759
HCL	HCL	80	91.7% - 98.8%	0.917 - 0.987	0.865 - 0.949
HCL	HCL	42	90.7% - 97.1%	0.906 - 0.972	0.862 - 0.942

selected at each step using accuracy as the performance measure. As shown in the figure, Subtype 6 has been selected as the first node, since the classifier yielded 98.8% accuracy for discriminating Subtype 6 from other subtypes, which was the highest accuracy among all subtypes. After removing the samples corresponding to subtype 6 and evaluating other subtypes in a one-against-all fashion, Subtype 2 yielded 98.6% accuracy, which was the highest among the nine remaining subtypes. This procedure continues until there is no subtype to classify. Figures 5.2 and 5.3 show the trees obtained in the same way, but using F-measure and AUC as the performance measures respectively.

Table 5.1 shows the performance of the model with different setups. In the first hierarchical setup, only 80 genes are used to predict the subtypes that the patient belongs to. The hierarchical model is able to increase the accuracy from around 72% to values between 92 and 99%, depending on the subtype, by using a new subset of genes based on this hierarchical method, which contains nine genes less than the original method.

To decrease the maximum number of genes per node, we tested different numbers of genes and a maximum of five genes per node yielding a very high performance to number of genes ratio. Using the aforementioned hierarchical tree and five probes for each node, the

Table 5.2: Genes corresponding to the top 100 probe IDs ranked by the Chi-Squared attribute evaluation algorithm to classify samples as one of the ten subtypes.

PGAP3	FOXC1	GRB7	CKS1B	UCHL5
ERBB2	PSMD3	CDCA7	CHEK1	BIRC5
WHSC1L1	INTS4	ERLIN2	PDSS1	CLNS1A
BRF2	MCM10	RRNAD1	CCNB2	NAT1
MIEN1	RAB11FIP1	GTPBP4	HEATR1	PPAPDC1B
NME3	TPX2	KIF2C	ZMYND10	WHSC1L1
GRB7	TRIP13	CEP55	FAM83D	CDC45
STARD3	CENPA	CA12	KIF14	REEP6
TCAP	MLPH	CA12	RAB26	ABAT
GRB7	FAM171A1	TRMU	HAPLN3	UBE2T
MELK	PDSS1	HID1	SLC7A8	CIRBP
GSDMB	CDCA8	RAD51AP1	SKP2	FAM64A
LSM1	NOSTRIN	FOXM1	ROGDI	DEGS2
MELK	PROSC	SLC52A2	CCNA2	
PSAT1	EXO1	CAPN8	CCNE1	
ORMDL3	UCK2	CDC20	ZG16B	
FOXA1	MIR3658	MCM10	NDUFC2	
GSDMB	CENPA	ILF2	NDUFC2	
ALG8	ASH2L	ASPM	COG2	
DDHD2	WWP1	LPIN1	IQCK	

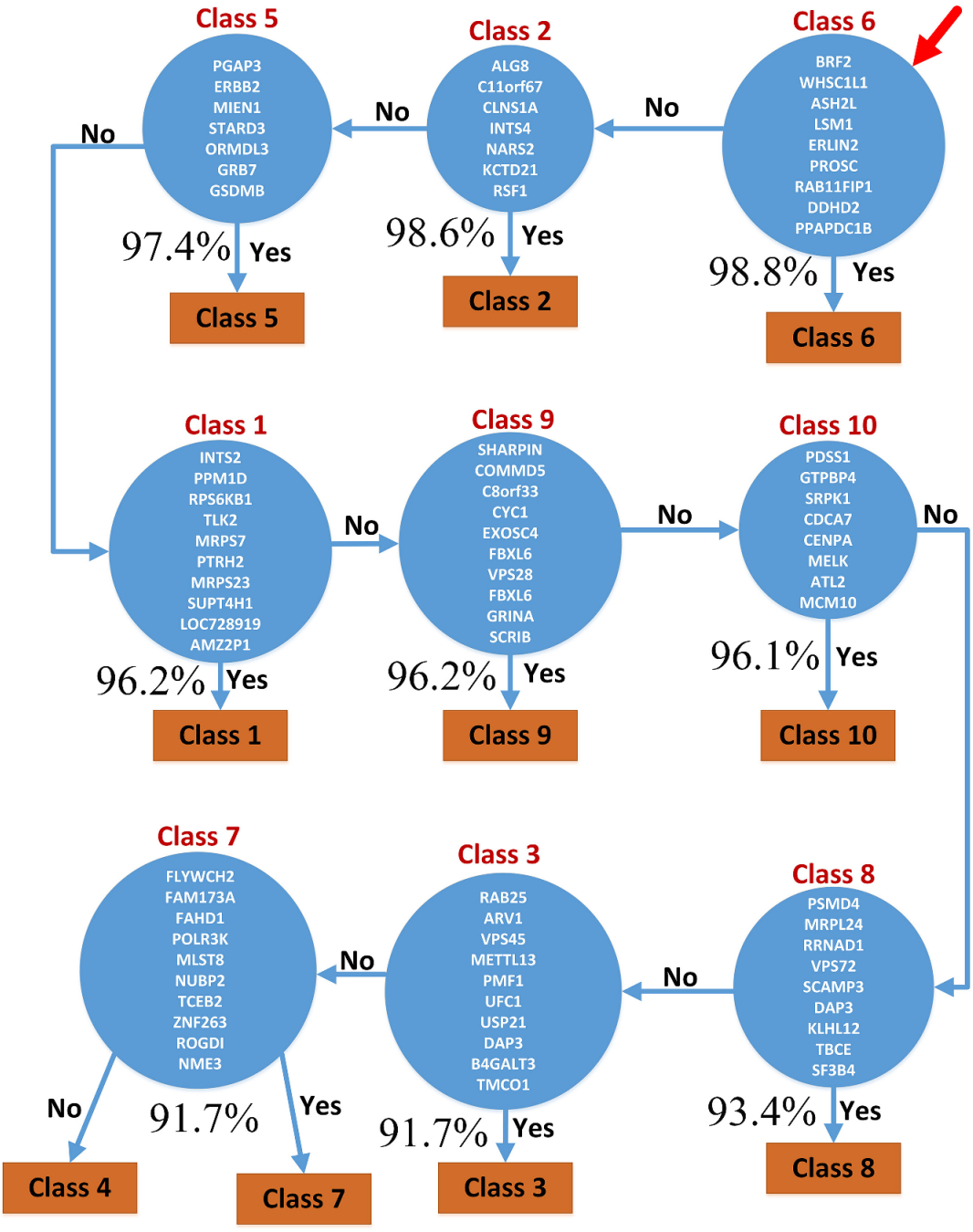

Figure 5.1: Hierarchical decision tree for determining breast cancer type using selected genes using accuracy as performance measure.

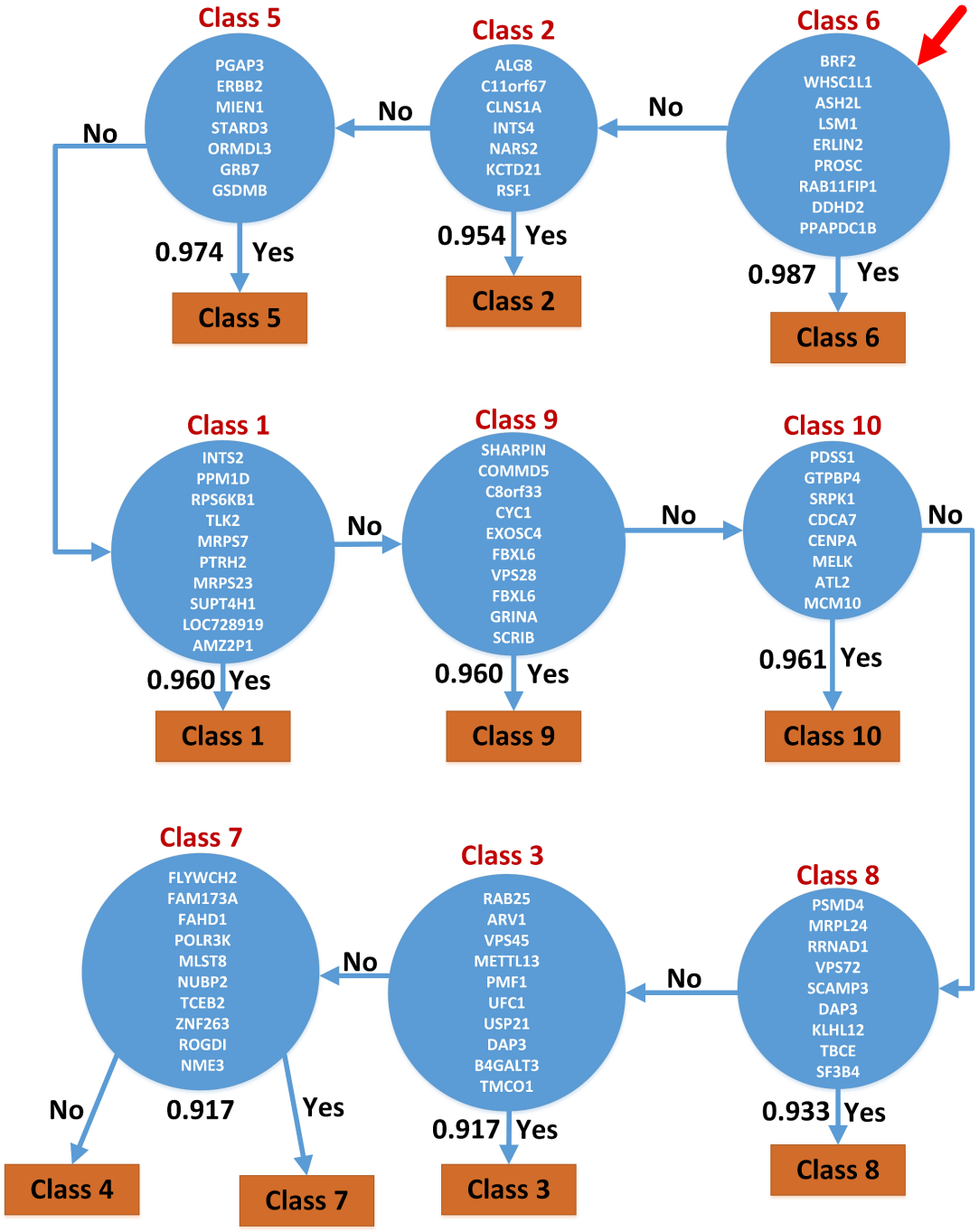

Figure 5.2: Hierarchical decision tree for determining breast cancer type using selected genes using F-measure as performance measure.

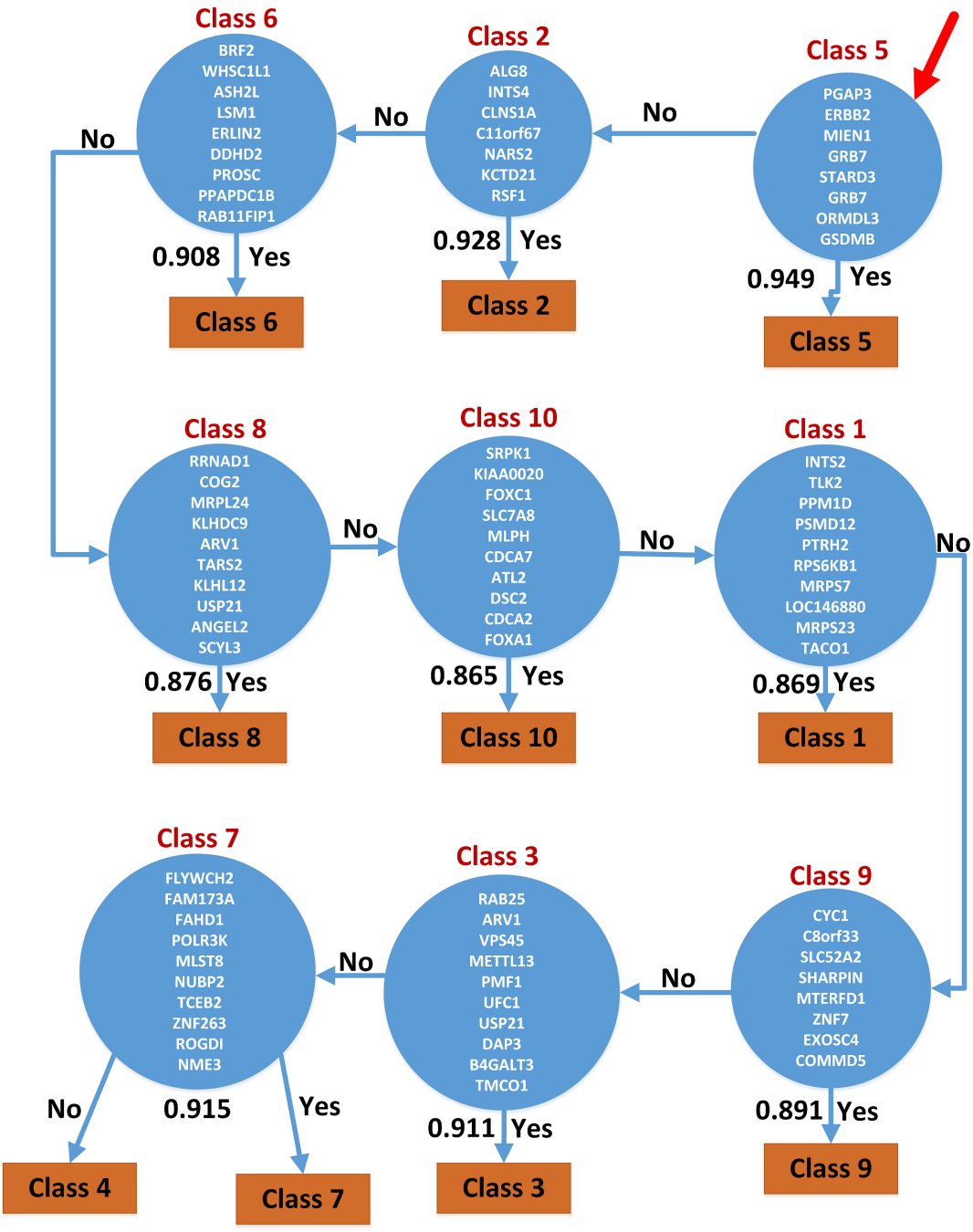

Figure 5.3: Hierarchical decision tree for determining breast cancer type using selected genes using AUC as performance measure.

Table 5.3: Top 20 genes ranked by the Chi-Squared attribute evaluation algorithm to classify samples as one of the ten subtypes.

PGAP3	MELK
ERBB2	GSDMB
WHSC1L1	LSM1
BRF2	MELK
MIEN1	PSAT1
NME3	ORMDL3
GRB7	FOXA1
STARD3	GSDMB
TCAP	ALG8
GRB7	DDHD2

proposed model yields a prediction performance very close to that of the previous model with ten probes for each node. These results are very interesting considering that we used only 42 genes instead of 82 genes. Figure 5.4 shows the tree learned in the training phase and the set of genes selected at each step using AUC as the performance measure.

From Table 5.1, we observe that using 80 to 82 genes, the proposed method achieves the highest performance in terms of accuracy, F-measure and AUC. Using only 42 genes in the last experiment also yields very high performance.

In a third experiment, considering accuracy as the performance measure, the proposed model performs very well using different numbers of genes per node (see Figure 5.5). In this comparison, we used Chi-Squared as the feature selection method and LibSVM as the classification algorithm. As can be seen in the figure, while in most of the cases the

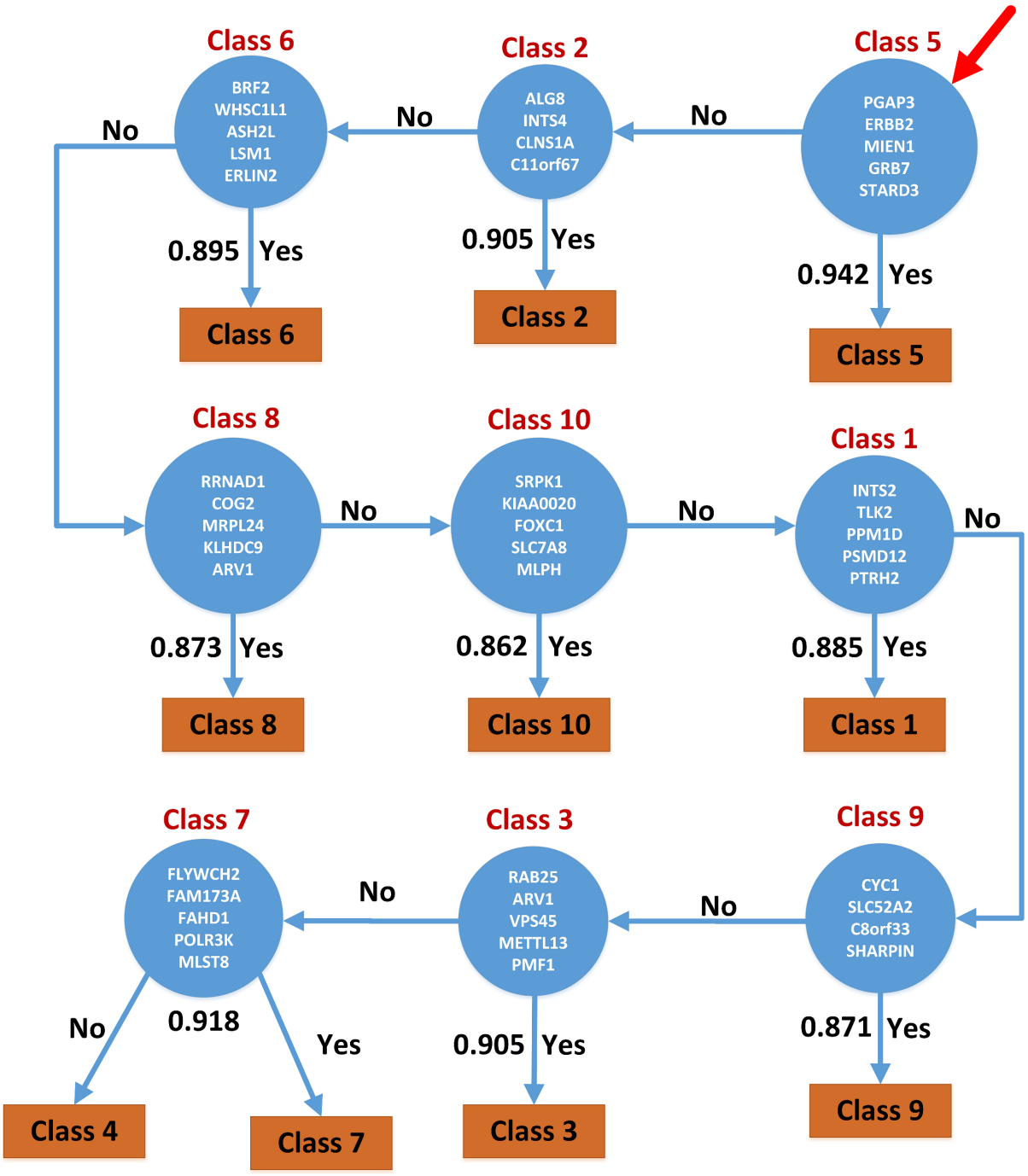

Figure 5.4: Hierarchical decision tree for determining breast cancer types using a maximum of 5 selected genes in each node and AUC as performance measure.

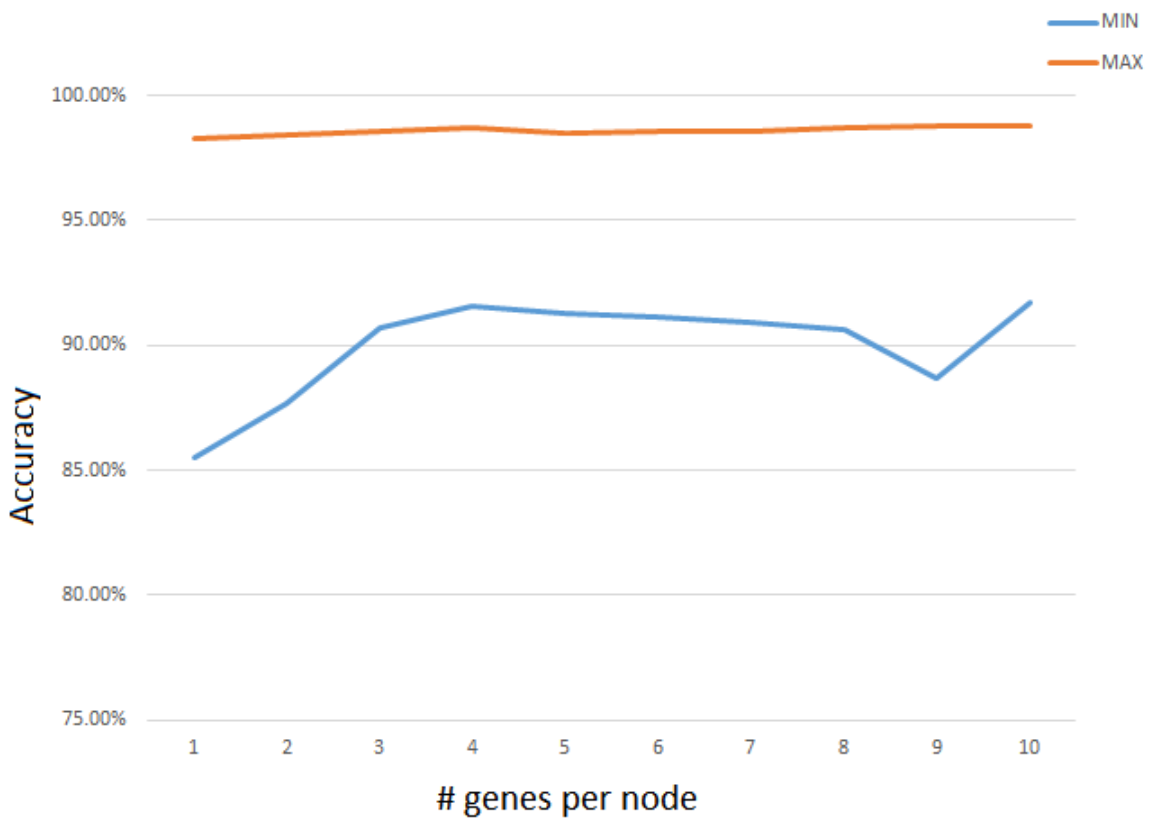

Figure 5.5: Comparison between minimum and maximum accuracy of the model using different number of genes per each node.

maximum accuracy of the model is very similar and close to 98%, the minimum accuracy varies from 85% to 92%. The reason behind this variation is that in most of the cases, we have at least one node in the tree, which is able to achieve very high accuracy for classifying its corresponding subtype even using one or two genes. In contrast, in some other nodes, using less than three genes can not provide more than 90% accuracy. Thus, for those nodes we have to increase the number of genes to keep the performance of the classifier corresponding to that node at a high level.

Moreover, we compared several classification methods using the same Chi-Squared fea-

ture selection algorithm. Table 5.4 shows the result of this comparison based on the three performance measures. For this comparison, we used Chi-Squared feature selection and 42 genes, based on the setup that corresponds to the last row of the Table 5.1. As shown in the table, LibSVM performs better than the other classifiers. In terms of performance measures, accuracy and F-measure provide a fair assessment, since they are prone to introduce biased results in imbalanced classification problems, which is the case of the dataset we use. Using AUC as the performance measure one can solve this issue and produce an unbiased evaluation.

Table 5.4: Comparison between different classification methods using the HCL method featuring chi-squared as feature selection and three performance measures.

Classification Method	Accuracy	F-measure	AUC
LibSVM	90.7%-97.1%	0.906-0.972	0.862-0.942
Decision Tree	88.4%-94.9%	0.884-0.949	0.819-0.916
JRip	89.4%-95.3%	0.891-0.952	0.821-0.926
KNN	90.3%-96.6%	0.904-0.967	0.870-0.931

To provide an additional insight to the hierarchical classification model used in this study, we have conducted the same gene selection and classification procedure for another dataset from the study of Hu et al. [31]. Selecting a total of 18 genes for that dataset yields very high prediction accuracy. Running the hierarchical model on the METABRIC dataset yields very high accuracy too, though lower than those of Hu's dataset (see Figure 5.6). This is possibly due to the fact that the subtypes in the METABRIC dataset were originally obtained using an unsupervised classification model, while those of Hu's dataset were pathologically verified. After comparing the genes selected for both datasets, it is notice-

able that only a few genes are in common for the corresponding subtypes. We then carried out the same gene selection and classification on both datasets, extending the number of genes up to 50 per node. The genes in common for both datasets are noticeable for the subtypes compared: 9 for Basal, 6 for Normal, 12 for Her2 and 19 for Lum A/B. The large number of common genes for Lum A/B (almost 40%) is consistent with the fact that these two subtypes are quite similar, and hence the lower accuracy of the model for classifying these two subtypes. It is also interesting to see that there are a very small number of genes in common for the Normal subtype. This is possibly due to the presence of some heterogeneity in the normal sample tissues. It is also encouraging to see several genes associated with specific subtypes. In particular, FoxA1 is common in the Basal subtype, which is consistent with previous studies [7]. These genes are important targets for further studies, as an extension of this work.

In addition, we found an interesting gene set within the luminal subtypes. It is shown that the mis-regulation of both BUB1 and CENPM genes have been implicated in different forms of cancer [12]. CDCA5 which encodes the protein Soronin, required for cohesin binding to the chromatin [57]. CDKN3 is also implicated in cell cycle regulation [28] and has been shown that, in several forms of cancer, CDKN3 is deleted or mutated [74].

5.1.2 Bottom-Up Approach

This section covers the results of the bottom-up approach for clustering the ten subtypes in an agglomerative fashion. Several distance evaluation methods such as single linkage, complete linkage, average linkage and the variance-based method using Ward's distance

measure have been applied. Figure 5.7 shows the tree obtained using the single linkage method.

We repeated the same experiment using the average linkage and complete linkage distance methods. Figures 5.8 and 5.9 show the trees obtained using the average linkage and complete linkage methods, respectively.

As shown in the figures, the complete linkage method provides a more balanced tree, whereas the average linkage and single linkage provide a completely linear tree in terms of subtype relationships.

We repeated the same experiment using the Ward's distance method as well. Since

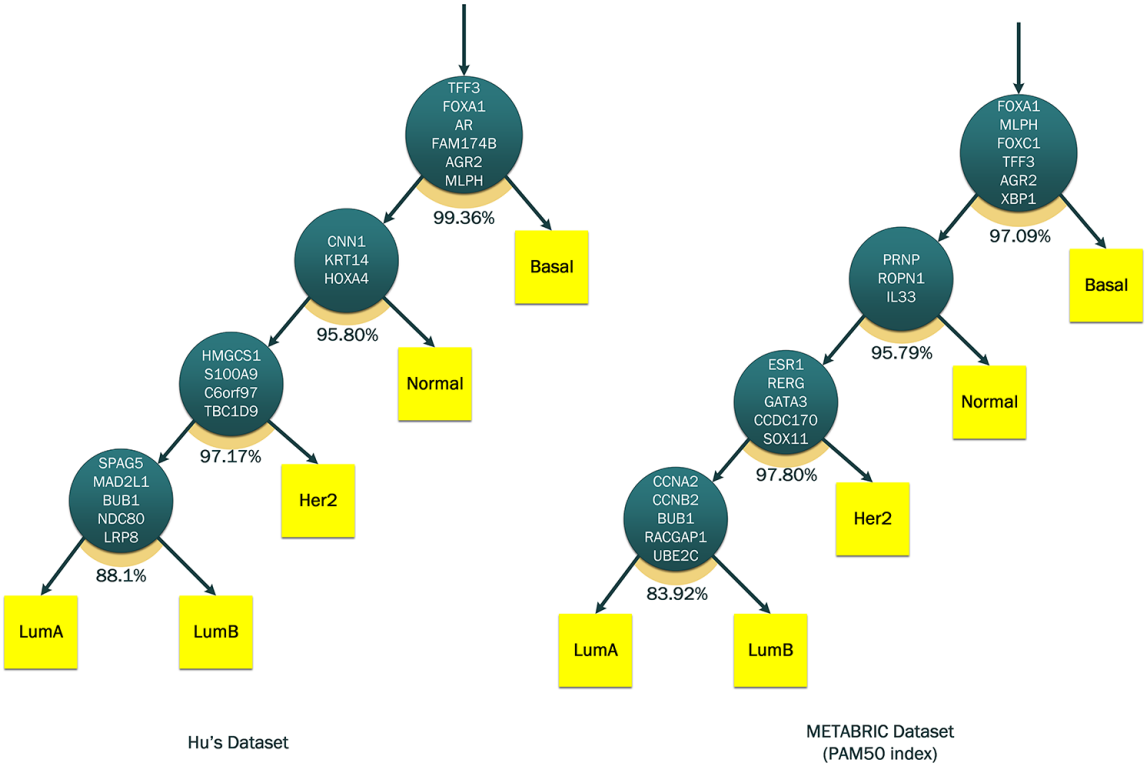

Figure 5.6: Schematic visualization of the trees obtained for Hu's dataset (left) and the METABRIC dataset (right).

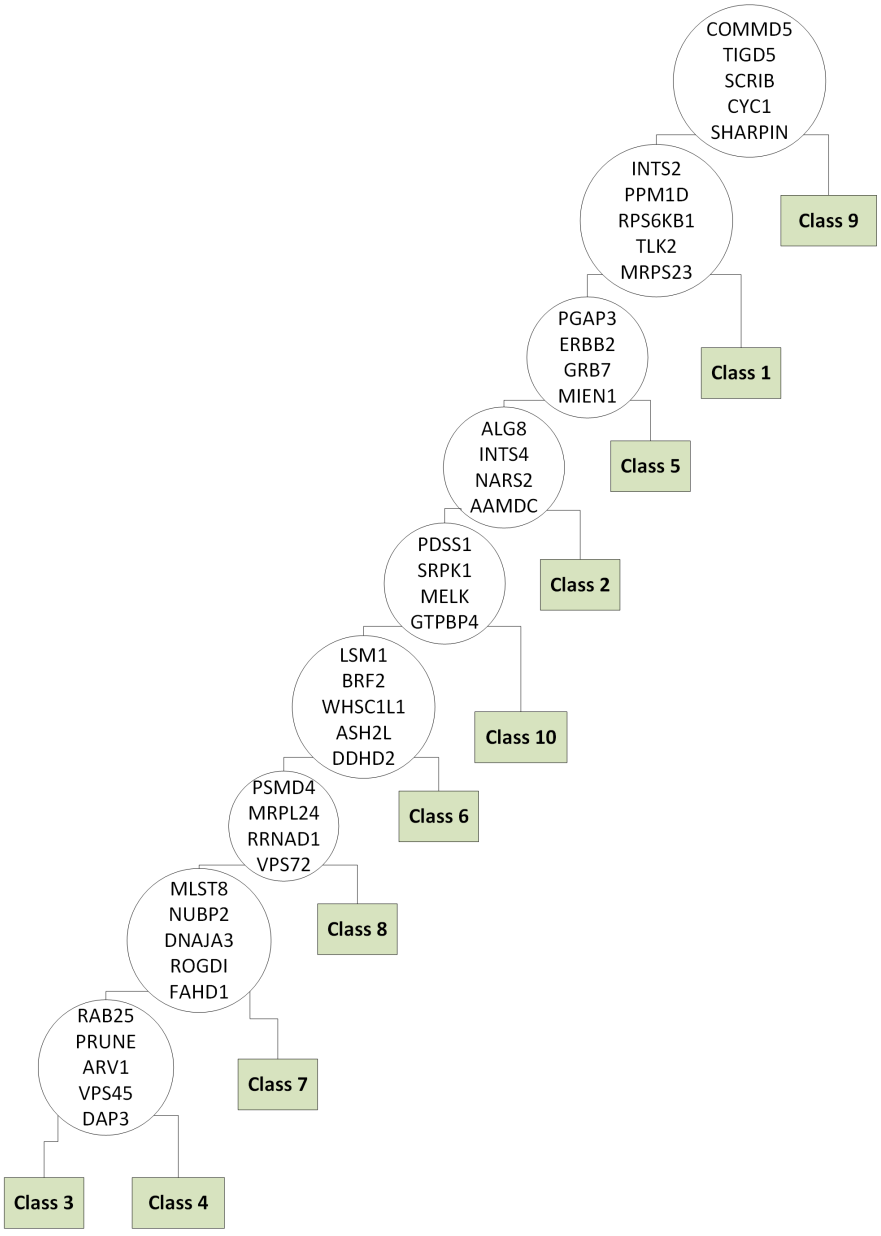

Figure 5.7: The hierarchical tree obtained using agglomerative clustering and single linkage as the distance method.

Ward's method takes the size of each cluster into account (i.e. the number of samples in each subtype), it tends to merge smaller clusters first. Figure 5.10 shows the tree obtained using the single linkage method.

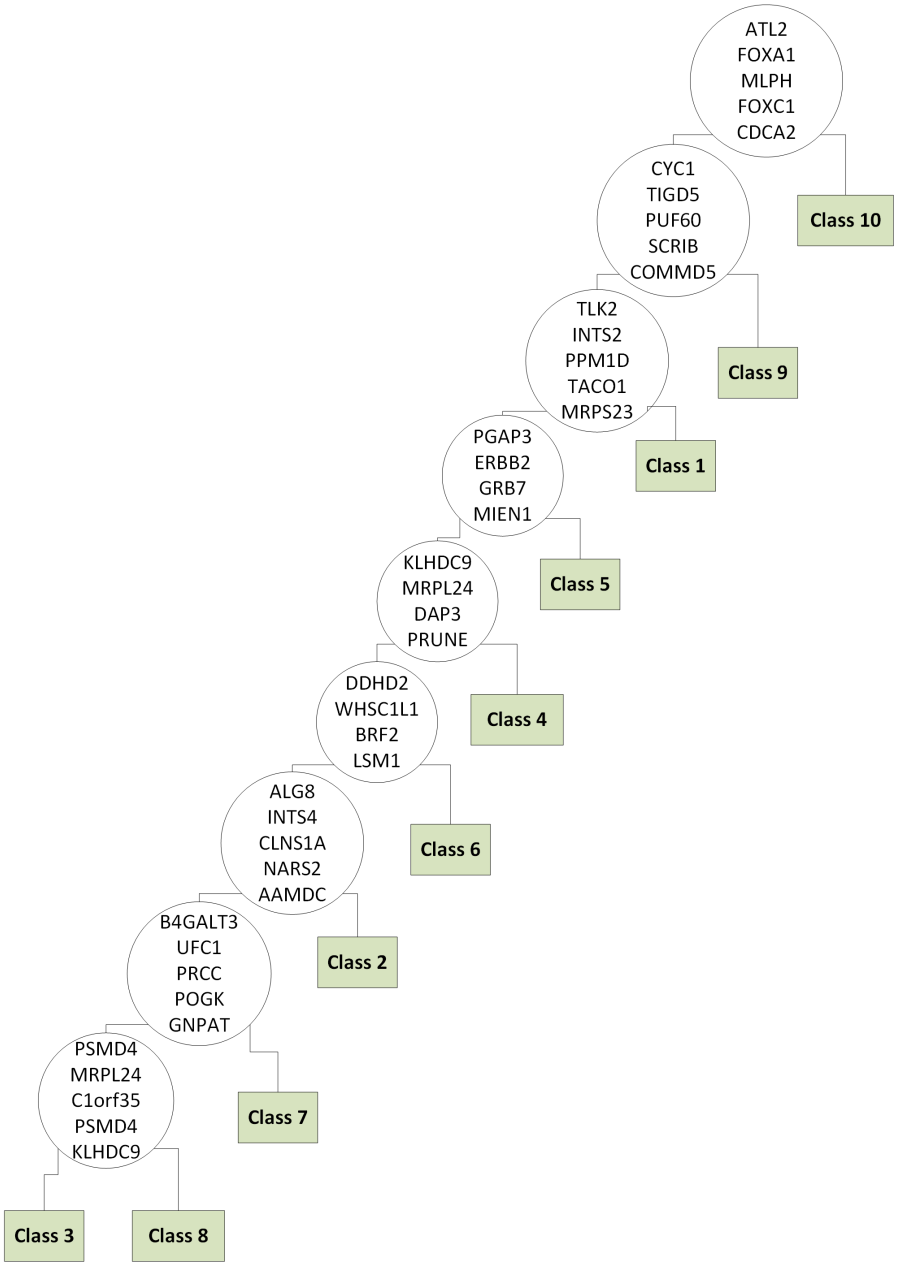

Figure 5.8: The hierarchical tree obtained using agglomerative clustering and average linkage as the distance method.

Moreover, we used different performance measures such as accuracy, F-measure and AUC. To achieve a more comprehensive comparison, we used three different classifiers:

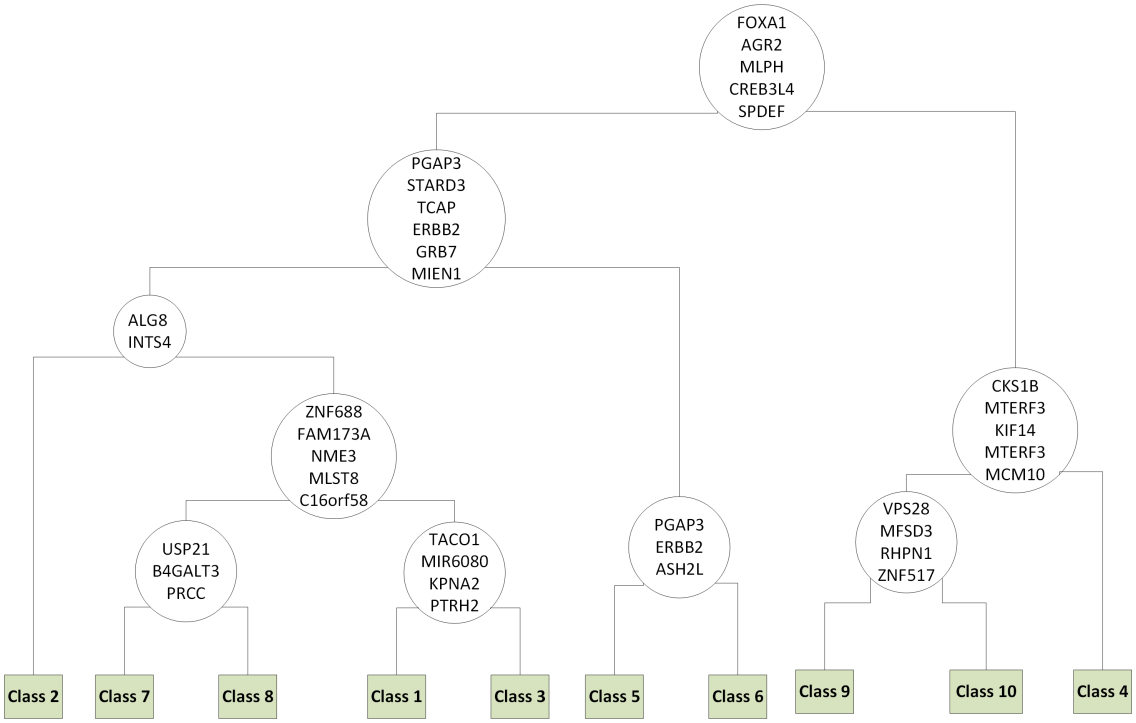

Figure 5.9: The hierarchical tree obtained using agglomerative clustering and complete linkage as the distance method.

Decision Tree, Random Forest and LibSVM. Table 5.5 shows the comparison between different distance methods based on the average of the performance measures in the tree. We repeated the same experiment and evaluated the minimum value of the performance measures across all trees. Table 5.6 shows the comparison between the methods in the worst case. As shown in Table 5.5, random forest performed better than decision tree and SVM using any distance method in terms of achieving a greater AUC. Using accuracy as the performance measure can be misleading since the classes (subtypes) contain an unbalanced number of samples. For example in Figure 5.7, the number of samples in the root of the tree is 67 for Subtype 9 versus 930 for the remaining subtypes. Using accuracy as the performance measure and LibSVM as the classifier provides almost 95% accuracy, while

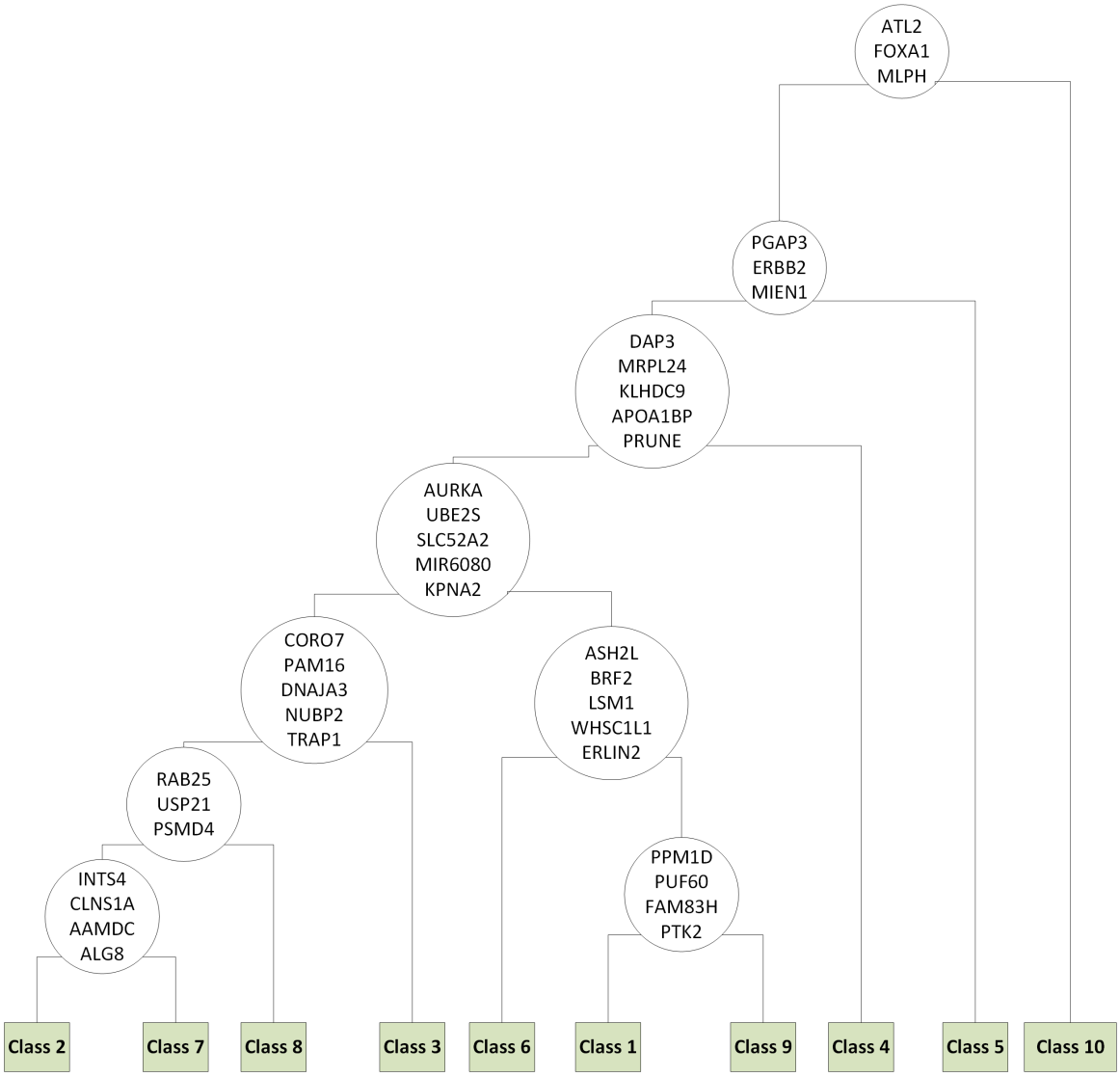

Figure 5.10: The hierarchical tree obtained using agglomerative clustering and Ward's method as the distance method.

43 samples out of a total 67 samples in subtype 9 are misclassified. Using AUC, on the other hand, yields 0.675 out of 1, which reflects the good performance of the model.

In addition, since Subtype 10 has a large overlap with Basal in the PAM50 index (see Figure 4.1), creating a separate branch for this subtype at the top of the tree makes more

Table 5.5: Comparison for using different distance methods to obtain the tree. The measures are obtained using the average of performances in all nodes of the tree.

Distance Method	Classification Method	# of Genes	Accuracy	F-measure	AUC
Single Linkage	Decision Tree	41	93.5%	0.933	0.875
	Random Forest		94.5%	0.944	0.957
	LibSVM		95.0%	0.948	0.872
Average Linkage	Decision Tree	42	92.4%	0.923	0.878
	Random Forest		92.7%	0.926	0.942
	LibSVM		93.9%	0.937	0.862
Complete Linkage	Decision Tree	37	89.8%	0.897	0.881
	Random Forest		90.1%	0.899	0.928
	LibSVM		91.5%	0.912	0.890
Ward's method	Decision Tree	37	89.1%	0.890	0.874
	Random Forest		90.8%	0.907	0.935
	LibSVM		91.7%	0.916	0.881

sense, biologically. This behavior can be seen in trees generated with *average linkage* and *Ward's method* only. Moreover, as Figure 4.1 shows, most of the samples from class *Her2* fall in Subtype 5 and there is a similar relationship between class *Normal* and Subtype 4. The only model out of the four compared trees (single linkage, complete linkage, average linkage and Ward's method) that supports these observations, is the tree based on *Ward's* method. Also the genes corresponding to class 10 (see Figure 5.10), are known to be involed in Basal subtype [7, 29]. *ERBB2* gene, which is known as *Her2* as well, is one of the genes related to subtype 5. Moreover, both PGAP3 and MIEN1 genes, which correspond to class

5, are shown to be over-expressed in *Her2* tumor cells [51].

5.2 Discussion

The results demonstrate an impressive accuracy to predict these new breast cancer types using only 40 to 80 genes. This could allow for the small number of genes to be utilized on a custom array, reducing costs of the array as well as significantly reducing data processing time and costs. It would be valuable to determine if these changes are reflected at the protein

Table 5.6: Comparison for using different distance methods to obtain the tree. The measures are obtained using the lowest performance in all nodes of the tree.

Distance Method	Classification Method	# of Genes	Accuracy	F-measure	AUC
Single Linkage	Decision Tree	41	87.2%	0.870	0.811
	Random Forest		89.1%	0.890	0.909
	LibSVM		89.8%	0.895	0.675
Average Linkage	Decision Tree	42	82.5%	0.823	0.797
	Random Forest		82.8%	0.828	0.877
	LibSVM		87.3%	0.873	0.683
Complete Linkage	Decision Tree	37	72.9%	0.729	0.732
	Random Forest		75.8%	0.758	0.790
	LibSVM		78.7%	0.765	0.699
Ward's method	Decision Tree	37	77.9%	0.778	0.784
	Random Forest		81.2%	0.813	0.858
	LibSVM		81.0%	0.809	0.784

level. This may lend itself to screening for subtypes based on tissue microarray technology, enabling the use of paraffin embedded patient samples. This would significantly reduce the technical problems inherent in obtaining quality RNA from routine clinical samples.

In the bottom-up approach, instead of using the whole model as a unified prediction model, we aimed to compute the level of similarity between the ten subtypes. Moreover, we aimed to determine which genes are the most responsible ones for separating similar subtypes. The results presented here show that using different distance methods can lead to different tree structures, which may or may not be balanced. Out of different distance methods, we found that Ward's method provides the most biologically promising three structure where class 10 and 5, which consist most of the Basal and Her2 samples, are separated from the rest at the top of the tree.

Resolving the biological roles for the individual genes that have emerged as unique identifiers for each subclass may reveal novel directions for therapeutic intervention. Furthermore, applying this method to more refined subtypes characterized using a combination of platforms, or incorporating patient response to therapy could further decrease the number of selected genes for each subtype and could reveal important aspects of breast cancer biology. Some preliminary results of this study have been published in [24] and an extended version of the article is under review, to be published in BMC bioinformatics journal.

Chapter 6

Conclusions

In this research, our goal was to build a predictive model using a small subset of genes that can accurately predict the newly discovered ten breast cancer subtypes using the METABRIC dataset. Considering 997 samples, more than 48,000 features and 630 Gigabytes of raw data, made the analysis of the METABRIC dataset truly challenging. Another challenge that we faced was creating a balance between the number of selected genes and the performance of the model.

Unlike the other models, which try to find a gene signature that classifies the type of breast cancer in an all-against-all fashion, using the top-down hierarchical model helped us to achieve a very high level of accuracy with a smaller subset of genes. Moreover, using the bottom-up approach, we were able to identify the level of similarity between different subtypes and yield the most significant genes that differentiate these subtypes.

6.1 Contributions

The contributions of this thesis can be summarized as follows:

- Using a hierarchical top-down approach for developing a prediction model for the newly discovered ten subtypes of breast cancer.
- Proposing a bottom-up approach based on different similarity measures for clustering subtypes and producing a tree-based model with semi-balanced topology.
- Using different classification methods and in-depth comparison of their performances on the METABRIC dataset.
- Evaluating the proposed models using different performance measures and comparing their performances for various cases.

6.2 Future Work

Although the METABRIC dataset consists of a large number of samples, using bigger datasets can help us better understand the features that predict these subtypes [3]. In this study, we targeted gene expression levels as key features for subtype prediction and identification, achieving very high performance. Using other biological information such as copy number variations (CNVs) and aberrations (CNAs) as well as single nucleotide polymorphism (SNP) can help build a more comprehensive model for identification of breast cancer subtypes.

Moreover, the development of breast cancer is usually caused by mutations of a small number of genes, called *driver genes*, whose changes deregulate many biological processes and, therefore, lead to initiation and progression of breast cancer. Finding those driver genes can lead to a deeper biological insight of the relationships among these ten breast cancer subtypes.

Bibliography

[1] Shigeo Abe. *Support vector machines for pattern classification*. Springer, 2010.

[2] Asmaa Al-Allak, Gianfilippo Bertelli, and Paul Lewis. Random forests: The new generation of machine learning algorithms to predict survival in breast cancer. *International Journal of Surgery*, 11(8):607, 2013.

[3] Hamid R Ali, Oscar M Rueda, Suet-Feung Chin, Christina Curtis, Mark J Dunning, SAJR Aparicio, and Carlos Caldas. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology*, 15:431, 2014.

[4] Adam L Asare, Zhong Gao, Vincent J Carey, Richard Wang, and Vicki Seyfert-Margolis. Power enhancement via multivariate outlier testing with gene expression arrays. *Bioinformatics*, 25(1):48–53, 2009.

[5] Nuno L Barbosa-Morais, Mark J Dunning, Shamith A Samarajiwa, Jeremy FJ Darot, Matthew E Ritchie, Andy G Lynch, and Simon Tavaré. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Research*, 38(3):e17–e17, 2010.

[6] Asa Ben-Hur and Jason Weston. A users guide to support vector machines. In *Data Mining Techniques for the Life Sciences*, pages 223–239. Springer, 2010.

[7] Gina M Bernardo, Gurkan Bebek, Charles L Ginther, Steven T Sizemore, Kristen L Lozada, John D Miedler, Lee A Anderson, Andrew K Godwin, Fadi W Abdul-Karim, Dennis J Slamon, et al. FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene*, 32(5):554–563, 2012.

[8] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006. ISBN 9780387310732.

[9] Fiona M Blows, Kristy E Driver, Marjanka K Schmidt, Annegien Broeks, Flora E Van Leeuwen, Jelle Wesseling, Maggie C Cheang, Karen Gelmon, Torsten O Nielsen, Carl Blomqvist, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Medicine*, 7(5):e1000279, 2010.

[10] Gianluca Bontempi and Patrick E. Meyer. A model-based relevance estimation approach for feature selection in microarray datasets. In *Artificial Neural Networks-ICANN 2008*, pages 21–31. Springer, 2008.

[11] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[12] Daniel P Cahill, Christoph Lengauer, Jian Yu, Gregory J Riggins, James KV Willson, Sanford D Markowitz, Kenneth W Kinzler, and Bert Vogelstein. Mutations of mitotic checkpoint genes in Human cancers. *Nature*, 392(6673):300–303, 1998.

[13] Yudong Cai, Tao Huang, Lele Hu, Xiaohe Shi, Lu Xie, and Yixue Li. Prediction of lysine ubiquitination with mrmr feature selection and analysis. *Amino Acids*, 42(4): 1387–1395, 2012.

[14] Lisa A Carey, Charles M Perou, Chad A Livasy, Lynn G Dressler, David Cowan, Kathleen Conway, Gamze Karaca, Melissa A Troester, Chiu Kit Tse, Sharon Edmiston, et al. Race, breast cancer subtypes, and survival in the carolina breast cancer study. *The Journal of American Medical Association*, 295(21):2492–2502, 2006.

[15] M. Emre Celebi, Hassan A. Kingravi, and Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210, 2013.

[16] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[17] Maggie CU Cheang, Stephen K Chia, David Voduc, Dongxia Gao, Samuel Leung, Jacqueline Snider, Mark Watson, Sherri Davies, Philip S Bernard, Joel S Parker, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute*, 2009.

[18] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[19] Michiel JL de Hoon, Seiya Imoto, John Nolan, and Satoru Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[20] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[21] Mark J Dunning, Mike L Smith, Matthew E Ritchie, and Simon Tavaré. beadarray: R classes and methods for illumina bead-based data. *Bioinformatics*, 23(16):2183–2184, 2007.

[22] Stephen B. Edge and Carolyn C. Compton. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of Surgical Oncology*, 17(6):1471–1474, 2010.

[23] Abbas Toloie Eshlaghy, Ali Poorebrahimi, Mandana Ebrahimi, Amir R Razavi, and Leila Ghasem Ahmad. Using three machine learning techniques for predicting breast cancer recurrence. *Journal of Health and Medicine Information*, 4(2):124, 2013.

[24] Forough Firoozbakht, Iman Rezaeian, Lisa Porter, and Luis Rueda. Breast cancer subtype identification using machine learning techniques. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference on*, pages 1–2. IEEE, 2014.

[25] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition (2nd Ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990. ISBN 0-12-269851-7.

[26] A. Goldhirsch, WC. Wood, AS. Coates, RD. Gelber, B. Thürlimann, H.J. Senn, et al. Strategies for subtypesdealing with the diversity of breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2011. *Annals of Oncology*, page mdr304, 2011.

[27] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[28] Gregory J Hannon, David Casso, and DKAP Beach. KAP: a dual specificity phosphatase that interacts with cyclin-dependent kinases. *Proceedings of the National Academy of Sciences*, 91(5):1731–1735, 1994.

[29] S. Hasmuller, R. Wirtz, M. Lenhard, S. Riickert, S. Kahlert, I. Bauerfeind, N. Ditsch, I. Ruhl, P. Fasching, and M. Untch. Response of basal-like tumors defined by ESR Her-2/neu, MLPH and MMP7 to neoadjuvant chemotherapy. *European Journal of Cancer Supplements*, 6(7):190, 2008.

[30] Jason I. Herschkowitz, Karl Simin, Victor J. Weigman, Igor Mikaelian, Jerry Usary, Zhiyuan Hu, Karen E. Rasmussen, Laundette P. Jones, Shahin Assefnia, Subhashini Chandrasekharan, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology*, 8 (5):R76, 2007.

[31] Zhiyuan Hu, Cheng Fan, Daniel S. Oh, JS. Marron, Xiaping He., Bahjat F. Qaqish, Chad Livasy, Lisa A. Carey, Evangeline Reynolds, Lynn Dressler, et al. The molecular

portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7(1):96, 2006.

[32] Chukwuemeka U Ihemelandu, LaSalle D Leffall Jr, Robert L Dewitty, Tammey J Naab, Haile M Mezghebe, Kepher H Makambi, Lucile Adams-Campbell, and Wayne A Frederick. Molecular breast cancer subtypes in premenopausal and postmenopausal african-american women: age-specific prevalence and survival. *Journal of Surgical Research*, 143(1):109–118, 2007.

[33] Anil K Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[34] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[35] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3): 231–240, 2011.

[36] Bi-Qing Li, Tao Huang, Lei Liu, Yu-Dong Cai, and Kuo-Chen Chou. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS one*, 7(4):e33393, 2012.

[37] Yifeng Li, Alioune Ngom, and Luis Rueda. A framework of gene subset selection using multiobjective evolutionary algorithm. In *Pattern Recognition in Bioinformatics*, pages 38–48. Springer, 2012.

[38] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pages 388–391, 1995.

[39] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, pages 388–388. IEEE Computer Society, 1995.

[40] Qingzhong Liu, Andrew H Sung, Zhongxue Chen, Jianzhong Liu, Xudong Huang, and Youping Deng. Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *PloS one*, 4(12):e8250, 2009.

[41] Xiaoxing Liu, Arun Krishnan, and Adrian Mondry. An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics*, 6(1):76, 2005.

[42] H. Lodish, A. Berk, C.A. Kaiser, M. Krieger, A. Bretscher, H. Ploegh, A. Amon, and M.P. Scott. *Molecular Cell Biology*. W. H. Freeman, 2012. ISBN 9781429234139.

[43] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. California, USA, 1967.

[44] Michael P. Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H. Benes, Pedro J. Ballester, and Julio Saez-Rodriguez. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One*, 8(4):e61318, 2013.

[45] Patrick E. Meyer, Frederic Lafitte, and Gianluca Bontempi. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9(1):461, 2008.

[46] Mohd Saberi Mohamad, Sigeru Omatu, Safaai Deris, and Michifumi Yoshioka. Particle swarm optimization for gene selection in classifying cancer classes. *Artificial Life and Robotics*, 14(1):16–19, 2009.

[47] Y.L. Pavlov. *Random Forests*. VSP, 2000. ISBN 9789067643146.

[48] Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[49] Charles M. Perou and Anne-Lise Børresen-Dale. Systems biology and genomics of breast cancer. *Cold Spring Harbor Perspectives in Biology*, 3(2):a003293, 2011.

[50] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[51] Gregory E. Plautz, Arun Modi, and Li-Xin Wang. ERBB2 amplicon passenger genes: A novel class of breast cancer antigens. *Cancer Research*, 74(19 Supplement):2897–2897, 2014.

[52] Jorge S Reis-Filho and Lajos Pusztai. Gene expression profiling in breast cancer:

classification, prognostication, and prediction. *The Lancet*, 378(9805):1812–1823, 2011.

[53] Iman Rezaeian, Yifeng Li, Martin Crozier, Eran Andrechek, Alioune Ngom, Luis Rueda, and Lisa Porter. Identifying informative genes for prediction of breast cancer subtypes. In *Pattern Recognition in Bioinformatics*, pages 138–148. Springer, 2013.

[54] L. Rokach. *Data Mining with Decision Trees: Theory and Applications*. Series in machine perception and artificial intelligence. World Scientific Publishing Company, Incorporated, 2007. ISBN 9789812771728.

[55] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, 1991.

[56] C. Sammut and G.I. Webb. *Encyclopedia of Machine Learning*. Encyclopedia of Machine Learning. Springer, 2011. ISBN 9780387307688.

[57] Julia Schmitz, Erwan Watrin, Péter Lénárt, Karl Mechtler, and Jan-Michael Peters. Sororin is required for stable binding of cohesin to chromatin and for sister chromatid cohesion in interphase. *Current Biology*, 17(7):630–636, 2007.

[58] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel Methods in Computational Biology*. MIT press, 2004.

[59] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[60] Therese Sørlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19): 10869–10874, 2001.

[61] Christos Sotiriou, Soek-Ying Neo, Lisa M McShane, Edward L Korn, Philip M Long, Amir Jazaeri, Philippe Martiat, Steve B Fox, Adrian L Harris, and Edison T Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, 100(18): 10393–10398, 2003.

[62] R Core Team et al. R: A language and environment for statistical computing. 2012.

[63] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Science, 2008. ISBN 9780080949123.

[64] H. Ulrich and G. KreBel. Advances in kernel methods. pages 255–268. 1999. ISBN 0-262-19416-3.

[65] SEER Program (National Cancer Institute (U.S.)), C.H. Johnson, and National Cancer Institute (U.S.). Cancer Statistics Branch. *The SEER Program Coding and Staging Manual 2004*. NIH publication. Cancer Statistics Branch, Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Public Health Service, U.S. Department of Health and Human Services, 2004.

[66] Laura J. van't Veer, Hongyue Dai, Marc J. Van De Vijver, Yudong D. He, Augustinus AM. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002.

[67] V. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer, 2000. ISBN 9780387987804.

[68] V. Vapnik and S. Kotz. *Estimation of Dependences Based on Empirical Data*. Information Science and Statistics. Springer, 2006. ISBN 9780387342399.

[69] Vladimir Vapnik and Corinna Cortes. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[70] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

[71] V.N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998. ISBN 9780471030034.

[72] K David Voduc, Maggie CU Cheang, Scott Tyldesley, Karen Gelmon, Torsten O Nielsen, and Hagen Kennecke. Breast cancer subtypes and the risk of local and regional relapse. *Journal of Clinical Oncology*, 28(10):1684–1691, 2010.

[73] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17 (4):395–416, 2007.

[74] L Wang, L Sun, J Huang, and M Jiang. Cyclin-dependent kinase inhibitor 3 (CDKN3) novel cell cycle computational network between human non-malignancy associated hepatitis/cirrhosis and hepatocellular carcinoma (HCC) transformation. *Cell Proliferation*, 44(3):291–299, 2011.

[75] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[76] A.R. Webb and K.D. Copsey. *Statistical Pattern Recognition*. Wiley, 2011. ISBN 9781119952961.

[77] John L Young. *SEER summary staging manual 2000: codes and coding instructions*. Number 1. National Cancer Institute, National Institutes of Health, 2001.

[78] Tao Zeng and Juan Liu. Mixture classification model based on clinical markers for breast cancer prognosis. *Artificial Intelligence in Medicine*, 48(2):129–137, 2010.

Appendix A

Table A.1: Performance of the *decision tree* classifier along with the *single linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
3 vs 4	87.9%	0.879	0.903
3,4 vs 7	87.2%	0.870	0.811
3,4,7 vs 8	89.7%	0.897	0.884
3,4,7,8 vs 6	97.7%	0.977	0.891
3,4,7,8,6 vs 10	94.9%	0.950	0.901
3,4,7,8,6,10 vs 2	97.7%	0.978	0.851
3,4,7,8,6,10,2 vs 5	98.5%	0.985	0.981
3,4,7,8,6,10,2,5 vs 1	93.1%	0.927	0.811
3,4,7,8,6,10,2,5,1 vs 9	94.5%	0.935	0.847

Table A.2: Performance of the *random forest* classifier along with the *single linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
3 vs 4	91.6%	0.916	0.970
3,4 vs 7	89.1%	0.890	0.927
3,4,7 vs 8	91.1%	0.911	0.953
3,4,7,8 vs 6	98.2%	0.982	0.986
3,4,7,8,6 vs 10	95.7%	0.956	0.961
3,4,7,8,6,10 vs 2	97.4%	0.974	0.991
3,4,7,8,6,10,2 vs 5	98.5%	0.985	0.996
3,4,7,8,6,10,2,5 vs 1	94.3%	0.943	0.925
3,4,7,8,6,10,2,5,1 vs 9	94.3%	0.941	0.909

Table A.3: Performance of the *LibSVM* classifier along with the *single linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
3 vs 4	92.3%	0.923	0.922
3,4 vs 7	89.8%	0.895	0.838
3,4,7 vs 8	91.3%	0.912	0.870
3,4,7,8 vs 6	98.2%	0.982	0.906
3,4,7,8,6 vs 10	96.5%	0.964	0.905
3,4,7,8,6,10 vs 2	98.0%	0.981	0.927
3,4,7,8,6,10,2 vs 5	98.1%	0.982	0.985
3,4,7,8,6,10,2,5 vs 1	95.4%	0.952	0.819
3,4,7,8,6,10,2,5,1 vs 9	95.0%	0.941	0.675

Table A.4: Performance of the *decision tree* classifier along with the *average linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
3 vs 8	83.9%	0.840	0.856
3,8 vs 7	89.7%	0.897	0.911
3,8,7 vs 2	97.4%	0.974	0.962
3,8,7,2 vs 6	98.3%	0.984	0.937
3,8,7,2,6 vs 4	82.5%	0.823	0.820
3,8,7,2,6,4 vs 5	98.5%	0.986	0.979
3,8,7,2,6,4,5 vs 1	93.3%	0.932	0.797
3,8,7,2,6,4,5,1 vs 9	93.6%	0.924	0.850
3,8,7,2,6,4,5,1,9 vs 10	94.7%	0.945	0.789

Table A.5: Performance of the *random forest* classifier along with the *average linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
3 vs 8	83.9%	0.840	0.922
3,8 vs 7	91.4%	0.914	0.955
3,8,7 vs 2	96.9%	0.969	0.982
3,8,7,2 vs 6	97.8%	0.978	0.986
3,8,7,2,6 vs 4	82.8%	0.828	0.877
3,8,7,2,6,4 vs 5	98.4%	0.984	0.993
3,8,7,2,6,4,5 vs 1	94.6%	0.944	0.938
3,8,7,2,6,4,5,1 vs 9	93.7%	0.934	0.896
3,8,7,2,6,4,5,1,9 vs 10	94.7%	0.946	0.933

Table A.6: Performance of the *LibSVM* classifier along with the *average linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
3 vs 8	87.3%	0.873	0.873
3,8 vs 7	92.9%	0.928	0.893
3,8,7 vs 2	97.4%	0.974	0.936
3,8,7,2 vs 6	97.4%	0.973	0.904
3,8,7,2,6 vs 4	85.8%	0.855	0.790
3,8,7,2,6,4 vs 5	98.2%	0.982	0.985
3,8,7,2,6,4,5 vs 1	95.8%	0.956	0.829
3,8,7,2,6,4,5,1 vs 9	94.7%	0.938	0.683
3,8,7,2,6,4,5,1,9 vs 10	95.6%	0.956	0.869

Table A.7: Performance of the *decision tree* classifier along with the *complete linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
1 vs 3	95.3%	0.953	0.959
2 vs 7,8,1,3	96.8%	0.969	0.894
5 vs 6	99.3%	0.993	0.989
7 vs 8	91.7%	0.917	0.926
9 vs 10	92.0%	0.920	0.932
1,3 vs 7,8	72.9%	0.729	0.732
9,10 vs 4	90.3%	0.903	0.900
2,7,8,1,3 vs 5,6	90.3%	0.898	0.806
2,7,8,1,3,5,6 vs 9,10,4	79.5%	0.787	0.789

Table A.8: Performance of the *random forest* classifier along with the *complete linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
1 vs 3	96.0%	0.961	0.988
2 vs 7,8,1,3	96.8%	0.969	0.965
5 vs 6	98.6%	0.986	0.988
7 vs 8	91.3%	0.912	0.966
9 vs 10	90.2%	0.902	0.971
1,3 vs 7,8	75.8%	0.758	0.840
9,10 vs 4	93.3%	0.933	0.976
2,7,8,1,3 vs 5,6	91.3%	0.908	0.867
2,7,8,1,3,5,6 vs 9,10,4	77.1%	0.760	0.790

Table A.9: Performance of the *LibSVM* classifier along with the *complete linkage* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
1 vs 3	96.9%	0.970	0.961
2 vs 7,8,1,3	97.2%	0.972	0.934
5 vs 6	100%	1.0	1.0
7 vs 8	94.0%	0.941	0.941
9 vs 10	93.8%	0.939	0.934
1,3 vs 7,8	78.8%	0.787	0.787
9,10 vs 4	92.4%	0.924	0.924
2,7,8,1,3 vs 5,6	91.6%	0.911	0.827
2,7,8,1,3,5,6 vs 9,10,4	78.7%	0.765	0.699

Table A.10: Performance of the *decision tree* classifier along with *Ward's* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
1 vs 9	86.0%	0.860	0.880
2 vs 7	94.8%	0.948	0.936
1,9 vs 6	94.1%	0.942	0.940
2,7 vs 8	87.2%	0.872	0.908
2,7,8 vs 3	77.9%	0.778	0.784
2,7,8,3 vs 6,1,9	83.9%	0.833	0.830
2,7,8,3,6,1,9 vs 4	86.2%	0.860	0.807
2,7,8,3,6,1,9,4 vs 5	96.6%	0.965	0.977
2,7,8,3,6,1,9,4,5 vs 10	95.3%	0.953	0.805

Table A.11: Performance of the *random forest* classifier along with *Ward's* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
1 vs 9	89.5%	0.895	0.934
2 vs 7	95.4%	0.954	0.969
1,9 vs 6	95.7%	0.958	0.988
2,7 vs 8	89.9%	0.899	0.958
2,7,8 vs 3	81.2%	0.813	0.858
2,7,8,3 vs 6,1,9	86.1%	0.858	0.923
2,7,8,3,6,1,9 vs 4	88.5%	0.882	0.911
2,7,8,3,6,1,9,4 vs 5	96.6%	0.966	0.981
2,7,8,3,6,1,9,4,5 vs 10	93.9%	0.939	0.890

Table A.12: Performance of the *LibSVM* classifier along with *Ward's* method for building the tree.

Involved Subtypes	Accuracy	F-measure	AUC
1 vs 9	90.2%	0.902	0.901
2 vs 7	96.1%	0.961	0.953
1,9 vs 6	96.3%	0.962	0.936
2,7 vs 8	91.6%	0.916	0.916
2,7,8 vs 3	81.0%	0.809	0.784
2,7,8,3 vs 6,1,9	88.0%	0.876	0.832
2,7,8,3,6,1,9 vs 4	89.7%	0.893	0.811
2,7,8,3,6,1,9,4 vs 5	97.0%	0.971	0.946
2,7,8,3,6,1,9,4,5 vs 10	95.0%	0.950	0.851

Appendix B

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line 2011 IEEE. 2) In the case of illustrations or tabular material, we require that the copyright line [Year of original publication] IEEE appear prominently with each reprinted figure and/or table. 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior authors approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication] 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line. 3) In placing the thesis on the author's university website, please display the

following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Vita Auctoris

Forough Firoozbakht was born in 1981 in Shiraz, Iran. She graduated from the Azad University of Shiraz in 2009 with a Bachelor of Science degree in Computer Engineering. She joined the University of Windsor's School of Computer Science in September 2013 and earned her Master of Science degree in Computer Science in December 2014.