# Effects of Gene Selection and Data Sampling on Prediction of Breast Cancer Treatments

by

Brian Heredia

A Thesis Submitted to the Faculty of

The College of Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Florida Atlantic University

Boca Raton, Florida

December 2014

**Effects of Gene Selection and Data Sampling on Prediction of Breast Cancer Treatments**

by

Brian Heredia

This thesis was prepared under the direction of the candidate's thesis advisor, Dr. Taghi M. Khoshgoftaar, Department of Computer and Electrical Engineering and Computer Science, and has been approved by the members of his supervisory committee. It was submitted to the faculty of the College of Engineering and Computer Science and was accepted in partial fulfillment of the requirements for the degree of Master of Science.

SUPERVISORY COMMITTEE:

Taghi M. Khoshgoftaar, Ph.D.
Thesis Advisor

Xingquan Zhu, Ph.D.

Bassem Alhalabi, Ph.D.

Nurgun Erdol, Ph.D.
Chair, Department of Computer and Electrical Engineering and Computer Science

Mohammad Ilyas, Ph.D.
Dean, College of Engineering and Computer Science

Deborah L. Floyd, Ed.D.
Interim Dean, Graduate College

11/6/2014
Date

**Acknowledgements**

I would like to take the time here to acknowledge a couple of people. First of all I want to thank my parents, Mr. Fransisco Heredia and Mrs. Rosana Heredia, without their love and support I would have never made it this far.

I would like to give special acknowledgement to my thesis advisor, Dr. Taghi M. Khoshgoftaar. Throughout my tenure as a Master's student he has been there to make the transition from learning to researching as smooth as possible. He inspired me into the field I am researching today and continues to provide support in my current endeavors. I would like to thank both Dr. Zvi S. Roth and Dr. Perambur S. Neelakanta for their help as academic advisors.

I would like to thank Mr. David Dittman for his help with the preparation of my thesis and my survey paper and a thank you as well for all of the members in the Data Mining and Machine Learning Laboratory of Florida Atlantic University.

## Abstract

Author: Brian Heredia

Title: Effects of Gene Selection and Data Sampling on Prediction of Breast Cancer Treatments

Institution: Florida Atlantic University

Thesis Advisor: Dr. Taghi M. Khoshgoftaar

Degree: Master of Science

Year: 2014

In recent years more and more researchers have begun to use data mining and machine learning tools to analyze gene microarray data. In this thesis we have collected a selection of datasets revolving around prediction of patient response in the specific area of breast cancer treatment. The datasets collected in this paper are all obtained from gene chips, which have become the industry standard in measurement of gene expression. In this thesis we will discuss the methods and procedures used in the studies to analyze the datasets and their effects on treatment prediction with a particular interest in the selection of genes for predicting patient response. We will also analyze the datasets on our own in a uniform manner to determine the validity of these datasets in terms of learning potential and provide strategies for future work which explore how to best identify gene signatures.

## Dedication

To my loving parents, Fransisco Heredia and Rosana Heredia, who have made all of this possible and who have supported me through all my decisions.

**Effects of Gene Selection and Data Sampling on Prediction of Breast Cancer Treatments**

## List of Tables

# 1 Introduction

Current efforts in the biomedical field are very focused on the understanding and treatment of cancer. Cancer is a broad term describing many diseases involving an unprecedented and uncontrolled amount of cell growth, usually referred to as a malignant tumor. The complex part of studying cancer is that cancer seems to have no single underlying cause; different types of cancer stem from different sources, even the same type of cancer isn't guaranteed to originate from the same source.

The current treatments for breast cancer include chemotherapy, radiation therapy, gene therapy, targeted therapies, and tumor resection through surgery, sometimes even two or three of these types of treatment at the same time. It's not uncommon to see chemotherapy or radiation therapy used after a surgery to ensure the tumor doesn't grow back; this is known as adjuvant therapy [2]. Chemotherapeutic drugs used to try to eradicate the tumor before surgery is called neoadjuvant chemotherapy. There are different approaches to therapies, for example the “shotgun” approach. The shotgun approach is used when the origin of the cells found in the tissue biopsy cannot be identified. It consists of using one chemotherapeutic drug after another in an attempt to see what works. The idea behind the shotgun approach is slowly becoming replaced by the “Gatling Gun” method. This method involves targeting multiple pathways, or sections within a pathway, at once using targeted therapies [3]. What this shows is that due to the complex nature of breast cancer, and cancer in general, not all neoadjuvant therapies have the same effect, especially since every person differs genetically.

The current standard in the biomedical research field for genomic analysis is called a Gene Microarray [4] (aka gene chips). This technology revolutionized modern research; a small chip that has thousands of wells which can each contain a chemical reaction. These chemical reactions are used to test expression levels of genes in each tissue sample. These microarrays work though using mRNA to measure expression values, the amount of mRNA expressed corresponds to how strongly a gene is being expressed in that well [4]. This is measured through the binding of the cDNA located in the wells to the mRNA from the gene placed into the well. Gene microarrays come with one down side, they produce tons of data. With the increase in data biologists had to reach out and learn the techniques used in data mining to analyze the data, thus the field of bioinformatics was born [5]. The idea behind data mining is to discover the hidden patterns in the data, when applied to the gene microarray datasets this can be used for multiple purposes such as patient response prediction, or tumor classification. The datasets produced from a gene microarray have tons of information but the issue is that not all this information is required or important; to combat this data pre-processing is used. Data pre-processing is the act of removing or ignoring certain attributes, or features, in the dataset for each instance. For example high dimensionality [6] and class imbalance [7] are both factors that will affect the outcome of your model in a negative way, data pre-processing combats both. Data pre-processing is applied and a new dataset is formed, this new dataset is used to build a model. Key decisions are then derived from the results of the model created.

## 1.1 Motivation

The primary goal when using data mining techniques is classification. Classification consists of using the independent attributes for each instance to predict what class the instance falls into, usually in a binary form. Classification is a very vital part of data mining and has many different far reaching applications, for example classification can be used in computer science when looking through code to predict whether a module would be faulty or not, it can be used in biology to determine patient response to a certain drug, and it can be used in finance to determine financial fraud. The power of classification lies in the patterns found in the data, the biggest issue when determining these patterns are class imbalance and high dimensionality.

High dimensionality occurs when there are a very large number of features for each instance in a dataset. The large amount of features leads to irrelevant features and redundant features [35]. Irrelevant features have no influence on the outcome of the classification; these features are useless from a classification standpoint. Redundant features are comprised of the same information found in other features throughout the dataset; these features are not necessary and simply increase computational time. To combat these factors we use a method called feature selection, which culls the irrelevant and redundant features from the dataset. The features selection techniques determine the meaningful features in the dataset and build the model using only those features found to be meaningful. The models are built from smaller datasets with less features but result in a more accurate and efficient classifier [8]. Since these classifiers are built using only the specific features selected the computational cost is also decreased [8].

Feature selection methods are split into two categories, filter-based and wrapper-based. Filter-based feature selection uses only the inherent values of the dataset to determine importance of features [9]. Normally a score is calculated for each feature and low-scoring features are removed from the dataset while high-scoring features are retained. A subset size is determined before running the feature selection technique and the top numbers of features equal to that subset size are used for creating the model. The advantages to filter-based feature selection include low computational cost and no requirement of a classifier. The disadvantages of filter-based feature selections are that they do not take into account the features that are dependent on other features, or the classifier interaction. Wrapper-based feature selection uses a classifier to test a subset of features. Wrapper-based feature selection techniques are very computationally intensive as they go through the features using different subsets and create a model for each subset [9]. In a high dimensional dataset the wrapper-based feature selection techniques are extremely costly. These feature selection techniques normally provide the optimum subset sizes and features but sacrifice the speed of filter-based feature selection.

Class imbalance occurs when instances in one class greatly outnumber instances in the other classes. Class imbalance becomes an issue because most classifiers are built using accuracy as a performance metric. Accuracy as a performance metric with imbalanced data usually leads to a higher classification rate towards the majority class. For example if a simple binary classification dataset consists of 90% negative instances and 10% positive instances the classifier will have a 90% accuracy rate by classifying everything as negative. However, in most studies the positive class is the class of interest and this misclassification of positive instances (known as false negative) makes the

problem worse. To counteract this class imbalance we use data sampling techniques [7]. Data sampling techniques attempt to overcome the class imbalance issue by either adding or removing instances from the dataset. The models will be built using only the data selected after the data sampling techniques have been applied, which will be a more balanced dataset.

### 1.2 Contributions

In 2013 fifteen datasets (including six breast cancer datasets) from works involving the use of microarrays were gathered and analyzed [1]. The fifteen datasets were thoroughly examined and the results made available to future researchers, the study went into detail about the original analysis of the datasets and provided their own analysis. Of the fifteen datasets six of them were breast cancer. Since then research into whether or not a specific cancer will respond to a specific treatment is still scarce but we have found three new works [10, 12, 22] which use five breast cancer microarray datasets. Our case study consists of five of the eleven total breast cancer datasets found, a combination of the new datasets and the ones analyzed in the study. These five datasets all suffer from high dimensionality, while three of them also suffer from class imbalance. This thesis will focus on the effects of feature ranking techniques and class imbalance techniques on the learning ability of these datasets. The contributions of this thesis will be:

1. A comparison of five breast cancer patient response datasets run through four classifiers using no feature selection, using feature selection only, and then using feature selection and data sampling.

2. The effect of subset sizes on feature selection and their effect on feature selection plus data sampling.
3. The effectiveness of feature selection and data sampling on high dimensional, class imbalanced breast cancer patient response prediction datasets.

### 1.3 Organization

This thesis will be ordered in the following chapters:

- Chapter 2 presents related works and background on the datasets obtained for this study.
- Chapter 3 provides the information on the feature ranking techniques, the data sampling techniques, and the classifiers used in this study. The information in this chapter is used in chapter 4.
- Chapter 4 investigates the results of analyzing the datasets with no feature selection, with only feature selection, and with feature selection and data sampling. This chapter also contains the comparison of the three methods.
- Chapter 5 contains the conclusions and future works.

# 2 Related Works

As this thesis falls under an interdisciplinary field some background on the original biological studies and certain bioinformatics studies is required.

## 2.1 Dataset Origins

The eleven breast cancer datasets mentioned in this thesis come from seven different studies; two of the datasets come from the same study but from different groups of patients and this happens in two of the studies. All of these studies focus on patient response prediction to breast cancer treatments. Patient response datasets usually involve an instance (the patient) which is labeled on whether or not they responded to a treatment (positive or negative). These studies are a good example of the different approaches data mining and machine learning have in the domain of bioinformatics. Table 2.1 below shows the different treatment options described in this section, the original study of which these treatments were done, the datasets provided by those studies, and finally whether or not those datasets were used in this thesis. Of the datasets we selected five total datasets from three studies, the two stemming from the Ma et al. study, the two from the Thuerigen et al. study, and one of the Haury et al. study, GSE1456 which shall be referred to as Haury, to use in our experiments.

Table 2.1: Data Information for the Related Studies

| Treatment Type | Source Study | Datasets | Used in this thesis |
|---|---|---|---|
| Neoadjuvant | Hallet et al. [10] | GSE25055 [10], GSE25065 [10] | No |
| FAC and T/FAC | Tabchy et al. [12] | GSE20271 [12] | No |
| Tamoxifen | Ma et al. [13] | Microdissections [13], Whole tissue [13] | Yes<br>Yes |
| | Chanrion et al. [15] | chanrion2008-majorityPositive[15] | No |
| | Pawitan et al. [20] | Pawitan et al. [20] | No |
| Docetaxel, Epirubicin, and Gemcitabine | Thuerigen et al. [21] | thuerigen2006Cy3[21], thuerigen2006Cy5[21] | Yes |
| Relapse | Haury et al. [22] | GSE1456 [22], GSE2034 [22] | Yes<br>No |

### 2.1.1 Neoadjuvant Chemotherapeutic

In Hallett et al. [10] two datasets were generated using a target based approach to identify genomic predictors of breast cancer response to chemotherapy. The instances in these datasets were classified as pathological complete response (pCR) or residual disease (RD); positive and negative respectively. In the first dataset from this paper, GSE 25055 [10], there were a total of 306 instances with 249 negative (RD) instances and 57 positive (pCR) instances. The second dataset, GSE 25065 [10], has a total of 182 instances with 140 negative and 42 positive instances. Both these datasets are imbalanced and suffer from high dimensionality.

The study focuses on the genes TOP2A and β-tubulin and how the expression levels of these genes determine response to chemotherapeutic agents. They found that there is a positive correlation between expression of TOP2A and β-tubulin; they actually found that the drug trastuzumab had a 10% success rate with patients who didn't have the two genes of interest expressed and a rate as high as 50% for those patients who showed gene expression in the two genes of interest [10]. This discovery led to the testing of neoadjuvant chemotherapy to see if it yielded the same results. Neoadjuvant chemotherapy uses multiple chemotherapeutic remedies instead of using a single drug. The neoadjuvant chemotherapy used in the study was a mix of anthracycline and taxane (AT). The data suggests that TOP2A expression is linked to any kind of anthracycline treatment [10]. β-tubulin was also tested for an association between expression level and treatment response. They found that β-tubulin expression levels were positively correlated with a positive response to the chemotherapeutic drug docetaxel.

The results indicate a connection between TOP2A and anthracycline based chemotherapeutic drugs and a connection between β-tubulin and taxane based drugs [10]. Once these patterns were identified they were then compared to the DLDA30 predictor. This DLDA30 predictor is the standard 30 gene expression signature used to test for patient response to a certain chemotherapeutic drug [10]. When looking for a TOP2A index in the DLDA30, using anathracycline based drugs; they only found that one of the 30 genes contained the same TOP2A probe. The same was done for β-tubulin using a taxane drug and they found 42 β-tubulin associated probe sets, 28 of these probes showed a positive response to β-tubulin levels while 14 of them showed a negative response. Using a new dataset the researchers tested the association between β-tubulin and

docetaxel. This new dataset contained 14 patients, all treated with docetaxel. The patients who achieved complete pathological response had a higher expression of β-tubulin. Overall the findings of this research can lead to a different approach in looking at chemotherapeutic responses. The authors found the β-tubulin prediction was relatively accurate but the TOP2A prediction method was not as significant.

The identification of target related genes was done based on a Pearson distance function [11]. The results were analyzed using the Receiver Operator Characteristic curve (ROC) analysis. A univariate logistic regression analysis was done to compare the clinical factors with the combination index score of β-tubulin and TOP2A. The clinical factors that were compared with the index scores were ER status, tumor grade, nodal status, and patient age. Only ER status (AUC = .68, $p < .001$), tumor grade (AUC = .69, $p < .001$), and the combination index (AUC .76, $p < .001$) were shown to predict a pCR or RD case. A multivariate logistic model was also created and the combination index remained significant to prediction with a p value of $< .001$; the two clinical factors that were also shown to predict pCR and RD were also significant (ER status: $p = .014$ and tumor grade: $p = .016$). The odds ratio for the combination index for the multivariate model was 1.33; meaning for each single unit increase in the index there was a 1.33 increase in chances of pCR. Welch's correction [32] was used when the variance was unequal in the two groups, responder's vs.. non-responders. ANOVA [31] and Tukey's multiple comparison tests were used to test for differences, a p value of .05 was used as the threshold of significance.

### 2.1.2 FAC and T/FAC Treatment

In Tabchy et al. [12] datasets were used to evaluate the 30 gene prediction panel using the chemotherapeutic drugs paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide. The classification system used in the datasets was pathological complete response, meaning the cancer went into remission, or residual disease, meaning the cancer is still there. The data had two separate treatments mixed into one dataset so we split the two sets of instances based on the treatment received. The first set was using a Fluorouracil (F), Doxorubicin (A), and Cyclophosphamide (C) treatment (FAC) while the second set was using Paclitaxel (T) in addition to the FAC (T/FAC). The first set had 7 positive (pCR) and 80 negative (RD) instances while the second set had 19 positive (pCR) and 72 negative (RD) instances. These datasets suffer from class imbalance and high dimensionality.

The treatments were given weekly, paclitaxel (80 mg/wk), followed by 5-fluorouracil (500 mg), doxorubicin (50 mg), and cyclophosphamide (500 mg) all on day 1 repeated in 21-day cycles for the T/FAC group. For the FAC group the same drugs as above were taken with the exception of paclitaxel at the same doses and schedule. Expression results for the 30 gene signature were entered into a class prediction algorithm [12], each instance was assigned a classification of either pCR or RD before the actual result was determined. Using the chemotherapy response calculator located on www.mdanderson.org/pcr a nomogram was created. A nomogram takes the basic information; patient age, tumor size, ET status, etc, and predict the probability of pCR after T/FAC therapies. Using the assumptions that the prevalence of DLDA-30 [33] marker positive patients is 30% [12], pCR rate to T/FAC treatment in the marker positive

group is >60% and pCR rate to FAC treatment is between 20% and 40%, repeated fitting of a logistic regression model was done with 10,000 iterations, using pCR as the dependent variable. A multivariate logistic regression model was also created to calculate odds ratio for pCR.

The cohort of patients consisted of 138 T/FAC and 135 FAC patients, all randomly chosen, although 20 of the T/FAC and 16 of the FAC patients were excluded from the analysis due to eligibility violations. Eleven of the patients in the FAC trial were actually given T/FAC to maximize response. The pCR rate for the T/FAC group was 19% (24 out of 129) while the rate in the FAC group was significantly lower at 9% (10 out of 113). The DLDA-30 genomic predictor had a positive prediction value (PPV) of 38%, a negative prediction value (NPV) of 88% and a sensitivity and specificity of 63% and 72% in the T/FAC arm. The AUC was .711 meaning the data wasn't that hard to learn from. In the FAC arm the PPV was 9% and the NPV was 92%, the specificity and sensitivity were 75% and 29% respectively. The AUC was 0.584, significantly lower than the T/FAC arm. In the multivariate logistic regression model the only independent predictors of pCR were ER status (P = 0.008), tumor size (P=0.018), and treatment arm (P = 0.022). There is a significant difference in pCR rate from the patients who received T/FAC and those who received FAC chemotherapy, with T/FAC neoadjuvant chemotherapy being the more successful.

Three class prediction algorithms were used in this study, Support Vector Machines (SVM) with linear, radial, and polynomial kernels, Diagonal Linear Discriminant Analysis (DLDA), and K-nearest neighbor (k-NN) using Euclidean distance. Cross validation [37] , or CV, was also used in this study. K-Nearest Neighbor

was varied using k = 3, 5, 7, 9, 11, 13,15,17,19, 21, 23, 25, 27, 29, 31, 33. The informative variables for the DLDA-30 gene signatures were found by using a logistic regression algorithm on the training set. A total of nine different training sample sizes were used, these ranged from 20 to 100 in increments of ten, no specification of what feature selection algorithm was used. In the multivariate logistic regression model created only ER status, treatment arm, and tumor size were significant independent predictors of pCR.

The data found in this study showed that the 30 gene signature used was in fact predictive to response in the T/FAC arm and not as strong a predictor in the FAC arm. One of the important things to take away from this study is that not all breast cancer treatments will be the same for different kinds of breast cancers. ER-positive and ER-negative breast cancers have large gene expression differences and cannot be considered to have the same response to a gene prediction signature. The DLDA-30 gene signature was created using both ER-positive and ER-negative breast cancer samples. Thus the gene signature contains gene expressions that are associated with both phenotypes. The factors tested in the nomogram; ER status, histological grade, basal-like vs. luminal B, were all used to predict general chemosensitivity.

### 2.1.3 Tamoxifen Studies

The following three studies all focus on a chemotherapeutic drug called tomaxifen. Tomoxifen is an antiestrogen agent prescribed for women in early-stage and metastatic hormone receptor-positive breast cancer [13]. The first paper by Ma et al. focuses on a two-gene expression ratio that is used to predict clinical outcome for breast cancer patients using the chemotherapeutic drug tamoxifen [13]. Breast cancer can be

classified based off two important factors; their estrogen receptor (ER) and their progestrogen receptor (PR) classification. These can be classified as ER+/PR+, ER+/PR-, or ER-/PR+, this study examines tamoxifen in ER+ breast cancer patients. Approximately 40% of ER+ cases fail to respond positively to the tamoxifen and eventually develop a tamoxifen resistance [13]. Before this study there were limited ways to predict clinical response to tamoxifen, and of the few ways available none were accurate enough to be considered useful. This study prosvides a two-gene ratio, HOXB13:IL17BR, which accurately predicts patient response to tamoxifen.

The study uses data from Massachusetts General Hospital where 103 ER+ women were administered adjuvant tamoxifen. These 103 patients had tumor samples snap-frozen and out of these 103 patients 60 were chosen for this study. Cohort 1 had 28 women who developed distant metastasis and 32 women who had not seen nonoccurence of the tumor after a follow up of 10 years. Gene expression profiling was performed using a gene microarray analysis; this profiling used a total of 22,000 genes of which to test the sample against. RNA was extracted from the snap-frozen biopsies and used for the gene expression profiling. The resulting dataset was then filtered based on overall variance, a total of 5,475 high variance genes were chosen for future analysis. A t-test was then performed on the reduced dataset. A cutoff value of $p = .001$ was used and 19 total differentially expressed genes were chosen based off the cutoff value. The same cohort was then reanalyzed using laser-capture microdissection (LCM) of the tumor cells to further validate the results of the first analysis. Variance testing and t-tests were performed on the LCM dataset in the same form as they were in the whole-tissue dataset. This was done to rule out contamination due to stromal cells and to further confirm the

information they obtained during the first whole-tissue data analysis. After the tests were done there were a total of 9 differentially expressed genes that met the p = .001 cutoff value. Comparing the 9 genes from the LCM and the 19 genes from the whole tissue analysis they found three genes that were differentially expressed in both datasets; the homeobox gene HOXB13, the interleuken 17B receptor IL17BR, and EST Al240933. The HOXB13 gene was found to be overexpressed in the tumor recurrence cases while the Al240933 and IL17BR genes were found to be overexpressed in the nonrecurrence cases.

Using the Area Under the ROC Curve (AUC) [14] as a performance metric the three genes were analyzed in terms of specificity and sensitivity as a biomarker of clinical outcome. The AUC value stands for the area under the receiver operator characteristic (ROC) curve. This value indicates how difficult it is to learn from that specific dataset. The values range from 0 to 1.00 and the higher the number the easier the dataset is to learn from. For the data obtained from the whole-tissue dataset the AUC for HOXB13, Al240933, and IL17BR were .67, .81, and .79 respectively. For the data obtained from the LCM tissues the AUC for HOXB13, Al240933, and IL17BR were .8, .76, and .76 respectively. A null model was created to test the significance of the AUC values; the null model, which predicted clinical outcomes at random, had an AUC of .5. From that result they concluded that the three genes were indeed significant and had potential for predicting clinical outcomes. ROC analysis was also done on the signal pathways EGFR and ERBB, which are associated with negative regulation of estrogen-dependent signaling. The AUC values for the genes ranged from .59 to .69. From the AUC values

they found that the HOXB13, IL17BR, and Al240933 are better predictors of tomaxifen response than the EGFR and ERBB pathways.

Due to the nature of the HOXB13 and the IL17BR genes being overexpressed in an opposite manner from each other the researchers thought that a ratio between the two genes would work as a predictor for tamoxifen response. Both t-tests and ROC analysis were performed on the ratio between the two genes and found that it does indeed work as a predictor for tamoxifen response. The AUC value for the ratio in the whole-tissue dataset was .81 and the value for the LCM dataset was .84. Further testing was done to see whether or not HOXB13:Al240933 ratio or all three genes together had a better prediction value; these tests all came up with a lower AUC value than the HOXB13:IL17BR ratio.

An independent testing of the ratio was performed on cohort 2. Cohort 2 consisted of 10 women who had reoccurring disease and 10 women who had remained disease free for a median follow up of 9 years. These tests were performed on formalin-fixed and paraffin-embedded (FFPE) specimens; these specimens are used in routine clinical testing and were chosen to see the effect of the HOXB13:IL17BR ratio in clinical tests. RNA was extracted from the FFPE tissue sections and then amplified using linear amplification and placed through RT-QPCR analysis. In agreement with the information they found in cohort one the expression ratio was highly correlated with clinical outcome; in this case there was a higher HOXB13 expression which was associated with poor outcomes. To test the prediction of the ratio RT-QTPCR was used from the frozen tissue samples, the patients in cohort 1, and the data was then used to create a logistic regression model. The overall accuracy for the model was 81%, while the positive and negative prediction rates

were 81% and 82% respectively. As to not be biased by using the dataset which they built the model to evaluate the model they also decided to use cohort 2, the FFPE group to test the model. The model predicted 16 out of the 20 patients correctly and had an accuracy rate of 80%. The positive and negative rates were 87% and 75% respectively, and the probability of the model predicting 16 out of 20 of the cases correctly was .01. The authors believe this two gene ratio is an ideal method of measure treatment response due to the ease of clinical application and the cheap and quick laboratory techniques.

The second study revolving around tamoxifen is the paper by Chanrion et al. [15] this paper focuses on a gene expression signature that predicts the recurrence of tamoxifen-treated breast cancer. The study addresses multiple studies where a gene expression signature is created and proposes a new 36-gene molecular signature that is highly predictive of clinical outcomes. The data was obtained from 155 patients at the Cancer Research Center of Val d'Aurelle in Montpellier, the Bergone Institute in Bordeaux, or the department of Obstetrics and Gynecology of Turin. A total of 132 ER+ and/or PR+ patients treated with adjuvant tamoxifen formed the training set and the 23 extra tumors were used as a validation set to test the models that will be built using the training set. Fresh tissues were formalin-fixed and paraffin-embedded (FFPE) while the remainder of the tumor was snap-frozen and stored at -80 degrees Celsius [15]. Using radioligand binding assays or immunohistochemistry all 155 patients were tested for ER+/PR+; only eight of the patients came up as ER-, with six of those eight being PR+. All patients were treated with tamoxifen at a dose of 20 mg daily for a 5 year period after surgery. Out of all the patients, one hundred and twenty-one patients also received adjuvant radiotherapy. Recurrence was found in 52 patients, with 48 of them being

distant metastasis and 4 of them being local recurrences [15]. Tumors from the recurrence patients were denoted as R while tumors from the patients who had no recurrence were labeled RF. Gene expression profiling was done using oligonucleotide microarrays; a total of 21,398 specific genes were used in the profiling. Due to the high dimensionality of the microarray datasets genes which had lower than 2-fold the mean expression of negative control spots in at least 40% of the samples were thrown out [15]. These genes were then further filtered by selecting 5,415 genes which varied by at least 3-fold from the median in at least 1% of the samples. A thousand runs of the Significance analysis of microarrays (SAM) [16] was then performed to identify genes whose expression levels best classified patients as R or RF. After running SAM 301 genes showed differences in the expression levels. From the 301 gene set the top 48 most discriminating genes were chosen for further analysis. Those 48 genes consisted of 17 that were overexpressed (positive SAM score) and 31 that were underexpressed (negative SAM score).

Using the 5,415 gene set, Prediction Analysis of Microarrrays (PAM) [17] was done and a classifier that predicts recurrence in tamoxifen treatments was created using the training set. Resampling was done on the training dataset to create two separate datasets; a learning dataset consisting of 85 patients and a testing dataset consisting of 47 patients. The learning set was then used to build a model and the testing set to evaluate that model. This resampling process was done 100 times and its performance was evaluated by the average proportion of misclassification for each associated test set. To determine the gene signature different signature lengths were used. They found that when changing the gene signature length from 26 to 36 there was an error decrease from 41% to 26%. This value remained steady from a signature size of 36-71. The 36 genes were

selected from the 100 PAM iterative signatures. Using the 36 gene signature the researchers found that the model classified the training set tumors with 80% sensitivity, 78% specificity, and 79% accuracy [18].

This process was redone using the k-Nearest Neighbor classifier (k-NN) [19] to confirm the signature created by PAM. A predictive signature was created with each run of the k-NN classifier, the optimum consensus signature was found to be a 52 gene signature found in >47% of the 100 k-NN signatures. Using this 52 gene signature the training set was classified with 83% sensitivity, 74% specificity, and 80% accuracy. Out of the 52 gene signature a total of 29 genes were also found in the PAM signature. Hierarchical clustering was then performed using the 36 gene signature and the training set. The clustering showed two main clusters, one for relapse, containing 34 out of the 46 R instances and the other cluster being for the RF instances, consisting of 71 out of the 86 RF tumors. The 36 gene signature was then validated by the 23 independent tumors from the original 155 patient dataset. Of this test dataset 6 of the patients were classified as relapse (R) and 17 of the patients were considered relapse free (RF). Using the 36 gene signature the model predicted 4 of the 6 R patients and fourteen of the 17 RF patients; a sensitivity rate of 82%, a specificity rate of 62%, and an accuracy rate of 78%. Overall the 36-gene signature was found to be the most statistically-significant factor, and the authors are very confident in the clinical relevance of this signature when identifying patient response to tamoxifen.

The third study used tamoxifen and/or goserelin as treatment options. This study, conducted by Pawitan et al. [20] used tamoxifen and/or goserelin patient response information to develop a 64 gene signature used to predict patient outcomes. In addition

to the tamoxifen/goserelin treaments, a combination of cyclophosphamide, nethotrexate, and 5-fluorouracil (CMF) was used to reduce the incidence of recurrence.The dataset used was obtained from the Karolinska Hospital; it was composed of 524 breast cancer patients who had been operated on. Of these 524 patients a total of 159 were selected for analysis. The remaining groups of patients were not analyzed due to either the size of the tumor or the amount of affected lymph nodes. These 159 patients compose the training dataset, while three separate independent datasets were used as validation. Of the 159 patients, 126 received adjuvant therapy while only 104 received a form of tamoxifen treatment. Out of the 159 total patients, 40 had a relapse (R) while the remaining 119 did not (RF). The first of the three validation sets contained 76 instances all derived from treated Swedish patients, the second held 135 untreated Swedish patients, while the last had 78 Dutch patients.

For the primary analysis of the dataset the binary classification model was created using good prognosis and poor prognosis. Poor prognosis was defined as the reoccurance of the disease due to breast cancer.. The gene expression analysis measured expression levels of a total 44,792 genes. The global mean method and a filter method were applied to remove useless genes; the first being removing the genes found in less than 10% of the total patients. This left them with 25,728 genes, which were then filtered down to 6,573 genes based off variability. A ranked list was created of these 6,573 genes using a two sample t-test value. Using multiple models built using diagonal linear discriminant analysis (DLDA) and leave-one-out cross validation the optimal set of predictors was chosen. Gene sizes ranging from 20 to 100 were tested, the optimum gene signature size was found to be 64. Using a 64 gene signature the results show an accuracy rate of 66%, a

sensitivity rate of 74%, and a specificity rate of 64% for the primary analysis data. The process was repeated using the secondary analysis data the model, this model performed with an accuracy rate of 69%, a sensitivity of 84% and a specificity of 65%. Unsupervised hierarchical clustering of the training set (Uppsala) was used to determine the risk groups; using Euclidean distance with complete linkage as a similarity measure. For the validation set (van't Veers) supervised hierarchical clustering was done with assignments based on samples near cluster with the closest centroid. This process yielded three main clusters; each of which was found to represent a specific risk group, low, medium, or high. The high risk group was associated with more poor outcomes than the other two groups; this was validated by the other cancer studies used to test the models. The authors believe this method to be used to identify patients who are at great risk of relapse and need additional therapies. Most patients in the low-risk group did not need any further therapy after the surgical treatment.

### 2.1.4 Docetaxel, Epirubicin, and Gemcitabine

The following breast cancer study did not use tamoxifen in their experiment but instead used a primary systemic therapy including the drugs gemcitabine, docetaxel, and epirubicin. The study conducted by Thuerigen et al. [21] used two therapy regimens; either gemcitabine and epirubicin for 10 weeks followed by docetaxel for 8 weeks or all three drugs for a total of 18 weeks. These two groups were then split into two datasets; the first group became the training set while the second group, all three drugs, became the test set. Gene expression analysis was then done using gene microarray chips. The RNA was extracted from tumor biopsies done 4 weeks after the completion of the drug regimen. A total of 21,329 gene probes were used in the gene expression analysis. Patient

response in this study was classified based off the amount of residual tumor located in the surgically-removed tissue; pathological complete response (pCR, no tumor residuals) and non-pCR were the two classes. According to the paper 100 patients comprise the dataset used although the published dataset online contains only 90 instances; 23 of them in the pCR class and 67 in the non-pCR class [21].

To create a gene signature for prediction of patient response two things were looked at. First, every probe was looked at in terms of its experimental quality. Any probes found in at least 80% of the instances were removed from all instances. A support vector machine (SVM) model was then built using recursive feature elimination (RFE) and cross validation to find the optimum gene signature; in this case that signature was 512 genes. When using the SVM model with the 512 gene signature an accuracy rate of 88% was obtained, with sensitivity and specificity scores of 78% and 90% respectively. After statistical significance testing it was determined that the gene signature is indeed an effective predictor of the pCR class. The authors believe the results of this study show an accurate correlation between the gene signature and the drug regimen of gemcitabine, epirubicin, and docetaxil when looking at complete pathological response in patients.

### 2.1.5 Relapse Study

The final study involving breast cancer was conducted by Haury et al. [22] this study had more of a machine learning view than a biological view to the datasets. The purpose of the study was to examine 32 feature selection techniques in terms of predictive performance on 4 breast cancer datasets. Due to redundancy issues only two of the four datasets were used in this study. The two datasets were for prediction of metastatic relapse in breast cancer on different cohorts. The first of the two had 159

instances, 40 of which were positive for relapse and the remainder which were negative (119 instances). The second dataset had 286 instances, 107 of which were positive and 179 which were negative instances. The methods used were the following: four filter techniques used the student's t-test, Wilcoxon sum-rank test, relative entropy, and the Bhattacharyya distance. For wrappers they used SVM recursive feature elimination (RFE and, greedy forward selection (GFS). The final two feature selection methods were embedded; which perform feature selection in the process of training the data. Lasso regression and elastic net were the two feature selection techniques used in this category.

Three ensemble aggregation techniques were also applied to "stabilize" the results. The aggregation methods were done for each feature selection technique. First they bootstrapped the training sample B = 50 to get ranking of all features by applying the method to each sample. Ensemble mean, ensemble –stability selection, and ensemble-exponential were the three ensemble methods applied to the data. Five total classification algorithms were tested on the data to measure accuracy; nearest centroid (NC), k-nearest neighbor (k-NN, k = 9), linear SVM, linear discriminant analysis, and Naïve bayes. AUC was used as the performance metric in two settings, one where 10-fold cross validation is used and the AUC is applied to the 10% used as a test.

The second setting is where they estimated the signature on one dataset and assess its accuracy on other datasets using 10-fold CV. Stability of the feature selection methods was also analyzed by comparing the signatures on different samples in various settings. Soft-perturbation, hard-perturbation, and between-dataset setting were used to test stability. Soft-perturbation is the act of sampling each dataset into pairs of subsets with 80% overlap. Estimating the gene signature on each subset and finding the overlap

between two signatures in a pair the fraction of shared genes. The random sampling of subsets is repeated 20 times on each subset and the values over all samples are averaged. Strong perturbation is the same method but with no overlap between two subsets of samples. The final process is between-datasets setting, which estimates signatures on each dataset independently using all the samples in that dataset.

Taking random feature selection as a baseline, the only feature selection techniques to perform better than the baseline, in accuracy was the students T-test, Lasso, and elastic net but the only one to outperform random feature selection in both accuracy and stability was the student t-test. Both Bhattacharyya distance and relative entropy were more stable but less accurate than random feature selection; which was not expected. Of the wrapper methods used, although more computationally heavy, none of them perform well when compared to the random feature selection method. The study also showed that when training a classifier the most accurate classifier was the nearest centroid; although simple it was the most effective. The authors found that when using a soft-perturbation approach the result may be misleading, the most effective method was using a hard-perturbation setting.

### 2.2 Bioinformatics Studies

As mentioned earlier gene microarray datasets consist of thousands of features and instances. These eleven datasets we found mostly contain the same class imbalance and high dimensionality issues. The following studies relate to methods used to combat class imbalance and high dimensionality.

### 2.2.1 Class Imbalance

Class imbalance occurs in most bioinformatics cancer datasets due to the nature of biological diseases. The minority class is usually the class that reacts well with the treatment which ends up being the class of interest. This causes the majority instances to outnumber the minority instances, leading to a model that is ineffective at classifying instances from the minority class. Due to class imbalance the normal classification methods produce a model that tends to favor classification of the majority class. The study in 2010 by Seiffert et al. titled "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance" elaborates on some techniques that can be used to combat class imbalance.

The study begins by differentiating the two techniques, data sampling and boosting. Data sampling uses oversampling, adding to the minority class, or undersampling, removing from the majority class, to balance out the class distribution. Oversampling and undersampling can both be done randomly or "intelligently" [7]. For the purpose of this study, and this thesis, we will only consider the random methods. Random undersampling (RUS) and random oversampling (ROS) both have their advantages and disadvantages. In ROS we add data to the minority class to even it out with the majority class; this additional data is created by taking the original minority class and randomly duplicating values. In RUS the data in the majority class is randomly removed to match closer the data in the minority class. Boosting is a technique that can amplify the classification ability of any weak classifier. Boosting doesn't alter the data in any way, leaving it imbalanced. The boosting method commonly used is called AdaBoost. Seiffert's work combined RUS and boosting into a hybrid algorithm,

RUSBoost, to combat class imbalance and tested it against an already known hybrid method SMOTEBoost [23].

The RUSBoost method is used in this thesis as a method to combat class imbalance partially due to the results found in this study. The study used four different performance metrics to evaluate the performance of RUSBoost and SMOTEBoost. These four metrics were the Receiver Operating Characteristic (ROC) curves, the Precision-Recall Characteristic (PRC) curves, the Kolmogorov-Smirnov (K-S) statistic, and the F-Measure. The ROC curve is a very common performance metric used in evaluating the learners. It graphs the true positive rate versus the false positive rate to show the performance of a classifier. The area under the ROC curve (AUC) is a single valued metric that quantifies the models performance and the ability to learn from the dataset. The PRC curve plots the recall, which is the true positive rate, against the precision. The precision is the amount of total correct positive predictions over total positive predictions. The K-S statistic calculates the difference between the empirical distribution function of the sample and the distance function of the reference distribution [7]. The distance determines how good the learner is at separating the two classes, the larger the distance the better the learner. The final performance metric is the F-measure. The F-measure uses a threshold based decision method that is derived from the recall and precision. These performance metrics were tested against four different classifiers; the C4.5D decision tree, the C4.5N decision tree, Naïve Bayes, and the RIPPER classifier.

The results show that the RUSBoost and SMOTEBoost both perform significantly better than the AdaBoost method, the general boosting method. The results show no significant difference between RUS and SMOTE in terms of the performance metric

AUC. SMOTE does significantly outperform RUS using the other three performance metrics. When combining RUS with boost the RUSBoost method is very similar to the SMOTEBoost method, with no significant difference between the performances. RUSBoost ends up being a better method due to the fact that it is an undersampling method. This allows for less computational time and a simpler technique.

### 2.2.2 High Dimensionality

In his study in 2009 [6] titled "Capturing best practice for microarray gene expression data analysis", Gregory Piatetsky-Shapiro delved into the realm of high dimensionality. Microarray datasets present a challenge to current standard data mining and machine learning techniques because of their high dimensionality. Not only does high dimensionality increase the computational cost but it also leads to various "false positives". A false positive in the domain of cancer treatment can be very detrimental to the patient. In the study two methods are described to deal with high dimensionality; feature selection and randomization. The randomization technique permutes the features many times and compares the strength of correlation with the original features [6]. Feature selection, as mentioned earlier in this paper, consists of eliminating useless or redundant features from the dataset before running a classifier.

Microarray datasets usually consist of a large number of features and a low number of instances. This leads to a problem when separating the instances into a training set and a validation set. This problem is addressed by using cross-validation. Cross-validation requires no supervised splitting of the data into training and validation sets [6].

Data pre-processing techniques are usually applied to microarray datasets in steps. The first step is usually a low-level form of data pre-processing. This includes thresholds, normalization, and filtering; these methods are all independent of the class. Thresholds are set and if the value for the features does not meet that threshold then they will not be included. Normalization simply helps make the data easier to read across all features; the usual method is by making the mean zero and the standard deviation between values one. Filtering is usually done with respect to variability across the samples. If a sample has a very low variation, meaning it is the same across most samples, then only one sample is required and the others can be removed. Once those steps are taken feature selection is finally applied to the dataset in order to tackle high dimensionality.

# 3 Methodology

The experiments done in this thesis use the methodology described in this section. This section will describe the feature ranking techniques, data sampling techniques, and the classifiers used in the experiments.

## 3.1 Dataset Characteristics

In our experiments we used a total of five breast cancer patient response datasets. These datasets stem from the studies in the related works section, we used both the datasets from Ma et al., both the datasets from Thuerigen et al., and one of the datasets from Haury et al. Table 3.1 contains some basic information on the five datasets. All five of the datasets suffer from high dimensionality, while only three of the five suffer from class imbalance. The two Ma et al. datasets do not have class imbalance since there is almost an equal amount of instances in the minority class when compared to the majority class. The number of features and number of minority/majority instances provide some insight into the size of these datasets.

Table 3.1: Attributes of the Five Breast Cancer Treatment Response Datasets

| Name | # Minority Instances | # Majority Instances | Total # of Instances | % Minority Instances | # Of Features |
|---|---|---|---|---|---|
| Thuerigen 2006 (Cy3/Cy5) [21] | 23 | 73 | 96 | 23.96% | 21881 |
| Thuerigen 2006 (Cy5/Cy3) [21] | 24 | 70 | 94 | 25.53% | 21881 |
| Ma 2004 (Microdissections) [13] | 28 | 32 | 60 | 46.67% | 22576 |
| Ma 2004 (Whole tumor) [13] | 28 | 32 | 60 | 46.67% | 22576 |
| Haury [22] | 40 | 119 | 159 | 25.16% | 12066 |

### 3.2 Classifiers

The four classifiers used in this thesis are commonly used in the gene microarray analysis field. These four classifiers are Naïve Bayes [27], Multilayer Perceptron (MLP) [28], Support Vector Machines (SVM) [29], and k-Nearest Neighbor (k-NN) [30]. These four classifiers were all implemented using the Weka software [24, 36]. These four classifiers were selected to test different classifying methods on the same data to get a more robust result. All methods were done under supervised learning. Supervised learning trains the classifier on a training set and then feeds the model unknown values and lets it classify these values based on the classifiers learning. All these classifiers were run using cross validation.

Cross validation is a method used to split the data into *n* folds and uses *n*-1 of those folds as the training set and the remaining fold as a test set. In the case of our experiment we used four runs of five-fold cross validation. This means we used a five-fold split, taking the dataset and splitting it into five folds, using four of them to train the data and then using the final one to test the model. Cross validation can be biased because the data is split randomly so there may be a lucky or unlucky split. To counteract this we ran four runs of this five-fold cross validation to eliminate the bias due to chance of a lucky, or unlucky, split.

The first of the four classifiers, Naïve Bayes, is a very simple learner. The Naïve Bayes classifier is based off probability; more specifically the prior probability that an instance is a member of a specific class and the likelihood of a value with respect to its surroundings. The prior probability is calculated for each class by taking the total amount of instances in a class and dividing it by the total amount of instances. The likelihood for each class is obtained by previously selecting a range of values around the new unknown and dividing the amount instances stemming from one class in that range by the total amount of instances of that same class. The final step is to find the posterior probability, which is the ratio of the prior probability multiplied by the likelihood over the evidence:

$$p(C|F_{1,\dots,}F_n) = \frac{p(C)p(F_{1,\dots,}F_n|C)}{p(F_{1,\dots,}F_n)}$$

The evidence is considered constant for the specific datasets so it can be ignored for the purpose of classification. Thus the final formula is shown below:

$$p(C|F_{1,\dots,}F_n) = p(C)\prod_n p(F_n|C)$$

Where $\prod_n p(F_n|C)$ is the simplified version of the likelihood formula.

The second of the classifiers mentioned is the Multilayer Perceptron. The MLP is a type of artificial neural network. Neural networks are designed to attempt to replicate how the brain works. In the case of the MLP this is done by have intermediate steps, called layers, that are used to combine basic attributes into higher-level concepts. These layers are composed of nodes and each node in a layer has a connection coming from every node in the layer before it and also has multiple connections stemming to every node in the layer after it. This provides a non-linearity to this classifier, making it more flexible when learning. Each node receives the weighted sum of the values of the previous nodes in the layer before it. This weighted sum is not the only information the node holds, an activation function is applied to the node once the weighted sum is received. This activation function results in a more clearly defined result. This is done because this activation function will separate the instances even more so than they already are. Although these neural networks seem to be robust and very redundant take caution because they are very prone to overfitting [25]. Overfitting occurs when a model created from a classifier begins memorizing the values in the training set instead of learning the actual trend between the data. This leads to weak predictive models that cannot classify unknown data because it doesn't understand the trend.

The next classifier used is the Support Vector Machine classifier. The SVM classifier is a very commonly used classifier in the bioinformatics domain. It is very efficient at producing accurate models with gene microarray datasets. This classifier assumes a linear separation between both classes. This allows a discriminant to split the instances into two classes based off distance. The linear discriminant uses the following formula:

$$g(x|w, \omega_0) = w^T x + \omega_0$$

With respect to the formula above, for the linear discriminant the only values that need to be learned are the weight vector, w and the bias, $\omega_0$. Since the SVM classifier uses a linear discriminant there may be more than one discriminant that accurately predicts the class of the instance. SVM assumed the best discriminant is the one that maximizes the distance between both classes. The distance between the classes is measured as the distance from the discriminant to the samples of both classes [26].

The final classifier used in our experiments is the k-Nearest Neighbor classifier. This classifier is an instance based classifier and also considered a lazy learning algorithm. This classifier uses only the instances in the training set and their data to classify unknowns; it does not generate statistical values from the actual instances. Instead the k-NN classifier uses a distance measure to map the distances between the instance of interest (the unknown test sample) and the k-nearest neighboring instances. In our experiment the *k* is set to a value of 5, meaning the classifier will map the closest five instances to the instance of interest. When the test sample needs to be classified in the k-NN classifier each of the *k* closest training samples are calculated and weighted. The weight is determined by the measurement of $\frac{1}{Distance}$ where the *distance* is the distance from that training sample to the unknown test sample. Once the distances and the weights are calculated the weights are added up for each class, positive and negative. The prediction of the unknown class is decided by the higher weight value between the positive or negative class. One of the flexible things about the k-NN classifier is its ability to use any distance metric to calculate the distance between the instances. In our

experiment, and as is the standard, the distance metric used was the Euclidean distance. The Euclidean distance is defined as

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

### 3.2.1 Performance Metric

The classifiers were all evaluated by a specific performance metric, the area under the Receiver Operator Characteristic (ROC) curve (AUC). The ROC is a graphical representation of the true positive rate on the y-axis and the false positive rate on the x-axis. The true positive rate is calculated as the amount of instances correctly classified as positive over the total amount of instances classified as positive. The false positive rate is considered the amount of incorrectly classified instances in the positive class over the total amount of instances classified as positive. The area under this ROC curve is a way to quantify model performance with a single value [41]. The area under the ROC curve varies from values of .1 to 1.0, with .5 being the considered a failing value for the predictive performance and 1 being a perfect value. The AUC measures the discrimination of the model, meaning it measures how well the model correctly differentiates between the classes.

## 3.3 Feature Selection Techniques

In our experiments we used three different feature selection techniques on all five of the datasets. Each of these techniques is from a different family of feature selection but all fall under the category of stable feature selection techniques [38]. The first, Receiver Operator Characteristic (ROC) curve, is a threshold based feature selection (TBFS)

technique [34]. The second, Information Gain (IG) [42], is a commonly used filter method. Finally Signal-to-noise (S2N) is a different filter method that is not commonly used in the context of feature selection; this method is part of the first order statistic family (FOS) [35] of feature selection. Feature selection techniques are split into two categories, filter based feature selection and wrapper based feature selection. All the feature ranking techniques used in our experiments fall into the filter based category, IG and S2N are both robust feature selection techniques [39]. A total of four different subset sizes were used for the feature selection techniques in our experiments; 25, 50, 100, and 200. We did see increased performance above 200 but not enough significance in the values to warrant using a subset size higher than 200, inversely we did not use a subset size lower than 25 since we found a threshold of performance at a subset size of 10, with no significance between the size of 10 and 25.

### 3.3.1 Area under the Receiver Operator Characteristic curve (ROC)

The area under a ROC curve is used as a threshold-based feature selection technique in our experiments. This is similar to the information gain feature selection method. The TBFS method is a bivariate procedure; each attribute is tested against the class, with no dependencies on other features in the dataset. Each feature in the dataset is normalized to have values in the range of 0 to 1; simple classifiers are then built for each value $t \in [0; 1]$ according to two different classification rules. This method is still considered a filter method and not a wrapper method due because no real classifiers are being built. The normalized values are then treated as the posterior probability. The first classification rule has features with normalized values greater than *t* classified as *P* and features with a value less than *t* classified as *N*. The second classification rule has features

with a normalized value higher than *t* classified as *N* and features with a normalized value less than *t* classified as *P*. These two classification rules are derived because for some attributes the large values may correlate with the positive class more than the negative class and vice versa. The TBFS algorithm is a general algorithm that works based off different metrics. The metric $\omega$ is decided on beforehand and calculated through the algorithm either at each threshold *t*, or across all thresholds. In our experiments the metric used for the TBFS algorithm is the AUC. This is the same performance metric used to evaluate the classifiers except we are applying it to the TBFS algorithm and the features in the dataset.

### 3.3.2 Information Gain (IG)

The Information Gain feature selection technique is a normal filter based feature selection technique. Using the entropy of uncertainty of each feature the IG feature selection technique determines the significance of each feature. The formula used for the IG feature selection technique uses entropy for each feature. The IG formula is as follows

$$\text{IG(t)} =$$

$$-\sum_i \Pr(c_i)\, log\, \Pr(c_i) + \Pr(t) \sum_i \Pr(c_i|t)\, log\, \Pr(c_i|t) + \Pr(t) \sum_i \Pr(c_i|t)\, log\, \Pr(c_i|t)$$

Using this formula we can measure the entropy across the data and compare it to the entropy of the feature.

### 3.3.3 Signal-to-Noise Ratio (S2N)

Signal-to-Noise ratio is normally used for determining the amount of signal to the level of background noise, as indicated in the name. Usually S2N is not used as a feature

selection technique, in the context of classification and feature selection, S2N indicates how well a feature separates two classes. The formula for S2N is as follows:

$$S2N = (\mu_p - \mu_N)/(\sigma_p + \sigma_N)$$

where the values of $\mu_p$ and $\mu_N$ are means for the specific attribute across all the instances in the dataset which belong to a specific class P, positive, or N, negative. The values of $\sigma_p$ and $\sigma_N$ are considered the standard deviation between values in that particular attribute as relating to the class. The S2N formula results in a single numerical value, this value is what determines the relevance of the feature towards the dataset; the higher the value the more relevant a feature is.

## 3.4 Data Sampling Technique

A total of three of the five datasets used in our experiments suffer from class imbalance. To counter the class imbalance issue found in these datasets a data sampling technique was applied. Data sampling techniques do nothing for already balanced datasets. Data sampling techniques range from oversampling to undersampling techniques, for our study the most efficient and effective option was the random under sampling technique.

### 3.4.1 Random Under Sampling (RUS)

The random undersampling (RUS) technique is a very simple and cost effective method to counteract class imbalance. The RUS technique, as implied by the name, uses random undersampling to decrease the amount of instances found in the majority class to equal a previously set ratio. In our experiments the minority/majority ratio used for the RUS technique was 50/50. These were found to be the optimal ratios for classifier

performance. RUS was chosen over the SMOTE and the ROS techniques due to two reasons. The first of which is that RUS can be performed with datasets that have missing values, in our experiments both Thuerigen datasets have missing values, the second reason being that RUS performs just as efficiently as SMOTE and ROS with a much lower computational cost due to the undersampling effect [7]. SMOTE and ROS were considered but rejected since the datasets would be enlarged and the results would not be significantly different [40].

# 4 Results

The result for the classification, feature selection, and data sampling techniques are described in this section. This chapter is split into two sections, classification and comparison. The classification section will talk about the parameters chosen for the learners, feature selection techniques, and data sampling techniques. The comparison will delve into the actual results for each dataset.

## 4.1 Classification

The experiments were split into three total parts. The first part of the experiment is simply running the classifiers across the whole dataset and measuring the performance metric, the AUC, for each classifier. The classifiers used in this part of the experiment were Support Vector Machines, Naïve Bayes, 5-Nearest Neighbor, and Multilayer Perceptron. These classifiers were run through a binary classification with the datasets. The two classes were complete pathological response (pCR), the minority class, and residual disease (RD), which was the majority class. The data was run through these four classifiers for binary prediction using 4 runs of 5-fold cross validation, this resulted in 4 values for the AUC. These values were averaged out for each classifier.The second part of the experiment was to apply feature selection to the datasets and then evaluate the performance of the new truncated datasets using the classifiers. Before applying the feature selection techniques to the dataset subset sizes must be chosen. For the purpose of our experiment we chose four different subset sizes, 25, 50, 100, and 200. The settings remain the same as the settings used in the first part of the experiment; binary

classification with the same classifiers and 4 runs of five-fold cross validation. The values were again averaged out over every cross validation run.

The third, and final, part of the experiment was to apply the RUS data sampling technique to the datasets that suffer from class imbalance. Only three of our five datasets suffer from class imbalance so only those three have been analyzed in this part of the experiment. The feature selection techniques are applied first and the data sampling technique is only applied before the model is trained on the data. Again the models are evaluated for binary classification using 4 runs of five-fold cross validation. The resulting AUC values are averaged out. In total we have 5 datasets x 3 feature rankers x 4 feature subset sizes x 1 data sampling method x 4 classifier x 4 runs x 5 folds = 4800 total models built.

## 4.2 Comparison

This section will compare the results of the three different parts of the experiment. It will compare the results within each part and from part to part.

### 4.2.1 No feature selection technique or Data Sampling

In this section we tested the ability of the datasets without the alteration of feature selection or data sampling. Table 4.1 below shows the results for the 4 runs of five-fold cross validation. The results are shown with respect to the AUC performance metric. In the tables the best result for each dataset has been bolded while the worst result has been italicized.

Table 4.1: Average AUC using full dataset

| | Classifier Used | | | |
|---|---|---|---|---|
| Datasets | NB | MLP | 5-NN | SVM |
| Thuerigen 2006 Cy3 | *0.5229* | 0.6938 | 0.5267 | **0.7715** |
| Thuerigen 2006 Cy5 | *0.5200* | 0.6650 | 0.5685 | **0.7432** |
| Ma 2004 (Micro dissections) | *0.5305* | 0.5331 | **0.5615** | 0.5488 |
| Ma 2004 (Whole tumor) | 0.7246 | **0.7273** | 0.7094 | *0.6749* |
| Haury | 0.6718 | 0.5988 | *0.5941* | **0.7122** |

The table above shows the average AUC value across all 4 runs of the cross validation. The data shows the performance on each dataset. For the first Thuerigen et al. dataset we see the best classifier is the support vector machine with an average AUC of .7715. This indicates the classifier was able to learn from the dataset with relative ease. The second Thuerigen et al. dataset also had support vector machine perform the best with an AUC of .7432. Again this shows how well the support vector machine work on both datasets from Thuerigen et al.

The micro dissection dataset from Ma et al. was difficult to learn from for all the classifiers. The best performance was found on the 5-NN classifier with an AUC of .5615. Although the 5-NN classifier outperformed the other classifiers it was not by a large margin, the second best classifier was the SVM classifier with an AUC of .5488. In the second Ma et al. dataset, the whole tumor dataset, the best performer was the MLP with an AUC of .7273. This whole tumor dataset was easier to learn from for all classifiers with the lowest AUC value coming in at .6749 for the SVM classifier.

The final dataset, from Haury et al. varied in difficulty depending on which classifier was used. The best classifier for this dataset was the SVM classifier with an AUC value of .7122 while the worst classifier was the 5-NN classifier with an AUC of .5941. This

dataset and both the Thuerigen et al. datasets suffer from class imbalance, we see that all three of these datasets also have SVM as the best classifier. There may be a correlation between class imbalance and performance with the SVM classifier.

### 4.2.2 Only Feature Selection applied

This section showed the performance of the classifiers on the dataset after using feature selection. The tables below show the results for the classifiers using the information gain feature selection technique for four subset sizes; 25, 50, 100, and 200. Again the performance metric is the AUC. The best subset size has been bolded for each dataset on each classifier while the worst subset size has been italicized.

Table 4.2: AUC values for NB with IG on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | **0.5767** | 0.5480 | 0.5452 | *0.5111* |
| Thuerigen 2006 Cy5 | *0.6266* | 0.6616 | 0.6874 | **0.7070** |
| Ma 2004 (Microdissections) | *0.6688* | 0.6790 | 0.6820 | **0.6917** |
| Ma 2004 (Whole tumor) | 0.6742 | **0.6841** | 0.6551 | *0.6508* |
| Haury | 0.7083 | **0.7140** | 0.7086 | *0.7027* |

Table 4.3: AUC values for MLP with IG on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | *0.5528* | 0.5685 | **0.5928** | 0.5866 |
| Thuerigen 2006 Cy5 | *0.6324* | 0.6619 | 0.6955 | **0.7013** |
| Ma 2004 (Microdissections) | *0.6387* | 0.6512 | 0.6849 | **0.7124** |
| Ma 2004 (Whole tumor) | *0.6439* | 0.6501 | 0.6717 | **0.6780** |
| Haury | 0.6207 | 0.5978 | *0.5974* | **0.6423** |

Table 4.4: AUC values for 5-NN with IG on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | *0.5861* | 0.6145 | **0.6301** | 0.5906 |
| Thuerigen 2006 Cy5 | *0.7144* | 0.7521 | 0.7503 | **0.7745** |
| Ma 2004 (Microdissections) | *0.6567* | 0.6935 | 0.7016 | **0.7114** |
| Ma 2004 (Whole tumor) | **0.6685** | 0.6358 | *0.6162* | 0.6467 |
| Haury | *0.6303* | 0.6304 | **0.6387** | 0.6315 |

Table 4.5: AUC values for SVM with IG on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | *0.5774* | 0.5901 | 0.5813 | **0.6076** |
| Thuerigen 2006 Cy5 | 0.6275 | *0.6059* | **0.6690** | 0.6571 |
| Ma 2004 (Microdissections) | 0.6535 | *0.6286* | 0.6508 | **0.7080** |
| Ma 2004 (Whole tumor) | *0.6203* | 0.6384 | 0.6323 | **0.6385** |
| Haury | 0.6460 | 0.6242 | *0.5772* | **0.6639** |

Tables 4.2 – 4.5 show the results of the classifiers after the information gain feature ranker has been applied to the datasets.

The results show that the first Thuerigen et al. dataset performs the best when using a subset size of 100 in conjunction with the 5-NN classifier, resulting in an AUC value of .6301. The first Thuerigen et al. dataset performs the worst when using the Naïve Bayes classifier with a subset size of 100. When compared with the performance shown using the full dataset with no feature selection shown in Table 4.1 the dataset performed the best when using the full dataset and the SVM classifier with an AUC of .7715.

The second Thuerigen et al. dataset performed the best when using a subset size of 200 and the 5-NN classifier with an AUC of .7745 for the information gain feature

ranker. This dataset had the worst performance when using a subset size of 50 and the SVM classifier with an AUC value of .6059. When compared to the full dataset run with no feature selection in Table 4.1 the second Thuerigen et al. dataset performed better when using the information gain feature ranker and a subset size of 200.

The first Ma et al. dataset performed the best when using a subset size of 200 and using the MLP classifier resulting in an AUC of .7124. This dataset performed the worst when using the subset size of 25 and the MLP classifier with an AUC of .6567. The full dataset without any ranking resulted in an AUC of .5615, which compared to the AUC value using the information gain ranker and the subset size of 200 shows that the performance was significantly increased by implementing a feature ranker.

The second Ma et al. dataset performed the best when using the Naïve Bayes classifier and the subset size of 50, which resulted in an AUC of .6841. This dataset performed the worst when using the 5-NN classifier and a subset size of 100, which resulted in an AUC of .6162. The full dataset run without the classifiers resulted in an AUC of .7273, outperforming all subset sizes and classifiers using the feature selection technique.

The final dataset, from Haury et al. showed the best performance when using the Naïve Bayes classifier and the subset size of 50 with an AUC of .7140. This dataset showed the worst performance when using the SVM classifier and the subset size of 100, with an AUC of .5772. The full dataset, without the IG feature selection, resulted in an AUC value of .7122. The feature selection did not significantly improve the performance

but it significantly lowered the computational time required by limiting the necessary features to 50.

Table 4.6: AUC values for NB with ROC on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | **0.6078** | 0.5955 | 0.5666 | *0.5502* |
| Thuerigen 2006 Cy5 | *0.6928* | 0.7134 | 0.7211 | **0.7240** |
| Ma 2004 (Microdissections) | 0.4806 | *0.4657* | 0.4952 | **0.5086** |
| Ma 2004 (Whole tumor) | *0.4668* | 0.4800 | 0.5058 | **0.5488** |
| Haury | 0.6921 | *0.6919* | **0.7059** | 0.7051 |

Table 4.7: AUC values for MLP with ROC on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | 0.5661 | **0.5705** | 0.5435 | *0.5335* |
| Thuerigen 2006 Cy5 | 0.7124 | **0.7179** | *0.7082* | 0.7112 |
| Ma 2004 (Microdissections) | *0.4246* | 0.5282 | 0.5236 | **0.6609** |
| Ma 2004 (Whole tumor) | *0.4112* | 0.4863 | 0.5621 | **0.6157** |
| Haury | **0.6128** | 0.5556 | *0.5366* | 0.6080 |

Table 4.8: AUC values for 5-NN with ROC on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | 0.5912 | **0.5965** | 0.5836 | *0.5726* |
| Thuerigen 2006 Cy5 | *0.7252* | 0.7425 | **0.7608** | 0.7521 |
| Ma 2004 (Microdissections) | *0.4890* | **0.5138** | 0.5036 | 0.5041 |
| Ma 2004 (Whole tumor) | **0.5460** | 0.5413 | 0.5248 | *0.5084* |
| Haury | 0.6566 | **0.6690** | *0.6412* | 0.6461 |

Table 4.9: AUC values for SVM with ROC on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | 0.5528 | **0.5870** | 0.5729 | *0.5422* |
| Thuerigen 2006 Cy5 | 0.6955 | **0.6969** | 0.6829 | *0.6646* |
| Ma 2004 (Microdissections) | *0.3956* | 0.4587 | 0.5200 | **0.6655** |
| Ma 2004 (Whole tumor) | *0.4008* | 0.5050 | 0.5725 | **0.6673** |
| Haury | 0.5943 | 0.5730 | *0.5652* | **0.5949** |

Tables 4.6 – 4.9 show the results of the classifiers after the ROC ranker has been applied to the datasets.

The results for the first Thuerigen et al. dataset when using the ROC ranker show the best method being the NB classifier with a subset size of 25, which has an AUC value of .6078. The worst result for this dataset occurs when running the ROC with a subset size of 200 and the MLP classifier. When compared with the IG ranker the results show that this dataset performs better when using IG with a subset size of 100 and the 5-NN classifier, resulting in an AUC of .6301. Although neither of these feature ranking techniques outperforms the full dataset, which has an AUC of .7715.

The second Thuerigen et al. dataset shows the best results for the ROC ranker when using a subset size of 100 and the 5-NN classifier with an AUC of .7608. This dataset performs the worst when using the SVM classifier and a subset size of 200. When compared to the IG ranker the results show the IG ranker with a subset size of 200 and the 5-NN classifier outperforms the ROC ranker with an AUC of .7745. Although the IG ranker outperforms the ROC for this dataset the subset size used in the ROC is 100 less features than the IG ranker. The best result when using the ROC when compared with the

full dataset in Table 4.1 shows that using the ROC ranker results in increased performance.

The third dataset, the microdissections dataset from Ma et al., had the best performance for the ROC ranker when using a subset size of 200 and the SVM classifier with an AUC of .6655. This dataset performed the worst when using a subset size of 25 and the SVM classifier, resulting in an AUC of .3956. When compared with the IG ranker the results show that this dataset performed better using the IG ranker, a subset size of 200, and the MLP classifier. When compared to the full dataset without a feature ranker there is evidence that the ROC ranker increases performance significantly.

The fourth dataset, the whole tissue dataset from Ma et al., showed the best performance for the ROC ranker using the subset size of 200 and the SVM classifier with an AUC value of .6673. This dataset performed the worst when using a subset size of 25 and the SVM classifier. This dataset mirrored the results of the first Ma et al. dataset. When compared to the AUC value for the best IG ranker option, an AUC of .6841, the performance shows that using the IG ranker outperforms the ROC ranker for this dataset. When comparing back to the full dataset with no rankers the results show a better performance using the full dataset.

The final dataset, Haury et al., shows the best performance for the ROC ranker when using a subset size of 100 and the NB classifier with an AUC of .7059. This dataset performs the worst when using a subset size of 100 and the MLP classifier with an AUC of .5366. When compared with the IG ranker the results show that the IG ranker with a subset size of 50 and the NB classifier results in an AUC value of .7140, outperforming

the ROC ranker method. When compared to the full dataset with no feature selection the results show that the full feature dataset outperforms the feature selection by a slight amount, an AUC of .7122, but this is not enough of a difference to be significant. The amount of computational time saved by using the ROC ranker results in a better overall performance for this dataset, even though the AUC for the full dataset is slightly higher.

Table 4.10: AUC values for NB using S2N on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | 0.6217 | **0.6388** | *0.5973* | 0.6036 |
| Thuerigen 2006 Cy5 | *0.6923* | 0.7020 | 0.7135 | **0.7185** |
| Ma 2004 (Microdissections) | **0.6739** | 0.6572 | 0.6555 | *0.6401* |
| Ma 2004 (Whole tumor) | **0.5945** | 0.5907 | 0.5645 | *0.5379* |
| Haury | *0.7007* | 0.7051 | **0.7211** | 0.7155 |

Table 4.11: AUC values for MLP using S2N on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | 0.6137 | **0.6402** | 0.5829 | *0.5742* |
| Thuerigen 2006 Cy5 | *0.6688* | 0.6723 | 0.6870 | **0.6954** |
| Ma 2004 (Microdissections) | 0.7145 | *0.7133* | 0.7470 | **0.7529** |
| Ma 2004 (Whole tumor) | *0.6616* | 0.6945 | 0.6954 | **0.7010** |
| Haury | **0.6388** | *0.5484* | 0.5606 | 0.6134 |

Table 4.12: AUC values for 5-NN using S2N on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | 0.6320 | **0.6755** | 0.6163 | *0.6123* |
| Thuerigen 2006 Cy5 | 0.7237 | *0.7225* | 0.7368 | **0.7452** |
| Ma 2004 (Microdissections) | *0.6891* | 0.7046 | **0.7245** | 0.7026 |
| Ma 2004 (Whole tumor) | **0.6157** | 0.5563 | 0.5745 | *0.5393* |
| Haury | *0.6327* | 0.6494 | 0.6668 | **0.6745** |

Table 4.13: AUC values for SVM using S2N on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen 2006 Cy3 | 0.6139 | **0.6642** | 0.6349 | *0.6023* |
| Thuerigen 2006 Cy5 | **0.6529** | 0.6526 | *0.6208* | 0.6506 |
| Ma 2004 (Microdissections) | 0.7236 | *0.6985* | **0.7497** | 0.7392 |
| Ma 2004 (Whole tumor) | *0.6777* | 0.6836 | 0.6986 | **0.7280** |
| Haury | **0.6058** | 0.5639 | *0.5234* | 0.6041 |

The Tables 4.10 – 4.13 show the results for the four classifiers with the Signal-to-Noise feature ranking technique applied to the datasets.

The first Thuerigen et al. dataset performs the best when using the S2N ranker at the subset size of 50 and the 5-NN classifier with an AUC of .6755. The worst result for this dataset occurs when running the S2N with a subset size of 200 and the MLP classifier. When compared with the IG ranker and the ROC ranker the results show that this dataset performs better when using S2N with a subset size of 50 and the 5-NN classifier. Although none of these features ranking techniques outperforms the full dataset, which has an AUC of .7715.

The second Thuerigen et al. dataset shows the best results for the S2N ranker when using a subset size of 200 and the 5-NN classifier with an AUC of .7452. This dataset performs the worst when using the SVM classifier and a subset size of 100. When compared to the IG ranker and the ROC the results show that both other rankers outperform the S2N ranker for this dataset. The best result when using the S2N when compared with the full dataset in Table 4.1 shows that using the S2N ranker results in increased performance.

The third dataset, the microdissections dataset from Ma et al., had the best performance for the S2N ranker when using a subset size of 200 and the MLP classifier with an AUC of .7529. This dataset performed the worst when using a subset size of 200 and the NB classifier, resulting in an AUC of .6401. When compared with the IG ranker and the ROC the results show that this dataset performed better than both other rankers. When compared to the full dataset without a feature ranker there is evidence that the S2N ranker increases performance significantly when compared by the AUC performance metric.

The fourth dataset, the whole tissue dataset from Ma et al., showed the best performance for the S2N ranker using the subset size of 200 and the SVM classifier with an AUC value of .7280. This dataset performed the worst when using a subset size of 200 and the NB classifier. When compared to the AUC value for the best IG ranker option, an AUC of .6841, and the AUC for the best ROC ranker option, an AUC of .6673, the S2N ranker shows a better performance than both other feature ranker techniques. When comparing back to the full dataset with no rankers the results show a better performance using the S2N ranker with a subset size of 200 and the SVM classifier.

The final dataset, Haury et al., shows the best performance for the S2N ranker when using a subset size of 100 and the NB classifier with an AUC of .7211. This dataset performs the worst when using a subset size of 100 and the SVM classifier with an AUC of .5234. When compared with the IG ranker and the ROC ranker the results show that the S2N ranker outperforms both other rankers with this dataset using the NB classifier and the subset size of 100. When compared to the full dataset with no feature selection

the results show that using the S2N feature selection technique outperforms the full dataset.

### 4.2.3 Data sampling and Feature Selection applied

This section will look at the performance of the classifiers when using the feature selection techniques and the RUS data sampling technique. This section will only consist of the three datasets that suffered from class imbalance, both the Thuerigen et al. datasets and the Haury et al. dataset.

Table 4.14: AUC values for NB using IG and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | *0.6621* | 0.6664 | 0.6806 | **0.6986** |
| Thuerigen2006Cy5 | 0.6716 | **0.6765** | *0.6486* | 0.6609 |
| Haury | **0.7161** | 0.7104 | *0.7097* | 0.7103 |

Table 4.15: AUC values for MLP using IG and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | 0.6299 | *0.6293* | 0.6725 | **0.7132** |
| Thuerigen2006Cy5 | *0.6496* | 0.6821 | 0.6832 | **0.6986** |
| Haury | **0.6881** | 0.6631 | 0.6700 | *0.6593* |

Table 4.16: AUC values for 5-NN using IG and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | *0.6604* | 0.6722 | 0.6950 | **0.7025** |
| Thuerigen2006Cy5 | 0.6771 | **0.6871** | *0.6570* | 0.6586 |
| Haury | *0.6718* | 0.6920 | 0.6948 | **0.7099** |

Table 4.17: AUC values for SVM using IG and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | 0.6015 | *0.5923* | 0.6593 | **0.7175** |
| Thuerigen2006Cy5 | 0.6559 | 0.6454 | *0.6373* | **0.6807** |
| Haury | **0.6926** | 0.6420 | *0.6391* | 0.6444 |

The Tables 4.14 – 4.17 show the average AUC values for the four classifiers on the datasets after the IG feature selection technique and the RUS data sampling technique were applied. The feature selection technique was applied first to the datasets, once the feature selection was finished the data sampling technique was applied before training the model.

Only three datasets were analyzed in this section, the first of which was the first Thuerigen et al. dataset. This dataset performed the best for the IG ranker and the RUS technique when using the SVM classifier and a subset size of 200, this combination resulted in an AUC of .7175. The dataset performed worst when using SVM and a subset size of 50. When compared to the best result just using the IG ranker there was a significant increase in performance, with a change in AUC from .6301 to .7175. Although when compared to the full dataset we still see a weaker performance.

The second of the three datasets was the second Thuerigen et al. dataset. This dataset performed the best for the IG ranker and the RUS technique when using the MLP classifier with a subset size of 200. When compared to the best result for just using the IG ranker there is a decrease of performance, from an AUC of .7745 to .6986, a significant drop in performance. When compared to the full dataset without IG or RUS there is a significantly better performance when using the full dataset.

The final dataset was the Haury et al. dataset. This dataset performed the best for the IG ranker and the RUS technique when using a subset size of 25 and the NB classifier, resulting in an AUC of .7161. When compared to the best result using just the IG classifier and the full dataset there is a slight increase of performance.

Table 4.18: AUC values for NB using ROC and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | 0.5130 | *0.5021* | 0.5101 | **0.5194** |
| Thuerigen2006Cy5 | *0.4658* | 0.4784 | 0.5111 | **0.5363** |
| Haury | *0.6987* | 0.7124 | 0.7122 | **0.7158** |

Table 4.19: AUC values for MLP using ROC and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | *0.4967* | 0.5448 | 0.5608 | **0.6372** |
| Thuerigen2006Cy5 | *0.4551* | 0.4959 | 0.5213 | **0.5789** |
| Haury | 0.6595 | *0.6568* | 0.6619 | **0.6680** |

Table 4.20: AUC values for 5-NN using ROC and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | 0.5172 | **0.5282** | 0.5172 | *0.4976* |
| Thuerigen2006Cy5 | 0.5123 | *0.4368* | 0.4945 | **0.5432** |
| Haury | *0.6694* | 0.6831 | 0.6937 | **0.7156** |

Table 4.21: AUC values for SVM using ROC and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | *0.4676* | 0.5911 | 0.5685 | **0.6311** |
| Thuerigen2006Cy5 | *0.4747* | 0.4979 | 0.5821 | **0.6214** |
| Haury | 0.6320 | *0.6293* | **0.6473** | 0.6471 |

Tables 4.18 – 4.21 show the results of the four classifiers when using both the ROC and the RUS techniques on the datasets for four subset sizes.

The first dataset looked at is the first of the two Thuerigen et al. datasets. This dataset performs the best for the ROC ranker and the RUS technique when using the MLP classifier with a subset size of 200. This dataset performs the worst when using the SVM classifier and a subset size of 25. When compared to the best result when just using the ROC ranker there is a slight increase in performance; the ROC and RUS combo had an AUC of .6372 while just the ROC had an AUC of .6078. Again when compared to the full dataset using the ROC and RUS combo resulted in a significantly lower performance. Although it is important to note that just using the ROC technique used a smaller subset size, 25, and a simpler learner, NB, so the computational time may be worth the decrease in performance.

The second dataset looked at was the second Thuerigen et al. dataset. This dataset performs best for the ROC and RUS combo when using a subset size of 200 and the SVM classifier with an AUC of .6214. This dataset performs the worst for this combo when using the 5-NN classifier and a subset size of 50; an AUC of .4368. When compared to the best result of using only the ROC there is a large decrease in performance, an AUC decrease of .7608 to .6214. When compared to the full dataset we see a significantly worse performance as well.

The final dataset used in this section was the Haury et al. dataset. This dataset showed the best performance for the ROC and RUS combo when using a subset size of 200 and the NB classifier. This dataset performed the worst when using a subset size of

50 and the SVM classifier. When compared to the best result for using only the ROC there is an increase of performance; an increase in AUC of .7059 to .7158. This combination of techniques also results in a better performance than the full dataset.

Table 4.22: AUC values for NB using S2N and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | **0.5469** | *0.5305* | 0.5364 | 0.5337 |
| Thuerigen2006Cy5 | *0.4865* | 0.5037 | 0.5236 | **0.5563** |
| Haury | **0.7139** | 0.7130 | 0.7077 | *0.7046* |

Table 4.23: AUC values for MLP using S2N and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | *0.5504* | 0.6131 | 0.6336 | **0.6698** |
| Thuerigen2006Cy5 | 0.5073 | *0.4884* | 0.5673 | **0.6166** |
| Haury | 0.6687 | *0.6456* | 0.6689 | **0.6751** |

Table 4.24: AUC values for 5-NN using S2N and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | 0.5133 | **0.5243** | *0.5098* | 0.5117 |
| Thuerigen2006Cy5 | *0.4204* | 0.4375 | 0.4863 | **0.5241** |
| Haury | 0.6652 | *0.6650* | **0.6926** | 0.6860 |

Table 4.25: AUC values for SVM using S2N and RUS on four subset sizes

| DataSets | Feature Subset Size | | | |
|---|---|---|---|---|
| | 25 | 50 | 100 | 200 |
| Thuerigen2006Cy3 | *0.5513* | 0.5897 | 0.6111 | **0.6629** |
| Thuerigen2006Cy5 | *0.5014* | 0.5030 | 0.6048 | **0.6366** |
| Haury | **0.6627** | 0.6288 | *0.6275* | 0.6456 |

The Tables 4.22 – 4.25 show the results of the four classifiers when using the Signal-to-Noise learner and the RUS technique on four different subset sizes.

The first of the datasets is the first Thuerigen et al. dataset. This dataset performed the best for the S2N and RUS combination when using a subset size of 200 and the SVM classifier. This dataset performed the worst for this combination when using the 5-NN classifier and a subset size of 100. When compared to the best result for just using the S2N ranker there is a decrease in performance when adding in the RUS technique; an AUC drop from .6755 to .6698. When compared to the full dataset there is a significant difference in performance.

The second of the datasets is the second Thuerigen et al. dataset. This dataset performed the best for the S2N and RUS combination when using a subset size of 200 and the SVM learner. The worst performance for this dataset was when the 5-NN learner was used with a subset size of 25. When compared to the best result for only using the S2N ranker there is a decrease in performance; a drop in AUC from .7452 to .6366. The decrease in performance is significant, using the combination of S2N and RUS also results in a significant decrease in performance from the full dataset performance.

The final dataset, from Haury et al., had the best performance for S2N and RUS when using the NB classifier and a subset size of 25. The worst performance for this dataset came when using the SVM classifier and a subset size of 100. When compared to the best result when just using the S2N ranker there was a slight decrease in performance when adding the RUS technique, a drop in the AUC from .7211 to .7139. Using the S2N ranker and the RUS classifier increases performance when compared to the results of the full dataset.

### 4.2.4 Three-way Comparison of Methods

Table 4.26: Best AUC Results for Each Dataset Across all three tests

| Datasets | No Alteration | FS Only | FS + DS |
|---|---|---|---|
| Thuerigen 2006 Cy3 | **0.7715** | *0.6755* | 0.7175 |
| Thuerigen 2006 Cy5 | 0.7432 | **0.7746** | *0.6871* |
| Ma 2004 (Microdissections) | *0.5616* | **0.7529** | N/A |
| Ma 2004 (Whole tumor) | *0.7273* | **0.7280** | N/A |
| Haury | *0.7122* | **0.7211** | 0.7161 |

The table 4.26 above shows the values for best performance across all three different experiments.

For the first Thuerigen et al. dataset the results show that using the full dataset results in the best performance. The filter only method results in the worst performance for this dataset. This dataset contained missing values and could have caused feature selection techniques to miss certain features due to those missing values. The combination of the filter selection and data sampling techniques resulted in a good performance, although not as good as the full dataset. The computational time saved by using this method may make it worthwhile to sacrifice a bit of performance.

For the remaining four datasets we see that the filter only technique increased performance and lead to a better result. In the Thuerigen et al. dataset the filter only method outperformed the full dataset even though this dataset suffered from missing values as well. When contrasted to the first Thuerigen et al. dataset the features selected increased performance while using both the feature selection and the data sampling techniques decreases performance. The Haury et al. dataset also saw the best improvement when using only the feature selection technique. The results for the full

dataset, the feature selection only, and the feature selection and data sampling for this dataset show only a small change in performance, although both the feature selection only and the combination of feature selection and data sampling result in better performance than using the full dataset.

For the two Ma et al. datasets there were only two experiments done because these datasets did not suffer from class imbalance. The results show for the first Ma et al. dataset there was a significant improvement when feature selection techniques was applied, for the second Ma et al. dataset the improvement was marginal.

# 5 Conclusion and Future Work

As the methods used to understand and treat cancer develop data will play a bigger part in determining what course of treatment each patient should be given. As analyzing the human genome becomes cheaper, personalized medicine and gene microarrays will become more and more common in medicine. The data mining tools used in bioinformatics will need to grow to accommodate the amount of data made available. Bioinformatics datasets will still suffer from the same problems they have always suffered from, high dimensionality and class imbalance. Possible solutions to these issues are feature selection and data sampling techniques. Feature selection techniques are commonly used in data mining to eliminate redundant features and limit useless features allowing learners to perform at optimum levels without useless features. Data sampling techniques are used often in data mining to balance out the classes to allow the learners to create a more realistic and usable model. Since a gene microarray dataset is composed of patients, the instances, and gene probes, the features, simply running feature selection and data sampling techniques can result in future areas of research. The experiments done in this thesis are to test the applications of those feature selection and data sampling techniques on the specific domain of breast cancer gene microarrays.

## 5.1 Conclusions

Feature selection and data sampling are very important when it comes to the analysis of gene microarray datasets. The subset sizes used to train the model range from 25 to 200, when compared to the smallest full dataset sizes the largest subset size is only 1.6% of the total full dataset. When compared to the results for using the full dataset using the smaller subset sizes resulted in models that ranged from almost as good to significantly better than using the full dataset. For the experiment SVM showed to be the superior classifier followed by 5-NN, then NB, and finally MLP.

The feature selection methods all seemed to have a positive effect on the datasets. Not only did they reduce the amount of computational time required for a model to be built but they also improved performance on some of the datasets. For every dataset except the first Thuerigen et al. dataset performance was increased when using the feature selection techniques. Overall the signal-to-noise ranker was the most efficient, increasing the performance of both Ma et al. datasets and the Haury et al. dataset the most. The information gain ranker increased performance of the second Thuerigen et al. dataset the most. The ROC performed the worst of the three rankers, although it did still slightly increase the performance of all four datasets.

When data sampling techniques were applied before training the model it seemed that both Thuerigen et al. datasets had a better performance across all four classifiers with the information gain feature ranker than just running the feature ranker alone. The datasets seemed to have a worse performance when using the ROC and S2N with RUS combinations over just using the ROC or S2N. For the Haury et al. dataset it seems the data sampling technique increased performance across all classifiers and all rankers.

Overall it seems the results show that when dealing with high dimensional and class imbalance breast cancer datasets the feature selection methods increase performance the most. It seems for two of the three class imbalanced datasets the performance was decreased using the feature selection and data sampling combination when compared to the full dataset. For the remaining dataset the performance was increased. The feature selection and data sampling methods used together seem to perform differently depending on the datasets, due to the limited number of class imbalance datasets used in this experiment more research into class imbalance in breast cancer datasets is required to have a definitive result.

### 5.2 Future Work

The experiments in this thesis have been localized to the bioinformatics domain and specifically breast cancer gene microarray datasets. There may be other related topics that could be the focus of future research, they are listed below.

- The datasets we had were all originating from breast cancer patient response studies. These datasets are very specific and it may do well to test these feature rankers and data sampling techniques on other cancerous datasets.

- More breast cancer patient response datasets could be gathered and analyzed using the same methodology but different feature rankers and data sampling techniques.

- The comparison between bioinformatics data and other domains can be done to see the effects of learners, rankers, and data sampling techniques on different data domains.

## Bibliography


[1] Wald, R., and Khoshgoftaar, T.M., Patient Response Datasets: Challenges and Opportunities. *Proceedings of the IEEE International Conference on Information Reuse and Integration-- IRI'13,* San Francisco, CA, , pp. 254-261, August 14-16, 2013.

[2] Davide Mauri, Nicholas Pavlidis, and John P. A. Ioannidis, Neoadjuvant Versus Adjuvant Systemic Treatment in Breast Cancer: A meta-analysis. *Journal of the National Cancer Institute,* Vol. 97, No. 3, pp. 188-194, February 2, 2005.

[3] Daniel M Keller, From shotgun to Gatling Gun: multi-barrelled approach to cancer treatment, *Oncology Times,* pp. 17-18, May 2009.

[4] Johnston, Mark, Gene Chips: Array of hope for understanding gene regulation. *Current Biology*, Vol. 8, Issue 5, pp. R171-R174, February 26, 1998.

[5] National Center for Biotechnology Information. Bioinformatics, 2004. http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html.

[6] G. Piatetsky-Shapiro, T. Khabaza, and S. Ramaswamy. Capturing best practice for microarray gene expression data analysis. In *KDD 03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 407–415, New York, NY, December 2009. ACM.

[7] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance." *IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans*, Vol. 40, No. 1, January 2010

[8] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald. "Feature selection with high dimentional imbalanced data." In *Proceedings of the 9th IEEE International Conference on Data Mining – Workshops (ICDM'09),* pages 507-514, Miami, FL, December 2009. IEEE Computer Society.

[9] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." *bioinformatics* 23, no. 19 (2007): 2507-2517.

[10] Robin M Hallet, Gregory Pond, John A Hassell, "A target based approach identifies genomicpredictors of breast cancer patient response to chemotherapy," *BMC Medical Genomics,* vol 5, 2012. [Online] http://www.biomedcentral.com/1755-8794/5/16

[11] J. L. Rodgers and W. A. Nicewander. "Thirteen ways to look at the correlation coefficient." *The American Statistician*, 42(1):59–66, February 1988.

[12] Adel Tabchy, Vicente Valero, Tatiana Vidaurre, et al. "Evaluation of a 30-Gene Paclitaxel, Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy Response Predictor in a Multicenter Randomized Trial in Breast Cancer," *Clin Cancer Res,* vol 16, pg 5351-5361, 2010. [Online] http://www.ncbi.nlm.nih.gov/pubmed/20829329

[13] X.-J. Ma, Z. Wang, P. D. Ryan, S. J. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, J. Tuggle, Y. Tran, D. Tran, A. Tassin, P. Amon, W. Wang, W. Wang, E. Enright, K. Stecker, E. Estepa-Sabal, B. Smith, J. Younger, U. Balis, J. Michaelson, A. Bhan, K. Habin, T. M. Baer, J. Brugge, D. A. Haber, M. G. Erlander, and D. C. Sgroi, "A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen," *Cancer Cell,* vol. 5, no. 6, pp. 607 – 616, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/B6WWK-4CM8XX0-D/2/c964821e0e62ebad0820f9237699257f

[14] Wray, Naomi R., Jian Yang, Michael E. Goddard, and Peter M. Visscher. "The genetic interpretation of area under the ROC curve in genomic profiling." *PLoS Genetics* 6, no. 2 (2010): e1000864.

[15] M. Chanrion, V. Negre, H. Fontaine, N. Salvetat, F. Bibeau, G. M. Grogan, L. Mauriac, D. Katsaros, F. Molina, C. Theillet, and J.-M. Darbon, "A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer," *Clinical Cancer Research*, vol. 14, no. 6, pp. 1744–1752, 2008. [Online]. Available: http://clincancerres.aacrjournals.org/content/14/6/1744.abstract

[16] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98:5116 ^21.

[17] Tibshirani R, Hastie T, Narasimhan B, et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression." *Proc Natl Acad Sci,* USA 2002;99:6567^72.

[18] Becker B., R. Kohavi, and D. Sommerfield. 2001. Visualizing the Simple Bayesian Classifier. In: *Information Visualization in Data Mining and Knowledge Discovery*,

U. Fayyad, G. Grinstein, and A. Wierse, eds. San Francisco: Morgan Kaufmann Publishers.

[19] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician* **46** (3): 175–185.

[20] Y. Pawitan, J. Bjohle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. Shaw, J. Smeds, L. Skoog, S. Wedren, and J. Bergh, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts," *Breast Cancer Research*, vol. 7, no. 6, pp. R953–R964, 2005. [Online]. Available: http://breast-cancer-research.com/content/7/6/R953

[21] O. Thuerigen, A. Schneeweiss, G. Toedt, P. Warnat, M. Hahn, H. Kramer, B. Brors, C. Rudlowski, A. Benner, F. Schuetz, B. Tews, R. Eils, H.-P. Sinn, C. Sohn, and P. Lichter, "Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer," *Journal of Clinical Oncology*, vol. 24, no. 12, pp. 1839–1845, 2006. [Online]. Available: http://jco.ascopubs.org/content/24/12/1839.abstract

[22] A. Haury, P. Gestraud, J.P. Vert, "The influence of Feature Selection methods on Accuracy, Stability, and Interpretability of Molecular Signatures." *Plos One*, Vol. 6, issue 12 (2011), e28210.

[23] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. Principles Knowl. Discov. Databases*, 2003, pp. 107–119.

[24] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2nd edition, 2005.

[25] S. Haykin. *Neural Networks: A Comprehensive Foundation 2nd edition*. Prentice Hall, 1998.

[26] Rish, Irina. "An empirical study of the naive Bayes classifier." In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, pp. 41-46. 2001.

[27] Murtagh, Fionn. "Multilayer perceptrons for classification and regression." *Neurocomputing* 2, no. 5 (1991): 183-197.

[28] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003).

[29] Cunningham, Padraig, and Sarah Jane Delany. "k-Nearest neighbour classifiers." *Multiple Classifier Systems* (2007): 1-17.

[30] de Haan, Jorn R., Ron Wehrens, Susanne Bauerschmidt, Ester Piek, René C. van Schaik, and Lutgarde MC Buydens. "Interpretation of ANOVA models for microarray data using PCA." *Bioinformatics* 23, no. 2 (2007): 184-190.

[31] Welch, B. L. "On the comparison of several mean values: an alternative approach." *Biometrika* (1951): 330-336.

[32] Hess, Kenneth R., Keith Anderson, W. Fraser Symmans, Vicente Valero, Nuhad Ibrahim, Jaime A. Mejia, Daniel Booser et al. "Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer." *Journal of clinical oncology* 24, no. 26 (2006): 4236-4244.

[33] Van Hulse, Jason, Taghi M. Khoshgoftaar, Amri Napolitano, and Randall Wald. "Threshold-based feature selection techniques for high-dimensional bioinformatics data." *Network modeling analysis in health informatics and bioinformatics* 1, no. 1-2 (2012): 47-61.

[34] Khoshgoftaar, Taghi, David Dittman, Randall Wald, and Alireza Fazelpour. "First order statistics based feature selection: A diverse and powerful family of feature seleciton techniques." In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 2, pp. 151-157. IEEE, 2012.

[35] Dittman, David. *Feature selection techniques and applications in bioinformatics*. Florida Atlantic University, 2011.

[36] Frank, Eibe, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H. Witten. "Data mining in bioinformatics using Weka." *Bioinformatics* 20, no. 15 (2004): 2479-2481.

[37] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In *IJCAI*, vol. 14, no. 2, pp. 1137-1145. 1995.

[38] Awada, Wael, Taghi M. Khoshgoftaar, David Dittman, Randall Wald, and Amri Napolitano. "A review of the stability of feature selection techniques for bioinformatics data." In *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on*, pp. 356-363. IEEE, 2012.

[39] Altidor, Wilker, Taghi M. Khoshgoftaar, and Jason Van Hulse. "Robustness of Filter-Based Feature Ranking: A Case Study." In *FLAIRS Conference*. 2011.

[40] Van Hulse, Jason, and Taghi Khoshgoftaar. "Knowledge discovery from imbalanced and noisy data." *Data & Knowledge Engineering* 68, no. 12 (2009): 1513-1542.

[41] Seliya, Naeem, Taghi M. Khoshgoftaar, and Jason Van Hulse. "A study on the relationships of classifier performance metrics." In *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on*, pp. 59-66. IEEE, 2009.

[42] Van Hulse, Jason, Taghi M. Khoshgoftaar, Amri Napolitano, and Randall Wald. "Feature selection with high-dimensional imbalanced data." In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pp. 507-514. IEEE, 2009.