Data Mining Techniques for Prognosis in Pancreatic Cancer

by

Stuart Floyd

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

May 2007

APPROVED:

_____

Professor Carolina Ruiz, Thesis Advisor

_____

Professor Sergio Alvarez (Boston College), Thesis Co-Advisor

_____

Professor Neil Heffernan, Thesis Reader

_____

Professor Michael A. Gennert, Head of Department

Abstract

This thesis focuses on the use of data mining techniques to investigate the expected survival time of patients with pancreatic cancer. Clinical patient data have been useful in showing overall population trends in patient treatment and outcomes. Models built on patient level data also have the potential to yield insights into the best course of treatment and the long-term outlook for individual patients. Within the medical community, logistic regression has traditionally been chosen for building predictive models in terms of explanatory variables or features. Our research demonstrates that the use of machine learning algorithms for both feature selection and prediction can significantly increase the accuracy of models of patient survival. We have evaluated the use of Artificial Neural Networks, Bayesian Networks, and Support Vector Machines. We have demonstrated ($p<0.05$) that data mining techniques are capable of improved prognostic predictions of pancreatic cancer patient survival as compared with logistic regression alone.

Table of Contents

Table of Figures

1 Introduction

With the increasing ability of health care centers to digitally organize clinical cancer patient data, new techniques now exist to explore treatment of cancer. This data has been useful in showing overall trends in patient treatment but has only until recently been applied to evaluating how to best treat individual patients. Models built on patient level data using data mining techniques have the potential to give insight into the best course of treatment, and the long-term outlook, for individual patients.

Our goal for this thesis has been to apply data mining techniques to patient level cancer data. Study into modeling of the relationships between a patient's history, medical record, disease stage, and outcome is central to the study of how to better treat cancer patients. Data mining techniques have the potential to improve existing models of these relationships [Hay06].

For this project, our focus is on pancreatic cancer. According to the American Cancer Society's 2007 statistics, pancreatic cancer is the 4th largest killer among all cancers in the United States [ACS07]. For those diagnosed with pancreatic cancer, there is a one year survival rate of 19% and a five year survival rate of 4%. If surgical resection of the tumor is performed, the five year survival jumps up to 40% [DHR05]. These numbers provide a convincing argument for the surgical removal of tumors from all patients but this is not always appropriate.

Pancreatic cancer is both very debilitating to the patient and very difficult to treat. For example a Whipple procedure, the most common surgical treatment for pancreatic cancer, can take over eight hours to perform and may take the patient several months to recover from. If the cancer has spread into the patient's arteries or other organs of the body, even this intense surgery is not able to cure the disease. In cases where it is not possible to remove the cancer, it is more appropriate to take steps to improve the patient's quality of life while living with the disease. For these reasons when deciding

whether surgical treatment of pancreatic cancer is appropriate, a trade off must always be made considering the patient's expected quality of life and survival time.

The pancreatic cancer data used in this study has been collected from The University of Massachusetts Medical School in collaboration with Dr. Jennifer Tseng and Dr. Giles Whalen. This sample includes the patients who were evaluated for surgical removal of a pancreatic tumor by the Department of Surgical Oncology between April 2002 and December 2005. Hayward's MS thesis, [Hay06], started the collection and analysis of this data. Within the dataset there are approximately two hundred attributes for each patient including attributes relating to the patient's cancer diagnosis, symptoms at diagnosis, relevant past history, family cancer history, lab and imaging scores, treatment, and follow up.

Our focus for this thesis is to build machine learning models using clinical cancer patient data. We want to use these models for prediction of survival time and to understand what factors influence this outcome.

2 Background

2.1 Data Mining Techniques

A wide variety of data mining techniques exist to model training data, also referred to as training instances, that can be used to predict the target value of an unseen instance. Specifically, these techniques construct models of the relationships between a set of input attributes, known as features, and a target concept.

Each of the attributes represents a dimension of the input space that is used to construct a model of the target concept. Feature selection algorithms exist to help reduce the number of dimensions in this input space. There are several techniques for feature selection including ones that select the best set of the original attributes and ones that transform the input space to obtain a better set of features to represent the data.

Models of the relationships between the input space and the target concept are constructed through use of machine learning algorithms.

An overview covering the data mining concepts and techniques that are used in this thesis is presented in this section.

2.1.1 Feature Selection

Feature selection algorithms, also commonly referred to as attribute selection algorithms, aim to reduce the dimensionality of the input space. Usually this is done by searching for the most relevant set of attributes. Reducing the dimensionality of the input space usually increases both the efficiency and the predictive accuracy of machine learning algorithms [WF05].

2.1.1.1 Gain Ratio Attribute Selection

$$\text{Entropy(TrainingInstance } I) \equiv \sum_{i=1}^{\text{number of target values}} -p_i \log_2 p_i$$

Where:
p_i is probability of classification target *i*
Formula 1: Entropy

$$\text{Gain}(\text{TrainingInstances } I, \text{SomeAttribute } A) \equiv \text{Entropy}(I) - \sum_{v \in Values(A)} \frac{|I_v|}{|I|} \text{Entropy}(I_v)$$

Where:
I_v is the subset of instances in I with attribute value v for attribute A
$|I|$ is the number of instances in set I
Formula 2: Information Gain

$$\text{SplitInformation}(\text{TrainingInstances } I, \text{SomeAttribute } A) \equiv -\sum_{v \in Values(A)} \frac{|I_v|}{|I|} \log_2 \frac{|I_v|}{|I|}$$

Where:
I_v is the subset of instances in I with attribute value v for attribute A
$|I|$ is the number of instances in set I
Formula 3: Split Information

$$\text{GainRatio}(\text{TrainingInstances } I, \text{SomeAttribute } A) \equiv \frac{\text{Gain}(I, A)}{\text{SplitInformation}(A)}$$

Formula 4: Gain Ratio

The formula for calculating the gain ratio is introduced in [Qui86] as a technique for evaluating attributes in the construction of decision trees. A critical component in calculating the gain ratio is entropy. Entropy is calculated in Formula 1 by summing over the negative product of the probability of each classification value times its logarithm. The probability of each classification value is calculated based on its frequency in the set of training instances I. If all of the classes have an equal probability of occurring, the entropy equation will return a higher value than if a small subset of possible values have higher probabilities than the rest. Entropy is therefore a measurement of the degree to which the classes are differentiated within the training instances.

Entropy is used to calculate the information gain for an attribute as shown in Formula 2. Note that I_v is the subset of the training data with attribute value v. The information gain measures the degree to which the attribute under consideration is able to increase the differentiation of the classes. Information gain is therefore looking for the attribute that, given its value, will most significantly decrease the entropy.

The way that entropy measures the degree to which the classes are differentiated within the training instances can be applied to an individual attribute using split information. Split information is shown in Formula 3 and is calculated the same as entropy just with the probability of the attribute in place of the probability of the classification target.

Information gain and split information are used to calculate the gain ratio for an attribute as shown in Formula 4 [Mit97].

The gain ratio is used for feature selection by running it for every attribute. The results are ranked and the designated number of attributes with the highest scores are returned [WF05].

2.1.1.2 Principal Components Analysis

$$C^{n\times n}=(\forall(c_{i,j})1\leqslant i\leqslant n\wedge 1\leqslant j\leqslant n, c_{i,j}=covariance(Attribute_i, Attribute_j))$$

Where:
n is the number of attributes
$Attribute_x$ is the x^{th} Attribute

Formula 5: Covariance Matrix [Smi02]

Principal Components Analysis transforms the input space from an n attribute input space to one represented by the patterns between the attributes. This transformation requires constructing an n by n covariance matrix as is defined in Formula 5. Eigenvectors and eigenvalues are then calculated for the covariance matrix. The eigenvectors are ranked by their eigenvalues where the highest eigenvalues are selected first [Smi02].

Note that for an input space with n attributes there will be an n by n covariance matrix which will produce n eigenvectors. Removing eigenvectors therefore reduces the number of dimensions in the input space while still giving every attribute some influence over the final classification.

2.1.1.3 ReliefF Attribute Selection

```
Set all weights W[A] = 0.0;
For i = 1 to Number_of_Instances do:
     R = Instance number i
     H = Find k nearest neighbors with same class value as R
     M = Find k nearest neighbors with different class value as R
    For A := 1 to Number_of_Attributes do:
         W[A] := W[A] + (difference(A,R,M) – difference(A,R,H)) / Number_of_Instances
End;

difference(A,R,X):
     number_different = 0;
    For j = 1 to k do:
          if Value(A,X[j]) != Value(A,R):
            number_different += absoluteValue(Value(A,R) - Value(A,X[j])) /
                          Max(Value(A,R), Value(A,X[j])) – Min(Value(A,R), Value(A,M[j]))
     return number_different / k;
End;
```

Figure 1: ReliefF Algorithm [RK97]

ReliefF attribute selection assigns a weight to each attribute based on how well that attribute is able to differentiate between nearby instances. The algorithm for ReliefF that is outlined in Figure 1 was originally presented in [Kon94]. For every instance, the k nearest neighbors of the same class value and the k nearest neighbors with different class values are found. These two groups of nearest neighbors are used to calculate the degree to which each attribute differentiates nearby instances. The weight for each attribute is adjusted by the difference between the value of the instance score and the value of its neighbor for that attribute. The weight is increased for neighbors having a different class and decreased for the neighbors that have the same class. Once the weights have been calculated, attributes are ranked from highest to lowest weights [Kon94].

2.1.1.4 Support Vector Machine Attribute Selection

This attribute evaluator uses the support vector machine (SVM) algorithm, discussed in the machine learning algorithms section, with a linear kernel function to evaluate the relevance of each attribute. Attributes with a greater influence over the final classification are assigned weights by the SVM algorithm that are further away from zero. The squares of these weights are used to rank the attributes [WF05].

2.1.2 Machine Learning Algorithms

Machine learning algorithms use training data to construct models of the relationships between a set of input attributes and a target attribute. These algorithms often are used for either regression or classification. Regression involves mapping of the input attributes to a numeric value while classification is the mapping of the input attributes to a nominal value. Algorithms designed for regression can often be modified to perform classifications.

2.1.2.1 ZeroR

ZeroR is the simplest of classifiers and can be thought of as the default classifier. It always predicts the majority target value. If multiple target values tie for majority value, it arbitrarily chooses one to predict [WF05].

2.1.2.2 Logistic Regression

Logistic regression is commonly used in the medical community to model relationships between sets of attributes. Its predictive power is trusted by the medical community so any other modeling approach used in this context should be compared against logistic regression to determine whether it either by constructs a more informative model or it increases predictive accuracy.

Basic logistic regression builds a model based on a set of attributes to predict the probability of

a binary classification target. A series of regression steps build a model where the inputs produce a probability representation the likelihood of the target value. For more complex classification targets, multiple logistic regressions can be performed and the resulting probabilities combined together to predict the most likely class [WF05].

2.1.2.3 Artificial Neural Networks

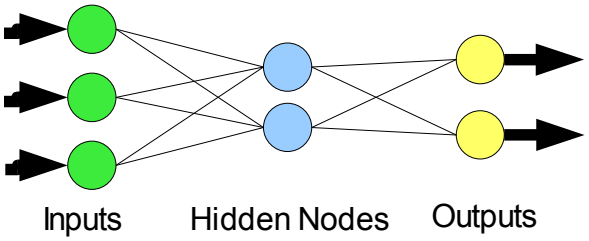

Figure 2: Example Artificial Neural Network

Artificial neural networks are a learning approach that builds a regression model as a series of interconnected computing nodes. The model used to generate a set of outputs from a set of inputs is crudely analogous to biological networks of neurons. Having multiple output nodes allows for classification by assuming the correct class is the output node with the highest regression value. An example of an artificial neural network is shown in Figure 2. This is an example of a feed-forward network, the artificial neural network architecture that is used in this thesis.

Feed-forward artificial neural networks are a learning approach that contain two or more layers of computing nodes. Every network has one layer of input nodes and one layer of output nodes. Most also have one or more hidden layers each with one or more hidden nodes. For both the input and the output layer of nodes, there is a node corresponding to every x_i in input vector X and to every y_j in output vector Y respectively. Hidden nodes increase the predictive power of neural networks by allowing for construction of more complex models.

The nodes in adjacent layers are fully inter-connected. Consider a network with n input nodes,

only one hidden layer with two hidden nodes, and one output node. In this network, there will be connections between all n input nodes and the two hidden nodes as well as between the two hidden nodes and the output node.

Each connection is assigned a weight which represents the strength of the connection. The major learning task involved with neural networks is the learning of these weights. The weights are trained by starting with small random weights and slowly changing them to reduce the network's error over the training data.

For all nodes, except the input nodes, a node's value is calculated by first finding the weighted sum of all its inputs. This value is then run through a sigmoid threshold unit to calculate the node's value [Mit97].

2.1.2.4 Bayesian Approaches

$$P(Target|InputVector)=\frac{P(InputVector|Target)\,P(Target)}{P(InputVector)}$$

Formula 6: Bayes' Theorem

Bayesian methods are based on Bayes' rule, detailed in Formula 6, which provides a way to calculate the probability of a classification target given the input vector from the probability of the classification target, the probability of the input vector, and the probability of the input vector given the classification target.

Bayesian networks use directed, acyclic, graphs to model the dependencies among variables. Variables are represented by nodes, each of which contains a table of probabilities. For every node, the table of probabilities provides the conditional probabilities of each of the variable's values, given each possible combination of the values of the variable's parents in the network.

One type of Bayesian Network is Naïve Bayes. Naïve Bayes assumes conditional independence

among all variables given the classification target. Although this approach loses conditional dependencies between variables, in practice it can be powerful for classification [RN03].

2.1.2.5 Decision Trees

Decision trees model a sequential set of choices that eventually result in a classification. Each choice represents an attribute, each possible value of that attribute leading to either another choice or a classification.

Construction of decision trees starts by selecting the attribute that is best able to increase the differentiation of the classes by splitting the attribute into all of its possible values. Each level of the tree is sequentially constructed in this manner based on the subset of training attributes with the values of the prior layers.

There are several metrics for the degree to which an attribute increases the differentiation of the classes including information gain and gain ratio which were discussed in the Gain Ratio Attribute Selection section, [Mit97].

2.1.2.6 Support Vector Machines

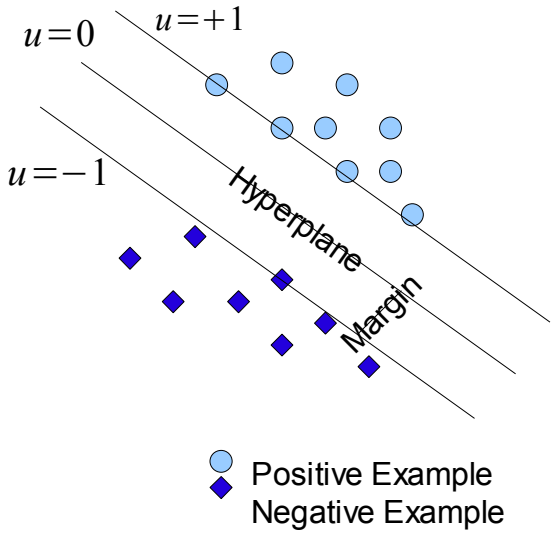

Figure 3: Support Vector Machine

Support Vector Machines, SVMs, are constructed by finding a hyperplane that divides the space of all possible instances into two classes. The hyperplane is constructed such that the two classes are maximally separated (i.e., the maximal margin is found). In Figure 3 there is an example of a one dimensional hyperplane separating the positive and the negative classes in a two dimensional input space.

A common problem in trying to find a hyperplane is that the two classes might not be linearly separable. SVMs use a kernel function to solve this problem by mapping the input space into one that is linearly separable by a hyperplane [CS00].

As an example, consider a two dimension input space divided by $f(x)=\ln(x)$. A line, a hyperplane in one dimension, can not be found that correctly separates this input space. However, if a kernel function $K(x)=e^x$ is used to transform the input space, $K(f(x))$ is clearly a hyperplane that separates the transformed positive and navigate instances when you consider that $K(f(x))=e^{\ln(x)}=x$.

2.1.3 Meta-learning

Meta-learning algorithms use multiple data mining techniques in constructing a model of the relationships between a set of input attributes and a target attribute.

[VD02] provides a literature review of several approaches to meta-learning and discusses the motivations for developing meta-learning algorithms.

2.1.3.1 Attribute Selected Classifier

The attribute selected classifier combines a feature selection algorithm and a machine learning algorithm into one classifier. This allows feature selection to be performed on only the training data, independent of the test data.

The attribute selected classifier first runs the specified feature selection algorithm over the training dataset. The selected features are then used in the construction of a model using the machine learning algorithm. The test data then uses the features selected by the training set to test this model [WF05].

2.1.3.2 Bagging

Bagging combines multiple models created from a machine learning algorithm for use in predicting a target. This can be particularly helpful in cases where there is only a small amount of training data available.

Bagging builds each of its models by randomly sampling a large subset of the training examples. It then uses a voting mechanism for classification where every model gets one vote towards the final prediction. This often results in a decreased variance [WF05].

2.1.3.3 Boosting

Boosting combines multiple models created from a machine learning algorithm for use in predicting a target. Boosting assigns every instance in the training set a positive weight. Initially all weights are the same. Models are built using the entire training dataset giving more influence to instances with higher weights. After each model has been constructed, the all instances that are misclassified by this model have their weights increased. Finally, models are combined by weighted voting. The accuracy of the model over the training data used for its construction is that model's weight in deciding the final classification. Boosting often is able to increase classification accuracy by focusing on the instances that are harder to classify. Focusing on the instances that are harder to classify might increase the problem of over fitting [WF05].

2.1.3.4 Stacking

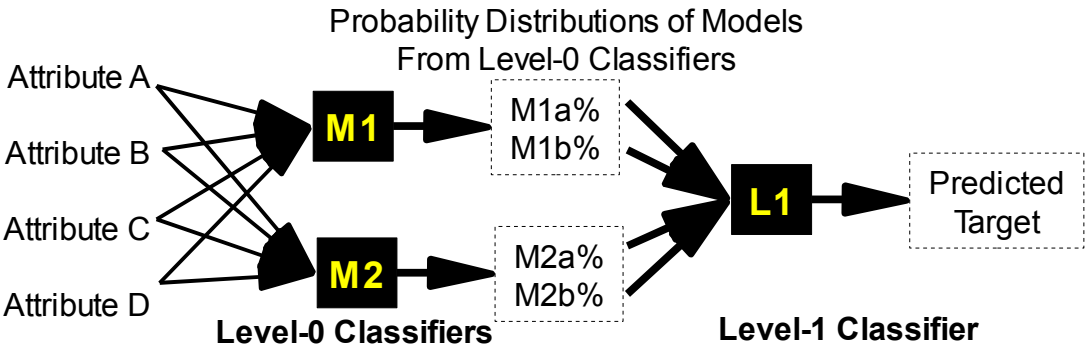

Figure 4: Stacking: Prediction of Unseen Instance

Stacking differs from Bagging and Boosting in that it combines the models constructed by several different machine learning algorithms when making a prediction about a target. Stacking combines the models generated by multiple machine learning algorithms using a classification meta-model. The machine learning algorithms are known as the level-0 algorithms and the meta-model is referred to as the level-1 model.

To train the level-1 model a dataset is constructed using cross validation, to be discussed in the Experimentation and Evaluation Techniques section, to generate the predicted target class probability distribution of each level-0 model for every instance in the training set. These probability distributions become the input attributes within the new dataset. Each instance in this new dataset keeps the target value of the instance that the probability distributions are generated from [WF05].

Figure 4 shows how the trained stacking model uses the level-0 and level-1 classifiers to predict an unknown target from a set of four input attributes. Note that this diagram assumes there are two possible target values, a and b. M1a% represents the probability of target value a predicted by M1 and M1b% represents the probability of target value b predicted by M1. Similarly for M2, M2a% represents the probability of target value a predicted by M2 and M2b% represents the probability of target value b predicted by M2.

Selection of the algorithm to construct the level-1 meta-learner has a large impact on the ability of stacking to increase the classification accuracy over that of the best level-0 model. A comparison of various algorithms for the level-1 model is presented in [DZ04].

2.1.4 Experimentation and Evaluation Techniques

The main goal of machine learning is to use training examples to construct a model that can make predictions about future examples. When constructing models, usually only a small subset of all possible examples are available. For this reason, constructed models are only an estimation of the actual distribution of the all possible examples.

We often want to be able to compare how well two machine learning algorithms are able to model the underlying distribution of the data when trained over the same subset of data. When, as is often the case, the underlying distribution is not fully understood there are several techniques that have been developed to estimate and compare this error [Mit97].

The methods discussed below are used to estimate the error of models built using machine learning algorithms and to compare the error of multiple models generated by different algorithms.

2.1.4.1 Cross Validation

Cross validation is a technique for estimating the error of a machine learning algorithm when used to model a dataset. It is a particularly helpful technique when only a small amount of data is available for both testing and training. Cross validation uses the entire dataset for both testing and training to improve error prediction over simply splitting the dataset into a subset for training and a subset for testing.

Cross validation works by randomly dividing a dataset with N instances into k bins of approximately N/k instances. The learning algorithm under consideration is then run k times, each

time using different bin for testing. The training set is composed of all of the instances not within the testing set each time.

Using cross validation results in every instance within the dataset being used once for testing and k-1 times for training. This gives us a good way to approximate the error of the machine learning model over the entire distribution of instances [WF05].

2.1.4.2 T-Test

Cross validation can be used to estimate the error of a machine learning algorithm's model of some unknown distribution. This still leaves the problem of how to compare two machine learning algorithms. This is the role of the t-test.

The t-test uses the standard deviation of the classification error of both algorithms which is calculated using the classification error and the number of attributes in the test set. Once standard deviation has been calculated, the two algorithms can be compared to determine if their difference in error is statistically significant at some confidence level. Note that the standard deviation decreases proportional to the square root of the size of the test set. This means to cut the standard deviation in half, the number of instances in the test set must be increased four fold [WF05].

2.1.4.3 ROC Curves

Receiver Operating Characteristic Curves, ROC Curves, are a way to evalutate machine learning algorithms. ROC Curves visually plot the rate of true positive classifications versus false positive classifications. This allows for a trade off to be made between the rate of false positives and the rate of true positives.

ROC curves are constructed by varying an underlying threshold parameter within the model constructed by the machine learning algorithm. One starts with the threshold set such that both the

number of true and false positives are zero (i.e., all instances are classified the negative class). This threshold is slowly incremented until a threshold is reached where the percent of false and true positives is one hundred percent (i.e., all instances are classified the positive class).

There are two main ways to compare ROC curves. The first useful metric when looking at ROC curves is the area under the ROC curve. The greater the percentage of the area that is under the ROC curve, the better the model. This number is expected to be at least 50%. The second way to compare ROC curves is to look for a point with a low false positive rate and a high true positive rate. This can be used to set a model, and a threshold, that best balances this trade off [Faw03].

2.2 Machine Learning in The Medical Domain

Over the past decade there has been an increase in the work done on applying machine learning algorithms to the medical domain. Artificial neural networks have been used to model survival in colon cancer [Ahm05], colorectal cancer [BCD05], and breast cancer [BGR97]. These papers all demonstrated that artificial neural networks can improve predictions about patient survival over traditional techniques. Bayesian networks have been used to identify the malignancy of breast cancer in [KRS97].

[Ahm05] discusses both the advantages and the disadvantages of artificial neural network techniques when applied to the medical domain. Among the advantages are their ability to model dependencies among attributes by minimizing assumptions about the underlying attributes prior to training. It goes on to credit such algorithms with the ability to reduce the number of false positives and with having greater classification power than regression algorithms. The major disadvantages that were discussed include being a 'black box' which produces result using an underlying model that is very difficult for a human to interpret; that they are often more difficult to use in the field due to the high computational cost of training; and that they are prone to overfitting the training data.

[KRS97] compares the performance of Bayesian networks and artificial neural networks, which the article reports as having been more commonly used in the past, on a dataset of breast cancer patients. Bayesian networks are found to be favorable for this classification as there is not a reliance on an incremental training process. This eliminates the problem of getting stuck in a locally optimal solution and reduces the problem of over-fitting the training data. Bayesian networks are also presented as constructing models that are easier for a human to understand than models constructed by artificial neural networks.

3 Our Approach

3.1 Data Mining Tool

The primary data mining tool used in this thesis is Weka [WF05]. Within Weka are Java implementations of many data preprocessing, machine learning, and meta-learning algorithms. Weka's code is licensed under the GNU General Public License Version 2 so the code is freely available. This allows the algorithms to be modified as needed and for exploration into the implementation details of each machine learning technique.

3.2 Research Question

As was presented in the introduction, decisions about the surgical removal of a tumor require consideration of the patient's expected survival time. For this reason, survival time is central to our research question:

- What is the expected survival time of the patient?

For this question, we have built models using a variety of machine learning algorithms.

3.3 Data

3.3.1 Source

The pancreatic cancer data has been collected from The University of Massachusetts Medical School. This sample includes the patients who were evaluated for surgical removal of a pancreatic tumor by the Department of Surgical Oncology between April 2002 and December 2005. John Hayward's thesis, [Hay06], started the collection and analysis of this data. Within the dataset there are 190 attributes for each patient including attributes relating to the patient's cancer diagnosis, symptoms at diagnosis, relevant past history, family cancer history, lab and imaging scores, treatment, and

survival. A summary of the categories of attributes and the number of attributes in each category is presented in Figure 5. Note that each of the groups of attributes are divided into three meta- categories: the pre-operative attributes, the peri-operative attributes, and the target attribute. The full list of attributes can be found in Appendix A.

Category	Number of Attributes	Description of Category
Pre-Operative Attributes		
Patient	6	Biographical information on patient
Presentation	21	Initial symptoms/information about patient at diagnosis
History	27	Patients past medical history, includes past cancer history
Serum	8	Lab scores of several tumor markers
Diagnostic Imaging	23	Details of imaging scans of patient
Endoscopy	25	Details of endoscopy of patient
Preliminary Outlook	1	Doctor's pre-surgical evaluation of patient
Total	**111**	
Peri-Operative Attributes		
Treatment	36	Details of treatment patient received
Resection	24	Details on surgical removal of tumor
Pathology	7	Details of tumor type after surgical removal
No Resection	11	Details on why tumor was not removed
Total	**78**	
Target Attribute		
Survival	**1**	Time from diagnosis to death
Grand Total	**190**	

Figure 5: Categories of Attributes in Survival Dataset

When a patient stops showing up for follow up appointments we can not simply assume the patient has died. There are many possibilities including that the patient has died, but it could also be that they are doing so well that they feel no need to keep coming back or have moved to a different part of the country. Death dates therefore are gleaned from many sources prior to being added to a patient's record. This makes collection of survival information challenging. Of the ninety seven patients we currently have records for, we only have death dates for sixty. It is these sixty patients that compose the dataset for this thesis.

3.3.2 Preprocessing

Survival time for all patients is recorded as the number of months between when the patient was

first seen at the hospital and the patient died. Overall, this attribute has a mean value of 10 months and a median value of 9 months.

Over all of the 60 patients and the 190 attributes there are 2,299 missing values. This means that 20% of the attributes are missing values. We rely on the machine learning algorithms we use to handle missing attributes and do no preprocessing to replace them.

With the help of the doctors at UMass we chose to construct a dataset by splitting the patients into three groups, each with an equal number of patients: those who survived less than six months, survived six months to twelve months, and survived more than twelve months. The rational for these categories is if, even with surgery, the expected survival time of the patient is less than six months the surgery is not worth performing. If the expected survival time is over twelve months then surgical removal of the tumor is more appropriate.

Models constructed using this dataset can be compared using a t-test but can not be compared using ROC Curves as they require a binary target. For this reason we constructed two additional datasets, one split at the median of nine months and the other split at six months.

In addition to predicting expected survival over the full dataset, we also wish to predict survival based only on the information known prior to surgery. For this reason we construct three pre-operative datasets by removing all of the peri-operative attributes, the attributes related to surgery, from each of the three datasets already created.

Figure 6 presents an overview of the six datasets constructed. The number of instance with each of the given target values is listed in the table. Note from Figure 5 that the dataset with all attributes has 190 attributes and the pre-operative dataset has 112 attributes.

Dataset	Target	Number of Instances with Target Value
All Attributes: Six and Twelve Month Split		
	<6 Months	20
	6-12 Months	20
	>12 Months	20
All Attributes: Nine Month Split		
	<9 Months	30
	>9 Months	30
All Attributes: Six Month Split		
	<6 Months	20
	>6 Months	40
Pre-Operative Attributes: Six and Twelve Month Split		
	<6 Months	20
	6-12 Months	20
	>12 Months	20
Pre-Operative Attributes: Nine Month Split		
	<9 Months	30
	>9 Months	30
Pre-Operative Attributes: Six Month Split		
	<6 Months	20
	>6 Months	40

Figure 6: Summary of Datasets Constructed

3.4 Experimental Approach

3.4.1 Baseline Algorithms

Ideally we would compare the accuracies of models constructed using machine learning algorithms with the doctor's predictions of the patients expected survival. Unfortunately our UMass collaborators know these patients well enough to identify many of them individually from this dataset. Therefore, asking them to predict an patient's survival given a (sub-)set of attributes from this dataset would likely not be representative of their ability to predict the expected survival of a new patient.

For this reason we must rely on other approaches to be the baseline algorithms for comparison with the models constructed using machine learning algorithms. ZeroR is an obvious benchmark algorithm as any we expect any useful algorithm to be able to predict a new patient's survival better than simply guessing the most likely choice. Logistic regression is trusted by the medical community and thus should be compared with any more advanced algorithm considered. Therefore, our two

benchmark algorithms are logistic regression and ZeroR.

3.4.2 Considerations

1. Feature Selection: For each patient in the dataset we have a large number of attributes and thus a highly dimensional feature space.
    - Various feature selection algorithms are evaluated to determine which selects the most relevant set of features.
2. Algorithm Consideration: Logistic regression is currently understood and trusted by doctors for use in predictive classification but more advanced algorithms exist that could improve classification accuracy.
    - Logistic regression and ZeroR give us benchmarks with respect to which the machine learning algorithms we use are compared.
3. Number of Patients: There are only a small number of patients within the database.
    - We evaluate the use of Meta-Learning algorithms to improve classification accuracy and reduce variance.

3.4.3 Experiments

Experiments are be run over all six datasets discussed in the data preprocessing section. In designing the experiments, each of the above consideration will be approached in the following manner:

1. We use the Attribute Selected Classifier to evaluate models built with different machine learning algorithms using the features selected by various feature selection algorithms. Specifically we investigate the use of the Gain Ratio, Principal Components, ReliefF, and Support Vector Machines for feature selection. All of these algorithms rank the most important

features so we run these algorithms several times, varying the condition on the number of features to return. Through these experiments we attempt to determine the optimal feature selection approach for a given machine learning algorithm.

2. We compare the baseline algorithms with several other machine learning algorithms, including artificial neural networks, Bayesian networks, decision trees, naïve Bayes networks, and support vector machines. For each dataset, we find the best combination of feature selection and machine learning algorithm. We compare these combinations with ZeroR and logistic regression.

3. We attempt to improve the classification accuracy by experimenting with both bagging, boosting, stacking, and our model selector, our own meta-learning algorithm. Our primary focus is on improving the classification accuracy by combining multiple machine learning models generated by the best pairs of feature selection and machine learning algorithm. We also investigate the use of bagging to decrease variance.

3.5 Design of Our Model Selector Meta-Classifier

The design of our model selector classifier is motivated by wanting to find subsets of instances that are predicted better by one machine learning model over another model. This is visually depicted in Figure 7 by showing an instance space with two models that each correctly cover only a subset of the instances. If we could correctly predict which model to use for each instance, this would increase the overall classification accuracy. A meta-model is used to learn the subsets of instances that will be best predicted by each model.

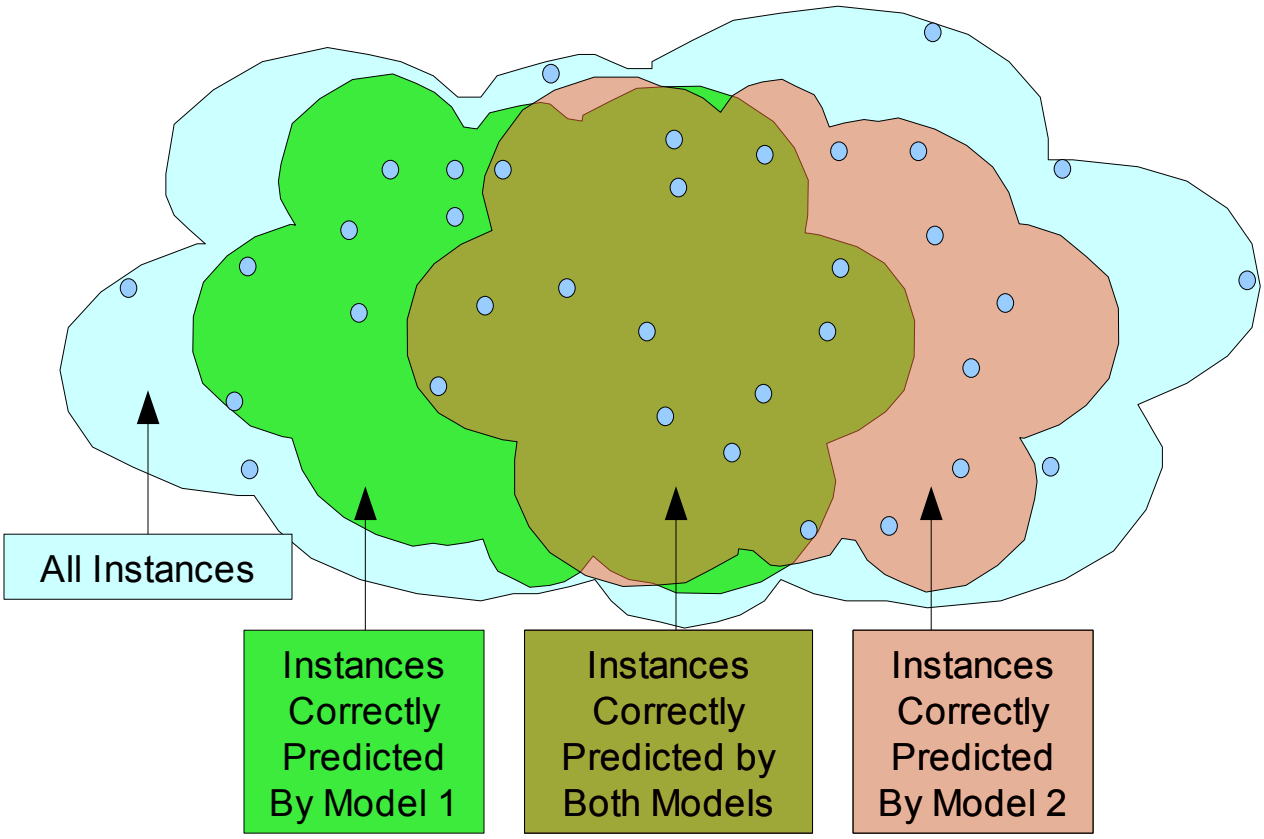

Figure 7: Our Model Selector: Motivation

Our model selector classifier is similar to stacking in that it uses the level-0 models constructed by several different machine learning algorithms to improve overall accuracy. The key difference between these two meta-learning algorithms is the function of their level-1 meta model. Stacking's level-1 classifier combines the target class probability distributions generated by running the unseen instance through each of the level-0 models while our model selector's level-1 classifier selects which of the level-0 models will perform best over the given test instance.

To train the level-1 model, a dataset is constructed using cross validation to determine which of the level-0 models does the best job of predicting the correct target of every instance in the training set. The best model is the one with the highest value for the actual target in its probability distribution. A new dataset is constructed from the original dataset by replacing the target of the original dataset with information about which model is best. The algorithm to train our model selector is outlined in Figure

8.

Inputs: Set of input instances
Set of Level-0 classifiers to use
Level-1 classifier to use
Output: Level-1 model predict which level-0 model is best
Level-0 models
Construct a new empty set of instances A that will be used to train the level-1 classifier
Divide input instances into k equal groups B_1, ... , B_k
Repeat for each group B_n**:**
Construct training set C composed of all k groups in B_1, ... , B_k except B_n
Train each level-0 classifier using set C
For each instance D in B_n:
Run each level-0 classifier on D - each will output a probability distribution of the target values
Among the output distributions, select one with highest probability of D's target value
Copy instance D, replacing target with the identifier of level-0 classifier with selected distribution
Add new instance to A
Train level-1 classifier using set of instances A
Rebuild each level-0 classifier over all original training instances

For a new instance:
run new instance through level-1 classifier
run new instance through level-0 classifier recommended by level-1 classifier

Figure 8: Our Model Selector: Algorithm

An example to illustrate the construction of the dataset to train the level-1 model is shown in Figure 9. The example patient in the figure with the given medical history as the input attributes is known to fall into the positive class. When M1 uses this set of symptoms to predict that there is a ninety percent chance that this patient is in the positive class and M2 predicts that there is a seventy percent chance that this patient is in the positive class. Since M1 has the highest confidence in the correct classification, this is the model said to best predict this instance. A new instances is then created using the patient's medical history as the input attributes and M1 as the target value. This new instance is added to the dataset used to train the level-1 model.

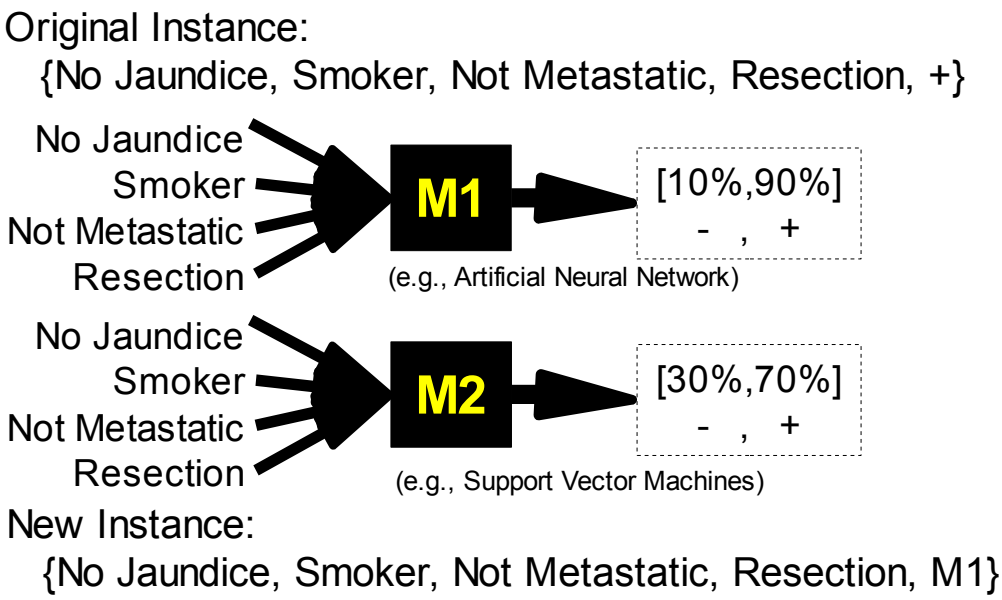

Figure 9: Our Model Selector - Construction of Dataset to Train Level-1 Model

Figure 10 shows how the trained stacking model uses the level-0 and level-1 classifiers to predict an unseen target from a set of four input attributes. Note that after the level-1 classifier predicts which model will perform best, the input attributes are run through the chosen model to make a prediction.

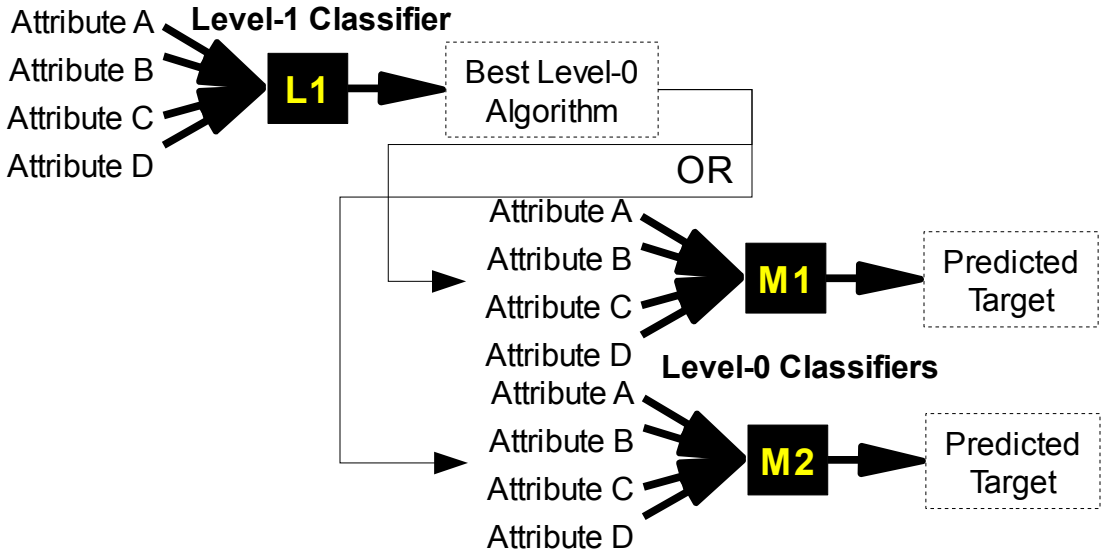

Figure 10: Our Model Selector: Prediction of Unseen Instance

4 Experimental Evaluation

The classification accuracy for all experiments is calculated by running ten repetitions, each repetition with a different initial random seed, of ten fold cross validation. Every set of experiments is run separately over each of the six datasets discussed in the data preprocessing section. These six datasets are summarized in Figure 6 on page 24.

We start off by finding for each dataset the combinations of feature selection and machine learning algorithm that result in the highest classification accuracies. The models with the highest classification accuracy are selected for further comparison. The classification accuracy of these models is compared to the classification accuracy of a model built using logistic regression. The t-test procedure implemented in Weka is be used for this comparison. We look for statistical significance at a $P<0.05$ level. The feature selection techniques used in this comparison, along with the details of the parameters used, are presented in Figure 11. The machine learning algorithms used in this comparison, along with the details of the parameters used, are presented in Figure 12.

Feature Selection Algorithm	Search Method	Parameters
GainRatioAttributeEval	Ranker	Missing Merge: True
PrincipalComponents	Ranker	Maximum Attribute Names: 5
		Normalize: True
		Transform Back To Original: False
		Variance Covered: 0.95
ReliefFAttributeEval	Ranker	Num Neighbors: 10
		Sample Size: -1
		Seed: 1
		Sigma: 2
		Weight By Distance: False
SVMAttributeEval	Ranker	Atts To Eliminate Per Iteration: 1
		Complexity Parameter: 1
		Epsilon Parameter: 1.0E-25
		Filter Type: Normalize Training Data
		Percent Threshold: 0
		Percent to Eliminate Per Iteration: 0
		Tolerance Parameter: 1.0E-10
Note that all feature selection run using the AttributesSelectedClassifier.		

Figure 11: Summary of Feature Selection Used

Machine Learning Algorithm	Parameters
ZeroR	n/a
Logistic Regression	Max Its: -1
	Ridge: 1.0E-8
SMO with Kernel of 0.9	Build Logistic Models: False
	C: 1.0
	Checks Turned Off: False
	Epsilon: 1.0E-12
	Filter Type: Normalize Training Data
	Kernel: PolyKernel -C 250007 -E 0.9
	Num Folds: -1
	Random Seed: 1
	Tolerance Parameter: 0.0010
SMO with Kernel of 0.9	Same as SMO with Kernel of 0.9 except:
	Kernel: PolyKernel -C 250007 -E 1.0
ANN with 1 Hidden Unit	GUI: False
	Auto Build: True
	Decay: False
	Hidden Layers: 1
	Learning Rate: 0.3
	Momentum: 0.2
	Nominal To Binary Filter: True
	Normalize Attributes: True
	Normalize Numeric Class: True
	Random Seed: 0
	Reset: True
	Training Time: 2000
	Validation Set Size: 0
	Validation Threshold: 20
ANN with 2 Hidden Units	Same as ANN with 1 Hidden Unit except:
	Hidden Layers: 2
Naïve Bayes	Use Kernel Estimator: False
	Use Supervised Discretization: False
J4.8	Binary Splits: False
	Confidence Factor: 0.25
	Min Num Obj: 2
	Num Folds: 3
	Reduced Error Pruning: False
	Save Instance Data: False
	Subtree Raising: True
	Unpruned: False
	Use Laplace: False
Bayesian Network: 1 Parent	Estimator: SimpleEstimator -A 0.5
	Search Algorithm: K2 -P 1 -S BAYES
	Use AD Tree: False
Bayesian Network: 2 Parents	Same as Bayesian Network: 1 Parent except:
	Search Algorithm: K2 -P 2 -S BAYES

Figure 12: Summary of Machine Learning Algorithms Used

Once we find the best combinations of feature selection and machine learning algorithm, we run further experiments using meta-learning algorithms. Bagging is evaluated on each of these combinations as a method to reduce the standard deviation of the classification accuracy, as decreasing the standard derivation may increase the statistical significance. Boosting is evaluated to increase the classification accuracy of each combination. Stacking and our model selector are used to combine the models constructed using the best combinations of feature selection and machine learning algorithm for each dataset. The goal of experimenting with stacking and bagging is that the combined models tend to have a higher classification accuracy than the original models. The parameters used for these meta-learning algorithms are detailed in Figure 13.

Meta Learning Algorithm	**Parameters**
Bagging	Bag Size Percent: 100
	Calc Out Of Bag: False
	Classifier: REPTree -M 2 -V 0.0010 -N 3 -S 1 -L -1
	Num Iterations: 10
	Seed: 1
Boosting: AdaBoostM1	Classifier: DecisionStump
	Num Iterations: 10
	Seed: 1
	Use Resampling: False
	Weight Threshold: 100
Stacking	Num Folds: 10
	Seed: 1
Our Model Selector	Num Folds: 10
	Seed: 1

Figure 13: Parameters Used For Meta-Learning Algorithms

Several combinations of machine learning algorithm and feature selection are evaluated for use as stacking and our model selector's level-1 learning method. The combinations evaluated for stacking are shown in Figure 14. The combinations evaluated for our model selector are shown in Figure 15. There are two algorithms evaluated for use as level-1 models that are not part of our overall investigation into the best combination of feature selection and machine learning algorithm so their parameters are not shown in Figure 12. These two machine learning algorithms are listed, with the

parameters used, in Figure 16.

Machine Learning Algorithm	Feature Selection Algorithm
ANN 1 Hidden Unit	n/a
ANN 2 Hidden Units	n/a
Linear Regression	n/a
LWL	n/a

Figure 14: Algorithms For Level-1 Model: Stacking

Machine Learning Algorithm	Feature Selection Algorithm
ANN 1 Hidden Unit	n/a
ANN 2 Hidden Units	n/a
ANN 1 Hidden Unit	ReliefF to Select 90
J48 & SVM_70	SVM to Select 70
Logistic	Principal Components to Select 70
NaiveBayes	Gain Ratio to Select 30
SMO with Kernel of 1.0	ReliefF to Select 40
J48	n/a
Logistic	n/a
LWL	n/a
NaiveBayes	n/a
SMO	n/a

Figure 15: Algorithms For Level-1 Model: Our Model Selector

Machine Learning Algorithm	Parameters
LWL	KNN: -1
	Classifier: DecisionStump
	Nearest Neighbor Search Algorithm: LinearNN -A EuclideanDistance
	Weighting Kernel: 0
Linear Regression	Attribute Selection Method: M5 Method
	Eliminate Colinear Attributes: True
	Ridge: 1.0E-8

Figure 16: Parameters for Machine Learning Methods Used Only to Train Level-1 Models

For the four datasets with a binary target, the combinations of feature selection and machine learning algorithm with the highest classification accuracies, along with logistic regression with no feature selection, are further compared using ROC curves. In comparing the ROC curves we look for models that both maximize the area under the curve and that are able to reach a high true positive rate while still maintaining a low false positive rate.

4.1 Results for Full Dataset with Six and Twelve Month Split

The dataset discussed in this section has all 190 attributes. The sixty patients in this dataset are split evenly into three groups based on the target of survival. These groups are are <6 month, 6-12 month, and >12 month survival.

4.1.1 Machine Learning Algorithms with No Feature Selection

Figure 17 shows the classification accuracies of models constructed using several different machine learning algorithms run over the dataset with a target split into three groups of zero to six, six to twelve and more than twelve month survival. The model constructed using a Bayesian network with two parents, highlighted in the figure, has the highest classification accuracy. This model is statistically significantly better ($p<0.05$) than ZeroR but not statistically different from logistic regression.

Machine Learning Algorithm	Classification Accuracy	Compare To ZeroR
ZeroR	33.3	No Statistically Significant Difference
Logistic Regression	41.7	
SMO with Kernel of 0.9	40.2	
SMO with Kernel of 1.0	39.8	
ANN with 1 Hidden Unit	40.0	
ANN with 2 Hidden Units	42.8	
Naïve Bayes	40.8	
J4.8	46.0	
Bayesian Network: 1 Parent	43.3	
Bayesian Network: 2 Parents	47.5	Better Than ZeroR

Figure 17: No Feature Selection: Six and Twelve Month Split

4.1.2 Combinations of Feature Selection and Machine Learning Algorithm

The graphs that follow show how the classification accuracies of the models built to predict a patient's expected survival time vary based on the feature selection algorithm used and the number of features selected. These models are constructed over the varying feature selection algorithms using several different machine learning algorithms. We use these graphs to find the combinations of feature

selection and machine learning algorithm that have the highest classification accuracy for this dataset.

Note that the highest classification accuracy obtained by constructing models with no feature selection is 47.5%. We look to use feature selection to improve this classification accuracy.

Figure 18 shows how varying the number of attributes selected by gain ratio attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 30. When 30 attributes are selected there are several models with classification accuracies above 47.5% including ones constructed using artificial neural networks with one hidden unit, artificial neural networks with two hidden units, Bayesian networks with one parent, and logistic regression. After this peak at 30 attributes most models show a gradual decrease in classification accuracy as the number of attributes selected increases. The notable exception to this is the model constructed using artificial neural networks with one hidden unit. The classification accuracy of this model continues to increase until it peaks when 60 attributes are selected.

Figure 19 shows how varying the number of features selected by principal components effects the classification accuracy of several machine learning algorithms. There is only a small overall increase in the classification accuracies as the number of features selected is increased. There are two of the algorithms that are able to construct models with classification accuracies above 47.5 percent when between 15 and 25 features are selected. This is the peak number of features to select with these two algorithms as when more or less features are selected, the classification accuracy decreases. These two algorithms are artificial neural networks with two hidden units and logistic regression.

Figure 20 shows how varying the number of features selected by ReliefF attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 40. Once 40

attributes have been selected, the classification accuracies remain relatively constant. This varies by the machine learning algorithm, however. Models constructed using artificial neural networks with two hidden units, for example, reach a peak accuracy that is over 47.5% at 30 attributes selected after which increasing the number of attributes selected decreases the classification accuracy. Artificial neural networks with one hidden unit show a slightly different behavior by reaching a classification accuracy of over 47.5% that only varies slightly as the number of attributes selected increases beyond 40 attributes.

Figure 21 shows how varying the number of features selected by support vector machine attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 100. This increase in classification accuracy is most obvious with the Bayesian network constructed using two parents which jumps twenty percentage points between when 40 and 60 attributes are selected. It takes 80 attributes to be selected before the models constructed by this algorithm consistently remain above 47.5%.

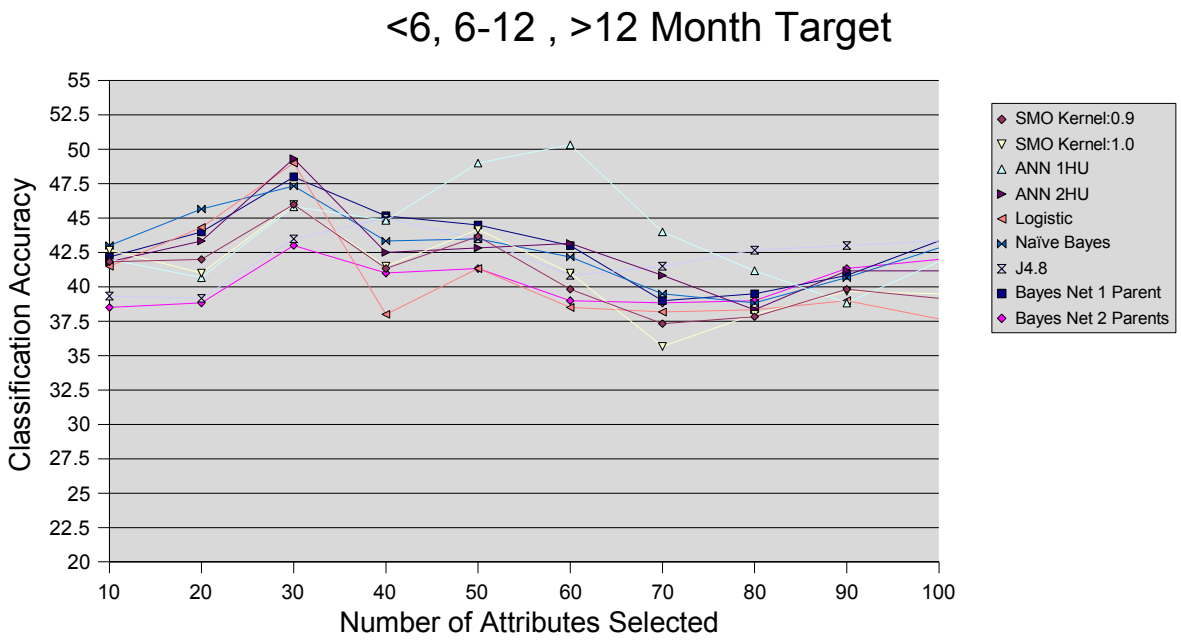

Figure 18: Gain Ratio Attribute Selection: Six and Twelve Month Split

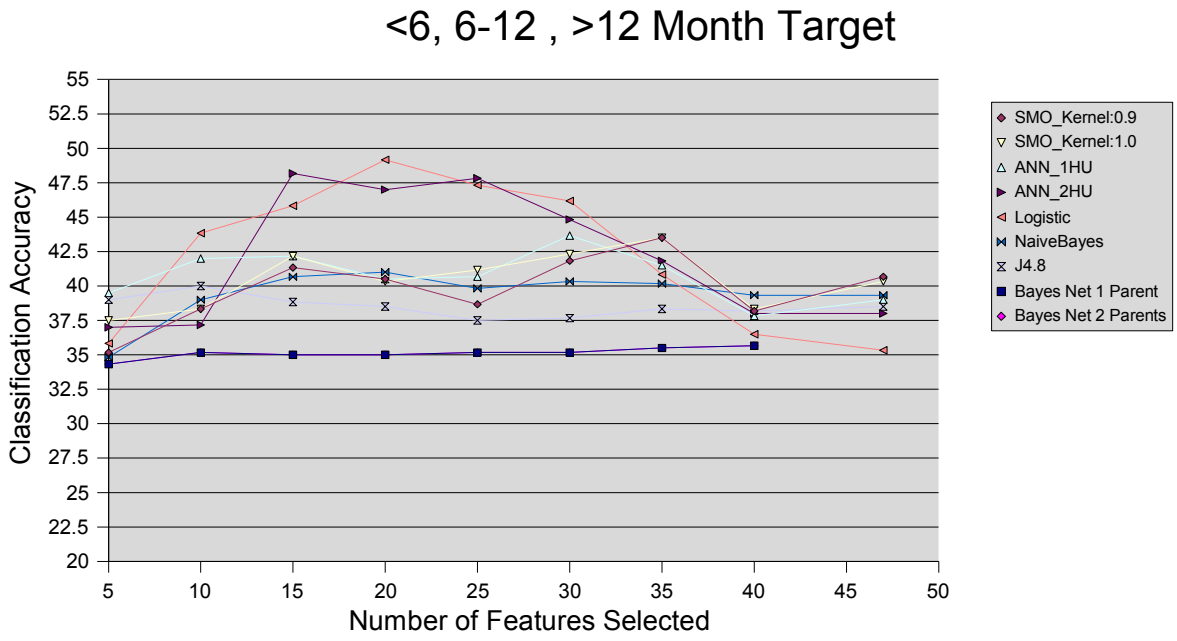

Figure 19: Principal Components: Six and Twelve Month Split

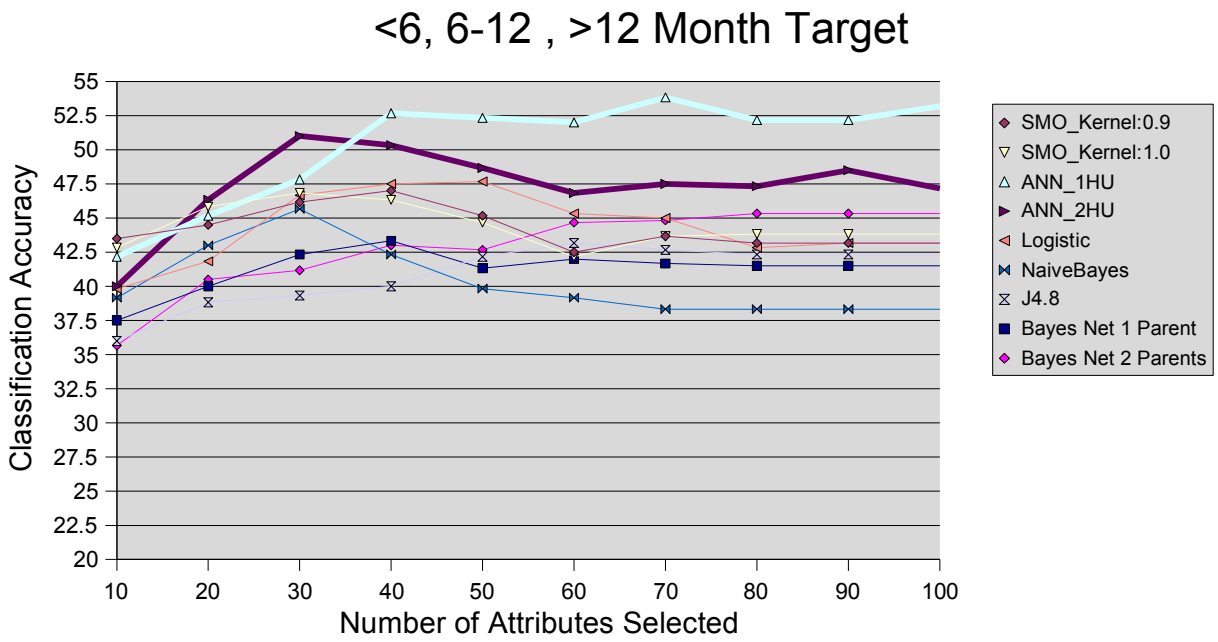

Figure 20: ReliefF Attribute Selection: Six and Twelve Month Split

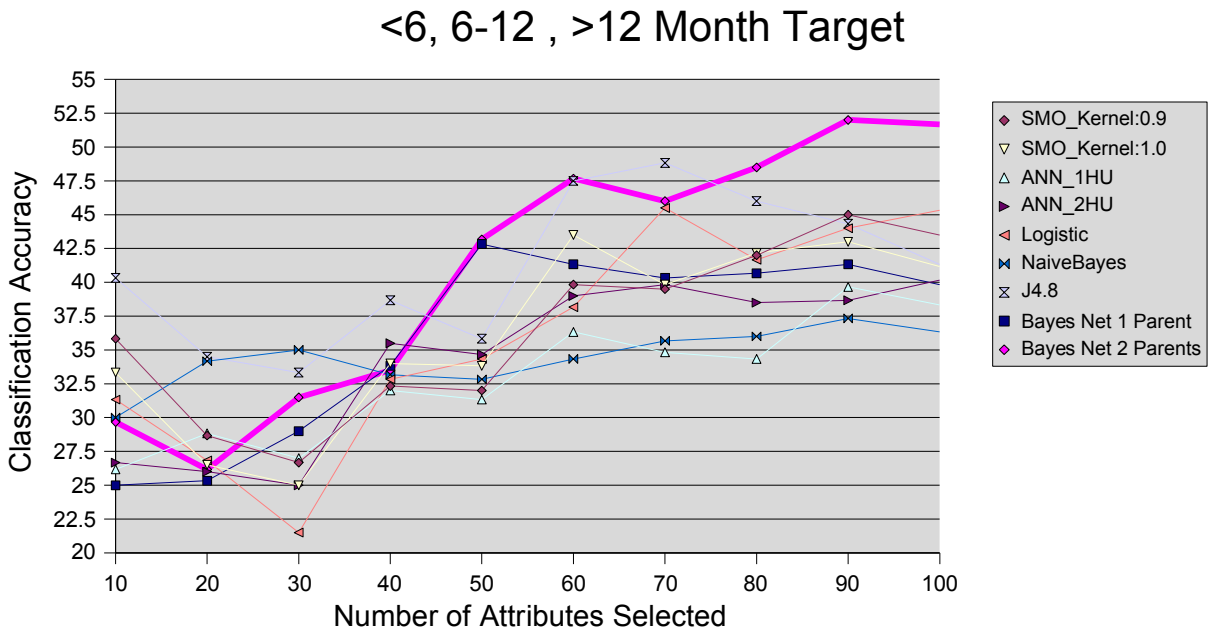

Figure 21: Support Vector Machine Attribute Selection: Six and Twelve Month Split

4.1.2.1 Baseline Models

Figure 17 shows that over this dataset the classification accuracy of logistic regression with no

feature selection is 41.7% and that of ZeroR is 33.3%. There is no statistically significant difference between logistic regression and ZeroR.

4.1.2.2 1st Noteworthy Combination: Artificial Neural Network with One Hidden Unit

The highest classification accuracy obtained over this dataset is 53.8% resulting from a model constructed using artificial neural networks with one hidden unit trained over 2,000 epochs. The top 70 attributes are selected to build this model using ReliefF attribute selection. Figure 20 shows this combination of feature selection and machine learning algorithm. This figure shows that once 40 attributes have been selected by ReliefF to build a model using this algorithm, there is only a small amount of variation in the classification accuracies of the resulting models by increasing the number of attributes selected. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.1.2.3 2nd Noteworthy Combination: Bayesian Network

The second highest classification accuracy obtained over this dataset is 52.0% resulting from a model constructed using a Bayesian network with K2 using a maximum of two parents. The top 90 attributes are selected to build this model using support vector machine attribute selection. Figure 21 shows this combination of feature selection and machine learning algorithm. This algorithm does not appear to stabilize in the the same way as observed around the model with the highest classification accuracy. Instead, Figure 21 shows a gradual increase in classification accuracy of models built using this Bayesian network algorithm as a greater number of attributes are selected using support vector machines for feature selection. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine

learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.1.2.4 3rd Noteworthy Combination: Artificial Neural Network with Two Hidden Units

The third highest classification accuracy obtained over this dataset is 51.0% resulting from a model constructed using artificial neural networks with two hidden units trained over 2,000 epochs. The top 30 attributes are selected to build this model using ReliefF attribution selection. Figure 20 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a peak classification accuracy when 30 attributes are selected using ReliefF where if more or less attributes are selected the classification accuracy decreases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.1.2.5 Summary of Noteworthy Combinations

- 53.8%: Artificial Neural Networks, One Hidden Unit, using ReliefF to select 70 attributes
- 52.0%: Bayesian Network, Two Parents, using Support Vector Machines to select 90 attributes
- 51.0%: Artificial Neural Networks, Two Hidden Units, using ReliefF to select 30 attributes

Note no statistically significant difference ($p<0.05$) between these three models.

4.1.3 Models Produced

4.1.3.1 Best Model with No Feature Selection

The best model with no feature selection resulted from a model constructed using a Bayesian network with a maximum of two parents. Since the second noteworthy combination is also a Bayesian network with a maximum of two parents, constructed with only 90 attributes selected by support vector

machines, and the second noteworthy combination has a higher classification accuracy we will not discuss the underlying Bayesian network model constructed without feature selection.

4.1.3.2 1st Noteworthy Combination: Artificial Neural Network with One Hidden Unit

The best combination of feature selection and machine learning algorithm resulted from a model constructed using artificial neural networks with one hidden unit trained over 2,000 epochs. The top 70 attributes are selected to build this model using ReliefF attribute selection.

4.1.3.2.1 Feature Selection

These 70 attributes, and the weights assigned to them by ReliefF, are listed in Figure 22. Note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

ReliefF Weight	Attribute Name		ReliefF Weight	Attribute Name
0.104	SurOncName		0.009	ResPOInfection
0.073	ResPODischStatus		0.008	CxPriorCancerSurgery
0.070	CxDiab		0.007	RadOncName
0.057	SHCigarette		0.007	ResTransfusion
0.047	SxOT		0.007	ResBloodLoss
0.045	EUSVascOmit		0.007	CTOtherNode
0.043	SxPru		0.007	ResPONG
0.038	PTCDx		0.007	ResPOLeak
0.038	PTCStent		0.006	ResPathR
0.038	TxChemo		0.006	EUSSMV
0.037	ResPOPulmComp		0.006	CxDiabDiet
0.033	Histology		0.005	CxPriorCancerRadiation
0.029	MedOncName		0.005	CxPriorCancer
0.027	SxBack		0.005	ResPODays
0.027	DemHeight		0.005	LabALT
0.024	EUSTumorSizeX		0.004	EUSCeliacNode
0.024	TxResect		0.003	CTTumorSizeX
0.023	TxPal		0.002	DemWeight
0.021	TxRadia		0.002	EUSStagingN
0.019	TxChemoGem		0.002	FamilyMotherDx
0.018	EUSNoNode		0.002	SxInd
0.018	SxSatiety		0.002	CxIHD
0.017	SHAlcohol		0.002	SxBC
0.016	EUSTumorSizeY		0.002	FamilyOther2
0.016	ResTFFP		0.001	CTPortalClass
0.015	CTNodeOmit		0.001	CxDiabOnset
0.015	EUSOtherNode		0.001	LabBili
0.014	LabAmylase		0.001	SxCCS
0.013	CxPriorCancerChemo		0.001	SxFati
0.012	ResPOCourse		0.001	TxPalRad
0.011	EUSDx		0.001	SxChola
0.009	SxWtloss		0.001	SxVom
0.009	ERCPDx		0.001	EUSStagingT
0.009	PreOutlook		0.001	FamilyOther1
0.009	TxChemoFlu		0.000	EUSCeliacClass

Figure 22: Top 70 Attributes Selected By ReliefF

Figure 23 shows how the progression of weights assigned by ReliefF decreases. The first ten features selected have weights assigned by ReliefF that quickly decrease from 0.1 to 0.040. The remaining sixty weights slowly level off as they approach zero. For the range between ten attributes selected and 30 attributes selected the rate of decrease in the ReliefF score is a somewhat consistent 0.0015 per additional feature selected. This rate levels off after 30 attributes have been selected with a slower decrease as the weight approaches zero. The weight for EUSCeliacClass, the seventieth

attribute selected which relates to details of celiac disease diagnosed through an endoscopic ultrasound, is $4.43 * 10^{-16}$ or very close to zero.

Figure 23: ReliefF Weights in Decreasing Order

The progression of weights displayed in Figure 23 indicate that after 30 attributes are selected, ReliefF anticipates very little additional information from additional attributes. This corresponds closely to Figure 20 where we can see that there is very little overall change in the classification accuracy of models generated using more than 30 attributes.

4.1.3.2.2 Machine Learning Model

The artificial neural network constructed over these attributes, with all but the the first 10 attributes omitted for readability, is shown in Figure 24. Artificial neural networks are always a challenge to decipher, this one is no different. For the input nodes, each non-binary attribute is split by its attribute values. For example, the attribute SurOncName, the name of the patient's primary cancer doctor, is split on each of the doctor's names. The first three sigmoid units, node 0, node 1, and node 2, represent the classification target values. The output node that results in the highest value is the target value that is predicted by the model.

For each output node, the combination of its weighted connection to node 3 and its threshold allows for interpretation of the which values of node 3 will trigger each of the target values. Node 3 with a very low negative value results in a classification of <6 months. Node 3 with a very high positive value results in a classification of >12 months. A value of closer to zero results in a classification of 6-12 months. Therefore, the weights that are positive between the input nodes and node 3 pull the classification towards predicting the patient will have an increased expected survival time while weights that are negative will decrease expected survival time.

Many of the weights can be interpreted accordingly. For instance, smoking is known to be a major contributing factor to pancreatic cancer [DHR05]. The weight between smoking, SHCigarette, and node 3 is negative indicating that smoking reduces the patient's expected survival. Similarly one would expect that a patient who is discharged to their home, ResPODischStatus=Home, to have a higher expected survival time than a patient who died in the hospital, ResPODischStatus=Died_in_Hospital. This is represented in this network by ResPODischStatus=Home having a positive weight and ResPODischStatus=Died_in_Hospital having a negative weight.

```
Classifier Model
Sigmoid Node 0          (<6 Months)
  Inputs  Weights
  Threshold  2.5275169924883842
  Node 3  -22.693258130991627
Sigmoid Node 1          (6-12 Months)
  Inputs  Weights
  Threshold  -1.1291974664851134
  Node 3  0.811973412063721
Sigmoid Node 2          (>12 Months)
  Inputs  Weights
  Threshold  -21.148512488693612
  Node 3  22.305942444459216
Sigmoid Node 3          (Hidden Unit)
  Inputs  Weights
  Threshold  -0.536992839455282
  Attrib SurOncName=Tada  0.49964854759393856
  Attrib SurOncName=Whalen  -0.749457220227305
  Attrib SurOncName=Andersen  0.7232550477678591
  Attrib ResPODischStatus=Home  1.9036354548752843
  Attrib ResPODischStatus=Died_in_Hospital  -1.2939439920064815
  Attrib ResPODischStatus=Acute_Rehab  -1.1364175007615474
  Attrib ResPODischStatus=Subacute_Rehab_Nursing_Facility  -1.3264765879128924
  Attrib CxDiab  -2.2911797748386884
  Attrib SHCigarette  -1.0331294017897965
  Attrib SxOT  2.121520582437838
  Attrib EUSVascOmit  -0.5618229827304928
  Attrib SxPru  -1.4291703431958664
  Attrib PTCDx  -1.534401191319715
  Attrib PTCStent  -1.5012022823311448
  Attrib TxChemo  -0.02166318622938725
  ...
```

Figure 24: Artificial Neural Network with One Hidden Unit

4.1.3.3 2nd Noteworthy Combination: Bayesian Network

The second best combination of feature selection and machine learning algorithm resulted from a model constructed using a Bayesian network constructed using two parents. The top 90 attributes are selected to build this model using support vector machine attribute selection.

4.1.3.3.1 Feature Selection

These 90 attributes are listed in Figure 25. Note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

Note that all of the attributes selected by support vector machines are pre-operative attributes.

They include many of the attributes relating to the patient's imaging studies, lab scores, medical history, and presentation symptoms.

PresumptiveDx	LabAmylase	PTCStentType	SxChola
CTHepatic	EUSHepaticClass	SxAbd	SHDrugUse
CTHepaticClass	EUSInferior	SxBack	SHOth
CTCeliacClass	EUSSMA	SxFati	SHExposure
CTSMAClass	EUSHepatic	SxInd	CxPriorCancerSurgery
CTSMA	EUSSMAClass	SxPru	SHAlcohol
CTSMVClass	EUSPortal	CxHF	SHCigarette
CTPortalClass	EUSCeliacNode	CxResp	FamilyOther1Dx
CTPortal	EUSPortalClass	CxIHD	FamilyOther2Dx
CTInferior	EUSInferiorClass	SxDyspha	FamilyOther2
CTSMV	EUSSMVClass	SxOT	FamilyFatherDx
CTInferiorClass	EUSSMV	SxSatiety	FamilyOther1
LabBili	CTTumorSizeX	SxWtloss	FamilyMotherDx
LabALT	PTCDx	SxJaun	CxDiabOnset
LabAlka	CTTumorSizeY	SxWtlossP	CxHyper
LabCEA	CTCeliacNode	DemECOG	CxRF
LabAlb	CTNodeOmit	DemWeight	CxDiab
LabCA19-9	CTOtherNode	DemHeight	CxDiabDiet
CTDx	EUSVascOmit	SxNau	CxDiabOral
CTCeliac	EUSCeliacClass	SxCCS	CxPriorCancer
CTVascOmit	EUSCeliac	SxVom	CxPriorCancerRadiation
LabAST	PTCStent	SxChole	

Figure 25: Top 90 Attributes Selected By Support Vector Machines

4.1.3.3.2 Machine Learning Model

An graphical overview of the Bayesian network constructed over these attributes is shown in Figures 26, 27, and 28. Note that the left-most part of the network is shown at the very top of Figure 26. Below the left-most part of the network is the part of the network which continues to its right. There is an overlap of CTVascOmit between these two parts. This pattern continues for the rest of the components in the network.

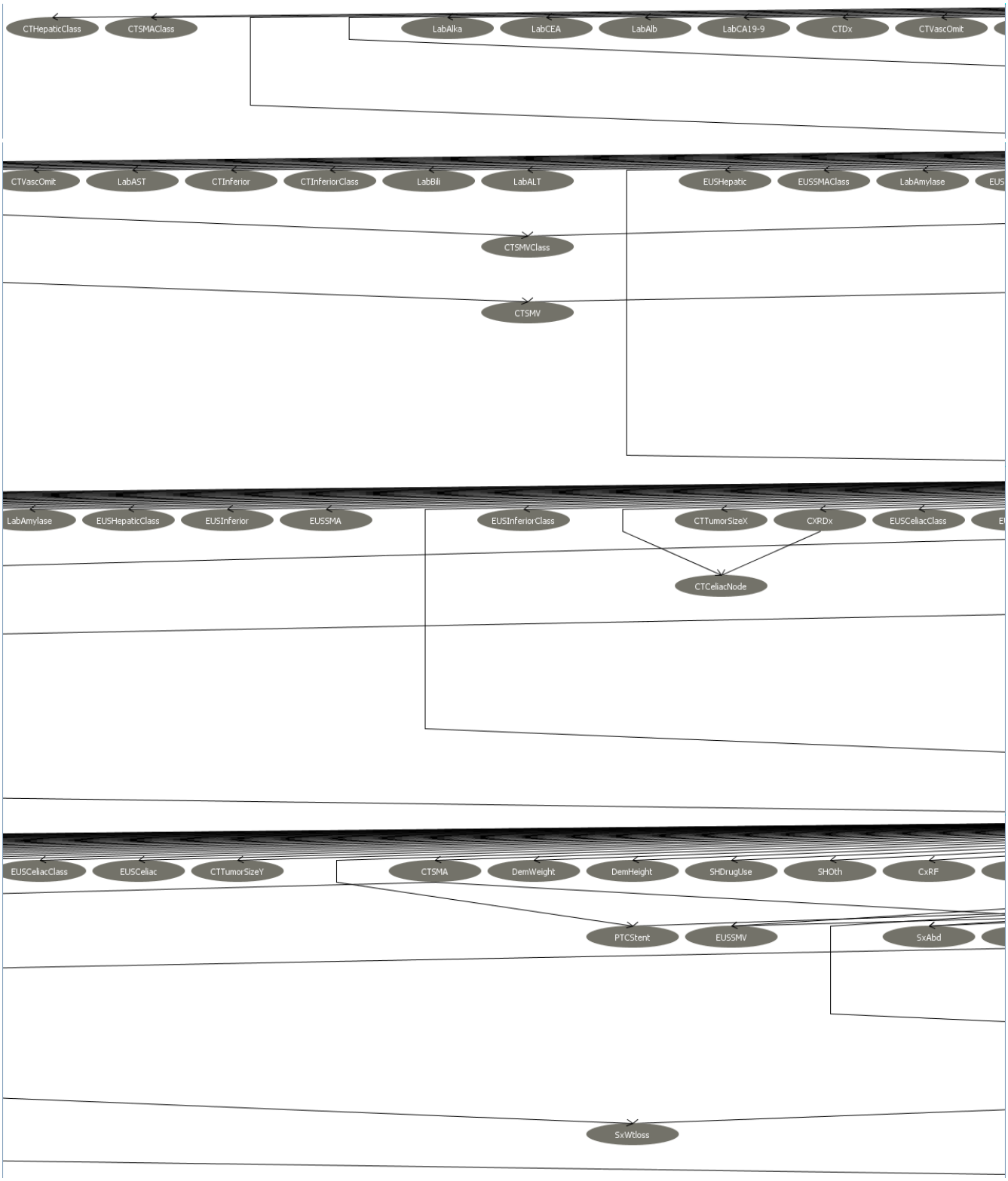

Figure 26: Bayesian Network: Overview Sections 1 of 3

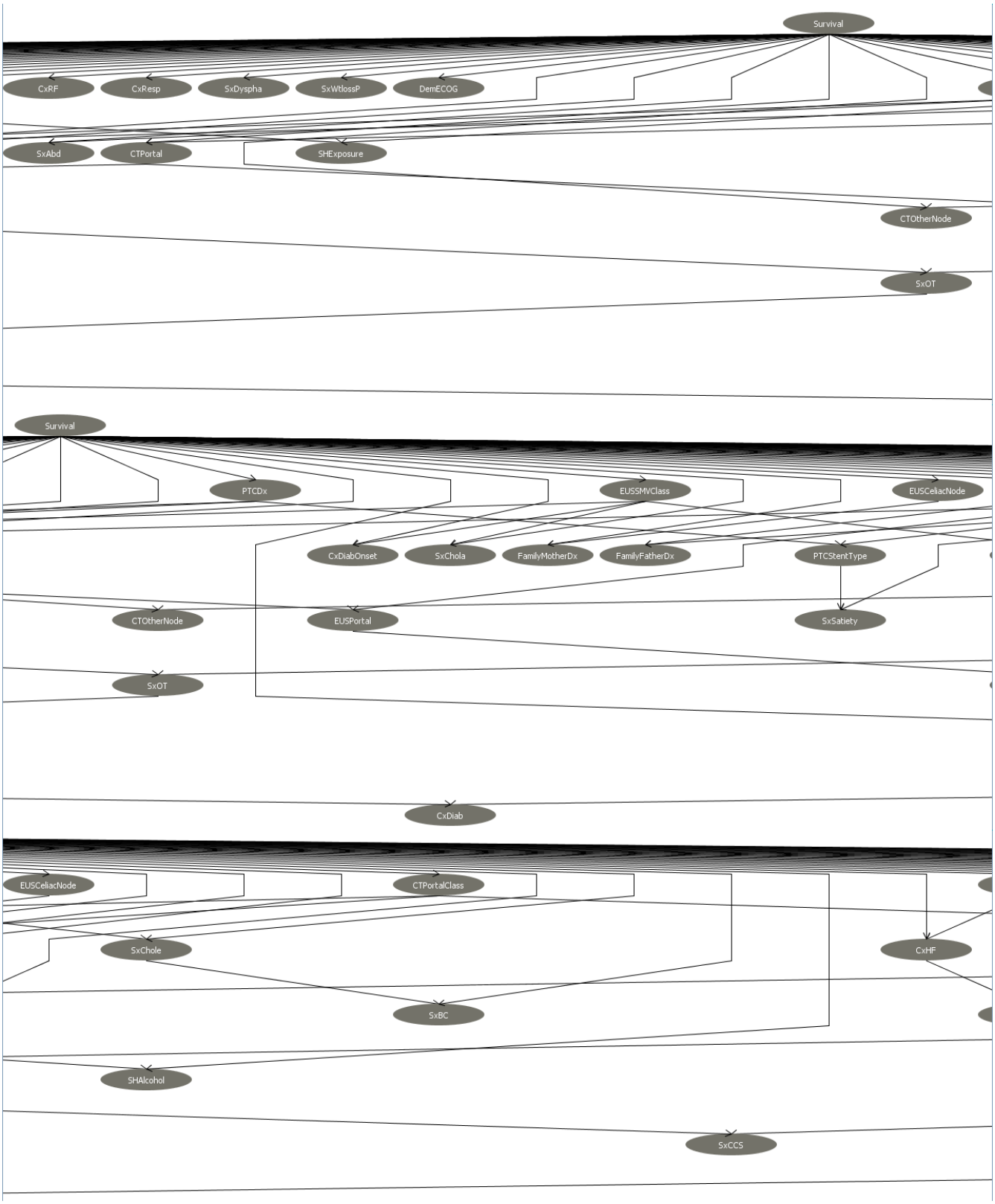

Figure 27: Bayesian Network: Overview Sections 2 of 3

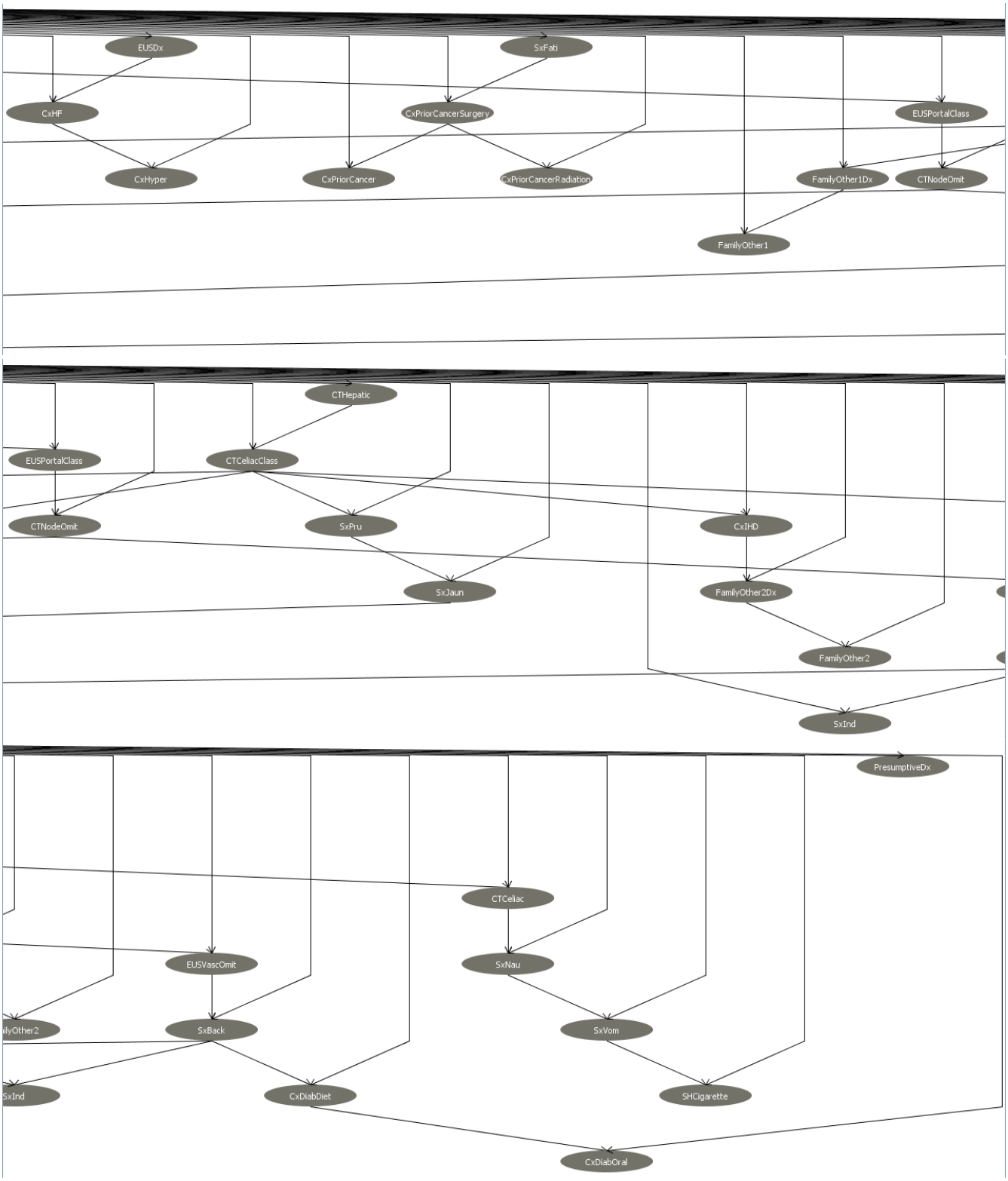

Figure 28: Bayesian Network: Overview Sections 3 of 3

Figure 29 provides the Weka output of this model, listing for each attribute its parents. Note

that the attributes with no parents, 43 of the 90 attributes, omitted from this output.

Classifier Model
Bayes Network Classifier
not using ADTree
#attributes=91 #classindex=90
Network structure (nodes followed by parents)

CTCeliacClass(2): Survival CTHepatic
CTSMVClass(3): Survival CTSMA
CTPortal(2): Survival CTPortalClass
CTSMV(2): Survival CTPortal
CTCeliac(2): Survival CTCeliacClass
EUSPortal(2): Survival CTPortal
EUSPortalClass(2): Survival CTPortalClass
EUSSMV(2): Survival EUSSMVClass
CTCeliacNode(2): Survival CXRDx
CTNodeOmit(2): Survival EUSPortalClass
CTOtherNode(2): Survival CTCeliacClass
EUSVascOmit(2): Survival CTNodeOmit
PTCStent(2): Survival PTCDx
PTCStentType(2): Survival PTCDx
SxAbd(2): Survival PTCDx
SxBack(2): Survival EUSVascOmit
SxInd(2): Survival SxBack
SxPru(2): Survival CTCeliacClass
CxHF(2): Survival EUSDx
CxIHD(2): Survival CTCeliacClass
SxOT(2): Survival CTNodeOmit
SxSatiety(2): Survival PTCStentType
SxWtloss(2): Survival SxOT

SxJaun(2): Survival SxPru
SxNau(2): Survival CTCeliac
SxCCS(2): Survival SxJaun
SxVom(2): Survival SxNau
SxChole(2): Survival EUSSMVClass
SxBC(2): Survival SxChole
SxChola(2): Survival EUSSMVClass
SHExposure(2): Survival CTSMA
CxPriorCancerSurgery(2): Survival SxFati
SHAlcohol(2): Survival EUSPortal
SHCigarette(2): Survival SxVom
FamilyOther1Dx(8): Survival CTCeliacClass
FamilyOther2Dx(4): Survival CxIHD
FamilyOther2(3): Survival FamilyOther2Dx
FamilyFatherDx(4): Survival CTPortalClass
FamilyOther1(7): Survival FamilyOther1Dx
FamilyMotherDx(9): Survival EUSCeliacNode
CxDiabOnset(2): Survival EUSSMVClass
CxHyper(2): Survival CxHF
CxDiab(2): Survival SxBack
CxDiabDiet(2): Survival SxBack
CxDiabOral(2): Survival CxDiabDiet
CxPriorCancer(6): Survival CxPriorCancerSurgery
CxPriorCancerRadiation(2): Survival CxPriorCancerSurgery

Survival(3):
LogScore Bayes: -1456.359899404217
LogScore BDeu: -619.8416930720299
LogScore MDL: -3043.9167522259627
LogScore ENTROPY: -1786.9529716237744
LogScore AIC: -2400.952971623777

Figure 29: Bayesian Network: Network Structure Details

There are four attributes that are parents for three or four nodes. They are CTCeliacClass, CTPortalClass, EUSSMVClass, and SxBack. CTCeliacClass is the parent of four nodes, the rest are parents of three. It is noteworthy that all of these except for SxBack, which indicates a presentation symptom of back pain, are read from diagnostic imaging scans. The other attributes represent details of the type of celiac disease detected by a CT scan, details of tumor involvement with portal vein as

detected by a CT scan, and details of tumor involvement with superior mesenteric vein as detected by endoscopic ultrasound respectively.

4.1.3.4 3rd Noteworthy Combination: Artificial Neural Network with Two Hidden Units

The third best combination of feature selection and machine learning algorithm resulted from a model constructed using artificial neural networks with two hidden units trained over 2,000 epochs. The top 30 attributes are selected to build this model using ReliefF attribute selection.

4.1.3.4.1 Feature Selection

These 30 attributes, and the weights assigned to them by ReliefF, are listed in Figure 30. They are identical to the first 30 attributes listed in Figure 22. Note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

Figure 23 shows how the progression of weights assigned by ReliefF decrease. We noted already that the rate of decline in the weights levels off after 30 attributes have been selected. Therefore it seems appropriate that 30 attributes performed particularly well.

4.1.3.4.2 Machine Learning Model

Some analysis was possible of the artificial neural network with only one hidden node but with the addition of the second hidden node, the model is no longer human interpretable. This is one of the disadvantages of artificial neural networks.

ReliefF Weight	Attribute Name	ReliefF Weight	Attribute Name
0.104	SurOncName	0.024	EUSTumorSizeX
0.073	ResPODischStatus	0.024	TxResect
0.070	CxDiab	0.023	TxPal
0.057	SHCigarette	0.021	TxRadia
0.047	SxOT	0.019	TxChemoGem
0.045	EUSVascOmit	0.018	EUSNoNode
0.043	SxPru	0.018	SxSatiety
0.038	PTCDx	0.017	SHAlcohol
0.038	PTCStent	0.016	EUSTumorSizeY
0.038	TxChemo	0.016	ResTFFP
0.037	ResPOPulmComp	0.015	CTNodeOmit
0.033	Histology	0.015	EUSOtherNode
0.029	MedOncName	0.014	LabAmylase
0.027	SxBack	0.013	CxPriorCancerChemo
0.027	DemHeight	0.012	ResPOCourse

Figure 30: Top 30 Attributes Selected By ReliefF

4.1.4 Meta-Learning

The following three Combinations of Feature Selection and Machine Learning Algorithm were found in the previous section to produce the highest classification accuracies over the dataset with a six and a twelve month split:

- 53.8%: Artificial Neural Networks, One Hidden Unit, using ReliefF to select 70 attributes
- 52.0%: Bayesian Network, Two Parents, using Support Vector Machines to select 90 attributes
- 51.0%: Artificial Neural Networks, Two Hidden Units, using ReliefF to select 30 attributes

4.1.4.1 Bagging

Figure 31 shows the effect that bagging has on each of the three best combinations of feature selection and machine learning algorithm selected for this dataset. Overall there is a slight decrease in the standard deviation but there is also a decrease in the classification accuracy. A positive effect of decreasing the standard deviation is that this tends to increase the statistical significance of the results. Bagging is not helpful over this dataset as the small decrease in the standard deviation is not enough to compensate for the decrease in classification accuracy.

Original		Bagging	
%Correct	σ	%Correct	σ
53.8	20.1	46.3	19.6
52.0	21.0	45.7	20.1
51.0	19.5	47.3	17.8

Figure 31: Bagging: Six and Twelve Month Split

4.1.4.2 Boosting

Figure 32 shows the effect boosting has on each of the three combinations of best feature selection and machine learning algorithm for this dataset. Overall there is a decrease in the classification accuracy when boosting is used. The standard deviation is also changed by boosting but only by a small amount. For these reasons, boosting is not helpful over this dataset.

Original		Boosting	
%Correct	σ	%Correct	σ
53.8	20.1	40.2	20.0
52.0	21.0	47.0	21.4
51.0	19.5	49.7	20.1

Figure 32: Boosting: Six and Twelve Month Split

4.1.4.3 Stacking

In this section we investigate the use of stacking to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Stacking is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

The results of the runs of Stacking over this dataset are presented in Figure 33. The far left column shows the level-1 classifier used to make the final target class prediction. Note that no result below has a classification accuracy greater than those of the initial models. Overall the best classification accuracies occur when stacking combines the three best combinations into one model.

Level-1 Classifier	Level – 0 Models			
	ANN 1HU, Bayes Net, ANN 2HU	Bayes Net, ANN 2HU	ANN 1HU, ANN2 HU	ANN 1HU, Bayes Net
ANN 1 Hidden Unit	36.7	35.7	36.0	35.5
ANN 2 Hidden Units	38.0	36.2	35.2	34.8
Linear Regression	47.2	48.8	49.0	46.8
LWL	40.3	42.3	40.3	42.5

Note: HU stands for Hidden Unit(s)

Figure 33: Stacking Results: Six and Twelve Month Split

4.1.4.4 Our Model Selector

In this section we investigate the use of our model selector to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Our model selector is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

The results of the runs of our model selector over this dataset are presented in Figure 34. The far left column shows the level-1 classifier used to determine the model that makes the final target class prediction. The classifiers with feature selection algorithms listed are run using the attribute selected classifier.

This meta-learning algorithm was able to slightly improve the classification accuracy in several cases by combining these algorithms. In all of the runs where there was an increase in the classification accuracy the algorithm with the highest classification accuracy, artificial neural networks with one hidden unit using ReliefF to select 70 attributes, was not used. Our model selector was therefore able to construct a level-1 model capable of selecting, for each instance, the best of the remaining two models to increase overall classification accuracy. The level-1 model with the highest classification accuracy in Figure 34, see highlighted row and column, was constructed using an artificial neural network with ReliefF to select the top 90 attributes. Unfortunately, there is still no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this model. This model has a classification accuracy that is a statistically significant improvement

($p<0.05$) over the classification accuracy of ZeroR.

Level-1 Classifier	Level – 0 Models			
	ANN 1HU, Bayes Net, ANN 2HU	Bayes Net, ANN 2HU	ANN 1HU, ANN2 HU	ANN 1HU, Bayes Net
ANN 1 Hidden Unit	51.8	52.3	49.5	49.7
ANN 2 Hidden Units	49.8	52.5	50.0	50.2
ANN 1 HU & ReliefF_90	52.0	54.8	50.7	51.0
J48 & SVM_70	51.5	51.0	51.2	49.8
Logistic & PrincComp_15	50.0	51.5	50.5	49.8
NaiveBayes & GainRatio_30	52.2	52.2	51.7	51.5
SMO & ReliefF_40	51.7	53.5	50.3	51.5
J48	50.2	51.7	50.5	50.3
Logistic	49.8	52.0	49.2	51.2
LWL	50.7	48.7	51.5	52.0
NaiveBayes	50.5	51.3	50.3	49.7
SMO	48.8	51.3	48.7	51.2

Note: HU stands for Hidden Unit(s)

Highlighted value has higher classification accuracy than level-0 models – not statistically significant ($p<0.05$)

Figure 34: Our Model Selector Results: Six and Twelve Month Split

4.1.4.4.1 Model Constructed

The highlighted classification accuracy in Figure 34 is constructed using a Bayesian network and the artificial neural network with two hidden units as level-0 models. The level-1 model used to predict which is best was trained using an artificial neural network with the top 90 attributes selected by ReliefF. Since this classification accuracy is higher, though not statistically significantly higher ($P<0.05$), than any of the best combinations of feature selection and machine learning algorithm we want to look into the resulting model.

Figure 35 shows the probability distributions of every instance over each model. The actual survival target is also in the table along with a label of which model is correct, if none of the models correct, or if all of the models correct. In twenty out of these sixty instances both models produce the correct classification. In fifteen neither model produces the correct prediction. This leaves twenty five instances where if we can predict the correct model, we can make the correct prediction and therefore increase classification accuracy.

Bayesian Network	Artificial Neural Network	Actual Target	Which Model
{0.2,0.08,0.72}	{0,0.59,0.41}	6-12 Months	ANN
{0.01,0.69,0.3}	{0,0.01,0.99}	>6 Months	ANN
{0.13,0.02,0.84}	{0.01,0.97,0.02}	6-12 Months	ANN
{0.07,0.6,0.33}	{4.93E-001,6.94E-007,0.51}	>6 Months	ANN
{0.82,0.17,0.01}	{5.20E-004,0.79,0.21}	6-12 Months	ANN
{0.27,0.12,0.61}	{0.97,0.03,7.03E-005}	<6 Months	ANN
{0.78,0.06,0.16}	{0.01,0.98,0.01}	6-12 Months	ANN
{0.11,0.65,0.24}	{0.96,0.04,9.58E-004}	<6 Months	ANN
{0.92,0.01,0.07}	{0.01,0.98,0.01}	6-12 Months	ANN
{0.08,0.54,0.38}	{0.01,0.18,0.82}	>6 Months	ANN
{0.06,0.7,0.24}	{0.61,0,0.39}	<6 Months	ANN
{0.01,0.02,0.97}	{4.80E-006,0.95,0.05}	6-12 Months	ANN
{0.04,0.08,0.88}	{0.04,0.96,4.12E-004}	6-12 Months	ANN
{0.01,0.48,0.51}	{0.03,0.97,6.29E-004}	>6 Months	Bayesian
{0.72,0.28,0}	{0.03,0.97,0}	<6 Months	Bayesian
{0.1,0,0.9}	{0.07,0.73,0.19}	>6 Months	Bayesian
{0.63,0.24,0.12}	{0.01,0.97,0.02}	<6 Months	Bayesian
{0.1,0.06,0.84}	{5.49E-004,0.76,0.23}	>6 Months	Bayesian
{1,0,0}	{0.02,0.94,0.04}	<6 Months	Bayesian
{0.02,0.05,0.93}	{0.04,0.96,0}	>6 Months	Bayesian
{0.04,0.75,0.21}	{0,0.02,0.97}	6-12 Months	Bayesian
{0.01,0.2,0.79}	{0.01,0.95,0.05}	>6 Months	Bayesian
{0.69,0.24,0.07}	{0.02,0.97,0.01}	<6 Months	Bayesian
{0.25,0.68,0.06}	{0,0.34,0.65}	6-12 Months	Bayesian
{0.16,0.65,0.2}	{0.49,0,0.51}	6-12 Months	Bayesian
{0.51,0.07,0.42}	{1,0,5.91E-004}	<6 Months	Both
{0.58,0.13,0.29}	{9.46E-001,1.61E-006,0.05}	<6 Months	Both
{0.77,0.21,0.02}	{0.98,0.01,0.01}	<6 Months	Both
{0.03,0,0.97}	{0,0.02,0.98}	>6 Months	Both
{0.92,0.08,3.84E-004}	{0.99,0.01,0}	<6 Months	Both
{0.09,0.26,0.65}	{6.30E-004,0.01,0.99}	>6 Months	Both
{0.01,0.94,0.05}	{0.02,0.97,0.01}	6-12 Months	Both
{0.97,0.01,0.03}	{0.98,0.01,0.01}	<6 Months	Both
{0.03,0.01,0.97}	{0,0,0.99}	>6 Months	Both
{0.03,0.97,2.72E-005}	{0.01,0.98,0.01}	6-12 Months	Both
{0.1,0.89,0.01}	{0.01,0.98,0.01}	6-12 Months	Both
{0.82,0.01,0.17}	{0.99,0.01,0.01}	<6 Months	Both
{0.01,0.07,0.93}	{4.25E-006,0.01,0.99}	>6 Months	Both
{9.83E-004,0,1}	{0,0,0.99}	>6 Months	Both
{0,1,0}	{0.01,0.98,0.01}	6-12 Months	Both
{0.01,0,0.99}	{1.18E-002,6.55E-006,0.99}	>6 Months	Both
{0.01,0.83,0.16}	{0.31,0.68,0.01}	6-12 Months	Both
{0.04,0.29,0.68}	{1.74E-002,4.48E-004,0.98}	>6 Months	Both
{0.74,0.24,0.02}	{0.98,0.01,0.01}	<6 Months	Both
{0.25,0.69,0.07}	{0.02,0.97,0.01}	6-12 Months	Both
{0.35,0.64,0.01}	{0.01,0.98,0.01}	<6 Months	Neither
{0.02,0.5,0.48}	{1.00E-005,0.95,0.05}	>6 Months	Neither
{0.28,0.02,0.7}	{9.99E-001,7.62E-004,2.01E-004}	6-12 Months	Neither
{0.01,0.5,0.49}	{0.01,0.88,0.11}	>6 Months	Neither
{0.03,0.97,0}	{4.67E-006,0.01,0.99}	<6 Months	Neither
{0,0.03,0.96}	{0.04,0.92,0.04}	<6 Months	Neither
{0.78,0.22,1.21E-004}	{0,0.02,0.98}	6-12 Months	Neither
{8.96E-004,0.38,0.62}	{0.01,0.98,0.01}	<6 Months	Neither
{0.4,0.52,0.08}	{0.46,0.53,0.01}	<6 Months	Neither
{0,7.67E-004,1}	{1.47E-002,5.23E-005,0.99}	6-12 Months	Neither
{0.01,0.58,0.41}	{0.01,0.94,0.05}	<6 Months	Neither
{0.28,0.62,0.1}	{0.03,0.97,4.30E-004}	>6 Months	Neither
{0.13,0.24,0.63}	{0.01,0.49,0.5}	6-12 Months	Neither
{0.01,0.99,0}	{0.01,0.96,0.04}	>6 Months	Neither
{0.94,0.06,3.84E-004}	{0.95,0,0.05}	>6 Months	Neither

Note that probability distribution {x,y,z} denotes the predicted probability of survival for <6 months, 6-12 months, and >12 months are x, y, and z respectively

Figure 35: Probability Distributions of Each Combined Model

Figure 36 lists the ninety attributes selected from the dataset to train the level-1 classifier by ReliefF. Note that only the first 41 attributes selected have ReliefF scores are above 0.005. This is an indication that perhapses too many features are being selected. We therefore re-ran this experiment using ReliefF to select the top 40 attributes. The resulting classification accuracy was 52.8, better than the two initial models that are selected between by our model selector but not as good as when the level-1 model is constructed over 90 attributes. Therefore, there is additional information contributed by the 50 attributes that get a lower score using ReliefF that help with predicting the best model for a given instance.

0.08	ResPxType	0.01	EUSTumorSizeX	0	TxPalBypass	0	CxMal
0.05	CTTumorSizeY	0.01	ResPathR	0	LabCA19-9	0	SHDrugUse
0.04	SHCigarette	0.01	Gender	0	PTCStentType	0	ResPOAbdominal
0.04	TxChemoIri	0.01	TxChemoFlu	0	Age	0	CxLiver
0.04	ResPathT	0.01	TxChemoGem	0	CTPortalClass	0	SHOth
0.03	CTTumorSizeX	0.01	EUSPortalClass	0	TxResect	0	FamilyOther2
0.03	GIMDName	0.01	CxIHD	0	NoResSMAInvolve	0	ResAttemptUn
0.03	EUSCyto	0.01	ResOrgans	0	EUSSMVClass	0	ResTCell
0.02	DemECOG	0.01	CTPortal	0	FamilyOther2Dx	0	NoResIVCInvolve
0.02	EUSTumorSizeY	0.01	TxPalStens	0	ResPOPulmComp	0	NoResHepaticInvolve
0.02	TxPal	0.01	SxInd	0	SxCCS	0	NoResCirrhosis
0.02	ResPOLeak	0.01	TxPalRad	0	TxChemoErb	0	NoResSMVInvolve
0.02	LabALT	0.01	LabBili	0	TxChemoFUDR	0	CxBleed
0.02	LabAST	0.01	LabAlb	0	EUSHepatic	0	ResPOLiverInsuf
0.02	NoResPVInvolve	0.01	ResVenRes	0	CTInferior	0	SxDyspha
0.02	PreOutlook	0.01	ResVenRec	0	CTHepaticClass	0	CxRF
0.02	ResPathM	0.01	ResTFFP	0	EUSSMA	0	CxResp
0.01	ResAttempt	0.01	NoResNoHandle	0	EUSCeliac	0	TxPalCeliac
0.01	ERCPStent	0	CTSMVClass	0	TxChemoCap	0	TxPalRes
0.01	EUSDx	0	ResPathV	0	EUSInferior	0	TxPalTho
0.01	PresumptiveDx	0	ERCPStentType	0	CTSMAClass	0	TxPalPara
0.01	EUSCeliacNode	0	NoResMagnitude	0	TxChemoAVA		
0.01	SxChola	0	EUSPortal	0	CTVascOmit		

Figure 36: Top 90 Attributes Selected by ReliefF for the Level-1 Classifier

4.1.5 Summary

Over the dataset with the six and twelve month split target, there is no statistically significant difference ($p<0.05$) between logistic regression and ZeorR. There is a statistically significant

difference ($p<0.05$) between the classification accuracy of ZeroR and the models constructed using the three noteworthy combinations of feature selection and machine learning algorithm. There is no statistically significant ($p<0.05$) difference between any of these models and logistic regression. There is also a statistically significant difference ($p<0.05$) between classification accuracy of the model constructed using the best machine learning algorithm without feature selection and ZeroR. There is no statistically significant ($p<0.05$) difference between any of these models and logistic regression. This is the only model run over the dataset with this target without feature selection that is statistically significantly ($p<0.05$) better than ZeroR.

The best attribute selection method over this dataset is RelieF attribute selection as it picked a sets of attributes that constantly resulted in high classification accuracies over a large portion of models constructed using machine learning algorithms. Support vector machine attribute selection and gain ratio attribute selection also both did a good job at selecting attributes. The attributes selected by gain ratio attribute selection are better at finding a set of features that improved the classification accuracy of models constructed with all of the machine learning algorithms while support vector machine attribute selection picked sets of attributes that were better for some of the algorithms over others.

The models constructed with the features selected by principal components consistently had slightly lower classification accuracies over the other three approaches to feature selection. However, principal components did better than any of the other feature selection algorithms at increasing the accuracy of a logistic regression. A model constructed using logistic regressing over the top 20 features selected by principal components has a classification accuracy of 49.17% which is statistically significantly better than ZeorR.

The best algorithms over this dataset are artificial neural networks and Bayesian networks constructed with a maximum of two parents. Artificial neural networks performed well over all of the

feature selection algorithms. Artificial neural networks with two hidden units performed better when fewer attributes, around 30, were selected while artificial neural networks with one hidden unit performed better when more attributes, around 60, were selected. Bayesian networks with a maximum of two parents performed better as more attributes were selected up to a peak classification accuracy when 90 attributes are selected.

Stacking, bagging, and boosting were not helpful over this dataset as they neither increasing clarification accuracy nor decrease the standard deviation. Our model selector is able to slightly increase the classification accuracy by selecting between using a model constructed using a Bayesian network and a model constructed using an artificial neural network by using an artificial neural network as the level-1 model to predict which model will make the best prediction for a given instance. The classification accuracy of the best model constructed using our model selector is statistically significantly ($p<0.05$) better than ZeroR. The classification accuracy of this model is not statistically significant different ($p<0.05$) from logistic regression.

4.2 Results for Full Dataset with Nine Month Split

The dataset discussed in this section has all 190 attributes. The sixty patients in this dataset are split evenly into two groups based on the target of survival. These groups are are <9 month and >9 month survival.

4.2.1 Machine Learning Algorithms with No Feature Selection

Figure 37 shows the classification accuracies of models constructed using several different machine learning algorithms run over the dataset with a target split into two groups of zero to nine and more than nine month survival. The model constructed using a Bayesian network with two parents, highlighted in the figure, has the highest classification accuracy. It is not statistically significantly better ($p<0.05$) than either ZeroR or logistic regression.

Machine Learning Algorithm	Classification Accuracy	Compare To ZeroR
ZeroR	50.0	No Statistically Significant Difference
Logistic Regression	57.5	
SMO with Kernel of 0.9	54.8	
SMO with Kernel of 1.0	54.5	
ANN with 1 Hidden Unit	56.7	
ANN with 2 Hidden Units	54.8	
Naïve Bayes	55.3	
J4.8	51.2	
Bayesian Network: 1 Parent	56.2	
Bayesian Network: 2 Parents	64.3	

Figure 37: No Feature Selection: Nine Month Split

4.2.2 Combinations of Feature Selection and Machine Learning Algorithm

The graphs that follow show how the classification accuracies of the models built to predict a patient's expected survival time vary based on the feature selection algorithm used and the number of features selected. These models are constructed over varying feature selection algorithms using several different machine learning algorithms. We use these graphs to find the combinations of feature selection and machine learning algorithm that have the highest classification accuracy for this dataset.

Note that the highest classification accuracy obtained by constructing models with no feature selection is 64.3%. We look to use feature selection to improve upon this classification accuracy.

Figure 38 shows how varying the number of attributes selected by gain ratio attribute selection effects the classification accuracy of several machine learning algorithms. There are no models with classification accuracies greater than 64.3% in this graph. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 40. The model with the highest classification accuracy plotted here, artificial neural networks with one hidden unit, has a classification accuracy that is highest when 40 attributes are selected by gain ratio attribute selection. After this peak at 40 attributes most models show a gradual decrease in classification accuracy as the number of attributes selected increases. The notable exception to this is the model constructed using a Bayesian network with two parents. The classification accuracy of this model continues to increase until it appears to peak when 90 attributes are selected. Since at this peak of 90 attributes the classification accuracy is still less than the best model constructed with no feature selection, we suspect that it would continue increasing as more attributes are selected since this is machine learning algorithm with the highest classification accuracy constructed with no feature selection.

Figure 39 shows how varying the number of features selected by principal components effects the classification accuracy of several machine learning algorithms. There is a small overall decrease in the classification accuracies as the number of features selected is increased. No model constructed using the features selected by principal components has a classification accuracy greater than 64.3%.

Figure 40 shows how varying the number of features selected by ReliefF attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 30. Once 30 attributes have been selected, the classification accuracies decrease slightly as more attributes are

selected. This varies by the machine learning algorithm, however. Models constructed using logistic regression and Bayesian networks with two parents continue to increase very slightly until their classification accuracies level off when around 60 attributes are selected. No model in this figure has a classification accuracy over 64.3%.

Figure 41 shows how varying the number of features selected by support vector machine attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 100. There are no models that have a classification accuracy of over 64.3% but there are two models that are particularly noteworthy. First, support vector machines with a linear kernel function's classification accuracy increases particularly quickly as the number of attributes selected increases from 30 to 60, reaching a maximum classification accuracy when 70 attributes are selected. Second, the model constructed using a Bayesian network with two parents increases slightly more gradually as more attributes are included in the model reaching a high classification accuracy when 100 attributes are selected. Since at the classification accuracy of this model constructed with 100 attributes is still very slightly less than be best model constructed with no feature selection, we suspect that it would continue increasing as more attributes are selected since this is machine learning algorithm with the highest classification accuracy constructed with no feature selection.

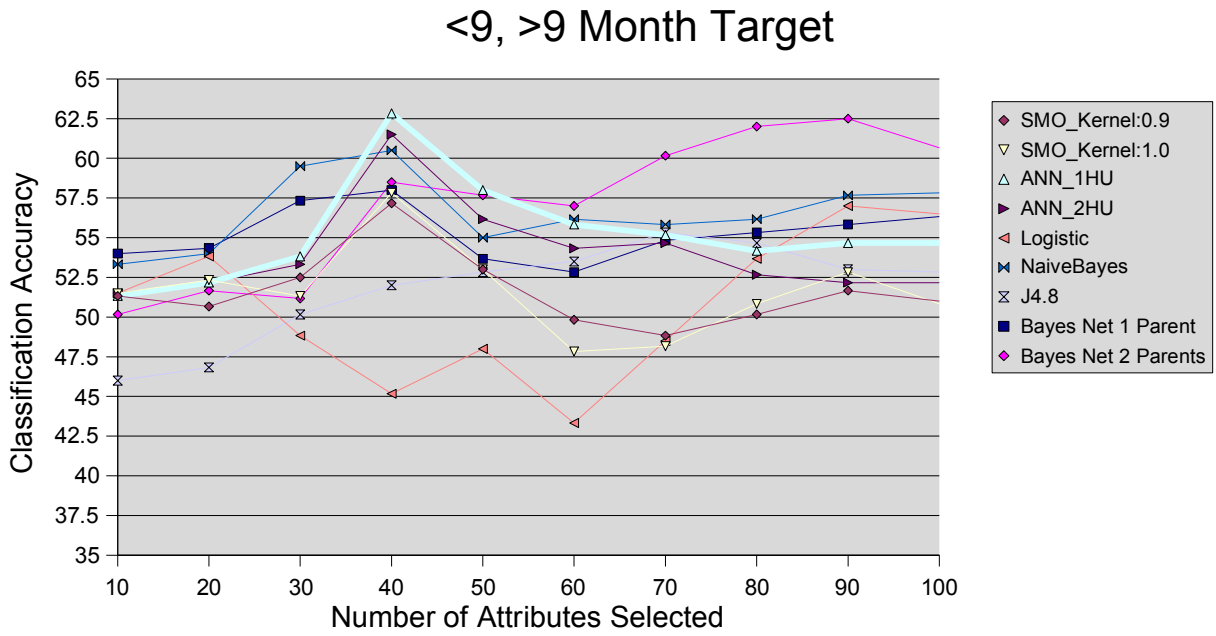

Figure 38: Gain Ratio Attribute Selection: Nine Month Split

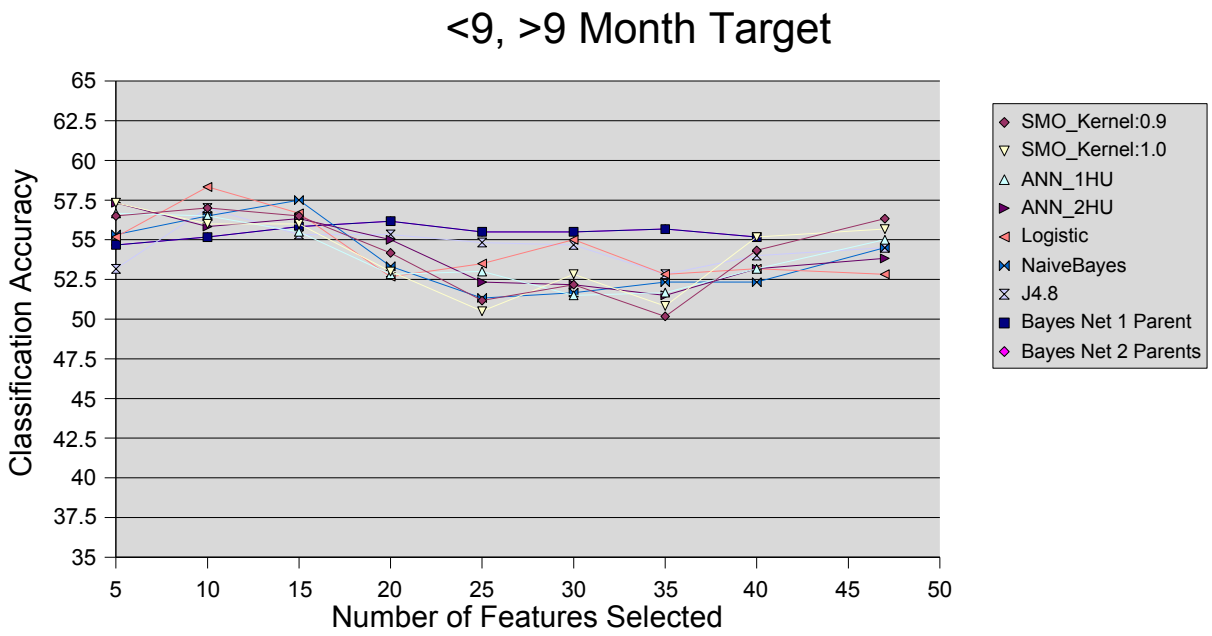

Figure 39: Principal Components: Nine Month Split

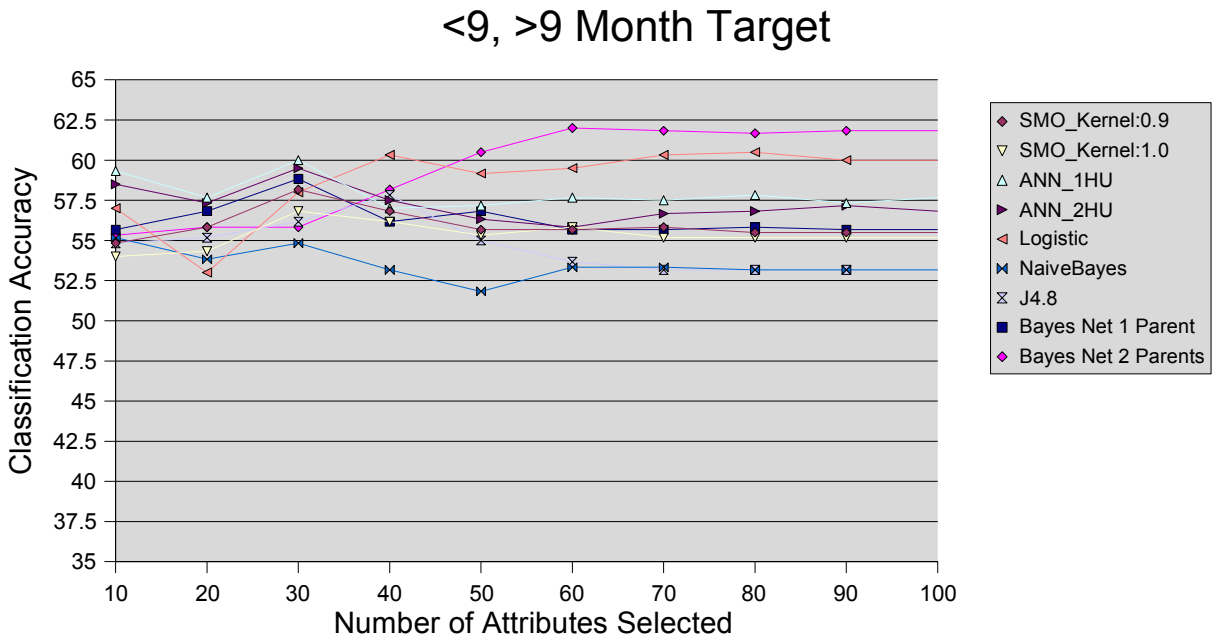

Figure 40: ReliefF Attribute Selection: Nine Month Split

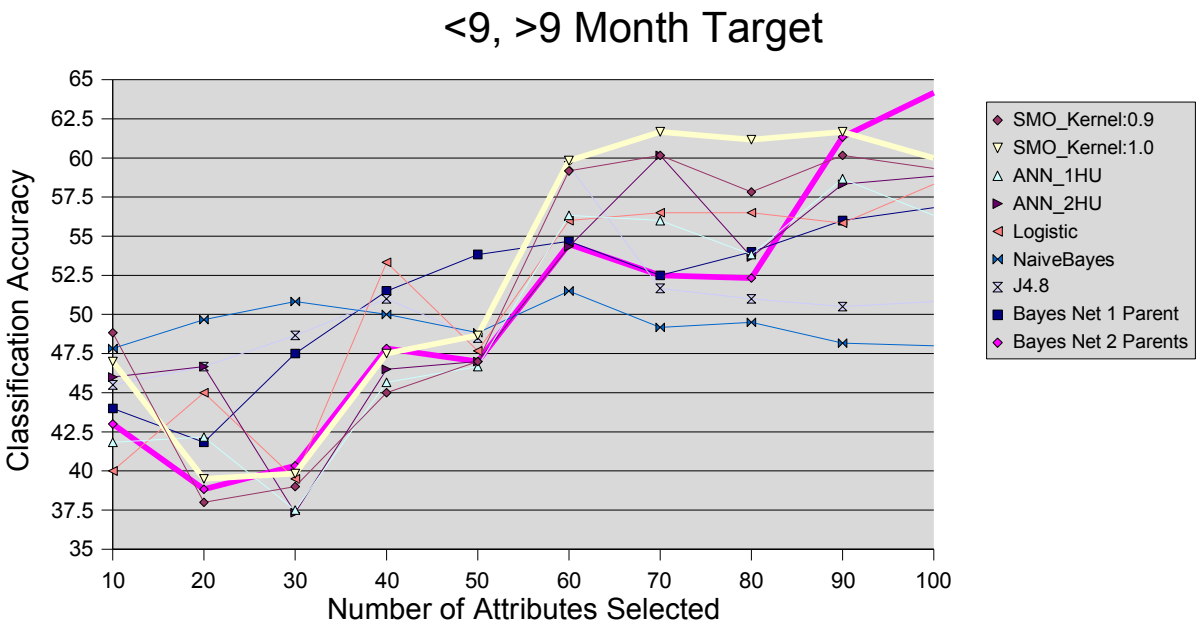

Figure 41: Support Vector Machine Attribute Selection: Nine Month Split

4.2.2.1 Baseline Models

Figure 37 shows that over this dataset the classification accuracy of logistic regression with no feature selection is 57.5% and that of ZeroR is 50.0%. There is no statistically significant difference between logistic regression and ZeroR.

4.2.2.2 1st Noteworthy Combination: Bayesian Network

The highest classification accuracy obtained over this dataset is 64.2% resulting from a model constructed using a Bayesian network with K2 using a maximum of two parents. The top 100 attributes are selected to build this model using support vector machine attribute selection. Note that this model has a smaller standard deviation than the best machine learning algorithm with no feature selection. Figure 41 shows this combination of feature selection and machine learning algorithm. This figure shows a gradual increase in classification accuracy of models built using this Bayesian network algorithm as a greater numbers of attributes are selected using support vector machines for feature selection. There is no statistically significant difference ($p<0.05$) between the classification accuracies

of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.2.2.3 2nd Noteworthy Combination: Artificial Neural Network with One Hidden Unit

The second highest classification accuracy obtained over this dataset is 62.8% resulting from a model constructed using artificial neural networks with one hidden unit trained over 2,000 epochs. The top 40 attributes are selected to build this model using gain ratio attribute selection. Figure 38 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a peak classification accuracy when 40 attributes are selected using gain ratio attribute selection and when more or less attributes are selected the classification accuracy decreases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. There is also no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.2.2.4 Model Similar to 1st Noteworthy Combination

The third highest classification accuracy obtained over this dataset is 62.5% resulting from a model constructed using a Bayesian network with K2 using a maximum of two parents. The top 90 attributes are selected to build this model using gain ratio attribute selection. Figure 38 shows this combination of feature selection and machine learning algorithm. As the first highest classifier was constructed using the same Bayesian network, we will not be using this combination of feature selection and machine learning algorithm for further analysis of this dataset.

4.2.2.5 3rd Noteworthy Combination: Support Vector Machines

The fourth highest classification accuracy obtained over this dataset is 61.7% resulting from a

model constructed using support vector machines with a linear kernel function. The top 70 attributes are selected to build this model using support vector machine attribution selection. Figure 41 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a maximal point when 70 attributes are selected using support vector machine attribute selection after which the addition of more attributes does not increase the classification accuracy. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. There is also no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.2.2.6 Summary of Noteworthy Combinations

- 64.2%:Bayesian Network, Two Parents, using Support Vector Machine to select 100 attributes
- 62.8%: Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 40 attributes
- 61.7%: Support Vector Machine, Linear Kernel, using SVM to select 70 attributes

Note no statistically significant difference ($p<0.05$) between these three models.

4.2.3 Models Produced

4.2.3.1 Best Model with No Feature Selection

The best model with no feature selection resulted from a model constructed using a Bayesian network with a maximum of two parents. Since the first noteworthy combination is also a Bayesian network with a maximum of two parents, constructed with only 100 attributes selected by support vector machines, and since the difference in accuracy between the two models is only 0.1% we will not discuss the underlying model constructed without feature selection.

4.2.3.2 1st Noteworthy Combination: Bayesian Network

The best combination of feature selection and machine learning algorithm resulted from a

model constructed using a Bayesian network constructed with a each node having a maximum of two parents. The top 100 attributes are selected to build this model using support vector machine attribute selection.

4.2.3.2.1 Feature Selection

These 100 attributes are listed in Figure 42. Note that all but six of the attributes selected are pre-operative. The other attributes selected are a wide mix of patient imaging tests, lab values, initial symptoms, and medical history. Note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

PresumptiveDx	EUSInferior	SxInd	SHCigarette
CTHepatic	EUSSMA	SxPru	FamilyOther1Dx
CTHepaticClass	EUSHepatic	CxHF	FamilyOther2Dx
CTCeliacClass	EUSSMAClass	CxResp	FamilyOther2
CTSMAClass	EUSPortal	CxIHD	FamilyFatherDx
CTSMA	EUSCeliacNode	SxDyspha	FamilyOther1
CTSMVClass	EUSPortalClass	SxOT	FamilyMotherDx
CTPortalClass	EUSInferiorClass	SxSatiety	CxDiabOnset
CTPortal	EUSSMVClass	SxWtloss	CxHyper
CTInferior	EUSSMV	SxJaun	CxRF
CTSMV	CTTumorSizeX	SxWtlossP	CxDiab
CTInferiorClass	PTCDx	DemECOG	CxDiabDiet
LabBili	CTTumorSizeY	DemWeight	CxDiabOral
LabALT	CTCeliacNode	DemHeight	CxPriorCancer
LabAlka	CTNodeOmit	SxNau	CxPriorCancerRadiation
LabCEA	CTOtherNode	SxCCS	CxPriorCancerChemo
LabAlb	EUSVascOmit	SxVom	CxBleed
LabCA19-9	EUSCeliacClass	SxChole	CxMal
CTDx	EUSCeliac	SxBC	CxLiver
CTCeliac	PTCStent	SxChola	ResPODays
CTVascOmit	EUSDx	SHDrugUse	ResPOInfection
LabAST	PTCStentType	SHOth	ResAttempt
CXRDx	SxAbd	SHExposure	ResPOCourse
LabAmylase	SxBack	CxPriorCancerSurgery	ResAttemptUn
EUSHepaticClass	SxFati	SHAlcohol	ResPOPulmComp

Figure 42:Top 100 Attributes Selected By Support Vector Machines

There are three attributes that are parents for more than three nodes. They are EUSSVMClass, PTCDx, and CTCeliacClass. EUSSVMClass and PTCDx are each the parents of four nodes and

CTCeliacClass is the parent of 6 nodes. These attributes are details of tumor involvement with superior mesenteric vein as detected by endoscopic ultrasound, details of the results of a percutaneous transhepatic cholangiography, and details of the type of celiac disease detected by a CT scan respectively. It is noteworthy that EUSSVMClass and CTCeliacClass both were parents of more than three nodes in the dataset with a 6 and a 12 month split.

4.2.3.2.2 Machine Learning Model

The Bayesian network constructed over these attributes is even larger than the one shown in Figures 26, 27, and 28 due to being constructed with even more attributes. Figure 43 provides the Weka output of this model, listing for each attribute its parents. Note that the attributes with no parents, 54 of the 90 attributes, omitted from this output.

Classifier Model
Bayes Network Classifier
not using ADTree
#attributes=101 #classindex=100
Network structure (nodes followed by parents)

CTCeliacClass(2): Survival CTHepatic
CTSMVClass(3): Survival CTSMA
CTPortal(2): Survival CTCeliacClass
CTSMV(2): Survival CTSMVClass
CTCeliac(2): Survival CTCeliacClass
EUSPortal(2): Survival CTPortal
EUSPortalClass(2): Survival CTPortalClass
EUSSMV(2): Survival EUSSMVClass
CTCeliacNode(2): Survival CTHepatic
CTNodeOmit(2): Survival EUSPortalClass
CTOtherNode(2): Survival CTCeliacClass
EUSVascOmit(2): Survival PTCDx
PTCStent(2): Survival PTCDx
PTCStentType(2): Survival PTCDx
SxAbd(2): Survival PTCDx
SxBack(2): Survival CTPortal
SxInd(2): Survival EUSSMV
SxPru(2): Survival CTCeliacClass
CxHF(2): Survival EUSDx
CxIHD(2): Survival CTCeliacClass
SxOT(2): Survival CTNodeOmit
SxSatiety(2): Survival PTCStentType
SxWtloss(2): Survival SxSatiety
SxJaun(2): Survival SxPru
SxNau(2): Survival CTCeliac
SxCCS(2): Survival SxJaun
SxVom(2): Survival SxNau
SxChole(2): Survival EUSSMVClass
SxBC(2): Survival SxChole
SxChola(2): Survival EUSSMVClass
SHExposure(2): Survival CTSMA
CxPriorCancerSurgery(2): Survival SxFati
SHAlcohol(2): Survival EUSPortal
SHCigarette(2): Survival EUSDx
FamilyOther1Dx(8): Survival CTCeliacClass
FamilyOther2Dx(4): Survival CxIHD
FamilyOther2(3): Survival FamilyOther2Dx
FamilyFatherDx(4): Survival CTPortalClass
FamilyOther1(7): Survival FamilyOther1Dx
FamilyMotherDx(9): Survival EUSCeliacNode
CxDiabOnset(2): Survival EUSSMVClass
CxHyper(2): Survival CxHF
CxDiab(2): Survival CxHyper
CxDiabDiet(2): Survival SxBack
CxDiabOral(2): Survival CxDiabDiet
CxPriorCancer(6): Survival CxPriorCancerSurgery
CxPriorCancerRadiation(2): Survival CxPriorCancerSurgery
CxPriorCancerChemo(2): Survival CxPriorCancerRadiation
ResPOInfection(2): Survival PTCStentType
ResAttempt(2): Survival CxIHD
ResPOCourse(2): Survival PTCDx
ResPOPulmComp(2): Survival ResPOCourse

Survival(2):
LogScore Bayes: -1427.08783745961
LogScore BDeu: -844.7675953615951
LogScore MDL: -2500.02819795567
LogScore ENTROPY: -1617.6969447968045

Figure 43: Bayesian Network: Network Structure Detail

4.2.3.3 2nd Noteworthy Combination: Artificial Neural Network with One Hidden Unit

The second best combination of feature selection and machine learning algorithm resulted from a model constructed using one hidden unit trained for 2000 epochs. The top 40 attributes are selected to build this model using gain ratio attribute selection.

4.2.3.3.1 Feature Selection

These 40 attributes are listed in Figure 44. Note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

Gain Ratio Weight	Attribute Name		Gain Ratio Weight	Attribute Name
0.18	PTCDx		0.14	CxIHD
0.18	PTCStent		0.14	CTSMA
0.18	ResPOInfection		0.14	SHExposure
0.18	NoResNoHandle		0.11	TxResect
0.18	SxChola		0.11	ResPODischStatus
0.18	NoResMagnitude		0.08	SxOT
0.18	TxPalStens		0.08	NoResPVInvolve
0.18	CTCeliacNode		0.08	PreOutlook
0.16	CxPriorCancerChemo		0.07	SxSatiety
0.16	CxDiabDiet		0.06	SxBack
0.16	NoResSMAInvolve		0.06	CxDiab
0.16	TxPalBypass		0.06	Histology
0.16	TxChemoIri		0.06	NoResMetastatic
0.14	TxPal		0.06	SxInd
0.14	CTHepatic		0.06	EUSDx
0.14	TxPalGasTube		0.06	CxPriorCancerSurgery
0.14	TxPalJejTube		0.06	NoResRefused
0.14	EUSCeliacNode		0.06	PresumptiveDx
0.14	TxChemoTax		0.05	EUSPortal
0.14	SxBC		0.04	ResAttempt

Figure 44: Top 40 Attributes Selected By Gain Ratio

Figure 45 shows how the progression of weights assigned by Gain Ratio decrease. The weights slowly decrease, running with the same value over several attributes before stepping to a lower value.

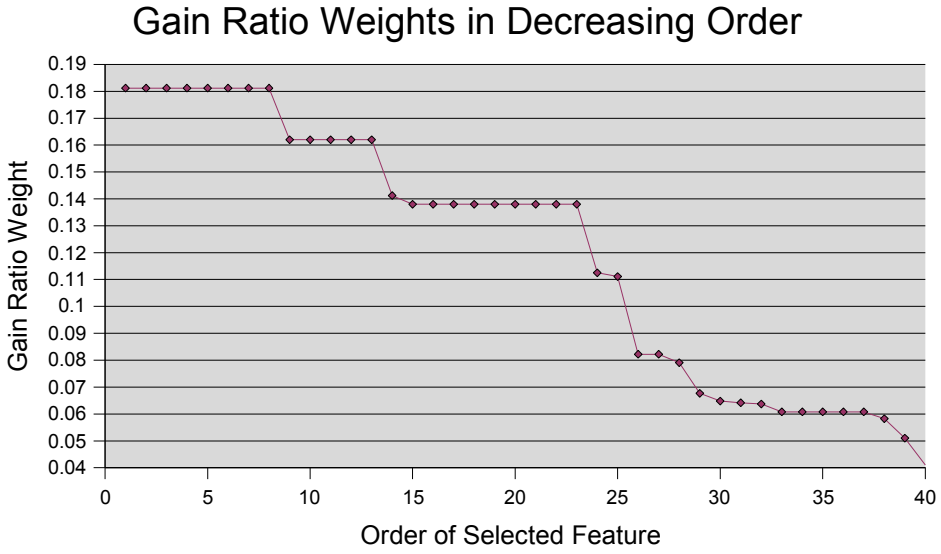

Figure 45: Gain Ratio Weights in Decreasing Order

The progression of weights displayed in Figure 45 indicate that the 30th attribute selected has a Gain Ratio that has decreased by over half from the first attributes gain ratio and the 40th attribute selected has decreased by nine times the first attributes gain ratio. This corresponds closely to Figure 38 where we can see that there is a peak in the classification accuracy of models generated when 30 attributes are selected. This can be seen with both artificial neural network models, the support vector machine models, and the model constructed by naïve Bayes. The model constructed using a Bayesian network with a maximum of two parents also shows this peak but goes on to obtain an even higher classification accuracy.

4.2.3.3.2 Machine Learning Model

The artificial neural network constructed over these attributes, with all but the the first 10 attributes omitted for readability, is shown in Figure 46. Artificial neural networks are always a challenge to decipher, this one is no different. The first two sigmoid units are the classification targets. The classification target that results in the highest value relative to that node's threshold is the target that is predicted by the model.

For each output node, the combination of its weighted connection to node 3 and its threshold thresholds allows for interpretation into the values of the node 3 will trigger each of the target values. If node 3 takes on a low negative value, the resulting classification will be >9 months. High positive values will result in a classification of <9 months. Therefore, the weights that are positive between the input nodes and node 3 pull the classification towards predicting the patient will have a decreased expected survival while weights that are negative will increase the expected survival.

Therefore attributes with a high positive weight such as ResPOInfection, indicating a post operative infection, reduce the patient's expected survival and attributes with a negative weight increase the patient's expected survival.

```
Classifier Model
Sigmoid Node 0          (<9 Months)
  Inputs    Weights
  Threshold    -2.7218271363833013
  Node 2    8.41621048296891
Sigmoid Node 1          (>9 Months)
  Inputs    Weights
  Threshold    2.721827073136636
  Node 2    -8.416207830334004
Sigmoid Node 2          (Hidden Unit)
  Inputs    Weights
  Threshold    0.3885385504638988
  Attrib PTCDx    1.0669319626611318
  Attrib PTCStent    1.0964211376227808
  Attrib ResPOInfection    4.481175484013518
  Attrib NoResNoHandle    0.5642637422341851
  Attrib SxChola    0.15421751169416648
  Attrib NoResMagnitude    -0.12216284281567234
  Attrib TxPalStens    -0.09391325940292407
  Attrib CTCeliacNode    2.029671015493038
  Attrib CxPriorCancerChemo    0.41910972760424237
  Attrib CxDiabDiet    -1.266612363821211
  ...
```

Figure 46: Artificial Neural Network with One Hidden Unit

4.2.3.4 3rd Noteworthy Combination: Support Vector Machine

The third best combination of feature selection and machine learning algorithm resulted from a model constructed using support vector machines constructed with linear kernel function. The top 70

attributes are selected to build this model using support vector machine attribute selection.

4.2.3.4.1 Feature Selection

These 70 attributes are listed in Figure 47. Note that all of the attributes are pre-operative including a mix of patient imaging tests, lab values, initial symptoms, and medical history. Also note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

PresumptiveDx	LabAlka	EUSSMAClass	EUSCeliacClass	SxOT
CTHepatic	LabCEA	EUSPortal	EUSCeliac	SxSatiety
CTHepaticClass	LabAlb	EUSCeliacNode	PTCStent	SxWtloss
CTCeliacClass	LabCA19-9	EUSPortalClass	EUSDx	SxJaun
CTSMAClass	CTDx	EUSInferiorClass	PTCStentType	SxWtlossP
CTSMA	CTCeliac	EUSSMVClass	SxAbd	DemECOG
CTSMVClass	CTVascOmit	EUSSMV	SxBack	DemWeight
CTPortalClass	LabAST	CTTumorSizeX	SxFati	DemHeight
CTPortal	CXRDx	PTCDx	SxInd	SxNau
CTInferior	LabAmylase	CTTumorSizeY	SxPru	SxCCS
CTSMV	EUSHepaticClass	CTCeliacNode	CxHF	SxVom
CTInferiorClass	EUSInferior	CTNodeOmit	CxResp	SxChole
LabBili	EUSSMA	CTOtherNode	CxIHD	SxBC
LabALT	EUSHepatic	EUSVascOmit	SxDyspha	SxChola

Figure 47:Top 70 Attributes Selected By Support Vector Machines

4.2.3.4.2 Machine Learning Model

The support vector machine constructed over these attributes is shown in Figure 48. Support vector machines are almost as difficult to decipher as artificial neural networks. This support vector machine has a linear kernel so its interpretation is much more straightforward. Since the kernel is linear, we can interpret the weights the very similar to the weights assigned by linear regression. It is important to keep in mind that it the models are not identical due to support vector machines use of a kernel function. In this model, positive weights pull the classification target closer to one class, here >9 months, and negative weights pull the target closer to the negative class, here <9 months.

```
Classifier Model
SMO
Kernel used:  Linear Kernel: K(x,y) = <x,y>
Classifier for classes: '(-inf-9.353425]', '(9.353425-inf)'

Machine linear: showing attribute weights, not support vectors.

       -0.5175 * (normalized) PresumptiveDx=Pancreatic_Tumor
 +      0.2895 * (normalized) PresumptiveDx=Periampullary_Tumor
 +     -1.0135 * (normalized) PresumptiveDx=Suspicious_Bile_Duct_Stricture
 +      0.327  * (normalized) PresumptiveDx=Other
 +      0.3746 * (normalized) PresumptiveDx=IPMT/IPMN
 +      0.5399 * (normalized) PresumptiveDx=Suspicious_Pancreatic_Cyst
 +     -0.6089 * (normalized) CTHepatic                    +     -0.915  * (normalized) PTCStent
 +     -0.0207 * (normalized) CTCeliacClass                +     -0.3629 * (normalized) EUSDx
 +      0.1775 * (normalized) CTSMVClass=Abuts             +     -0.6254 * (normalized) PTCStentType
 +     -0.1775 * (normalized) CTSMVClass=Occluded          +     -0.1407 * (normalized) SxAbd
 +      0.4108 * (normalized) CTPortal                     +     -0.9078 * (normalized) SxBack
 +      0.0046 * (normalized) CTSMV                        +     -0.6814 * (normalized) SxFati
 +     -0.3939 * (normalized) LabBili                      +      1.0905 * (normalized) SxInd
 +     -0.1741 * (normalized) LabALT                       +     -1.165  * (normalized) SxPru
 +     -0.5764 * (normalized) LabAlka                      +     -0.0403 * (normalized) CxHF
 +     -0.0166 * (normalized) LabCEA                       +      0.5882 * (normalized) CxIHD
 +      1.087  * (normalized) LabAlb                       +      0.3965 * (normalized) SxOT
 +     -0.2843 * (normalized) LabCA19-9                    +     -1.058  * (normalized) SxSatiety
 +     -0.0193 * (normalized) CTCeliac                     +      0.0265 * (normalized) SxWtloss
 +      0.3404 * (normalized) LabAST                       +      0.1468 * (normalized) SxJaun
 +      0.0843 * (normalized) LabAmylase                   +     -0.0684 * (normalized) SxWtlossP
 +     -0.4861 * (normalized) EUSPortal                    +     -1.0537 * (normalized) DemECOG
 +      0.3296 * (normalized) EUSCeliacNode                +     -0.3032 * (normalized) DemWeight
 +      1      * (normalized) EUSPortalClass               +      0.5318 * (normalized) DemHeight
 +      0.4368 * (normalized) EUSSMV                       +      0.0951 * (normalized) SxNau
 +     -0.3124 * (normalized) CTTumorSizeX                 +      0.6176 * (normalized) SxCCS
 +     -0.915  * (normalized) PTCDx                        +     -0.1266 * (normalized) SxVom
 +     -0.1128 * (normalized) CTTumorSizeY                 +     -0.6263 * (normalized) SxChole
 +     -0.6089 * (normalized) CTCeliacNode                 +     -0.6263 * (normalized) SxBC
 +      0.3611 * (normalized) CTNodeOmit                   +     -1      * (normalized) SxChola
 +      0.1201 * (normalized) CTOtherNode                  +      0.5306
 +      0.2804 * (normalized) EUSVascOmit

Number of kernel evaluations: 1535 (98.144% cached)
```

Figure 48: Support Vector Machines with Linear Kernel

4.2.4 ROC Curves

ROC curves for the combinations of feature selection and machine learning algorithm with the highest classification accuracy appear at the end of this section. The combinations will be presented

and discussed in the following order:

- 57.5%: Logistic Regression
- 64.2%:Bayesian Network, Two Parents, using Support Vector Machine to select 100 attributes
- 62.8%: Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 40 attributes
- 61.7%: Support Vector Machine, Linear Kernel, using SVM to select 70 attributes

4.2.4.1 Baseline Model: Logistic Regression

Figure 49 shows the ROC curve for logistic regression. The area under this curve is 0.61. This curve shows that to correctly predict 90% of the patients who will survive for more than nine months, you have to incorrectly predict that 80% of the patients who will not survive for nine months will survive for greater than nine months. To correctly predict 80% of the patients who will survive for more than nine months, you have to incorrectly predict that 66% of the patients who will not survive for nine months will survive for greater than nine months.

4.2.4.2 1st Noteworthy Combination: Bayesian Network

Figure 50 shows the ROC curve for the combination of feature selection and machine learning algorithm with the highest classification accuracy over this dataset. The area under this curve is 0.66. This curve shows that to correctly predict 90% of the patients who will survive for more than nine months, you have to incorrectly predict that 84% of the patients who will not survive for nine months will survive for greater than nine months. To correctly predict 80% of the patients who will survive for more than nine months, you have to incorrectly predict that 62% of the patients who will not survive for nine months will survive for greater than nine months. Overall, this curve slightly better than logistic regression as there is a slight improvement in both the area under the curve and the trade off between the rates of true and false positives.

4.2.4.3 2nd Noteworthy Combination: Artificial Neural Networks with One Hidden Unit

Figure 51 shows the ROC curve for the combination of feature selection and machine learning

algorithm with the second highest classification accuracy over this dataset. The area under this curve is 0.63. This curve shows that to correctly predict 90% of the patients who will survive for more than nine months, you have to incorrectly predict that 75% of the patients who will not survive for nine months will survive for greater than nine months. This curve shows that to correctly predict 80% of the patients who will survive for more than nine months, you have to incorrectly predict that 60% of the patients who will not survive for nine months will survive for greater than nine months. Overall, this curve is slightly better than both logistic regression and the first combination of feature selection and machine learning algorithm. This is due to the slightly better trade off between the true and false positive rates shown in Figure 51.

4.2.4.4 3rd Noteworthy Combination: Support Vector Machines

Figure 52 shows the ROC curve for the combination of feature selection and machine learning algorithm with the third highest classification accuracy over this dataset. The area under this curve is 0.64. This curve shows that to correctly predict 90% of the patients who will survive for more than nine months, you have to incorrectly predict that 67% of the patients who will not survive for nine months will survive for greater than nine months. To correctly predict 80% of the patients who will survive for more than nine months, you have to incorrectly predict that 50% of the patients who will not survive for nine months will survive for greater than nine months. Overall, this curve is the best of all the curves constructed over this dataset due to having the best trade off between the true and false positive rates.

4.2.4.5 Summary

The three combinations of feature selection and machine learning algorithm with the highest classification accuracies are ranked by their ROC curves as follows:

1. Support Vector Machine, Linear Kernel, using SVM to select 70 attributes

2. Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 40 attributes
3. Bayesian Network, Two Parents, using Support Vector Machine to select 100 attributes

It is noteworthy that this is the opposite ordering from that obtained from classification accuracies.

4.2.4.6 Curves

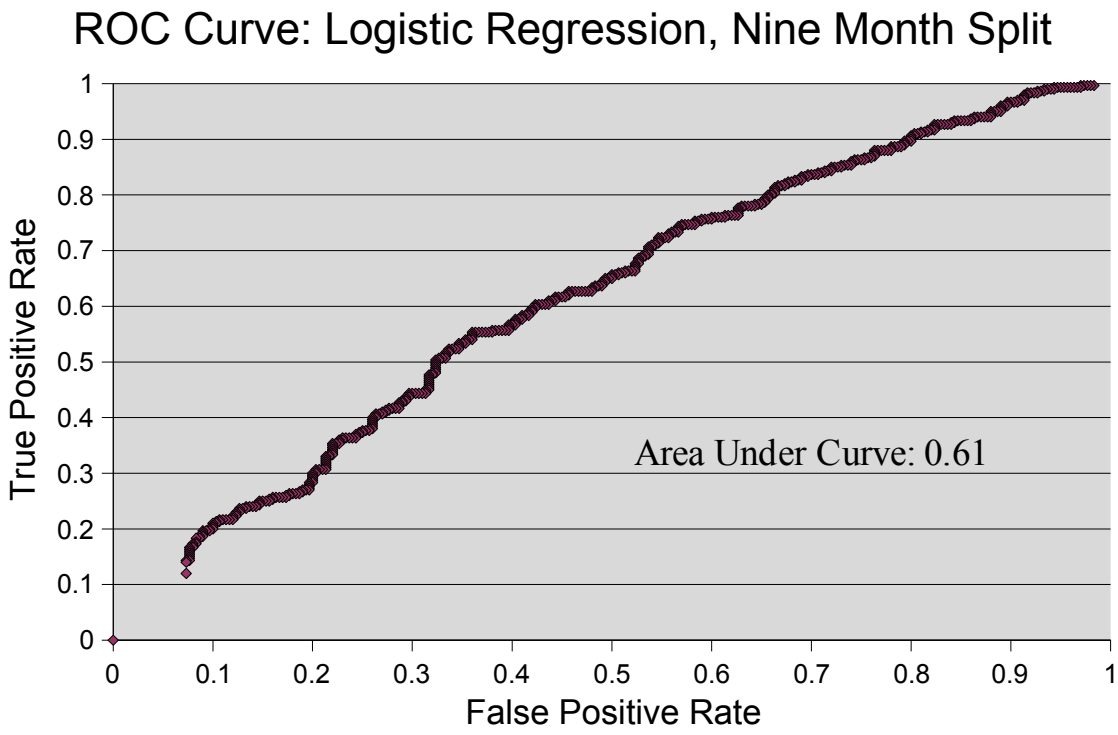

Figure 49: ROC Curve - Logistic Regression: Nine Month Split

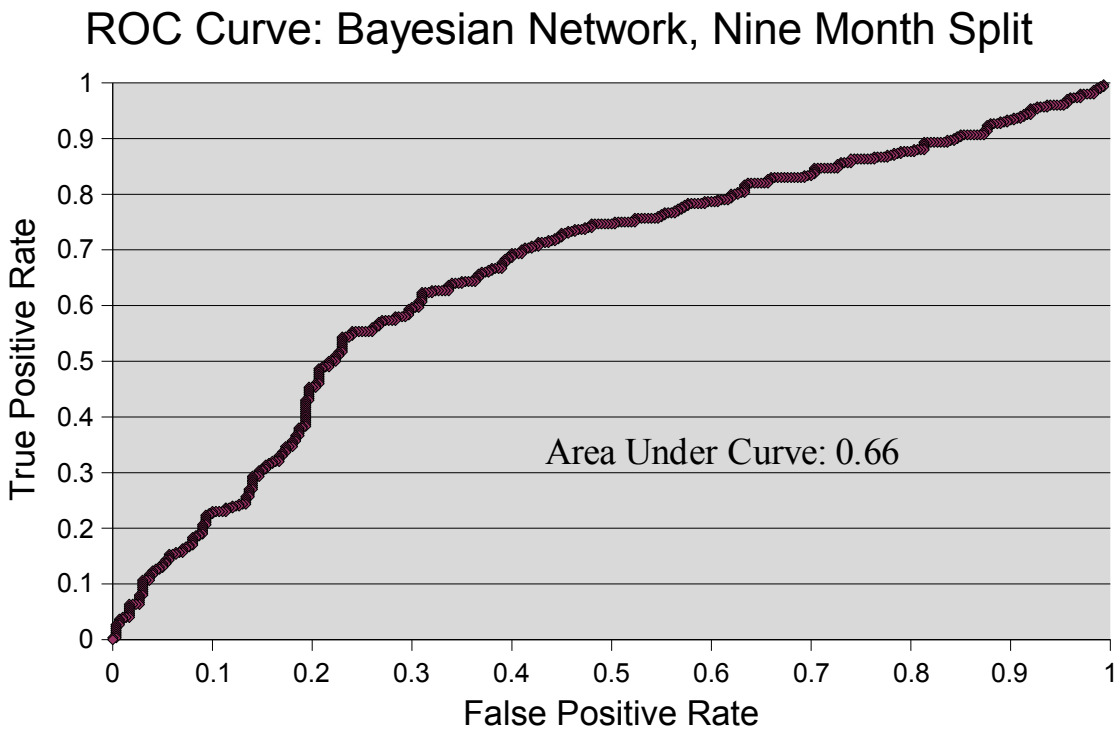

Figure 50: ROC Curve – Bayesian Network: Nine Month Split

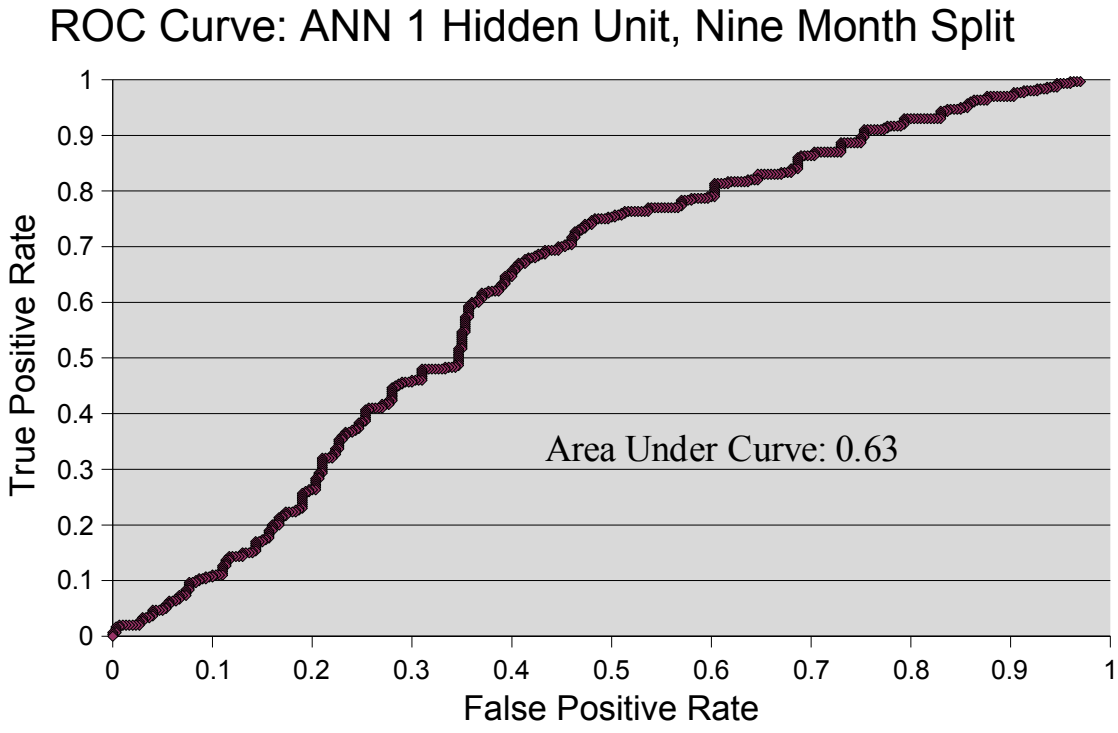

Figure 51: ROC Curve – Artificial Neural Network, One Hidden Unit: Nine Month Split

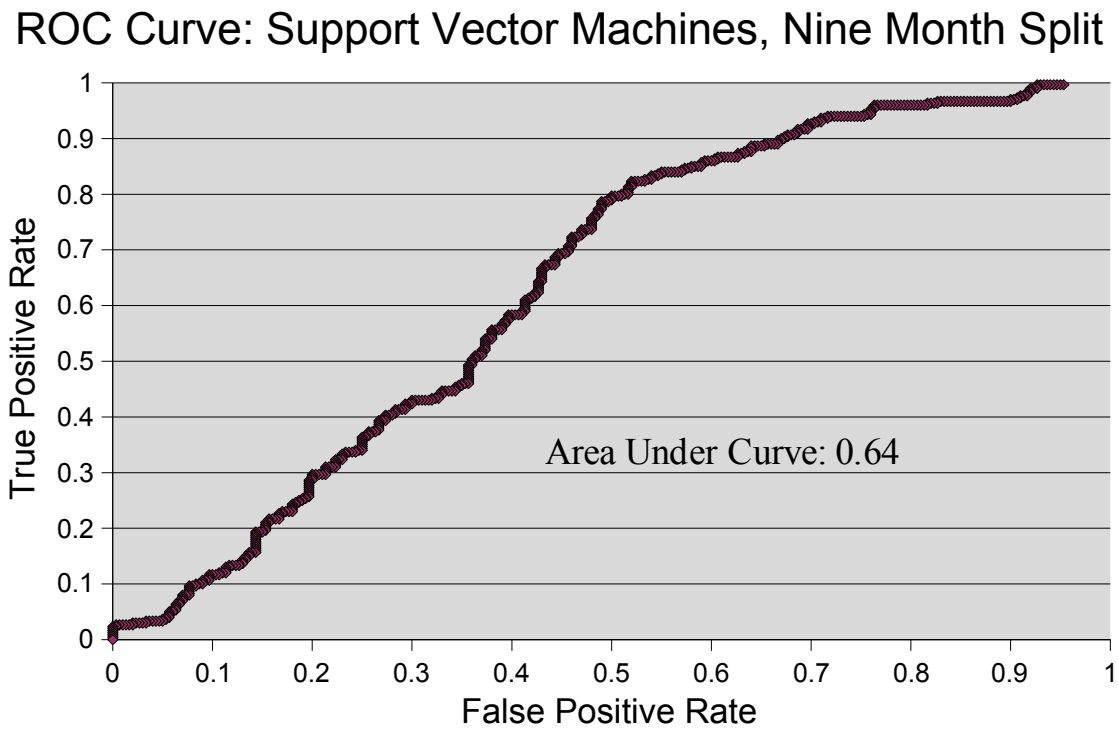

Figure 52: ROC Curve – Support Vector Machines: Nine Month Split

4.2.5 Meta-Learning

The following three Combinations of Feature Selection and Machine Learning Algorithm were found in the previous section to produce the highest classification accuracies over the dataset with a nine month split:

- 64.2%:Bayesian Network, Two Parents, using Support Vector Machine to select 100 attributes
- 62.8%: Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 40 attributes
- 61.7%: Support Vector Machine, Linear Kernel, using SVM to select 70 attributes

4.2.5.1 Bagging

Figure 53 shows the effect that bagging has on each of the three best combinations of feature selection and machine learning algorithm selected for this dataset. For each of the three combinations there is an decrease in the classification accuracy. The first two combinations have an increase in the standard deviation, which is the opposite of our goal for bagging. In the third case bagging does decrease the standard derivation but this is not helpful as the decrease in the standard deviation is not

enough to compensate for the decrease in classification accuracy.

Original		Bagging	
%Correct	σ	%Correct	σ
64.2	18.7	62.2	19.2
62.8	18.6	60.7	20.6
61.7	21.1	58.0	18.6

Figure 53: Bagging: Nine Month Split

4.2.5.2 Boosting

Figure 54 shows the effect boosting has on each of the three combinations of best feature selection and machine learning algorithm selected for this dataset. Over the first combination of feature selection and machine learning algorithm, boosting decreases the classification accuracy without having a large effect on the standard deviation. Unfortunately there is still not a statistically significant difference between logistic regression and boosting over the top combination of feature selection and machine learning. The other two combinations of feature selection and machine learning algorithm have a decrease in classification accuracy. The last combination does have a large decrease in its standard deviation but this is offset by the decrease in its classification accuracy. Due to the first result, this dataset does hint at the potential usefulness of boosting, even if it does not allow for a claim of statistical significance.

Original		Boosting	
%Correct	σ	%Correct	σ
64.2	18.7	65.5	18.2
62.8	18.6	60.2	19.2
61.7	21.1	56.2	17.7

Figure 54: Boosting: Nine Month Split

4.2.5.3 Stacking

In this section we investigate the use of stacking to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Stacking is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

Results from the stacking runs over this dataset are presented in Figure 55. The far left column shows the level-1 classifier used to make the final target class prediction. Note that no result below has a classification accuracy greater than any of the initial models. Overall the best classification accuracies occur when stacking combines all three of the best combinations into one model.

Level-1 Classifier	Level – 0 Models			
	Bayes Net, ANN 1HU, SVM	ANN 1HU, SVM	Bayes Net, SVM	Bayes Net, ANN 1HU
ANN 1 Hidden Unit	52.3	50.0	51.2	50.3
ANN 2 Hidden Units	52.7	51.2	51.3	50.5
Linear Regression	57.5	53.7	56.7	57.3
LWL	55.5	54.0	53.7	52.7

Note: HU stands for Hidden Unit

Figure 55: Stacking Results: Nine Month Split

4.2.5.4 Our Model Selector

In this section we investigate the use of our model selector to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Our model selector is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

The results of the runs of our model selector over this dataset are presented in Figure 56. The far left column shows the level-1 classifier used to determine the model that makes the final target class prediction. The classifiers with feature selection algorithms listed are run using the attribute selected classifier.

This meta-learning algorithm was not able to improve the classification accuracy by combining these combinations of feature selection and machine learning algorithm. Overall, the best set of runs are the ones where the model with the lowest classification accuracy is not included.

Level-1 Classifier	Level – 0 Models			
	Bayes Net, ANN 1HU, SVM	ANN 1HU, SVM	Bayes Net, SVM	Bayes Net, ANN 1HU
ANN 1 Hidden Unit	55.7	56.0	55.2	61.3
ANN 2 Hidden Units	56.7	55.3	54.2	61.7
ANN 1 HU & ReliefF_90	54.3	56.3	56.7	60.3
J48 & SVM_70	53.7	56.5	58.7	60.5
Logistic & PrincComp_15	53.3	55.5	56.8	62.0
NaiveBayes & GainRatio_30	55.0	54.7	57.0	62.8
SMO & ReliefF_40	55.8	55.7	56.2	61.7
J48	54.3	56.2	59.5	61.3
Logistic	58.3	55.0	55.7	61.7
LWL	52.8	56.0	58.2	61.5
NaiveBayes	55.2	52.0	55.5	61.0
SMO	56.2	57.3	56.0	61.0

Figure 56: Our Model Selector Results: Nine Month Split

4.2.6 Summary

Over the dataset with the nine month split target, there is no statistically significant difference ($p<0.05$) between the classification accuracy of logistic regression and ZeroR. There is a statistically significant difference ($p<0.05$) between the classification accuracy of ZeroR and one out of three of the models constructed using the noteworthy combinations of feature selection and machine learning algorithm found. There is no statistically significant ($p<0.05$) difference between any of these models and logistic regression. There is not a statistically significant difference ($p<0.05$) in the classification accuracy of the model constructed using the best machine learning algorithm without feature selection and ZeroR. There is no statistically significant ($p<0.05$) difference between the classification accuracy of logistic regression and the models constructed by running the machine learning algorithms without feature selection.

The best attribute selection method over this dataset is support vector machine attribute selection as it picked a sets of attributes that resulted in highest classification accuracy over a large portion of models constructed using machine learning algorithms. Support vector machine attribute selection had the highest accuracy when more than 60 attributes are selected. ReliefF attribute

selection and gain ratio attribute selection also both do a good job of selecting the best attributes for predicting this target. The attributes selected by gain ratio attribute selection are better at finding a set of features that improve the classification accuracy of models constructed with all of the machine learning algorithms while ReliefF attribute selection picked sets of attributes that were better for some of the algorithms than others. The models constructed with the features selected by principal components consistently had slightly lower classification accuracies than the other three approaches to feature selection.

The best algorithm over this dataset is Bayesian networks with a maximum of two parents. Bayesian networks had the highest classification accuracies when a large number of attributes were selected and performed well over all of the feature selection algorithms except for principal components. The ROC curve constructed for the Bayesian network had the largest area under the curve but did not have the best trade off in upper right part of the curve.

Artificial neural networks and support vector machines also performed well over this dataset but there was no statistically significant difference between the classification accuracy of the models they constructed and ZeroR. However, their ROC curves had a better trade off of true to false positives despite having a lower area under the cure when compared to the curve constructed for Bayesian networks.

Stacking, bagging, and our model selector were not useful over this dataset as they neither increased the classification accuracy nor decreased the standard deviation. Boosting was able to very sightly increase the classification accuracy over one of the best combinations of feature selection and machine learning algorithm.

4.3 Results for Full Dataset with Six Month Split

The dataset discussed in this section has all 190 attributes. The sixty patients in this dataset are split into two groups based on the target of survival. These groups are are <6 month and >6 month survival.

4.3.1 Machine Learning Algorithms with No Feature Selection

Figure 57 shows the classification accuracies of models constructed using several different machine learning algorithms run over the dataset with a target split into two groups of zero to six and more than six month survival. The model constructed using a Bayesian network with two parents, highlighted in the figure, has the highest classification accuracy. There is not statistically significantly difference ($p<0.05$) between this accuracy and that of both ZeroR or logistic regression.

Machine Learning Algorithm	Classification Accuracy	Compare To ZeroR
ZeroR	66.7	No Statistically Significant Difference
Logistic Regression	61.3	
SMO with Kernel of 0.9	62.2	
SMO with Kernel of 1.0	62.3	
ANN with 1 Hidden Unit	65.3	
ANN with 2 Hidden Units	64.8	
Naïve Bayes	64.2	
J4.8	68.0	
Bayesian Network: 1 Parent	69.3	
Bayesian Network: 2 Parents	71.5	

Figure 57: No Feature Selection: Six Month Split

4.3.2 Combinations of Feature Selection and Machine Learning Algorithm

The graphs that follow show how the classification accuracies of the models built to predict a patient's expected survival time vary based on the feature selection algorithm used and the number of features selected. These models are constructed over the varying feature selection algorithms using several different machine learning algorithms. We use these graphs to find the combinations of feature selection and machine learning algorithm that have the highest classification accuracy for this dataset.

Note that the highest classification accuracy obtained by constructing models with no feature selection is 71.5%. We look to use feature selection to improve upon this classification accuracy.

Figure 58 shows how varying the number of attributes selected by gain ratio attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall decrease in the classification accuracies as the number of attributes is increased from 10 to 100. Models constructed using naïve Bayes in particular do well when only 10 attributes are selected with a classification accuracy of over 71.5% but this classification accuracy gradually decreases as more attributes are selected. Artificial neural networks with two hidden units respond differently to varying numbers of attributes selected. This model has a peak classification accuracy of over 71.5% when 50 attributes are selected that decreases when more or less attributes are selected.

Figure 59 shows how varying the number of features selected by principal components effects the classification accuracy of several machine learning algorithms. There is a small overall decrease in the classification accuracies as the number of features selected is increased. There are two of the algorithms that are able to construct models with classification accuracies above 71.5% when 10 features are selected. This is the peak number of features to select with these two algorithms as when more features are selected the classification accuracy decreases. These two algorithms are artificial neural networks with one hidden unit and logistic regression.

Figure 60 shows how varying the number of features selected by ReliefF attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 40. Once 40 attributes have been selected, the classification accuracies remain relatively constant. This varies by the machine learning algorithm, however. Models constructed using artificial neural networks with one hidden unit and artificial neural networks with two hidden units have reach a peak accuracy that is over

71.5% with only 10 attributes selected after which increasing the number of attributes selected decreases the classification accuracy.

Figure 61 shows how varying the number of features selected by support vector machine attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall peak in the classification accuracies as the number of attributes is selected is between 50 and 60 with the classifications accuracies decreasing when more or less attributes are selected. This increase in classification accuracy is most predominate with the support vector machine constructed using a kernel function with an exponent of 0.9. This model has a peak classification accuracy when 60 attributes are selected at which point the accuracy is above 71.5%.

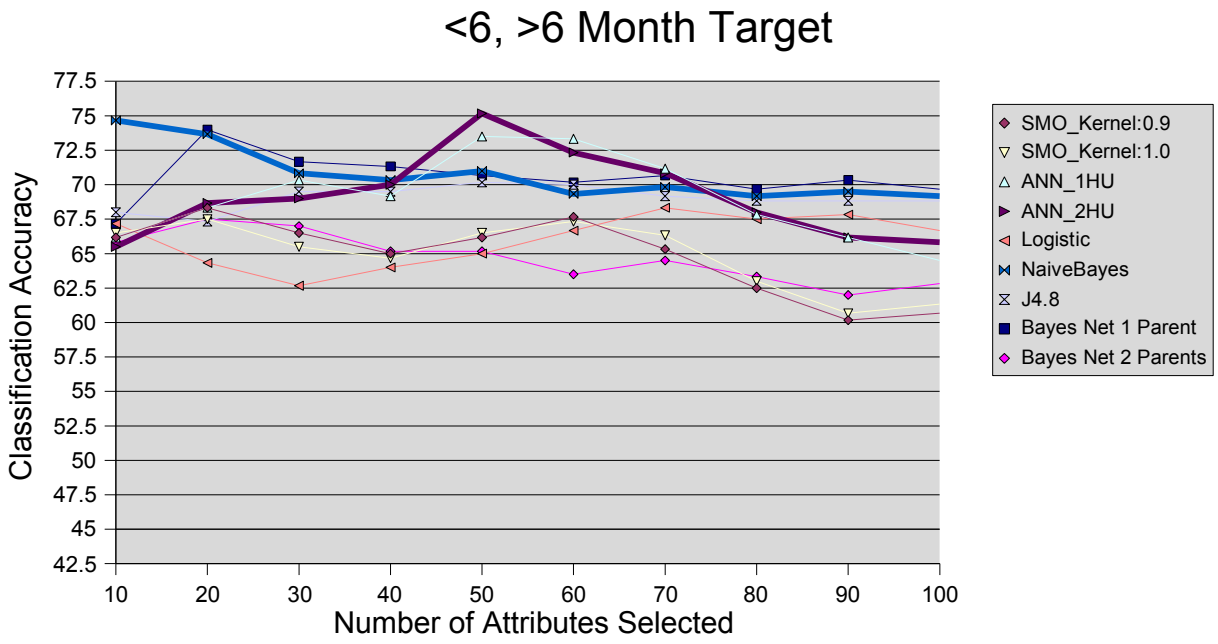

Figure 58: Gain Ratio Attribute Selection: Six Month Split

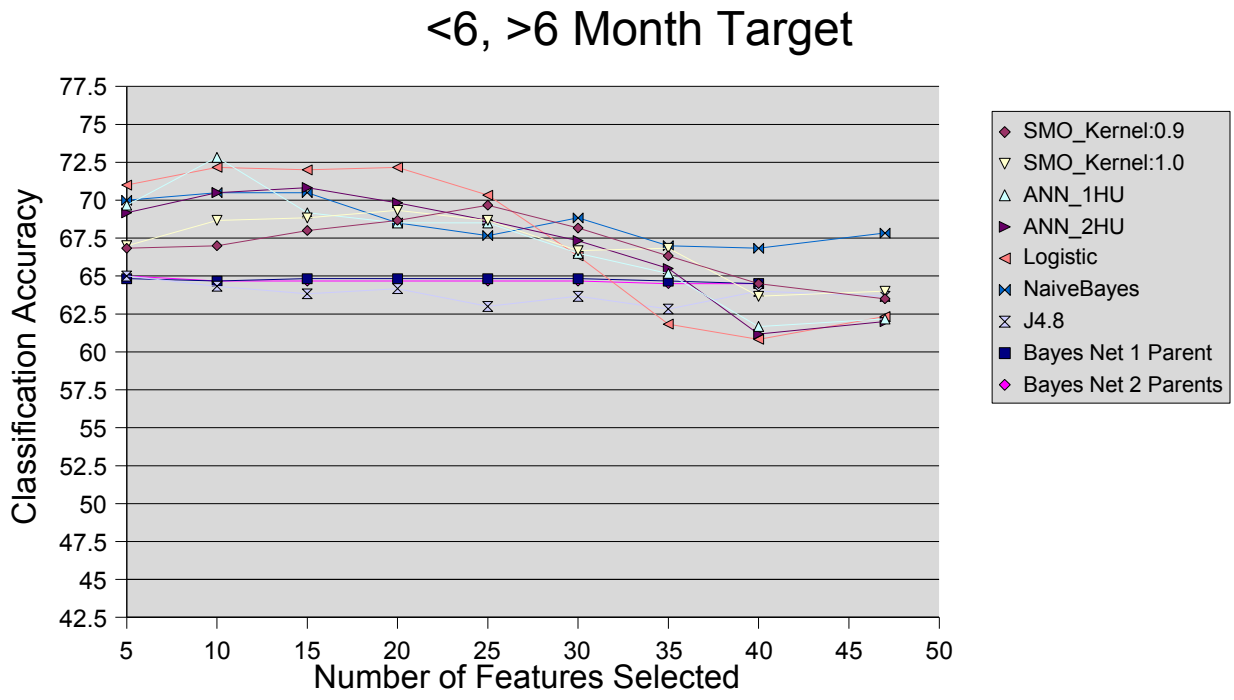

Figure 59: Principal Components: Six Month Split

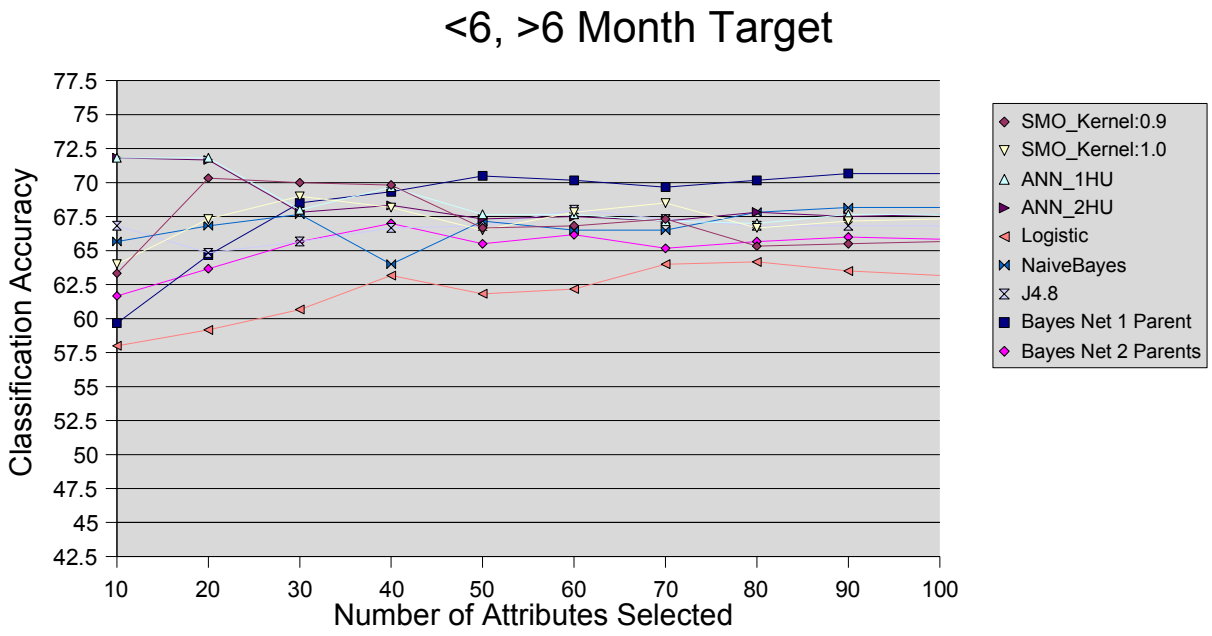

Figure 60: ReliefF Attribute Selection: Six Month Split

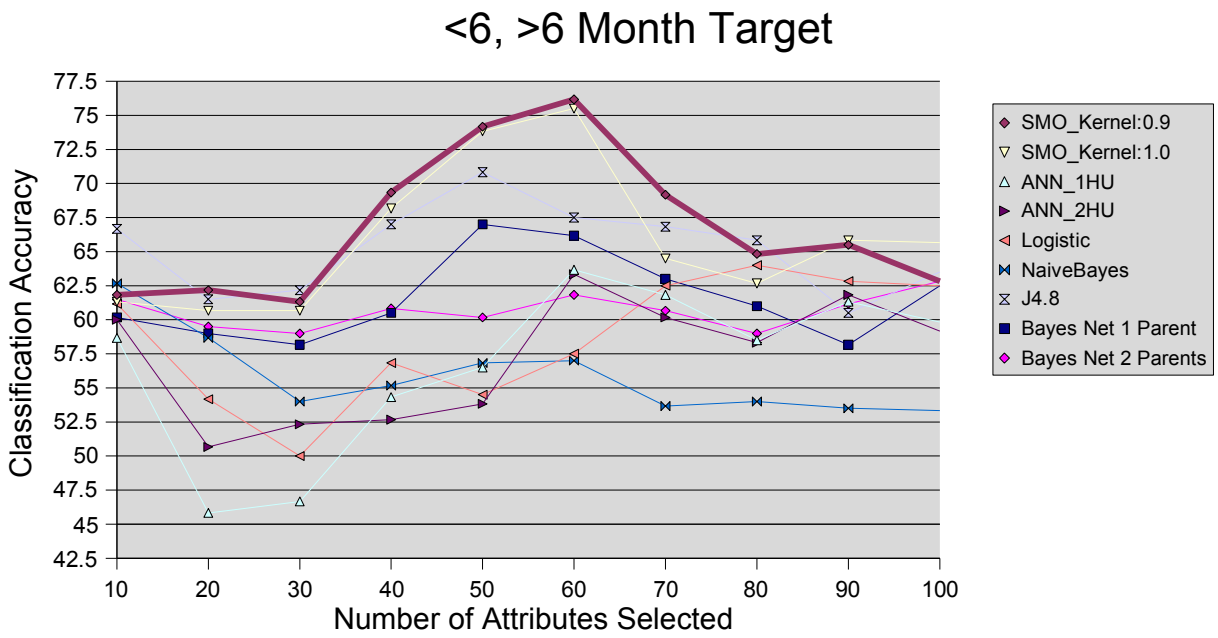

Figure 61: Support Vector Machine Attribute Selection: Six Month Split

4.3.2.1 Baseline Models

Figure 57 shows that over this dataset the classification accuracy of logistic regression with no feature selection is 61.3% and that of ZeroR is 66.7%. There is no statistically significant difference between logistic regression and ZeroR.

4.3.2.2 1st Noteworthy Combination: Support Vector Machines

The highest classification accuracy obtained over this dataset is 76.2% resulting from a model constructed using support vector machines with a kernel exponent of 0.9. The top 60 attributes are selected to build this model using support vector machine attribute selection. Figure 61 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm reaches a peak classification accuracy when 60 attributes are selected using support vector machine attribute selection where if more or less attributes are selected the classification accuracy decreases. This combination of feature selection and machine learning algorithm has a classification accuracy that

is a statistically significant improvement ($p<0.05$) over that obtained by Logistic Regression. There is no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.3.2.3 Model Similar to 1st Noteworthy Combination

The second highest classification accuracy obtained over this dataset is 75.5% resulting from a model constructed using support vector machines with a linear kernel function. The top 60 attributes are selected to build this model using support vector machine attribute selection. Figure 61 shows this combination of feature selection and machine learning algorithm. This machine learning algorithm is very similar to the one with the highest classification accuracy and uses the same feature selection. This combination also shows the same patter of reaching a peak when 60 attributes are selected. For these reasons, we will not be using this combination of feature selection and machine learning algorithm for further analysis of this dataset.

4.3.2.4 2nd Noteworthy Combination: Artificial Neural Network with Two Hidden Units

The third highest classification accuracy obtained over this dataset is 75.2% resulting from a model constructed using artificial neural networks with two hidden unit trained over 2,000 epochs. The top 50 attributes are selected to build this model using gain ratio attribution selection. Figure 58 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a peak classification accuracy when 50 attributes are selected using gain ratio attribute selection where if more or less attributes are selected the classification accuracy decreases. This combination of feature selection and machine learning algorithm has a classification accuracy that is a statistically significant improvement ($p<0.05$) over that obtained by Logistic Regression. There is no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.3.2.5 3rd Noteworthy Combination: Naïve Bayes

The fourth highest classification accuracy obtained over this dataset is 74.7% resulting from a model constructed using naïve Bayes. The top 10 attributes are selected to build this model using gain ratio attribute selection. Figure 58 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm results in the highest classification accuracy when 10 attributes are selected using gain ratio attribute selection and with a decreasing accuracy as the number of attributes that are selected increases. This combination of feature selection and machine learning algorithm has a classification accuracy that is a statistically significant improvement ($p<0.05$) over that obtained by Logistic Regression. There is no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.3.2.6 Summary of Noteworthy Combinations

- 76.2%: Support Vector Machines, Kernel:0.9, using SVM to select 60 attributes
- 75.2: Artificial Neural Networks, Two Hidden Units, using Gain Ratio to select 50 attributes
- 74.7%: Naïve Bayes, using Gain Ratio to select 10 attributes

Note no statistically significant difference ($p<0.05$) between these three models.

4.3.3 Models Produced

4.3.3.1 Best Model with No Feature Selection

The best model with no feature selection resulted from a model constructed using a Bayesian network with a maximum of two parents.

4.3.3.1.1 Machine Learning Model

The Bayesian network constructed over these attributes is much larger than the one shown in due to being constructed with over two times as many attributes. Figure 62 provides the Weka output of this model, listing for each attribute its parents. Note that the attributes with only survival as a parent, 88 of the 189 attributes, omitted from this output.

```
Bayes Network Classifier
not using ADTree
#attributes=190 #classindex=189
Network structure (nodes followed by parents)
SxJaun(2): Survival PresumptiveDx
SxChola(2): Survival SxChole
SxBC(2): Survival SxChole
SxNau(2): Survival SxChola
SxVom(2): Survival SxNau
SxCCS(2): Survival SxJaun
SxPru(2): Survival SxJaun
SxAbd(2): Survival SxJaun
SxBack(2): Survival SxJaun
SxOT(2): Survival SxJaun
CxDiab(2): Survival SxBack
CxDiabOral(2): Survival CxDiab
CxDiabDiet(2): Survival CxDiabOral
CxDiabOnset(2): Survival SxChola
CxHyper(2): Survival CxDiab
CxPriorCancer(6): Survival SxFati
CxPriorCancerChemo(2): Survival SxBack
CxPriorCancerRadiation(2): Survival CxPriorCancerChemo
CxPriorCancerSurgery(2): Survival CxPriorCancer
SHCigarette(2): Survival SxAbd
SHAlcohol(2): Survival SHCigarette
FamilyOther1Dx(8): Survival FamilyOther1
FamilyOther2(3): Survival CxDiabOnset
FamilyOther2Dx(4): Survival FamilyOther2
CXRDx(2): Survival FamilyFatherDx
CTCeliac(2): Survival CxIHD
CTCeliacClass(2): Survival CTCeliac
CTSMA(2): Survival SHExposure
CTHepatic(2): Survival CTCeliacClass
CTSMV(2): Survival SxVom
CTSMVClass(3): Survival SHExposure
CTPortal(2): Survival CxIHD
CTPortalClass(2): Survival CTPortal
CTCeliacNode(2): Survival CXRDx
CTOtherNode(2): Survival CxPriorCancer
CTNodeOmit(2): Survival SxVom
PTCDx(2): Survival CTNodeOmit
PTCStent(2): Survival PTCDx
PTCStentType(2): Survival SxSatiety
EUSDx(2): Survival CxHF
EUSVascOmit(2): Survival SxJaun
EUSSMV(2): Survival CxDiabOnset
EUSSMVClass(2): Survival CxDiabOnset
EUSPortal(2): Survival CTPortal
EUSPortalClass(2): Survival CTPortalClass
EUSOtherNode(2): Survival EUSSMV
EUSNoNode(2): Survival EUSVascOmit
EUSStagingT(4): Survival EUSPortal
EUSStagingN(3): Survival SHCigarette
EUSCyto(7): Survival PresumptiveDx
ERCPDx(2): Survival SxJaun
ERCPStent(2): Survival ERCPDx
ERCPStentType(2): Survival EUSStagingT
Histology(11): Survival PresumptiveDx
PreOutlook(3): Survival EUSStagingT
TxResect(2): Survival PreOutlook
TxLap(2): Survival Histology
TxRadia(2): Survival Histology
TxChemo(2): Survival TxRadia
TxChemoFlu(2): Survival TxChemo
TxChemoGem(2): Survival TxChemo
TxChemoIri(2): Survival EUSDx
TxChemoLeu(2): Survival CTSMVClass
TxChemoTax(2): Survival SxInd
TxChemoOthSpecify(1): Survival
TxPal(2): Survival PreOutlook
TxPalBypass(2): Survival CxHF
TxPalRad(2): Survival TxPal
TxPalStens(2): Survival CTCeliac
TxPalGasTube(2): Survival TxPalBypass
TxPalJejTube(2): Survival TxPalStens
ResPxType(7): Survival SxJaun
ResVenRes(2): Survival CxIHD
ResVenRec(2): Survival ResVenRes
ResOrgans(4): Survival FamilyOther2
ResTransfusion(2): Survival SxCCS
ResTUnits(2): Survival ResTransfusion
ResTFFP(2): Survival ResTUnits
ResAttempt(2): Survival CxIHD
ResPOCourse(2): Survival PresumptiveDx
ResPOInfection(2): Survival SHAlcohol
ResPOLeak(2): Survival EUSPortalClass
ResPONG(2): Survival ResPOLeak
ResPOPulmComp(2): Survival ResPOCourse
ResPODischStatus(4): Survival PTCDx
ResPathT(5): Survival ResPOLeak
ResPathN(2): Survival ResPathT
ResPathM(3): Survival EUSCeliacNode
ResPathR(3): Survival ResTransfusion
ResPathV(2): Survival ResVenRes
NoResNoHandle(2): Survival TxPalBypass
NoResRefused(2): Survival TxResect
NoResMagnitude(2): Survival CTCeliacNode
NoResCeliacInvolve(2): Survival CTCeliac
NoResSMAInvolve(2): Survival TxPalStens
NoResMetastatic(2): Survival TxPal
Gender(2): Survival EUSStagingN
MedOncName(5): Survival CxPriorCancerRadiation
SurOncName(3): Survival MedOncName
RadOncName(6): Survival SxInd
GIMDName(2): Survival SHExposure
Survival(2):
LogScore Bayes: -2853.4212887809313
LogScore BDeu: -1480.37707937041
LogScore MDL: -5385.451563589298
LogScore ENTROPY: -3328.0434210727
LogScore AIC: -4333.043421072699
```

Figure 62: Bayesian Network: Network Structure Detail

There are three attributes that are parents for more than three nodes. They are PresumptiveDx

are CxIHD are each the parents of four nodes and SxJaun is the parent of 8 nodes. These attributes encode information about the partient pre-operative diagnosis, co-morbidity of ischemic heart disease, and presence of jaundice when patient was admitted respectively.

4.3.3.2 1st Noteworthy Combination: Support Vector Machine

The best combination of feature selection and machine learning algorithm resulted from a model constructed using support vector machines constructed using a kernel function with an exponent of 0.9. The top 60 attributes are selected to build this model using support vector machine attribute selection.

4.3.3.2.1 Feature Selection

These 60 attributes are listed in Figure 63. Note that all of the attributes are pre-operative including a mix of patient imaging tests, lab values, initial symptoms, and medical history. Also note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

PresumptiveDx	LabBili	EUSHepaticClass	PTCDx	SxBack
CTHepatic	LabALT	EUSInferior	CTTumorSizeY	SxFati
CTHepaticClass	LabAlka	EUSSMA	CTCeliacNode	SxInd
CTCeliacClass	LabCEA	EUSHepatic	CTNodeOmit	SxPru
CTSMAClass	LabAlb	EUSSMAClass	CTOtherNode	CxHF
CTSMA	LabCA19-9	EUSPortal	EUSVascOmit	CxResp
CTSMVClass	CTDx	EUSCeliacNode	EUSCeliacClass	CxIHD
CTPortalClass	CTCeliac	EUSPortalClass	EUSCeliac	SxDyspha
CTPortal	CTVascOmit	EUSInferiorClass	PTCStent	SxOT
CTInferior	LabAST	EUSSMVClass	EUSDx	SxSatiety
CTSMV	CXRDx	EUSSMV	PTCStentType	SxWtloss
CTInferiorClass	LabAmylase	CTTumorSizeX	SxAbd	SxJaun

Figure 63: Top 60 Attributes Selected By Support Vector Machines

4.3.3.2.2 Machine Learning Model

The support vector machines constructed over these attributes is shown in Figure 64. Support

vector machines with a non linear kernel are very cryptic. Note that the 44 support vectors shown in the figure are the subset of transformed instances that made it into the model. Within each support vector is a column for each of the 60 selected attributes. This works out to 67 columns because the non -binary attributes need to be split into binary attributes. Each attribute within the support vectors is weighted from zero to one. Each support vector is likewise weighted from zero to one. The support vectors are summed together. If the resulting sum is greater than zero, the classification is positive. A positive classification in this example indicates the expected survival is >6 months. A negative classification therefore indicates an expected survival of <6 months.

```
Classifier Model
SMO

Kernel used:
  Poly Kernel: K(x,y) = <x,y>^0.9

Classifier for classes: '(-inf-5.753425]', '(5.753425-inf)'

BinarySMO

    0.9295 * <1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0.139851 0.086029 0.133692 0.003702 0.709677 0.113312 0 0 0 0.408115 0 0.185545
           0 0 0 0 0 0 0 1 0 0 0 0.33871 0 0.4 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 > * X]
-     1    * <1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0.068069 0.01594 0.047106 0.000185 0.645161 0.00425 0 0 0 0.124105 0 1 0 0 0 0 0 0
           0 1 0 0 0 0.305684 0 0.435714 0 0 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 > * X]
-     1    * <1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0.019802 0.021331 0.045985 0.000092 0.354839 0.071676 0 0 0 0.338902 0 0.096009
           0 0 0 0 0 1 0 1 0 0 0 0.305684 0 0.435714 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 > * X]
+     0.9646 * <1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0.02104 0.068448 0.015478 0.004754 0.225806 0.003542 0 0 0 0.272076 0 0.06041
           0 0 0 0 0 1 0 1 0 0 0 0.305684 0 0.435714 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 0 > * X]
+     0.6182 * <0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0.035891 0.034693 0.02961 0.039718 0.612903 0.003079 0 0 0 0.541766 1
           0.044229 0 0 0 0 0 0 0 1 0 0 0 0 0 0.1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 > * X]
...
+     0.4901 * <1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 1 0 0.27599 0.05579 0.095334 0 0.548387 0.00026 0 1 0 0.491647 0 0.039914 0 0 0 0 0
           0 0 1 0 0 0 0.290323 0 0.525 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 > * X]
+     0.3278 * <1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0.173267 0.014534 0.043293 1 0.645161 0.00014 0 0 0 0.121718 0 0.052859 0 0 0
           0 0 1 0 0 0 0 0 0.548387 0 0.6 0 1 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 > * X]
+     0.6065

Number of support vectors: 44

Number of kernel evaluations: 1802 (96.637% cached)
```

Figure 64: Support Vector Machines with 0.9 Exponent

4.3.3.3 2nd Noteworthy Combination: Artificial Neural Networks with Two Hidden Units

The second best combination of feature selection and machine learning algorithm resulted from a model constructed using artificial neural networks using two hidden units and trained over two thousand epochs. The top 50 attributes are selected to build this model using gain ratio attribute selection.

4.3.3.3.1 Feature Selection

These 60 attributes are listed in Figure 65. Note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

The progression of weights displayed in Figure 66 indicate that from the 1st attribute selected to the 20th attribute selected there is a quick decrease in the weights assigned by the gain ratio of around two thirds. However, the initial gain ratios presented here are higher than the gain ratios calculated for the top attributes selected over the dataset with the nine month split. This indicates that fewer attributes may be required over this dataset to reach the same predictive accuracy as was obtained in the dataset with the nine month split. In comparing Figure 58, Gain Ratio Attribute Selection: Six Month Split, and Figure 38, Gain Ratio Attribute Selection: Nine Month Split, we see that over this dataset we get a much higher classification accuracy with only 10 or 20 attributes than in the dataset with the nine month split.

Gain Ratio Weight	Attribute		Gain Ratio Weight	Attribute
0.29	NoResMagnitude		0.09	TxResect
0.29	PTCStent		0.08	EUSCeliacNode
0.29	PTCDx		0.08	TxChemoTax
0.29	TxPalStens		0.08	CxIHD
0.29	NoResNoHandle		0.08	ResPOPulmComp
0.26	CxPriorCancerChemo		0.07	Histology
0.26	TxPalBypass		0.06	ResTFFP
0.26	NoResSMAInvolve		0.06	CTCeliacNode
0.22	CTHepatic		0.06	CxPriorCancerRadiation
0.22	TxPalJejTube		0.06	SxChola
0.22	TxPalGasTube		0.06	ResPOInfection
0.22	CTSMA		0.05	SxSatiety
0.22	SxBC		0.05	CxDiab
0.22	SHExposure		0.05	NoResRefused
0.18	TxPal		0.05	CxPriorCancerSurgery
0.16	ResPODischStatus		0.05	NoResMetastatic
0.15	EUSDx		0.05	PresumptiveDx
0.12	SxInd		0.04	EUSNoNode
0.11	SurOncName		0.03	ResTUnits
0.11	CxHF		0.03	MedOncName
0.09	CxDiabDiet		0.03	SHCigarette
0.09	EUSSMV		0.02	SxOT
0.09	TxChemoIri		0.02	CTSMV
0.09	CXRDx		0.02	NoResPVInvolve
0.09	PreOutlook		0.02	CxDiabOral

Figure 65: Top 50 Attributes Selected By Gain Ratio

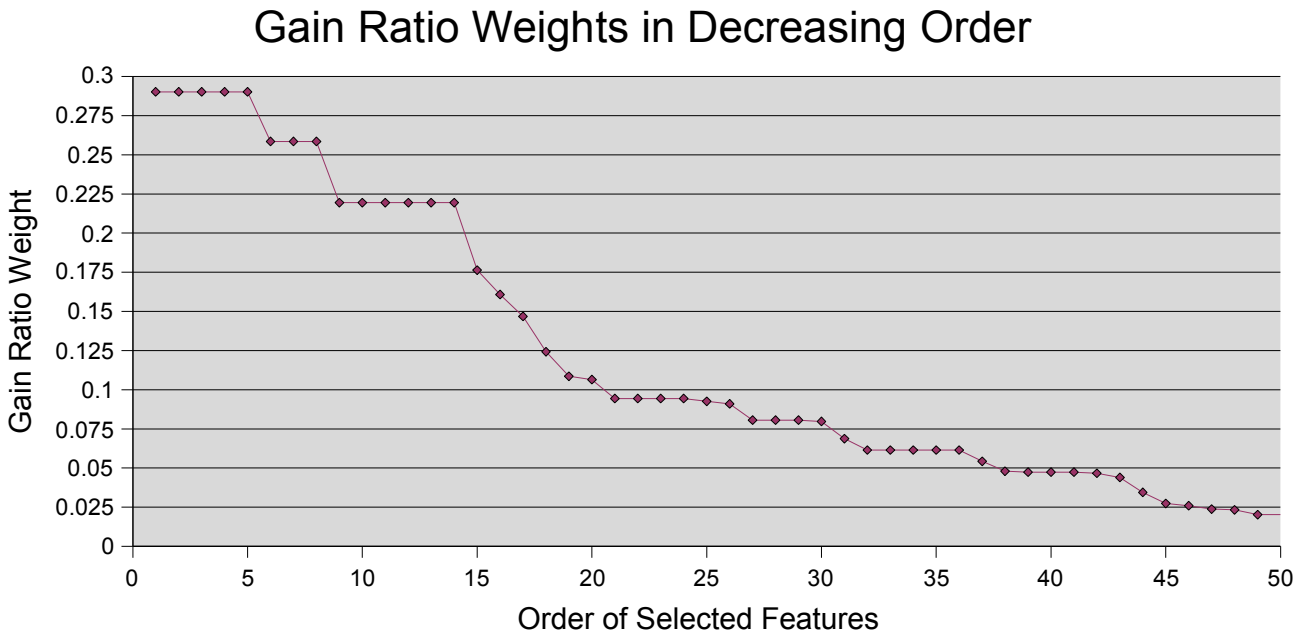

Figure 66: Gain Ratio Weights in Decreasing Order

4.3.3.3.2 Machine Learning Model

As has been discussed already, artificial neural networks with two hidden units are not human readable. The underlying model is cryptic to analyze and make any conclusions from.

4.3.3.4 3rd Noteworthy Combination: Naïve Bayes

The third best combination of feature selection and machine learning algorithm resulted from a model constructed using naïve Bayes. The top 10 attributes are selected to build this model using gain ratio attribute selection.

4.3.3.4.1 Feature Selection

These 10 attributes are listed in Figure 67. Note that a listing of all of the attributes in the datasets, along with a brief description about the information they capture, appears in Appendix A.

The weights calculated by gain ratio over this dataset have already been discussed and are displayed in Figure 66.

Gain Ratio Weight	Attribute
0.29	NoResMagnitude
0.29	PTCStent
0.29	PTCDx
0.29	TxPalStens
0.29	NoResNoHandle
0.26	CxPriorCancerChemo
0.26	TxPalBypass
0.26	NoResSMAInvolve
0.22	CTHepatic
0.22	TxPalJejTube

Figure 67: Top 10 Attributes Selected By Gain Ratio

4.3.3.4.2 Machine Learning Model

The naïve Bayes model constructed using the 10 selected attributes is shown in Figure 68. This model contains all of the required information to use Bayes' Theorem, Formula 6, to calculate the probability of each target given an input string.

Note that this table shows that for these attributes selected by the gain ratio, when the target is >6 months all of the attributes selected have the same value in the training instances. The instances with a target of <6 months often have the same value as the >6month class, but there is some diversity. This observation about the selected attributes is concerning. With so little variation, and when you consider the prior probability of the >6 month dataset is 0.66, this model is at risk of becoming ZeroR. In practice this combination does have a high overall classification accuracy.

```
Classifier Model
Naive Bayes Classifier

Class '(-inf-5.753425]': Prior probability = 0.34

NoResMagnitude:  Discrete Estimator. Counts =  18 4  (Total = 22)
PTCStent:  Discrete Estimator. Counts =  18 4  (Total = 22)
PTCDx:  Discrete Estimator. Counts =  18 4  (Total = 22)
TxPalStens:  Discrete Estimator. Counts =  18 4  (Total = 22)
NoResNoHandle:  Discrete Estimator. Counts =  18 4  (Total = 22)
CxPriorCancerChemo:  Discrete Estimator. Counts =  19 3  (Total = 22)
TxPalBypass:  Discrete Estimator. Counts =  19 3  (Total = 22)
NoResSMAInvolve:  Discrete Estimator. Counts =  19 3  (Total = 22)
CTHepatic:  Discrete Estimator. Counts =  20 2  (Total = 22)
TxPalJejTube:  Discrete Estimator. Counts =  20 2  (Total = 22)


Class '(5.753425-inf)': Prior probability = 0.66

NoResMagnitude:  Discrete Estimator. Counts =  41 1  (Total = 42)
PTCStent:  Discrete Estimator. Counts =  41 1  (Total = 42)
PTCDx:  Discrete Estimator. Counts =  41 1  (Total = 42)
TxPalStens:  Discrete Estimator. Counts =  41 1  (Total = 42)
NoResNoHandle:  Discrete Estimator. Counts =  41 1  (Total = 42)
CxPriorCancerChemo:  Discrete Estimator. Counts =  41 1  (Total = 42)
TxPalBypass:  Discrete Estimator. Counts =  41 1  (Total = 42)
NoResSMAInvolve:  Discrete Estimator. Counts =  41 1  (Total = 42)
CTHepatic:  Discrete Estimator. Counts =  41 1  (Total = 42)
TxPalJejTube:  Discrete Estimator. Counts =  41 1  (Total = 42)
```

Figure 68: Naive Bayes Model

4.3.4 ROC Curves

ROC curves for the combinations of feature selection and machine learning algorithm with the highest classification accuracy appear at the end of this section. The combinations will be presented and discussed in the following order:

- 61.3%: Logistic Regression
- 76.2%: Support Vector Machines, Kernel:0.9, using SVM to select 60 attributes
- 75.2: Artificial Neural Networks, Two Hidden Units, using Gain Ratio to select 50 attributes
- 74.7%: Naïve Bayes, using Gain Ratio to select 10 attributes

4.3.4.1 Baseline Model: Logistic Regression

Figure 69 shows the ROC curve for logistic regression. The area under this curve is 0.63. This curve shows that to correctly predict 90% of the patients who will survive for more than six months,

you have to incorrectly predict that 74% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 63% of the patients who will not survive for six months will survive for greater than six months.

4.3.4.2 1st Noteworthy Combination: Support Vector Machines

Figure 70 shows the ROC curve for the combination of feature selection and machine learning algorithm with the highest classification accuracy over this dataset. The area under this curve is 0.67. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 56% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 50% of the patients who will not survive for six months will survive for greater than six months. This curve is better than logistic regression due to both a larger area under the curve and having a better trade off between the true and false positive rates.

4.3.4.3 2nd Noteworthy Combination: Artificial Neural Networks with One Hidden Unit

Figure 71 shows the ROC curve for the combination of feature selection and machine learning algorithm with the second highest classification accuracy over this dataset. The area under this curve is 0.73. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 55% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 43% of the patients who will not survive for six months will survive for greater than six months. This curve is better than both logistic regression and the curve for the highest classification accuracy due to having both a larger area under the curve and having a better trade off between the true and false positive rates.

4.3.4.4 3rd Noteworthy Combination: Naïve Bayes

Figure 72 shows the ROC curve for the combination of feature selection and machine learning algorithm with the third highest classification accuracy over this dataset. Note that this curve is the most irregular of the curves for this dataset with the true positive rate and the false positive rate appear to increase at very close to the same rate until a false positive rate of 50% is reached. At this point, the true positive rate increases much faster than the false positive rate until it levels off with a true positive rate approaching 100%. The area under this curve is 0.59. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 63% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 58% of the patients who will not survive for six months will survive for greater than six months. Though this curve does has a slightly better trade off between the true and false positive rates when compared to the curve produced by logistic regression, the other two curves from the higher classification accuracies both have much better trade offs. In addition, this curve has the lowest area under the cure of any of the other curves for this dataset.

4.3.4.5 Summary

The three combinations of feature selection and machine learning algorithm with the highest classification accuracies are ranked by their ROC curves as follows:

1. Artificial Neural Networks, Two Hidden Units, using Gain Ratio to select 50 attributes
2. Support Vector Machines, Kernel:0.9, using SVM to select 60 attributes
3. Naïve Bayes, using Gain Ratio to select 10 attributes

4.3.4.6 Curves

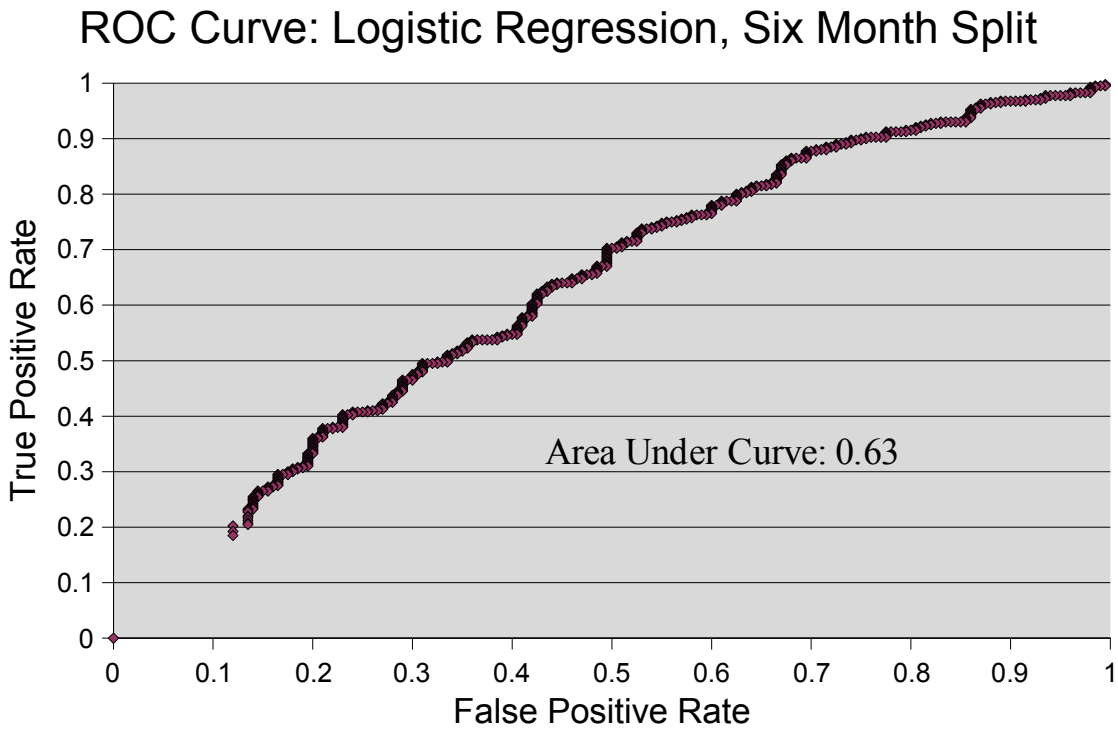

Figure 69: ROC Curve - Logistic Regression: Six Month Split

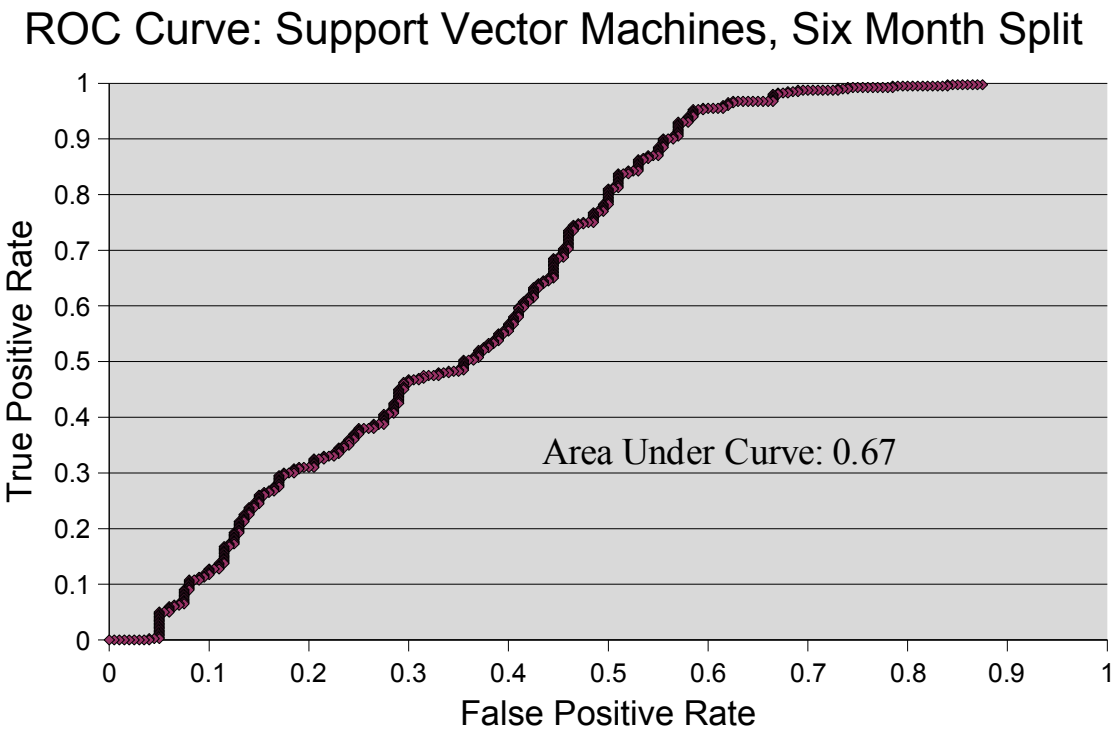

Figure 70: ROC Curve – Support Vector Machines: Six Month Split

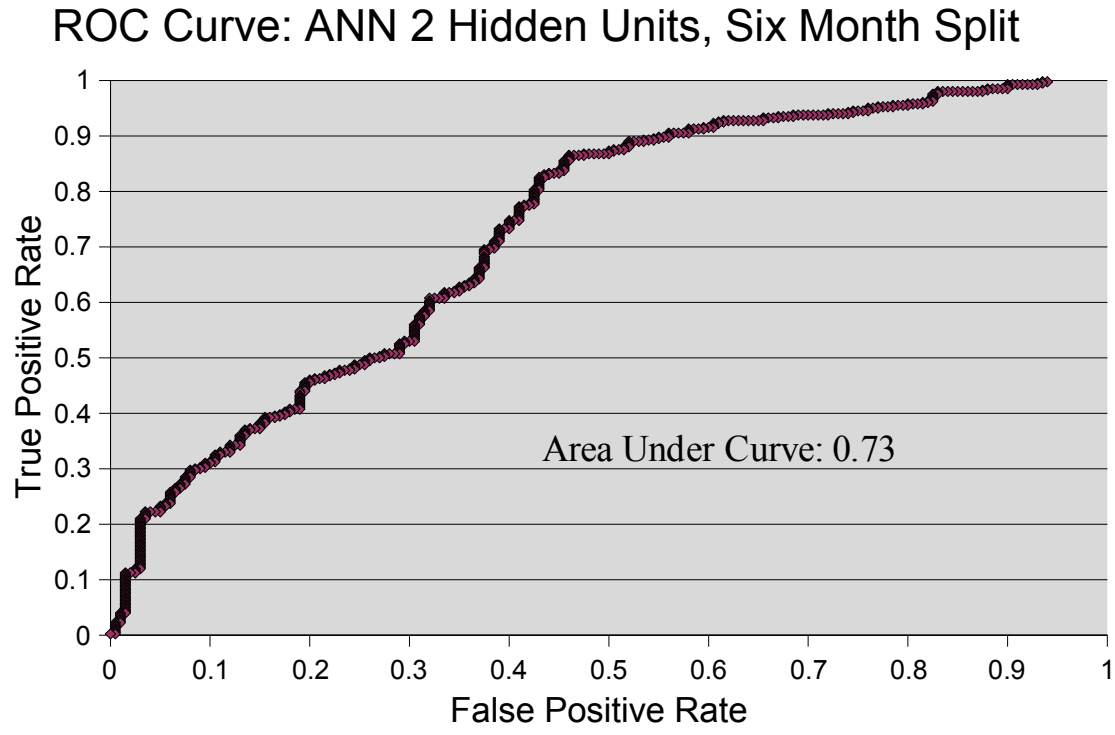

Figure 71: ROC Curve – Artificial Neural Network, Two Hidden Units: Six Month Split

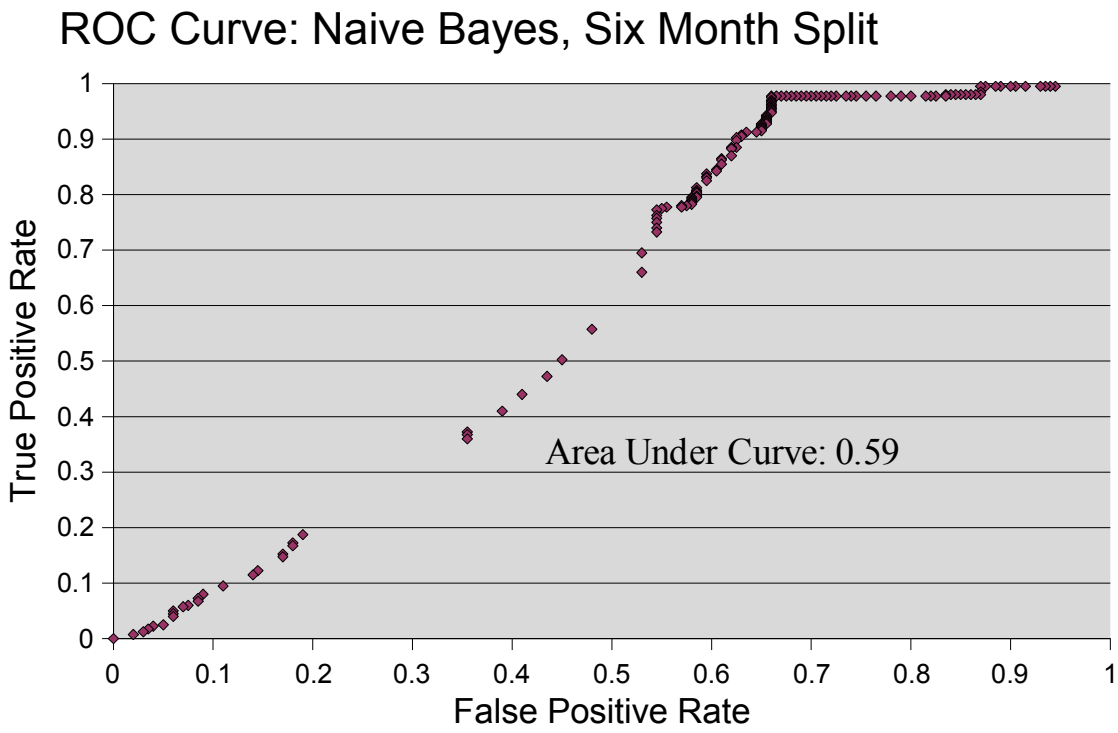

Figure 72: ROC Curve – Naïve Bayes: Six Month Split

4.3.5 Meta-Learning

The following three Combinations of Feature Selection and Machine Learning Algorithm were found in the previous section to produce the highest classification accuracies over the dataset with a six month split:

- 76.2%: Support Vector Machines, Kernel:0.9, using SVM to select 60 attributes
- 75.2: Artificial Neural Networks, Two Hidden Units, using Gain Ratio to select 50 attributes
- 74.7%: Naïve Bayes, using Gain Ratio to select 10 attributes

4.3.5.1 Bagging

Figure 73 shows the effect that bagging has on each of the three best combinations of feature selection and machine learning algorithm selected for this dataset. Overall there is almost no change or an increase in the standard deviation. On top of that there is also a decrease in the classification accuracy. For these reasons, bagging is not useful over this dataset.

Original		Bagging	
%Correct	σ	%Correct	σ
76.2	16.5	72.2	17.3
75.2	15.1	72.0	16.1
74.7	12.4	74.5	12.4

Figure 73: Bagging: Six Month Split

4.3.5.2 Boosting

Figure 74 shows the effect boosting has on each of the three combinations of best feature selection and machine learning algorithm for this dataset. Overall there is a decrease in the classification accuracy when boosting is used. The standard deviation is also increased by boosting in two out of the three cases. For these reasons, boosting is not useful over this dataset.

Original		Boosting	
%Correct	σ	%Correct	σ
76.2	16.5	64.2	18.9
75.2	15.1	65.2	15.1
74.7	12.4	67.0	18.5

Figure 74: Boosting: Six Month Split

4.3.5.3 Stacking

In this section we investigate the use of stacking to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Stacking is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

The results of the runs of Stacking over this dataset are presented in Figure 75. The far left column shows the level-1 classifier used to make the final target class prediction. Note that no result below has a classification accuracy greater than those of the initial models. The best classification accuracies are when the model with the second highest classification accuracy is left out.

Level-1 Classifier	Level – 0 Models			
	SVM, ANN 2HU, Naive Bayes	ANN 2HU, Naive Bayes	SVM, Naïve Bayes	SVM, ANN 2HU
ANN 1 Hidden Unit	61.8	61.5	63.2	58.2
ANN 2 Hidden Units	64.3	63.8	62.8	59.0
Linear Regression	69.8	69.8	71.0	71.0
LWL	68.8	70.8	72.5	69.3

Figure 75: Stacking Results: Six Month Split

4.3.5.4 Our Model Selector

In this section we investigate the use of our model selector to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Our model selector is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

The results of the runs of our model selector over this dataset are presented in Figure 76. The far left column shows the level-1 classifier used to determine the model that makes the final target class prediction. The classifiers with feature selection algorithms listed are run using the attribute selected classifier.

This meta-learning algorithm was not able to improve the classification accuracy by combining these algorithms. However, there were several models constructed where combination with the highest classification accuracy was not used that had classification accuracies very close to the models that were combined in the construction of the model. In particular classification accuracies of the models with a level-1 classifier using J4.8 selecting 70 attributes using support vector machine attribute selection and using LWL are statistically significantly higher than logistic regression. These are highlighed in Figure 76.

Level-1 Classifier	Level – 0 Models			
	SVM, ANN 2HU, Naive Bayes	ANN 2HU, Naive Bayes	SVM, Naïve Bayes	SVM, ANN 2HU
ANN 1 Hidden Unit	69.2	72.5	71.0	69.3
ANN 2 Hidden Units	69.5	72.2	71.0	68.7
ANN 1 HU & ReliefF_90	70.3	72.0	73.0	69.8
J48 & SVM_70	69.0	74.8	70.7	68.7
Logistic & PrincComp_15	68.5	73.3	69.0	69.2
NaiveBayes & GainRatio_30	69.3	72.8	72.5	70.2
SMO & ReliefF_40	70.0	72.3	71.2	69.2
J48	68.8	73.7	71.8	68.5
Logistic	69.5	72.7	73.0	72.0
LWL	70.2	74.2	71.5	71.0
NaiveBayes	67.5	71.0	69.7	68.3
SMO	70.0	71.8	71.2	70.0

Note: HU stands for Hidden Unit(s)

Figure 76: Our Model Selector: Six Month Split

4.3.5.4.1 Model Constructed

The model constructed by our model selector combined the models constructed using naïve Bayes and artificial neural network with two hidden units together using a J4.8 decision tree with support vector machines selecting the top 70 attributes. This result is shown in Figure 76, see first highlighted row and column. Since the classification accuracy of this model is statistically significantly great than logistic regression's classification accuracy, we want to look into how this model is constructed.

Figure 77 shows the probability distributions of every instance over each model. The actual target is also in the table along with a label of which model was correct, if none of the models are correct, or if all of the models are correct. In thirty six out of these sixty instances both models produce the correct classification. In eight neither instance produces the correct prediction. This leaves sixteen models where if we can predict the correct model, we can make the correct prediction. Note that when both artificial neural networks and naïve Bayes both predict the same target, artificial neural networks is overall much more certain of its prediction.

Artificial Neural Network	Naïve Bayes	Actual Target	Which Model
{0.72,0.28}	{0.12,0.88}	<6 Months	ANN
{0.91,0.09}	{0.32,0.68}	<6 Months	ANN
{0.95,0.05}	{0.08,0.92}	<6 Months	ANN
{0.88,0.12}	{0.39,0.61}	<6 Months	ANN
{0.99,0.01}	{0.11,0.89}	<6 Months	ANN
{0.05,0.95}	{0.09,0.91}	>6 Months	Both
{0,1}	{0.39,0.61}	>6 Months	Both
{0,1}	{0.11,0.89}	>6 Months	Both
{0,1}	{0.38,0.62}	>6 Months	Both
{0.99,0.01}	{0.78,0.22}	<6 Months	Both
{0.01,0.99}	{0.12,0.88}	>6 Months	Both
{0.01,0.99}	{0.09,0.91}	>6 Months	Both
{0.95,0.05}	{1,0}	<6 Months	Both
{0,1}	{0.38,0.62}	>6 Months	Both
{0.01,0.99}	{0.11,0.89}	>6 Months	Both
{0.99,0.01}	{0.76,0.24}	<6 Months	Both
{0.93,0.07}	{0.96,0.04}	<6 Months	Both
{0.01,0.99}	{0.45,0.55}	>6 Months	Both
{0,1}	{0.11,0.89}	>6 Months	Both
{0,1}	{0.11,0.89}	>6 Months	Both
{0,1}	{0.11,0.89}	>6 Months	Both
{0.01,0.99}	{0.11,0.89}	>6 Months	Both
{0,1}	{0.07,0.93}	>6 Months	Both
{0.01,0.99}	{0.07,0.93}	>6 Months	Both
{0,1}	{0.11,0.89}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0.99,0.01}	{0.85,0.15}	<6 Months	Both
{0.01,0.99}	{0.12,0.88}	>6 Months	Both
{0.01,0.99}	{0.12,0.88}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0.07,0.93}	{0.1,0.9}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0.02,0.98}	{0.1,0.9}	>6 Months	Both
{0.01,0.99}	{0.12,0.88}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0.01,0.99}	{0.12,0.88}	>6 Months	Both
{0,1}	{0.12,0.88}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0,1}	{0.1,0.9}	>6 Months	Both
{0.02,0.98}	{0.84,0.16}	<6 Months	Naive Bayes
{0.95,0.05}	{0.08,0.92}	>6 Months	Naive Bayes
{0.99,0.01}	{0.48,0.52}	>6 Months	Naive Bayes
{0.99,0.01}	{0.46,0.54}	>6 Months	Naive Bayes
{0.89,0.11}	{0.12,0.88}	>6 Months	Naive Bayes
{0.59,0.41}	{0.47,0.53}	>6 Months	Naive Bayes
{0.98,0.02}	{0.1,0.9}	>6 Months	Naive Bayes
{0.99,0.01}	{0.1,0.9}	>6 Months	Naive Bayes
{0,1}	{0.85,0.15}	<6 Months	Naive Bayes
{0.99,0.01}	{0.47,0.53}	>6 Months	Naive Bayes
{0.95,0.05}	{0.47,0.53}	>6 Months	Naive Bayes
{0,1}	{0.1,0.9}	<6 Months	Neither
{0.03,0.97}	{0.11,0.89}	<6 Months	Neither
{0,1}	{0.1,0.9}	<6 Months	Neither
{0,1}	{0.1,0.9}	<6 Months	Neither
{0,1}	{0.09,0.91}	<6 Months	Neither
{0,1}	{0.1,0.9}	<6 Months	Neither
{0.06,0.94}	{0.07,0.93}	<6 Months	Neither
{0,1}	{0.1,0.9}	<6 Months	Neither

Note that probability distribution {x,y} denotes the predicted probability of survival for <6 months and >6 months are x and y respectively

Figure 77: Probability Distributions of Each Combined Model

Figure 78 lists the seventy attributes selected from the dataset to train the level-1 classifier by support vector machine attribute selection. All of these attributes are pre-operative.

PresumptiveDx	LabAlka	EUSSMAClass	EUSCeliacClass	SxOT
CTHepatic	LabCEA	EUSPortal	EUSCeliac	SxSatiety
CTHepaticClass	LabAlb	EUSCeliacNode	PTCStent	SxWtloss
CTCeliacClass	LabCA19-9	EUSPortalClass	EUSDx	SxJaun
CTSMAClass	CTDx	EUSInferiorClass	PTCStentType	SxWtlossP
CTSMA	CTCeliac	EUSSMVClass	SxAbd	DemECOG
CTSMVClass	CTVascOmit	EUSSMV	SxBack	DemWeight
CTPortalClass	LabAST	CTTumorSizeX	SxFati	DemHeight
CTPortal	CXRDx	PTCDx	SxInd	SxNau
CTInferior	LabAmylase	CTTumorSizeY	SxPru	SxCCS
CTSMV	EUSHepaticClass	CTCeliacNode	CxHF	SxVom
CTInferiorClass	EUSInferior	CTNodeOmit	CxResp	SxChole
LabBili	EUSSMA	CTOtherNode	CxIHD	SxBC
LabALT	EUSHepatic	EUSVascOmit	SxDyspha	SxChola

Figure 78: Top 70 Attributes Selected By Support Vector Machines

Figure 79 shows the decision tree constructed by J4.8 using the seventy attributes selected by support vector machines. Of the seventy attributes selected, only seven of them are used in the construction of this decision tree. It is interesting the mix of attributes that are in the decision tree. The first attribute is a presentation symptom of back pain which the Bayesian networks also find to be the parent of several other attributes. This is followed by information relating to an imaging score and finished by a lab test.

```
Classifier Model
J48 pruned tree
------------------

SxBack = FALSE
|  SxJaun = TRUE
|  |  EUSVascOmit = FALSE
|  |  |  LabAlb <= 2.4: Naive Bayes (3.16/0.16)
|  |  |  LabAlb > 2.4: Artifical Neural Network (16.84/2.0)
|  |  EUSVascOmit = TRUE: Naive Bayes (6.0/1.0)
|  SxJaun = FALSE: Artifical Neural Network (22.0/3.0)
SxBack = TRUE
|  CTNodeOmit = FALSE
|  |  LabBili <= 0.9: Naive Bayes (3.43)
|  |  LabBili > 0.9: Artifical Neural Network (4.57/0.57)
|  CTNodeOmit = TRUE: Naive Bayes (4.0)

Number of Leaves  :         7

Size of the tree :  13
```

Figure 79: J4.8 Constructed as Level-1 Model

4.3.6 Summary

Over the dataset with the six month split target, there is no statistically significant difference ($p<0.05$) between the classification accuracy of logistic regression and ZeroR. There is a statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and the models constructed using the three noteworthy combinations of feature selection and machine learning algorithm found. There is no statistically significant ($p<0.05$) difference between the classification accuracies of these models and ZeroR. While it is great to have models that perform better than logistic regression, the models are not useful they don't perform better than ZeroR.

There is not a statistically significant difference ($p<0.05$) between classification accuracy of the model constructed using the best machine learning algorithm without feature selection and ZeroR. There is no statistically significant ($p<0.05$) difference between the classification accuracy of these models and logistic regression.

The best attribute selection method over this dataset is gain ratio attribute selection as it selects a sets of attributes that results in high classification accuracies over a large portion of models

constructed using machine learning algorithms. Gain ratio attribute selection consistently produces high accuracies in models constructed by several different machine learning algorithms. Support vector machine attribute selection also selects attributes for models with high classification accuracies but particularly when around 60 attributes are selected.

The two best algorithms over this dataset are artificial neural networks with two hidden units and support vector machines with a kernel exponent of 0.9. Although the support vector machines have a slightly higher classification accuracy, artificial neural networks have a greater area under their ROC curve and a better trade off between true and false positives.

Bagging, boosting, and stacking were not useful over this domain as they neither increase classification accuracy nor decreased the standard deviation. Our model selector was able to select between artificial neural networks and naïve bayes for each instance using a level-1 model constructed with either J4.8 or LWL to increase the classification accuracy over naïve Bayes alone. Unfortunately, the classification accuracy was still less than artificial neural networks alone.

4.4 Results for Pre-Operative Dataset with Six and Twelve Month Split

The dataset discussed in this section has all 112 attributes. The sixty patients in this dataset are split evenly into three groups based on the target of survival. These groups are are <6 month, 6-12 month, and >12 month survival.

4.4.1 Machine Learning Algorithms with No Feature Selection

Figure 80 shows the classification accuracies of models constructed using several different machine learning algorithms run over the dataset with a target split into three groups of zero to six, six to twelve and more than twelve month survival. The model constructed using a Bayesian network with two parents, highlighted in the figure, has the highest classification accuracy. This is statistically significantly better ($p<0.05$) than ZeroR but not statistically different ($p<0.05$) from logistic regression.

Machine Learning Algorithm	Classification Accuracy	Compare To ZeroR
ZeroR	33.3	No Statistically Significant Difference
Logistic Regression	45.0	
SMO with Kernel of 0.9	43.7	
SMO with Kernel of 1.0	43.7	
ANN with 1 Hidden Unit	43.3	
ANN with 2 Hidden Units	40.2	
Naïve Bayes	38.0	
J4.8	41.8	
Bayesian Network: 1 Parent	43.2	
Bayesian Network: 2 Parents	48.3	Better Than ZeroR

Figure 80: No Feature Selection: Six and Twelve Month Split

4.4.2 Combinations of Feature Selection and Machine Learning Algorithm

The graphs that follow show how the classification accuracies of the models built to predict a patient's expected survival time vary based on the feature selection algorithm used and the number of features selected. These models are constructed over the varying feature selection algorithms using several different machine learning algorithms. We use these graphs to find the combinations of feature selection and machine learning algorithm that have the highest classification accuracy for this dataset.

Note that the highest classification accuracy obtained by constructing models with no feature selection is 48.3%. We look to use feature selection to improve upon this classification accuracy.

Figure 81 shows how varying the number of attributes selected by gain ratio attribute selection effects the classification accuracy of several machine learning algorithms. Overall there is very little change in the classification accuracies as the number of attributes is increased from 10 to 100. The notable exception to this is artificial neural networks with one hidden unit which has a classification accuracy that peaks when 50 attributes are selected at above 48.3%. The classification accuracy of this model decreases as more or less attributes are selected.

Figure 82 shows how varying the number of features selected by principal components effects the classification accuracy of several machine learning algorithms. There is only a small overall increase in the classification accuracies as the number of features selected is increased. No algorithm is able to construct a model using the features selected by principal components that has a classification accuracy above 48.3%.

Figure 83 shows how varying the number of features selected by ReliefF attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall peak in the classification accuracies when 20 the number of attributes are selected. Once 20 attributes have been selected there is a slight decrease in the classification accuracy. Two algorithms that show this pattern are support vector machines with a linear kernel and Bayesian networks with a two parents. These two both have classification accuracies over 48.3% when 20 attributes are selected.

Figure 84 shows how varying the number of features selected by support vector machine attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 100. The only algorithm with a classification accuracy of above 48.3% was a model constructed with a

Bayesian network with two parents using the top 100 attributes selected.

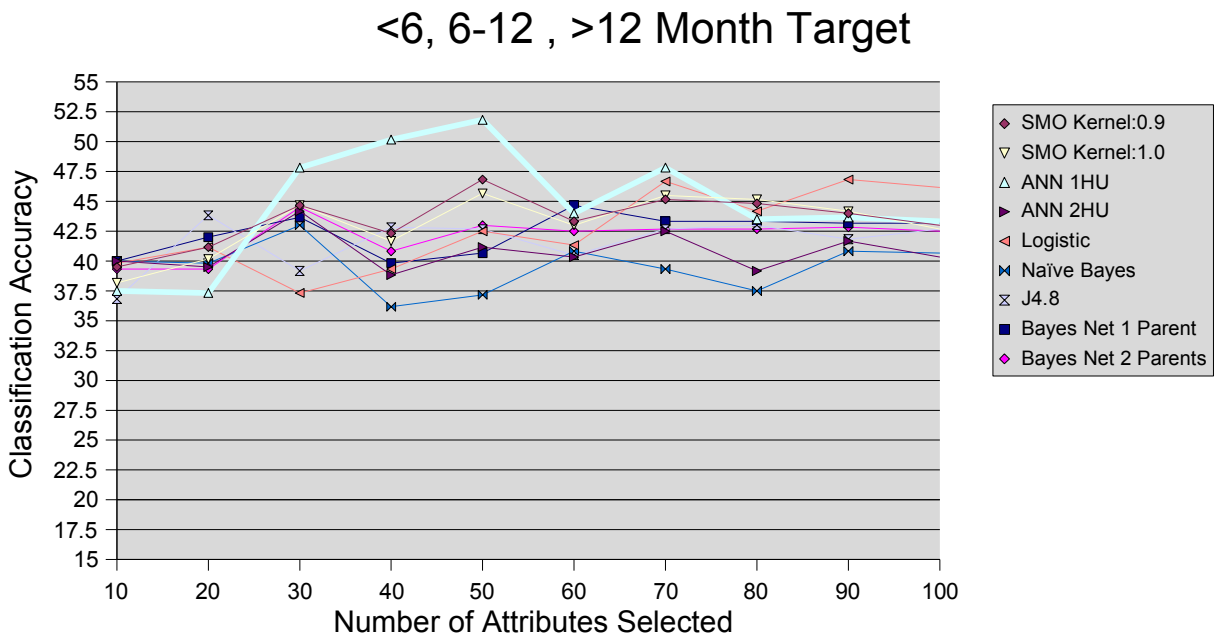

Figure 81: Gain Ratio Attribute Selection: Six and Twelve Month Split

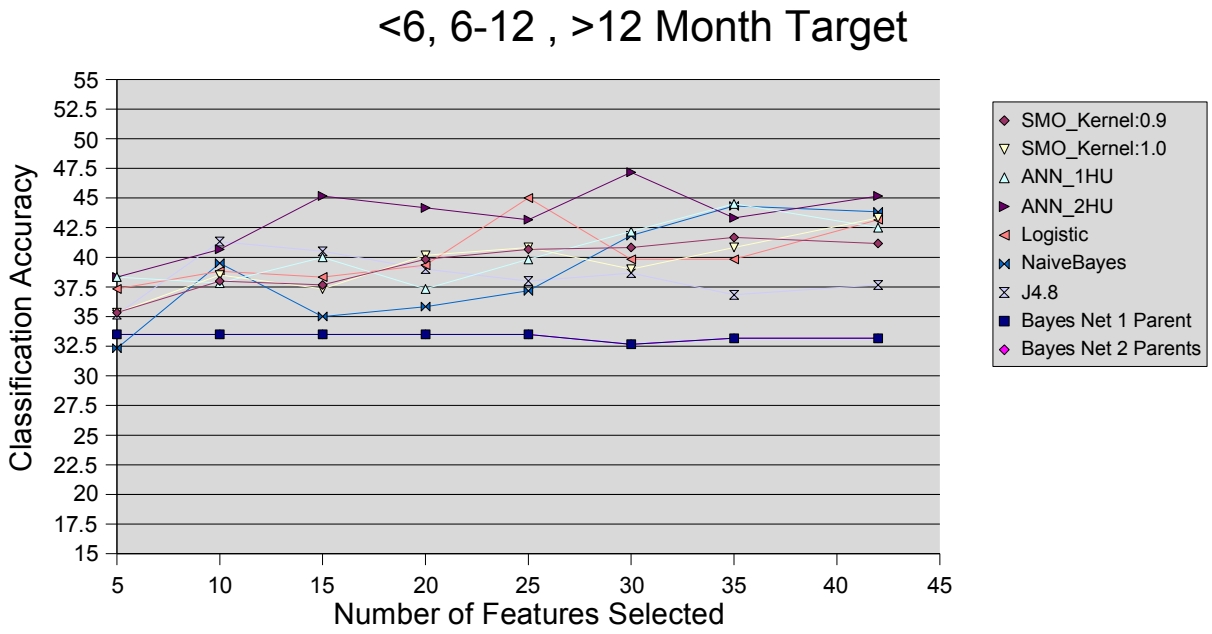

Figure 82: Principal Components: Six and Twelve Month Split

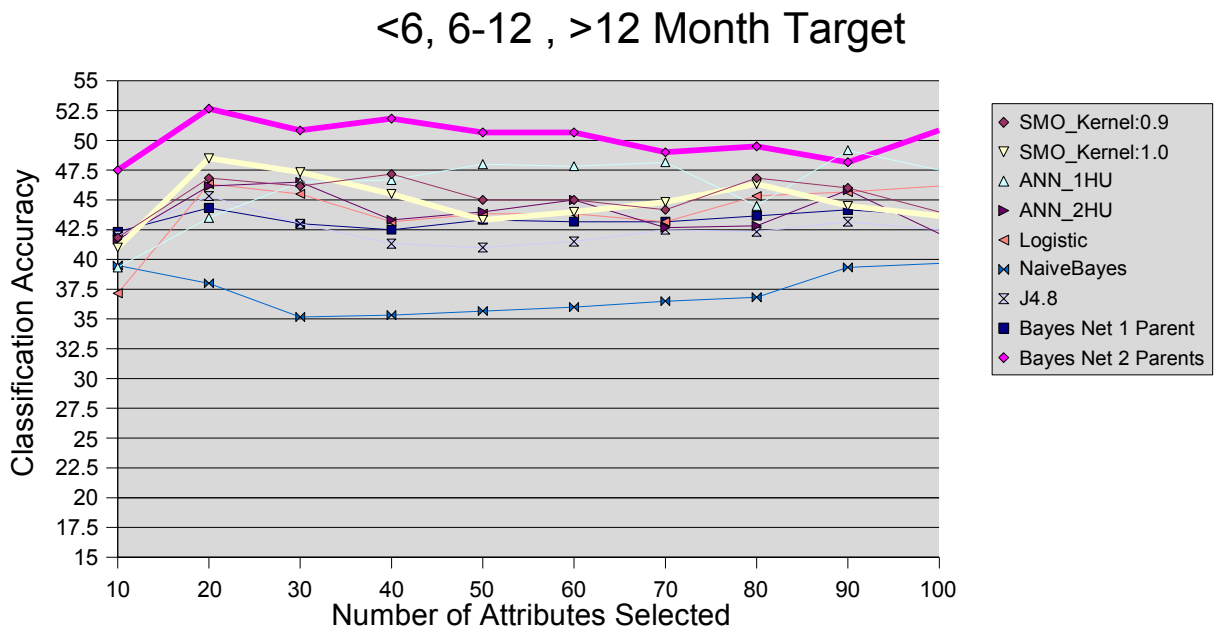

Figure 83: ReliefF Attribute Selection: Six and Twelve Month Split

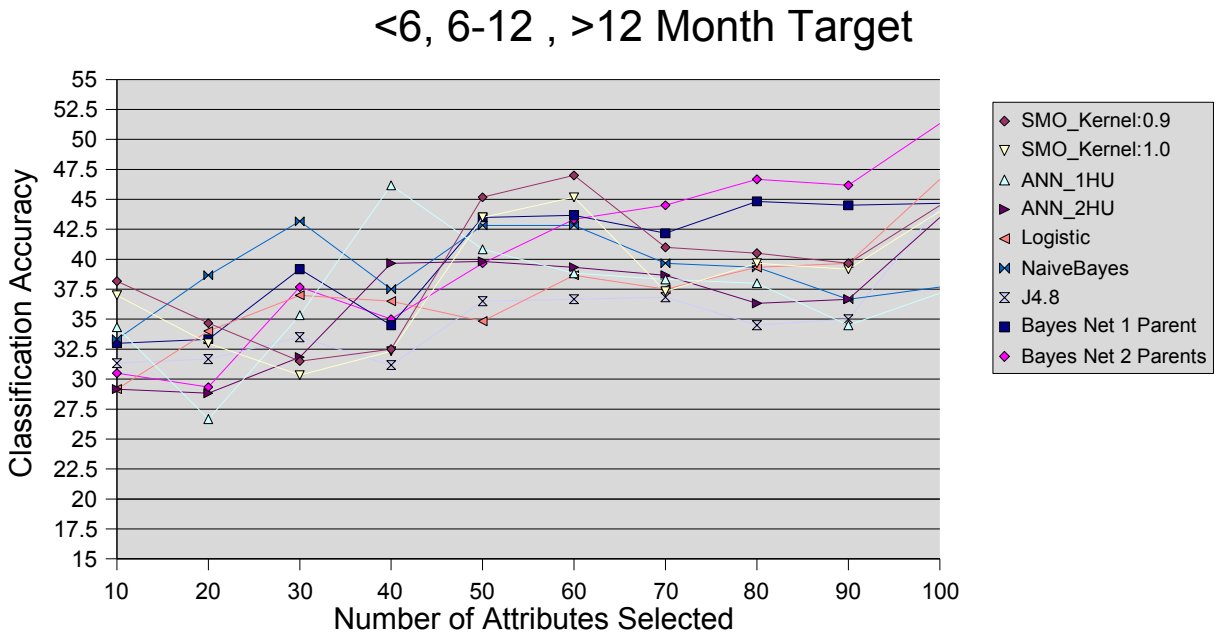

Figure 84: Support Vector Machine Attribute Selection: Six and Twelve Month Split

4.4.2.1 Baseline Models

Figure 80 shows that over this dataset the classification accuracy of logistic regression with no

feature selection is 45.0% and that of ZeroR is 33.3%. There is no statistically significant difference between logistic regression and ZeroR.

4.4.2.2 1st Noteworthy Combination: Bayesian Network

The highest classification accuracy obtained over this dataset is 52.7% resulting from a model constructed using a Bayesian network with K2 using a maximum of two parents. The top 20 attributes are selected to build this model using ReliefF attribute selection. Figure 83 shows this combination of feature selection and machine learning algorithm. This figure shows that once 20 attributes have been selected by ReliefF to build a model using this algorithm, there is a slow decrease in classification accuracies of the resulting models by a the number of attributes selected is increased. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.4.2.3 2nd Noteworthy Combination: Artificial Neural Network with One Hidden Unit

The second highest classification accuracy obtained over this dataset is 51.8% resulting from a model constructed using an artificial neural network with one hidden unit trained over 2,000 epochs. The top 50 attributes are selected to build this model using gain ratio attribute selection. Figure 81 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a peak classification accuracy when 50 attributes are selected using gain ratio where if more or less attributes are selected the classification accuracy decreases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification

accuracy of ZeroR.

4.4.2.4 Model Similar to 1st Noteworthy Combination

The third highest classification accuracy obtained over this dataset is 51.3% resulting from a model constructed using a Bayesian network with K2 using a maximum of two parents. The top 100 attributes are selected to build this model using support vector machine attribute selection. Figure 84 shows this combination of feature selection and machine learning algorithm. As the first highest classifier was constructed using the same Bayesian network, we will not be using this combination of feature selection and machine learning algorithm for further analysis of this dataset.

4.4.2.5 3rd Noteworthy Combination: Support Vector Machines

The third highest classification accuracy obtained over this dataset is 48.5% resulting from a model constructed using support vector machines with a linear kernel function. The top 20 attributes are selected to build this model using ReliefF attribution selection. Figure 83 shows this combination of feature selection and machine learning algorithm. This figure shows that once 20 attributes have been selected by ReliefF to build a model using this algorithm, there is a slow decrease in classification accuracies of the resulting models by a the number of attributes selected is increased. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.4.2.6 Summary of Noteworthy Combinations

- 52.7%: Bayesian Network, Two Parents, using ReliefF to select 20 attributes
- 51.8%: Artificial Neural Network, One Hidden Unit, using Gain Ratio to select 50 attributes
- 48.5%: Support Vector Machines, linear kernel, using ReliefF to select 20 attributes

Note no statistically significant difference ($p<0.05$) between these three models.

4.4.3 Meta-Learning

The following three Combinations of Feature Selection and Machine Learning Algorithm were found in the previous section to produce the highest classification accuracies over the dataset with a nine month split:

- 52.7%: Bayesian Network, Two Parents, using ReliefF to select 20 attributes
- 51.8%: Artificial Neural Network, One Hidden Unit, using Gain Ratio to select 50 attributes
- 48.5%: Support Vector Machines, linear kernel, using ReliefF to select 20 attributes

When we investigated meta-learning over the dataset with all attributes we found that bagging, boosting, and stacking are not useful. In several cases our model selector was able to very slightly improve classification accuracy so we continue to presents results on models it is used to construct.

4.4.3.1 Our Model Selector

In this section we investigate the use of our model selector to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Our model selector is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

The results of the runs of our model selector over this dataset are presented in Figure 85. The far left column shows the level-1 classifier used to determine the model that makes the final target class prediction. The classifiers with feature selection algorithms listed are run using the attribute selected classifier.

Level-1 Classifier	Level – 0 Models			
	Bayes Net, ANN, SVM	ANN, SVM	Bayes Net, SVM	Bayes Net, ANN
ANN 1 Hidden Unit	49.3	50.7	49.3	50.8
ANN 2 Hidden Units	47.8	50.3	49.3	50.2
ANN 1 HU & ReliefF_90	50.0	49.3	49.3	49.3
J48 & SVM_70	47.7	51.8	50.0	51.2
Logistic & PrincComp_15	49.5	50.8	50.2	50.8
NaiveBayes & GainRatio_30	48.3	50.2	48.5	50.2
SMO & ReliefF_40	48.5	50.0	48.3	52.0
J48	48.7	50.3	51.3	50.8
Logistic	47.5	51.3	50.2	51.3
LWL	49.3	51.2	50.2	49.5
NaiveBayes	50.5	53.3	50.5	51.3
SMO	48.8	51.3	50.8	52.8

Figure 85: Our Model Selector Results: Six and Twelve Month Split

This meta-learning algorithm was able to slightly improve the classification accuracy by combining these algorithms in several cases. In particular, when the models constructed by artifical neural networks and support vector machines are used as the level-0 model and the level-1 model used is naïve Bayes, the resulting accuracy is 1.5 points greater than the accuracy of either of the level-0 models. Unfortunately, there is still no statistically significant difference between logistic regression and this model. This model has a classification accuracy that is a statistically significant improvement over the classification accuracy of ZeroR.

4.4.4 Summary

Over the dataset with the six and twelve month split target, there is no statistically significant difference ($p<0.05$) between the classification accuracy of logistic regression and ZeroR. There is a statistically significant difference ($p<0.05$) between the classification accuracy of ZeroR and the models constructed using the three noteworthy combinations of feature selection and machine learning algorithm found. There is a statistically significant difference ($p<0.05$) between classification accuracy of ZeroR and the model constructed using the best machine learning algorithm without feature selection. There is no statistically significant difference ($p<0.05$) between the classification accuracy

of logistic regression and any of these model.

The best attribute selection methods over this dataset are ReliefF attribute selection and gain ratio attribute selection. Over this dataset, both of these feature selection algorithms are able to select sets of 50 or less attributes to construct models with high classification accuracies over a good portion of the machine learning algorithms. Support vector machine attribute evaluation also does a decent job selecting attributes over this dataset but only when it is asked to select the top 100.

The two best algorithms over this dataset are artificial neural networks with one hidden unit and Bayesian networks with a maximum of two parents. Bayesian networks construct a model with a classification accuracy that is statistically significantly better ($p<0.5$) than ZeroR on the dataset with no feature selection. A slightly more accurate model was constructed by selecting the 20 most relevant attributes using ReliefF. Artificial neural networks are among the models with the highest classification accuracy over the attributes selected by support vector machine attribute selection, gain ratio attribute selection and by ReliefF attribute selection.

Our model selector is able is able to slightly increase the classification accuracy by selecting between models constructed using a Bayesian network and support vector machines with a linear kernel by using a naïve Bayes model as the level-1 meta model to predict which model will make the best prediction for a given instance. The classification accuracy of this model constructed using our model selector is statistically significantly ($p<0.05$) better than ZeroR. The classification accuracy is not statistically significant different ($p<0.05$) from logistic regression.

4.5 Results for Pre-Operative Dataset with Nine Month Split

The dataset discussed in this section has all 112 attributes. The sixty patients in this dataset are split evenly into two groups based on the target of survival. These groups are are <9 month and >9 month survival.

4.5.1 Machine Learning Algorithms with No Feature Selection

Figure 86 shows the classification accuracies of models constructed using several different machine learning algorithms run over the dataset with a target split into two groups of zero to nine and more than nine month survival. The model constructed using a Bayesian network with two parents, highlighted in the figure, has the highest classification accuracy. This is statistically significantly better ($p<0.05$) than ZeroR but not statistically different ($p<0.05$) from logistic regression.

Machine Learning Algorithm	Classification Accuracy	Compare To ZeroR
ZeroR	50.0	No Statistically Significant Difference
Logistic Regression	58.8	
SMO with Kernel of 0.9	62.2	
SMO with Kernel of 1.0	62.5	
ANN with 1 Hidden Unit	58.5	
ANN with 2 Hidden Units	58.3	
Naïve Bayes	49.3	
J4.8	49.5	
Bayesian Network: 1 Parent	55.0	
Bayesian Network: 2 Parents	64.7	Better Than ZeroR

Figure 86: No Feature Selection: Nine Month Split

4.5.2 Combinations of Feature Selection and Machine Learning Algorithm

The graphs that follow show how the classification accuracies of the models built to predict a patient's expected survival time vary based on the feature selection algorithm used and the number of features selected. These models are constructed over the varying feature selection algorithms using several different machine learning algorithms. We use these graphs to find the combinations of feature selection and machine learning algorithm that have the highest classification accuracy for this dataset.

Note that the highest classification accuracy obtained by constructing models with no feature selection is 64.7%. We look to use feature selection to improve upon this classification accuracy.

Figure 87 shows how varying the number of attributes selected by gain ratio attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall low point when 20 attributes are selected from which the classification accuracies slowly increase as the number of attributes is increased from 20 to 100. There are three algorithms that construct models with classification accuracies above 64.7%. They are logistic regression, support vector machines with a linear kernel, and support vector machines with a kernel exponent of 0.9. The change in classification accuracies of the two support vector machine algorithms as the number of attributes selected is closely matched with the model with the linear kernel consistently performing better. These two algorithms both reach a high classification accuracy when between 80 and 90 attributes are selected. Logistic regression reaches a peak when 70 attributes are selected from which its classification accuracy decreases if more or less attributes are selected.

Figure 88 shows how varying the number of features selected by principal components effects the classification accuracy of several machine learning algorithms. There is overall the highest classification accuracies are when between 10 and 15 features are selected and when 40 features are selected. There are no models with a classification accuracy greater than 64.7%. Artificial neural networks with two hidden units consistently performs best with whatever number of features are selected.

Figure 89 shows how varying the number of features selected by ReliefF attribute selection effects the classification accuracy of several machine learning algorithms. There is a small overall increase in the classification accuracies as the number of attributes is increased from 10 to 50 and than from 70 to 100. The only model shown in this figure with a classification accuracy above 64.7% is

constructed with a Bayesian network using two parents. The highest classification accuracy for this model is when 100 attributes are selected.

Figure 90 shows how varying the number of features selected by support vector machine attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall increase in the classification accuracies as the number of attributes is increased from 10 to 100. The highest classification accuracies are when 100 attributes are selected. There are no models with a classification accuracy greater than 64.7%.

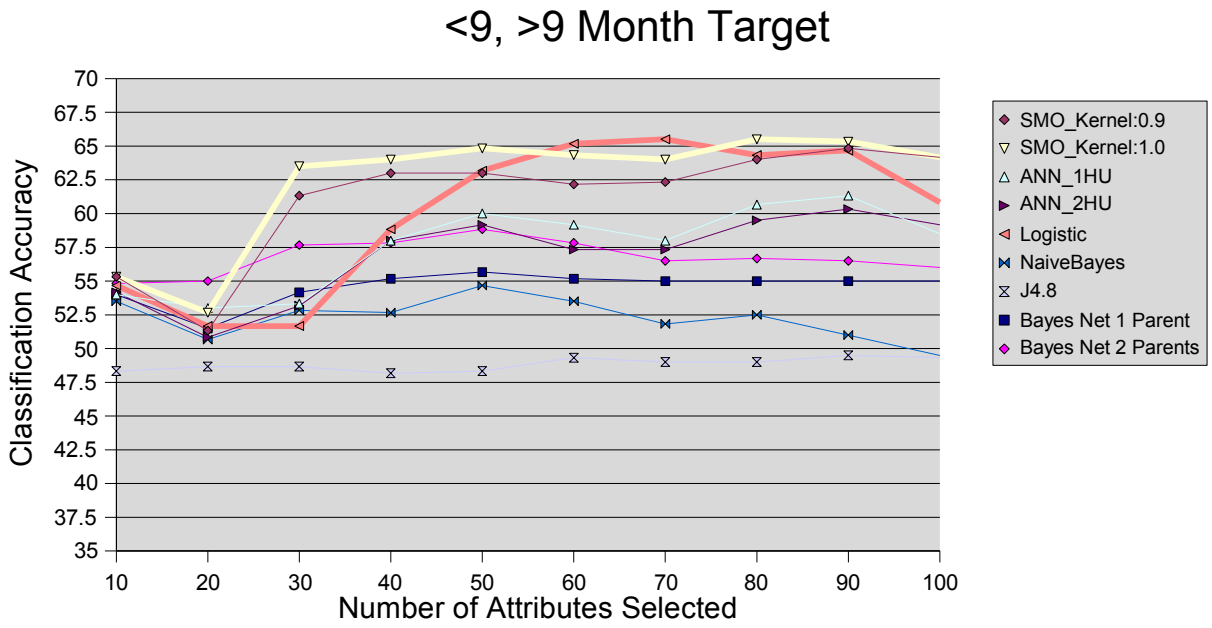

Figure 87: Gain Ratio Attribute Selection: Nine Month Split

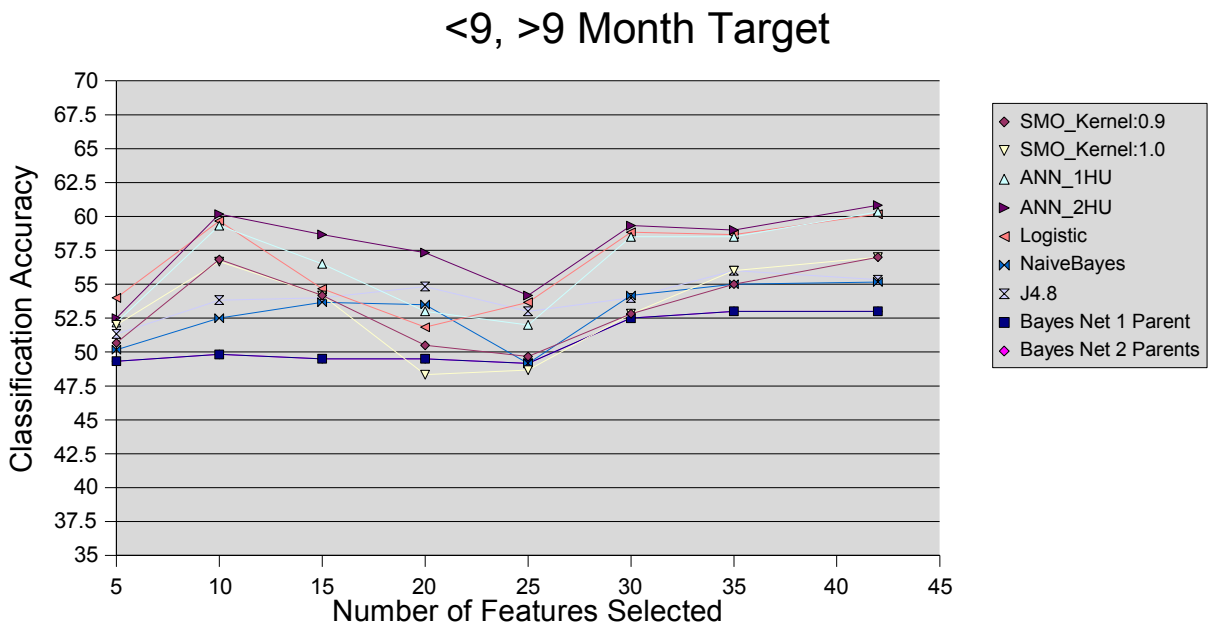

Figure 88: Principal Components: Nine Month Split

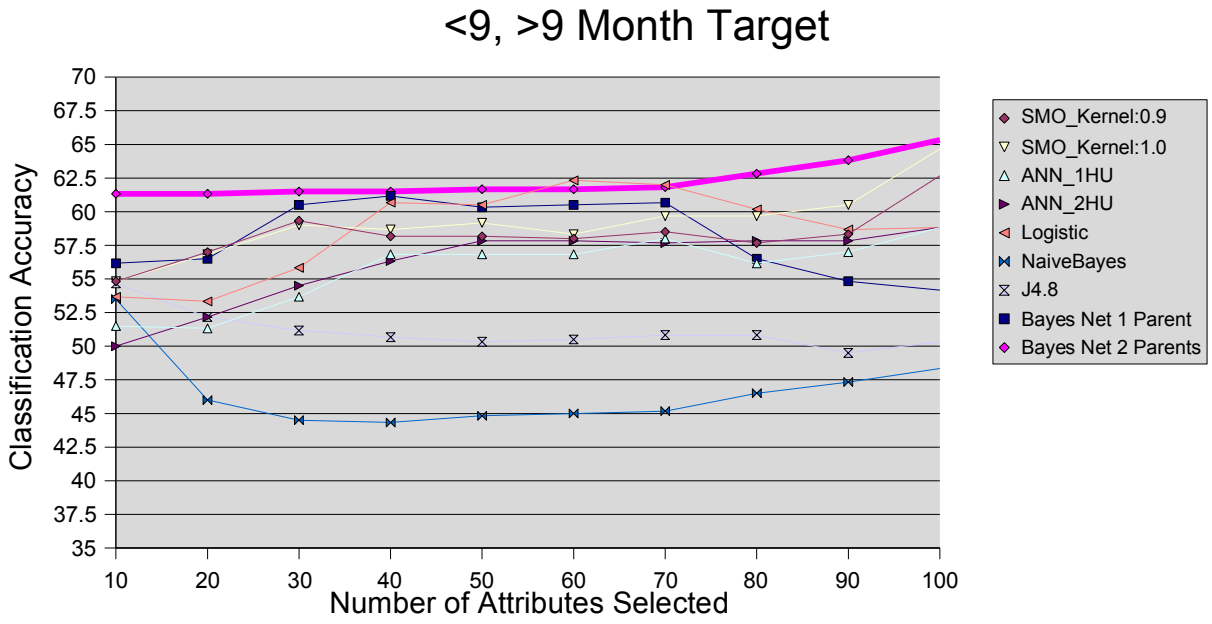

Figure 89: ReliefF Attribute Selection: Nine Month Split

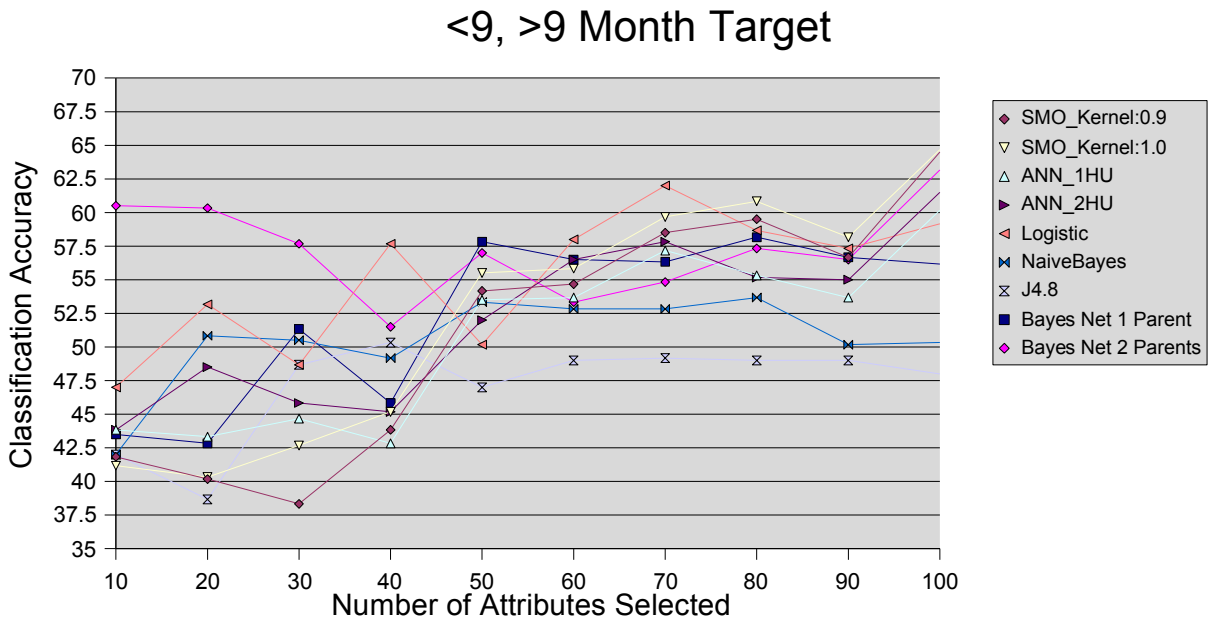

Figure 90: Support Vector Machine Attribute Selection: Nine Month Split

4.5.2.1 Baseline Models

Figure 86 shows this dataset the classification accuracy of logistic regression with no feature selection is 58.8% and that of ZeroR is 50.00%. There is no statistically significant difference between logistic regression and ZeroR.

4.5.2.2 1st Noteworthy Combination: Logistic Regression

The top two combinations of feature selection and machine learning model have the same classification accuracy of 65.5%. They are ranked by lowest standard deviation.

Of the two models with the highest classification accuracy, the first model was constructed using logistic regression. The top 70 attributes are selected to build this model using gain ratio attribute selection. Figure 87 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a peak classification accuracy when 70 attributes are selected using gain ratio attribute selection and when more or less attributes are selected the

classification accuracy decreases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.5.2.3 2nd Noteworthy Combination: Support Vector Machines

Of the two models with the highest classification accuracy of 65.5%, the first model was constructed using support vector machines with a linear kernel function. The top 80 attributes are selected to build this model using gain ratio attribution selection. Figure 87 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a peak classification accuracy when 80 attributes are selected using gain ratio where if more or less attributes are selected the classification accuracy decreases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.5.2.4 3rd Noteworthy Combination: Bayesian Network

The third highest classification accuracy obtained over this dataset is 65.3% resulting from a model constructed using a Bayesian network with K2 using a maximum of two parents. The top 100 attributes are selected to build this model using ReliefF attribute selection. Figure 89 shows this combination of feature selection and machine learning algorithm. This figure shows a gradual increase in classification accuracy of models built using this Bayesian network algorithm as a greater numbers of attributes are selected using support vector machines for feature selection. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. This combination has a classification

accuracy that is a statistically significant improvement ($p<0.05$) over the classification accuracy of ZeroR.

4.5.2.5 Summary of Noteworthy Combinations

- 65.5%: Logistic Regression using Gain Ratio to select 70 attributes
- 65.5%: Support Vector Machines, linear kernel, using Gain Ratio to select 80 attributes
- 65.3%: Bayesian Network, Two Parents, using ReliefF to select 100 attributes

Note no statistically significant difference ($p<0.05$) between these three models.

4.5.3 ROC Curves

ROC curves for the combinations of feature selection and machine learning algorithm with the highest classification accuracy appear at the end of this section. The combinations will be presented and discussed in the following order:

- 58.8%: Logistic Regression with no feature selection
- 65.5%: Logistic Regression using Gain Ratio to select 70 attributes
- 65.5%: Support Vector Machines, linear kernel, using Gain Ratio to select 80 attributes
- 65.3%: Bayesian Network, Two Parents, using ReliefF to select 100 attributes

4.5.3.1 Baseline Model: Logistic Regression with no Feature Selection

Figure 91 shows the ROC curve for logistic regression. The area under this curve is 0.63. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 75% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 63% of the patients who will not survive for six months will survive for greater than six months.

4.5.3.2 1st Noteworthy Combination: Logistic Regression with Feature Selection

Figure 92 shows the ROC curve for the combination of feature selection and machine learning algorithm with the highest classification accuracy over this dataset. The area under this curve is 0.69.

This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 70% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 53% of the patients who will not survive for six months will survive for greater than six months. This curve is better than logistic regression with no feature selection due to both a larger area under the curve and having a better trade off between the true and false positive rates.

4.5.3.3 2nd Noteworthy Combination: Support Vector Machines

Figure 93 shows the ROC curve for the combination of feature selection and machine learning algorithm with the second highest classification accuracy over this dataset. The area under this curve is 0.73. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 64% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 43% of the patients who will not survive for six months will survive for greater than six months. This curve is better than logistic regression with no feature selection due to both a larger area under the curve and having a better trade off between the true and false positive rates. This curve is also better than the curve for the model with the highest classification accuracy, logistic regression with feature selection, due to both a better trade off between the true and false positive rates and a larger area under the curve.

4.5.3.4 3rd Noteworthy Combination: Bayesian Network

Figure 94 shows the ROC curve for the combination of feature selection and machine learning algorithm with the third highest classification accuracy over this dataset. The area under this curve is 0.69. This curve shows that to correctly predict 90% of the patients who will survive for more than six

months, you have to incorrectly predict that 67% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 54% of the patients who will not survive for six months will survive for greater than six months. This curve is better than logistic regression with no feature selection due to both a larger area under the curve and having a better trade off between the true and false positive rates. This curve is comparable to the curve constructed using the combination with the highest classification accuracy because of the same area under the curve and comparable trade offs.

4.5.3.5 Summary

The three combinations of feature selection and machine learning algorithm with the highest classification accuracies are ranked by their ROC curves as follows:

1. Support Vector Machines, linear kernel, using Gain Ratio to select 80 attributes
2. Logistic Regression using Gain Ratio to select 70 attributes
3. Bayesian Network, Two Parents, using ReliefF to select 100 attributes

4.5.3.6 Curves

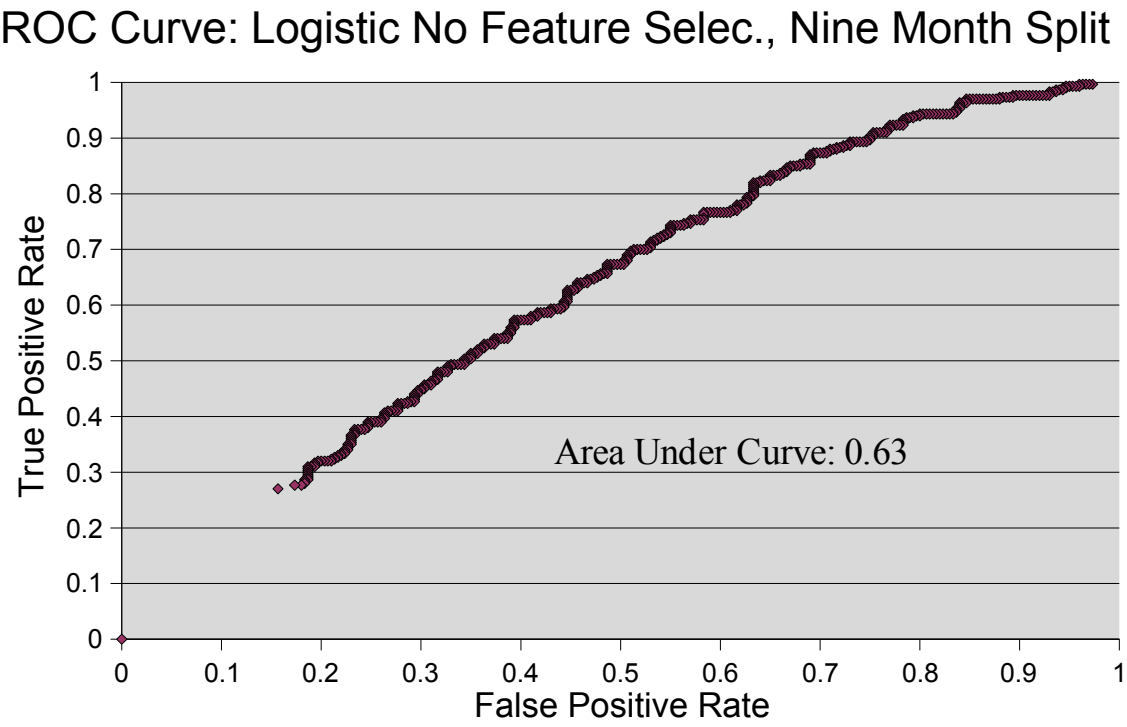

Figure 91: ROC Curve - Logistic Regression, No Feature Selection: Nine Month Split

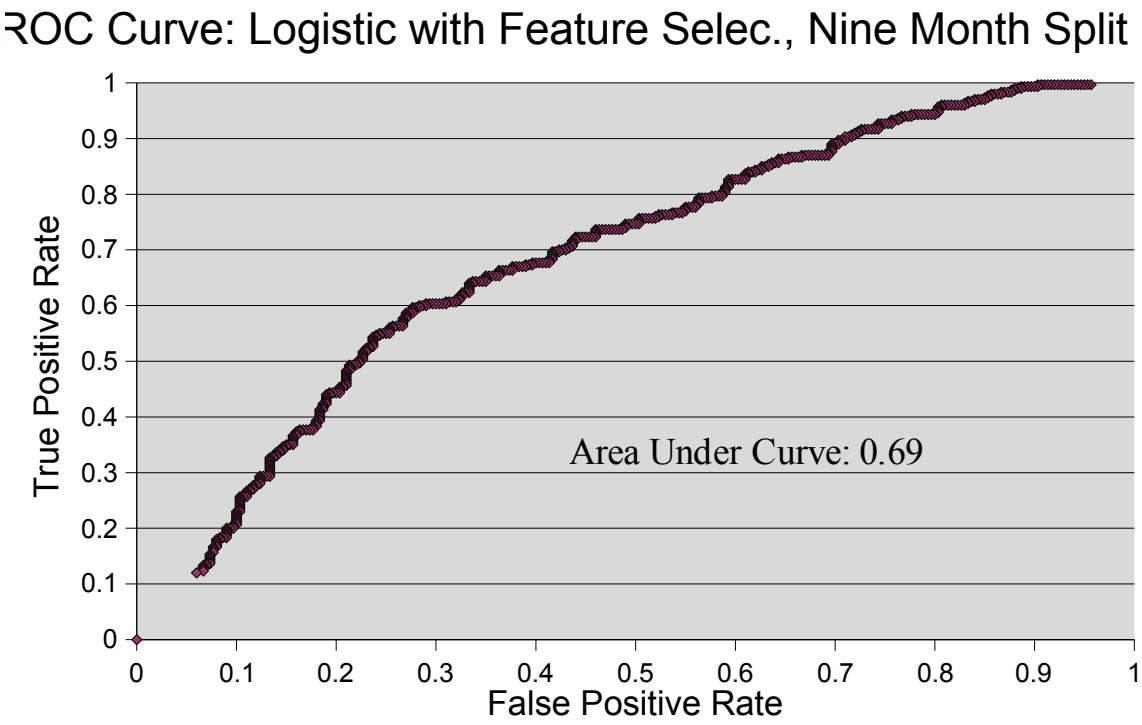

Figure 92: ROC Curve - Logistic Regression, with Feature Selection: Nine Month Split

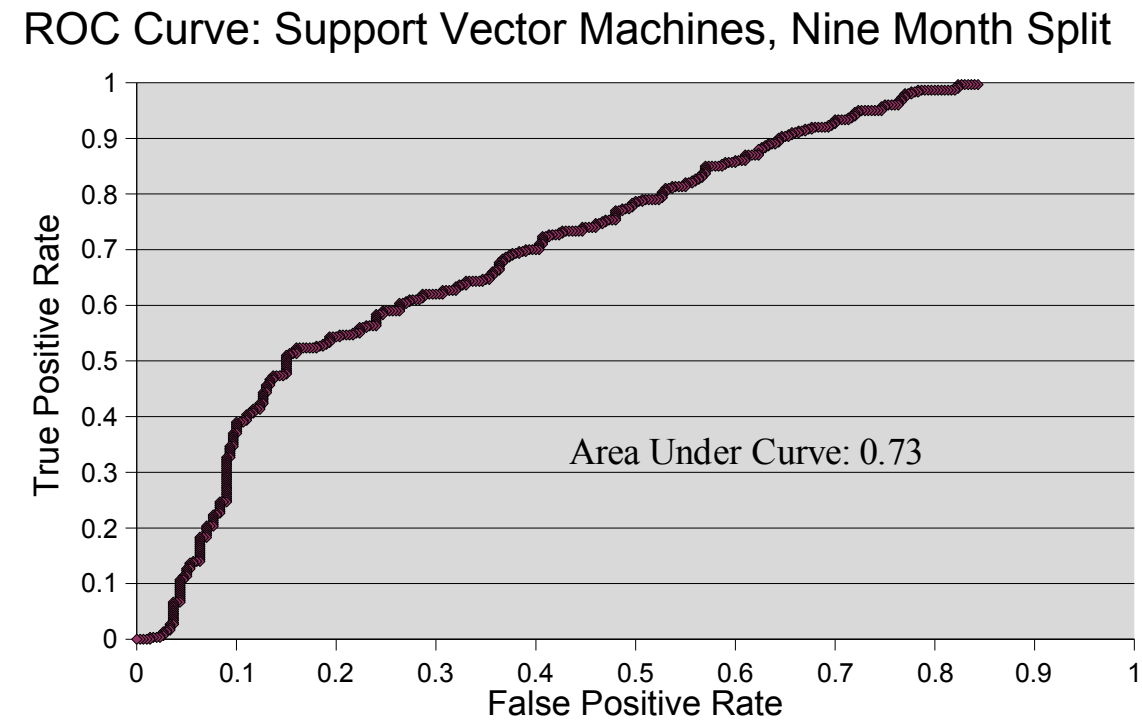

Figure 93: ROC Curve - Support Vector Machines: Nine Month Split

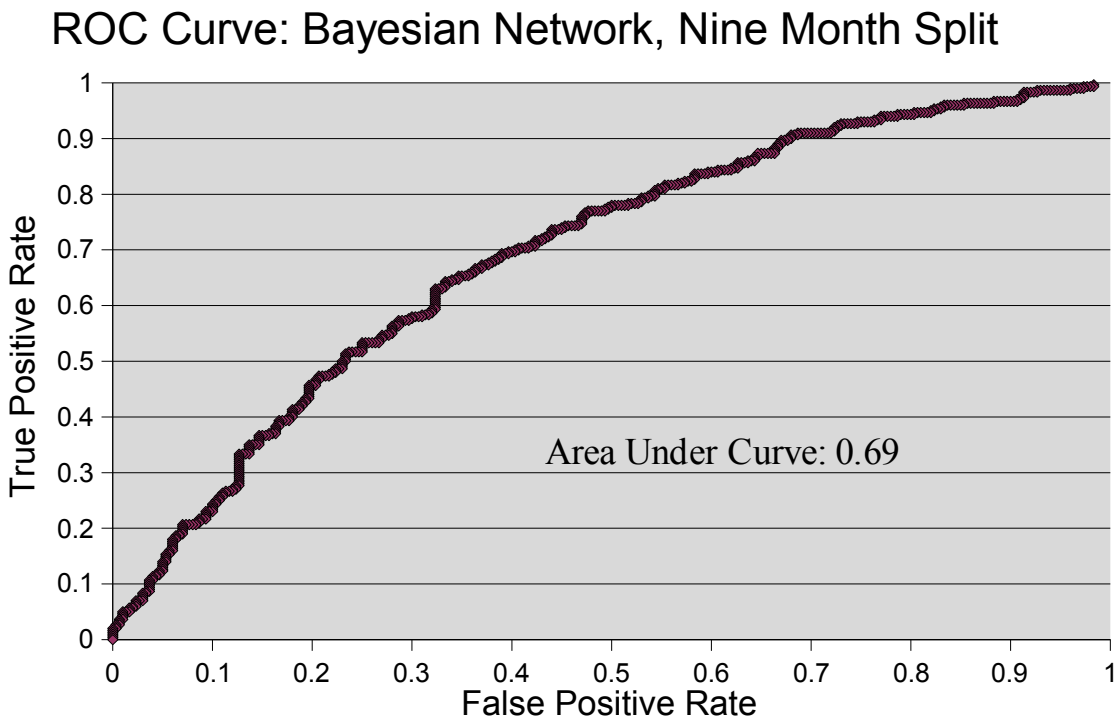

Figure 94: ROC Curve - Bayesian Network: Nine Month Split

4.5.4 Meta-Learning

The following three Combinations of Feature Selection and Machine Learning Algorithm were found in the previous section to produce the highest classification accuracies over the dataset with a nine month split:

- 65.5%: Logistic Regression using Gain Ratio to select 70 attributes
- 65.5%: Support Vector Machines, linear kernel, using Gain Ratio to select 80 attributes
- 65.3%: Bayesian Network, Two Parents, using ReliefF to select 100 attributes

When we investigated meta-learning over the dataset with all attributes we found that bagging, boosting, and stacking are not useful. Our model selector in several cases was able to very slightly improve classification accuracy so we continue analysis.

4.5.4.1 Our Model Selector

In this section we investigate the use of our model selector to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Our model selector is first run using all three of these algorithms followed by three runs where

one of these algorithms is removed.

The results of the runs of our model selector over this dataset are presented in Figure 95. The far left column shows the level-1 classifier used to determine the model that makes the final target class prediction. The classifiers with feature selection algorithms listed are run using the attribute selected classifier.

This meta-learning algorithm was able to slightly improve the classification accuracy by combining these algorithms in several cases. The best increase in accuracy occurred from a model constructed combining logistic regression and support vector machines using naïve Bayes. This model does almost 2 percentage points better than either of the models used for its construction. Unfortunately, there is still no statistically significant difference between logistic regression and this model. This model has a classification accuracy that is a statistically significant improvement over the classification accuracy of ZeroR.

Level-1 Classifier	Level – 0 Models			
	Logistic, SVM, Bayes Net	SVM, Bayes Net	Logistic, Bayes Net	Logistic, SVM
ANN 1 Hidden Unit	64.6	65.2	65.7	66.5
ANN 2 Hidden Units	64.3	65.2	65.8	66.2
ANN 1 HU & ReliefF_90	65.2	65.0	65.3	66.3
J48 & SVM_70	65.3	65.0	64.8	65.2
Logistic & PrincComp_15	64.0	65.0	62.2	65.7
NaiveBayes & GainRatio_30	63.2	63.2	64.3	65.0
SMO & ReliefF_40	64.5	65.5	65.2	66.3
J48	64.5	65.8	65.2	65.8
Logistic	65.0	66.7	65.3	65.5
LWL	64.3	64.2	63.7	64.5
NaiveBayes	64.8	64.2	62.8	67.3
SMO	65.3	66.2	65.0	65.7

Figure 95: Our Model Selector Results: Nine Month Split

4.5.5 Summary

Over the dataset with the nine month split target, there is no statistically significant difference ($p<0.05$) between the classification accuracy of logistic regression and ZeroR. There is a statistically

significant difference ($p<0.05$) between the classification accuracy of ZeroR and the models constructed using the three noteworthy combinations of feature selection and machine learning algorithm found. There is a statistically significant difference ($p<0.05$) between classification accuracy of ZeroR and the model constructed using the best machine learning algorithm without feature selection. There is no statistically significant ($p<0.05$) difference between any of these models and logistic regression.

The best attribute selection methods over this dataset are ReliefF attribute selection and gain ratio attribute selection. Over this dataset gain ratio did a better job of selecting attributes for logistic regressing and support vector machines while ReliefF did a better job of selecting attributes for Bayesian networks.

The support vector machines with a linear kernel, logistic regression, and Bayesian networks with a maximum of two parents all perform well over this dataset. The classification accuracies of these models are close to identical so they are compared based on their ROC curves. Support vector machines have the largest area under the curve and the best trade off between true and false positive rates. The area under the curve and trade offs between the Bayesian network and linear regression are very similar.

Our model selector is able is able to increase the classification accuracy by about 3 percentage points by selecting between models constructed using logistic regression and support vector machines with a linear kernel by using a naïve Bayes model as the level-1 meta model to predict which model will make the best prediction for a given instance. The classification accuracy of this model constructed using our model selector is statistically significantly ($p<0.05$) better than ZeroR. The classification accuracy is not no statistically significant different ($p<0.05$) from logistic regression.

This is the dataset that our model selector shines over. The majority of its results have a higher

classification accuracy than the classification accuracy of the initial models used in its construction.

4.6 Results for Pre-Operative Dataset with Six Month Split

The dataset discussed in this section has all 112 attributes. The sixty patients in this dataset are split into two groups based on the target of survival. These groups are are <6 month and >6 month survival.

4.6.1 Machine Learning Algorithms with No Feature Selection

Figure 96 shows the classification accuracies of models constructed using several different machine learning algorithms run over the pre-operative dataset with a target split into two groups of zero to six and more than six month survival. ZeroR, highlighted in figure, has the highest classification accuracy. This is not statistically significantly different from ($p<0.05$) logistic regression. The lowest classification accuracy resulted from a model constructed using naïve Bayes. This model is statistically significantly worse ($p<0.05$) than ZeroR and is not statistically different ($p<0.05$) from logistic regression.

Machine Learning Algorithm	Classification Accuracy	Compare To ZeroR
ZeroR	66.7	No Statistically Significant Difference
Logistic Regression	62.2	
SMO with Kernel of 0.9	65.2	
SMO with Kernel of 1.0	64.5	
ANN with 1 Hidden Unit	58.8	
ANN with 2 Hidden Units	59.0	
Naïve Bayes	52.2	Worse Than ZeroR
J4.8	62.3	No Statistically Significant Difference
Bayesian Network: 1 Parent	61.2	
Bayesian Network: 2 Parents	66.5	

Figure 96: No Feature Selection: Six Month Split

4.6.2 Combinations of Feature Selection and Machine Learning Algorithm

The graphs that follow show how the classification accuracies of the models built to predict a patient's expected survival time vary based on the feature selection algorithm used and the number of features selected. These models are constructed over the varying feature selection algorithms using

several different machine learning algorithms. We use these graphs to find the combinations of feature selection and machine learning algorithm that have the highest classification accuracy for this dataset.

Note that the highest classification accuracy obtained by constructing models with no feature selection is 66.7%. We look to use feature selection to improve upon this classification accuracy.

Figure 97 shows how varying the number of attributes selected by gain ratio attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall peak in the classification accuracies when the number of attributes selected is between 30 and 40. When more or less attributes are selected, the classification accuracy decreases. There several algorithms that produce models with classification accuracies above 66.7% including logistic regression, support vector machines with a kernel exponent of 0.9, and artificial neural networks with one hidden unit. These have the highest classification accuracies when 40, 30, and 30 attributes are selected respectively.

Figure 98 shows how varying the number of features selected by principal components effects the classification accuracy of several machine learning algorithms. There is an overall decrease in the classification accuracies as the number of features selected is increased. There are several exceptions to this including naïve Bayes which reaches a peak classification accuracy of above 66.7% when 35 features are selected.

Figure 99 shows how varying the number of features selected by ReliefF attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall peak classification accuracies when the number of attributes selected is 20. The classification accuracies decrease as more or less than 20 attributes are selected. There are four algorithms that construct models using these 20 attributes with classification accuracies above 66.7%. They are support vector machines with a linear kernel, support vector machines with a kernel exponent of 0.9, Bayesian

networks with two parents, and Bayesian networks with one parent.

Figure 100 shows how varying the number of features selected by support vector machine attribute selection effects the classification accuracy of several machine learning algorithms. There is an overall slight increase in the classification accuracies as the number of attributes is increased from 30 to 70. This not a very clear tread, however. The algorithm that consistently constructs the best performing models over this dataset is Bayesian networks with one parent which has the highest classification accuracy when 70 attributes are selected. At this point the classification accuracy is greater than 66.7%.

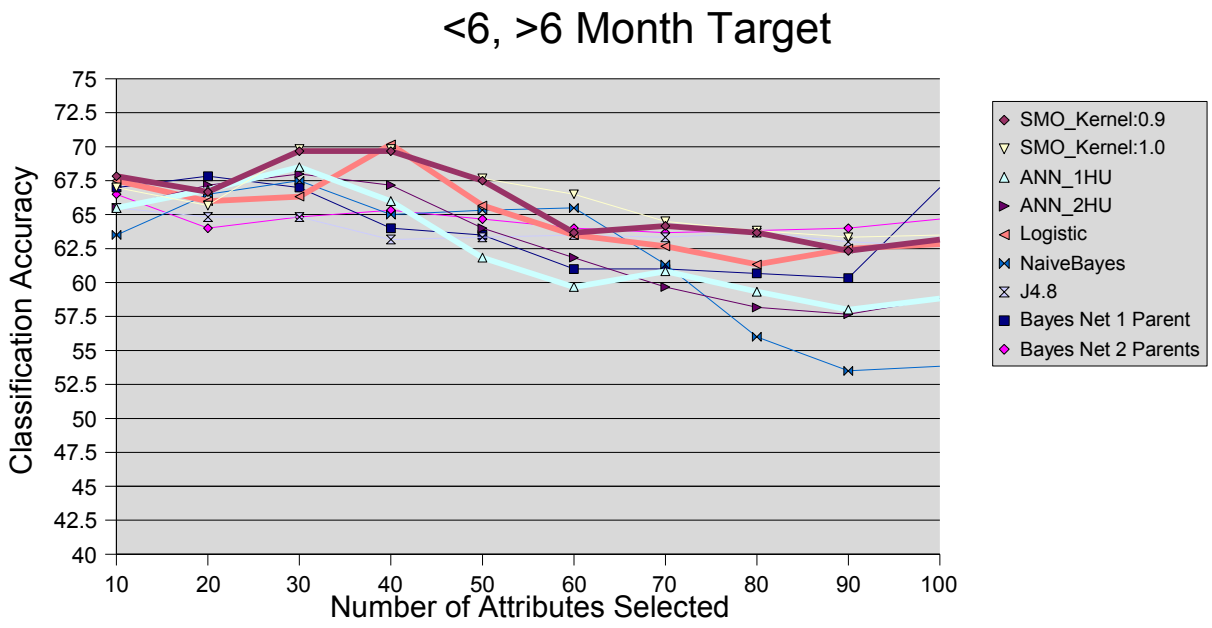

Figure 97: Gain Ratio Attribute Selection: Six Month Split

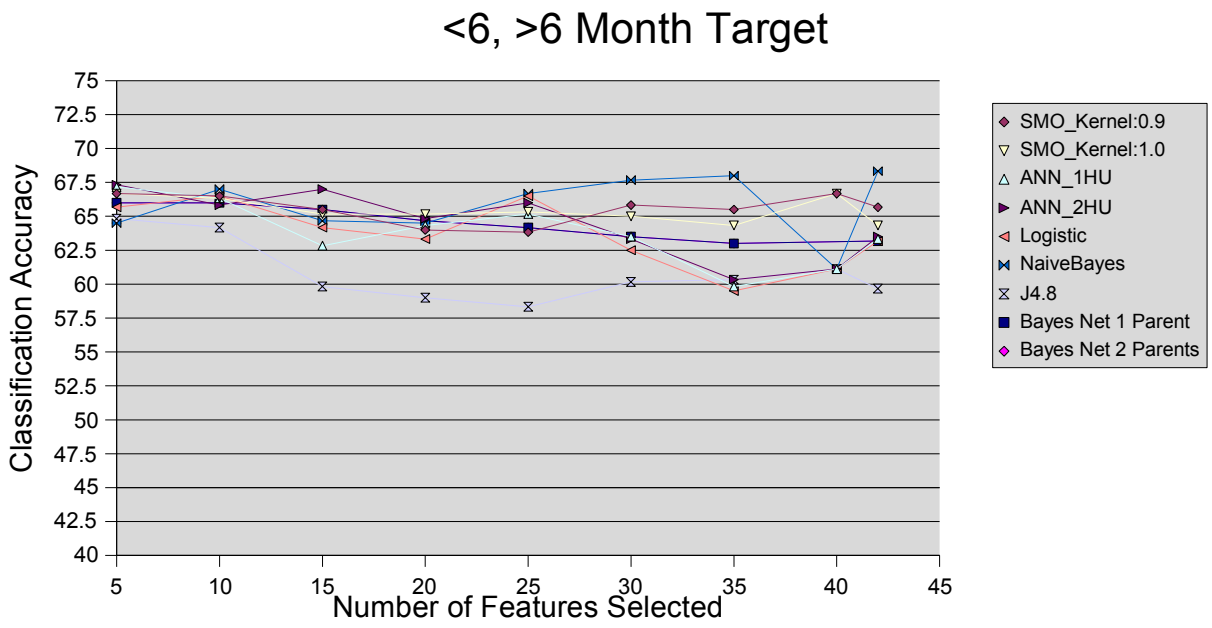

Figure 98: Principal Components: Six Month Split

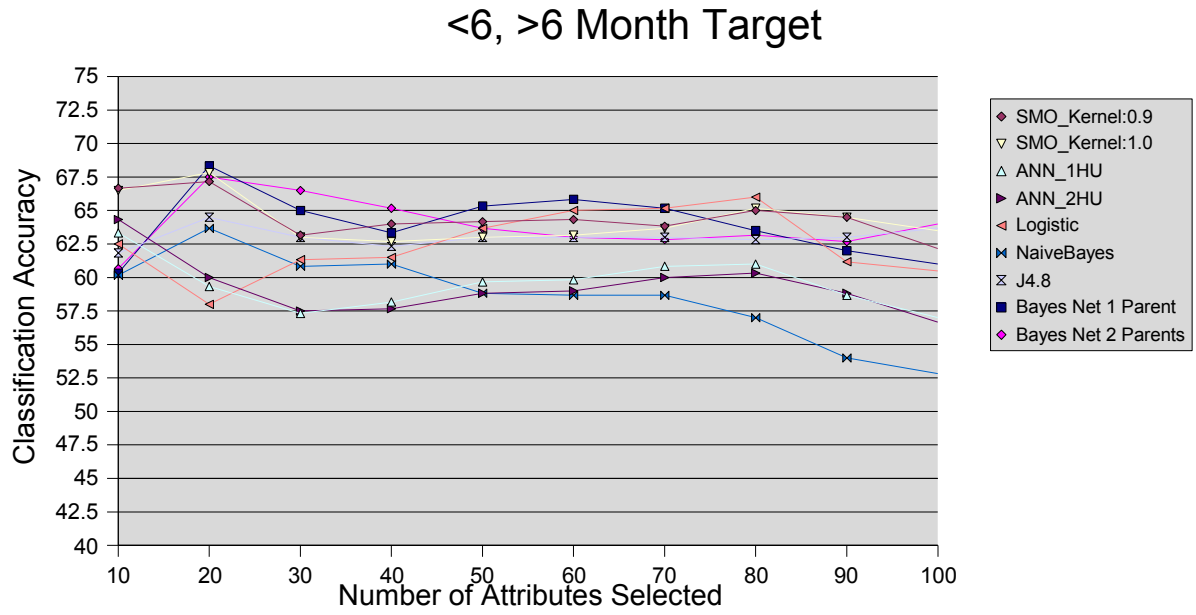

Figure 99: ReliefF Attribute Selection: Six Month Split

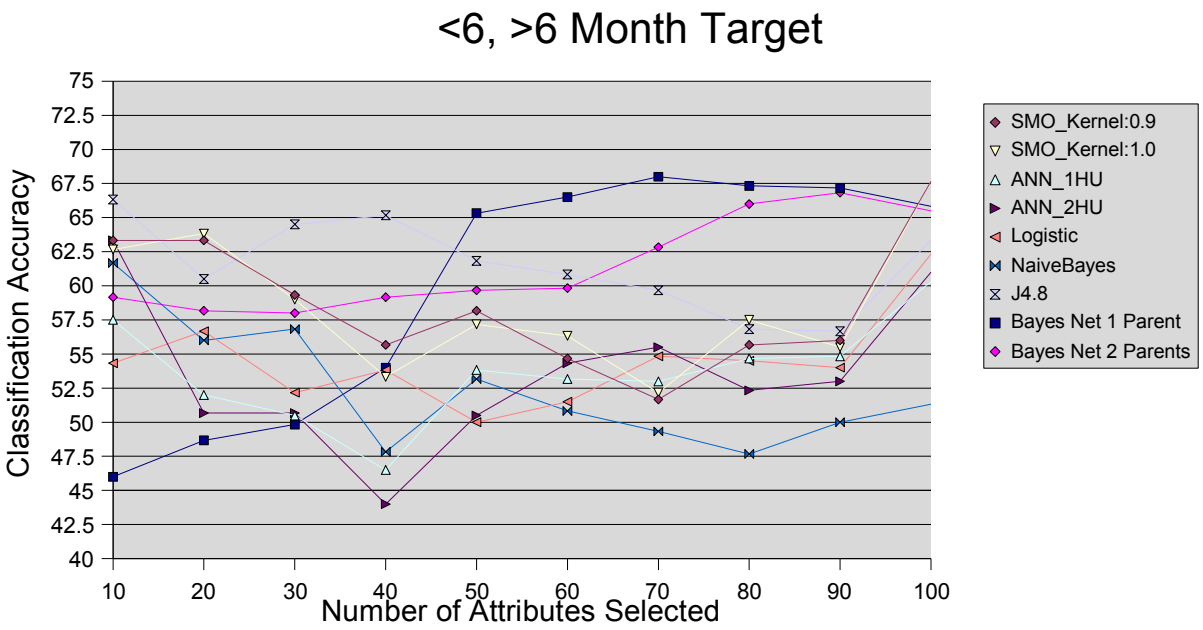

Figure 100: Support Vector Machine Attribute Selection: Six Month Split

4.6.2.1 Baseline Models

Figure 96 shows that over this dataset the classification accuracy of logistic regression with no feature selection is 62.2% and that of ZeroR is 66.7%. There is no statistically significant difference between logistic regression and ZeroR.

4.6.2.2 1st Noteworthy Combination: Logistic Regression

The highest classification accuracy obtained over this dataset is 70.2% resulting from a model constructed using logistic regression. The top 40 attributes are selected to build this model using gain ratio attribute selection. Figure 97 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm reaches a peak classification accuracy when 40 attributes are selected using support vector machine attribute selection where if more or less attributes are selected the classification accuracy decreases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature

selection and machine learning algorithm. There is also no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.6.2.3 2nd Noteworthy Combination: Support Vector Machines

The second highest classification accuracy obtained over this dataset is 69.8% resulting from a model constructed using support vector machines with a kernel exponent of 0.9. The top 30 attributes are selected to build this model using gain ratio attribution selection. Figure 97 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm appears to reach a peak classification accuracy when 30 attributes are selected using gain ratio attribute selection where if more or less attributes are selected the classification accuracy decreases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. There is also no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.6.2.4 Model Similar to 2nd Noteworthy Combination

The third highest classification accuracy obtained over this dataset is 69.7% resulting from a model constructed using support vector machines with a linear kernel function. The top 30 attributes are selected to build this model using gain ratio attribute selection. Figure 97 shows this combination of feature selection and machine learning algorithm. This machine learning algorithm is very similar to the one with the highest classification accuracy and uses the same feature selection. For these reasons, we will not be using this combination of feature selection and machine learning algorithm for further analysis of this dataset.

4.6.2.5 3rd Noteworthy Combination: Artificial Neural Network, One Hidden Unit

The fourth highest classification accuracy obtained over this dataset is 68.5% resulting from a

model constructed using an artificial neural network with one hidden unit. The top 30 attributes are selected to build this model using gain ratio attribute selection. Figure 97 shows this combination of feature selection and machine learning algorithm. The figure shows that this algorithm results in the highest classification accuracy when 30 attributes are selected using gain ratio attribute selection, with a decreasing accuracy as the number of attributes that are selected increases. There is no statistically significant difference ($p<0.05$) between the classification accuracies of logistic regression and this combination of feature selection and machine learning algorithm. There is also no statistically significant difference ($p<0.05$) between ZeroR and this combination.

4.6.2.6 Summary of Noteworthy Combinations

- 70.2%: Logistic Regression using Gain Ratio to select 40 attributes
- 69.8%: Support Vector Machines, 0.9 kernel, using Gain Ratio to select 30 attributes
- 68.5%: Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 30 attributes

Note no statistically significant difference ($p<0.05$) between these three models.

4.6.3 ROC Curves

ROC curves for the combinations of feature selection and machine learning algorithm with the highest classification accuracy appear at the end of this section. The combinations will be presented and discussed in the following order:

- 62.2%: Logistic Regression
- 70.2%: Logistic Regression using Gain Ratio to select 40 attributes
- 69.8%: Support Vector Machines, 0.9 kernel, using Gain Ratio to select 30 attributes
- 68.5%: Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 30 attributes

4.6.3.1 Baseline Model: Logistic Regression with no Feature Selection

Figure 101 shows the ROC curve for logistic regression. The area under this curve is 0.60. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 83% of the patients who will not survive for six months

will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 66% of the patients who will not survive for six months will survive for greater than six months.

4.6.3.2 1st Noteworthy Combination: Logistic Regression with Feature Selection

Figure 102 shows the ROC curve for the combination of feature selection and machine learning algorithm with the highest classification accuracy over this dataset. The area under this curve is 0.72. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 63% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 46% of the patients who will not survive for six months will survive for greater than six months. This curve is better than logistic regression with no feature selection due to both a larger area under the curve and having a better trade off between the true and false positive rates.

4.6.3.3 2nd Noteworthy Combination: Support Vector Machines

Figure 103 shows the ROC curve for the combination of feature selection and machine learning algorithm with the second highest classification accuracy over this dataset. The area under this curve is 0.68. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 69% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 54% of the patients who will not survive for six months will survive for greater than six months. This curve is better than logistic regression with no feature selection due to both a larger area under the curve and having a better trade off between the true and false positive rates. This curve performs similar to the curve constructed using the

combination with the highest classification accuracy because of the same area under the curve and comparable trade offs.

4.6.3.4 3rd Noteworthy Combination: Bayesian Network

Figure 104 shows the ROC curve for the combination of feature selection and machine learning algorithm with the third highest classification accuracy over this dataset. The area under this curve is 0.66. This curve shows that to correctly predict 90% of the patients who will survive for more than six months, you have to incorrectly predict that 76% of the patients who will not survive for six months will survive for greater than six months. To correctly predict 80% of the patients who will survive for more than six months, you have to incorrectly predict that 54% of the patients who will not survive for six months will survive for greater than six months. This curve is better than logistic regression with no feature selection due to both a larger area under the curve and having a better trade off between the true and false positive rates. This curve is not as good as the curve constructed using the combination with the highest classification accuracy because of the a decreased area under the curve and a worse trade off.

4.6.3.5 Summary

The three combinations of feature selection and machine learning algorithm with the highest classification accuracies are ranked by their ROC curves as follows:

1. Logistic Regression using Gain Ratio to select 40 attributes
2. Support Vector Machines, 0.9 kernel, using Gain Ratio to select 30 attributes
3. Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 30 attributes

4.6.3.6 Curves

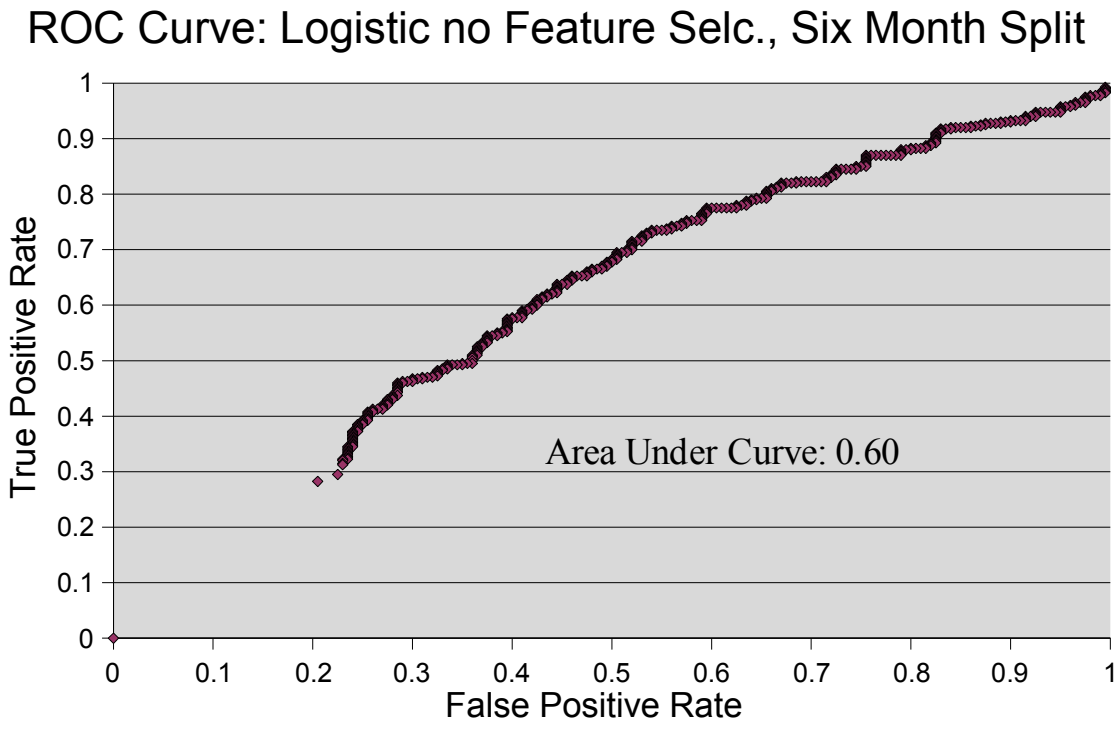

Figure 101: ROC Curve - Logistic Regression, No Feature Selection: Six Month Split

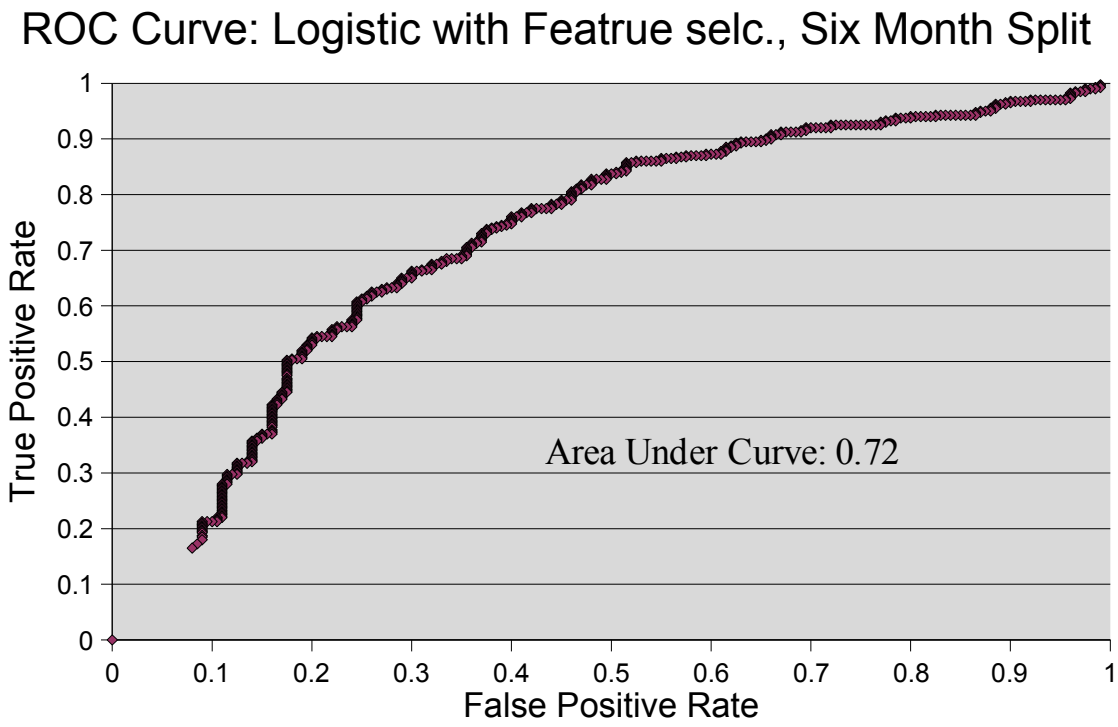

Figure 102: ROC Curve - Logistic Regression, With Feature Selection: Six Month Split

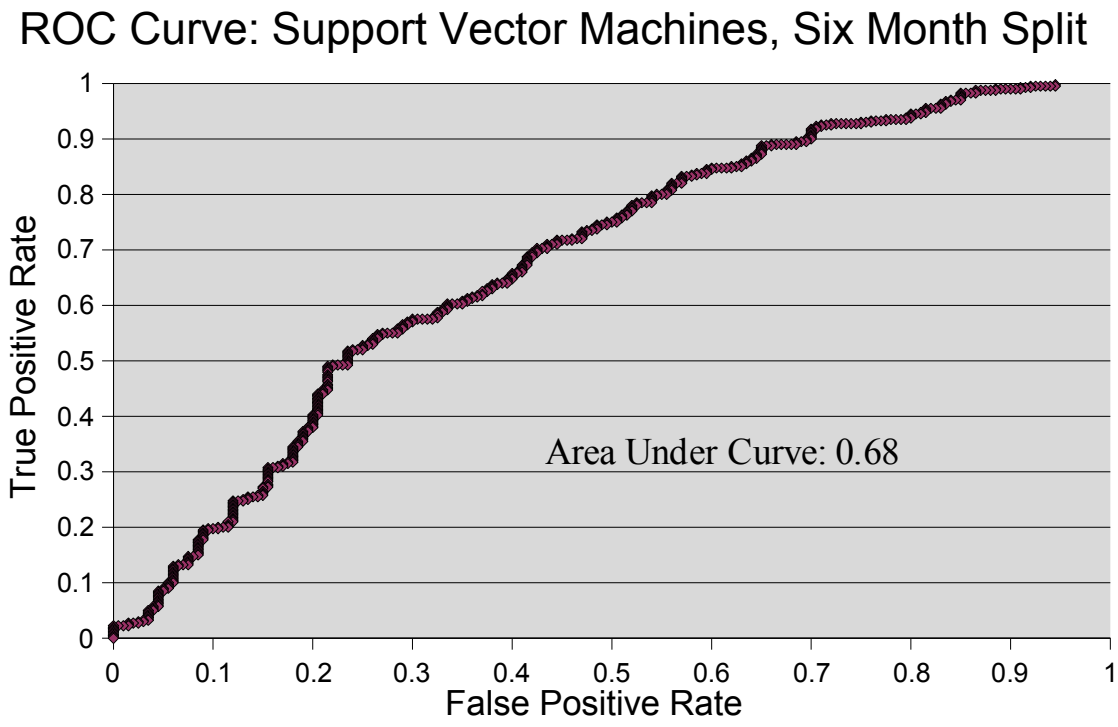

Figure 103: ROC Curve - Support Vector Machines: Six Month Split

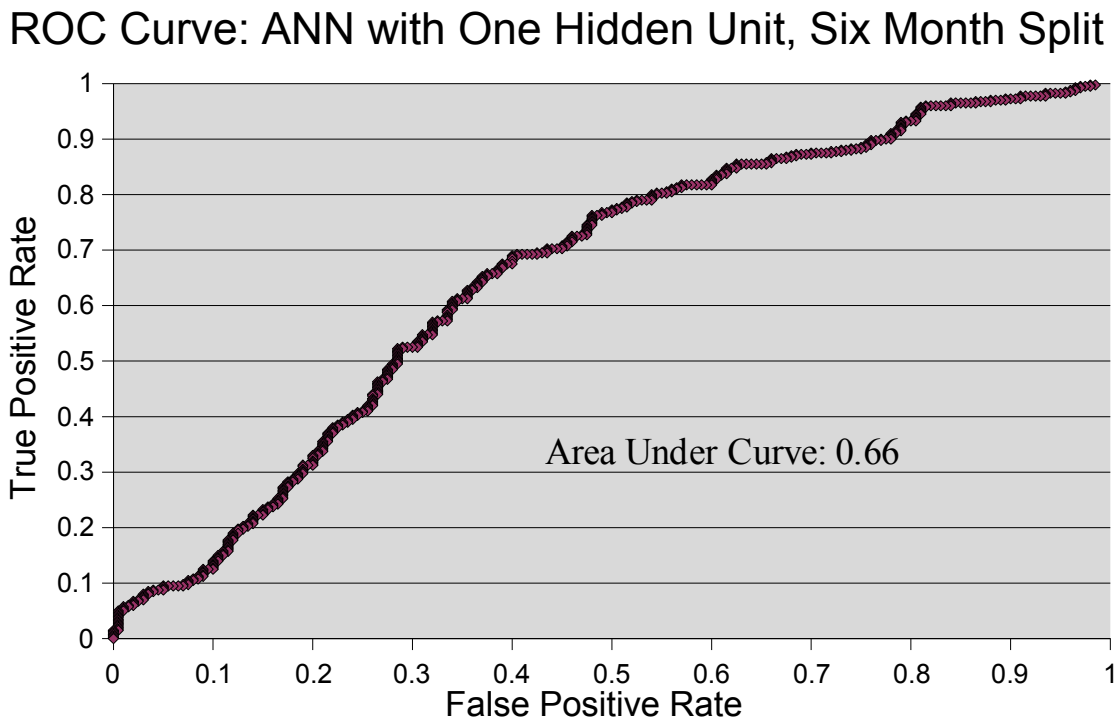

Figure 104: ROC Curve - Artificial Neural Network, One Hidden Unit: Six Month Split

4.6.4 Meta-Learning

The following three Combinations of Feature Selection and Machine Learning Algorithm were found in the previous section to produce the highest classification accuracies over the dataset with a nine month split:

- 70.2%: Logistic Regression using Gain Ratio to select 40 attributes
- 69.8%: Support Vector Machines, 0.9 kernel, using Gain Ratio to select 30 attributes
- 68.5%: Artificial Neural Networks, One Hidden Unit, using Gain Ratio to select 30 attributes

When we investigated meta-learning over the dataset with all attributes we found that bagging, boosting, and stacking are not useful. Our model selector in several cases was able to very slightly improve classification accuracy so we continue analysis.

4.6.4.1 Our Model Selector

In this section we investigate the use of our model selector to combine the models constructed by the three combinations of feature selection and machine learning algorithm found to be best for this dataset. Our model selector is first run using all three of these algorithms followed by three runs where one of these algorithms is removed.

The results of the runs of our model selector over this dataset are presented in Figure 105. The far left column shows the level-1 classifier used to determine the model that makes the final target class prediction. The classifiers with feature selection algorithms listed are run using the attribute selected classifier.

This meta-learning algorithm was able to very slightly improve the classification accuracy by combining these algorithms in several cases. The best increase in accuracy occurred from a model constructed combining logistic regression and support vector machines using naïve Bayes. This model does just over 0.5 percentage points better than either of the models used for its construction. When dealing with only 60 instances such a small increase is almost meaningless. As can be expected, there is still no statistically significant difference between logistic regression and this model. This model has a classification accuracy that is a statistically significant improvement over the classification accuracy of ZeroR.

Level-1	Level – 0 Models			
Model	Logistic, SVM, ANN	SVM, ANN	Logistic, ANN	Logistic, SVM
ANN 1 Hidden Unit	68.0	67.7	68.0	67.8
ANN 2 Hidden Units	67.5	67.2	68.0	67.7
ANN 1 HU & ReliefF_90	68.7	67.8	67.8	68.2
J48 & SVM_70	69.0	68.8	68.7	70.3
Logistic & PrincComp_15	69.8	68.5	67.8	70.8
NaiveBayes & GainRatio_30	68.5	68.2	66.7	70.2
SMO & ReliefF_40	67.8	67.0	67.2	70.2
J48	68.2	68.7	68.2	69.2
Logistic	67.8	67.5	68.3	69.2
LWL	68.3	68.2	69.8	70.0
NaiveBayes	66.8	68.8	67.8	69.0
SMO	66.3	67.7	67.2	69.0

Figure 105: Our Model Selector Results: Six Month Split

4.6.5 Summary

Over the dataset with the six month split target, there is no statistically significant difference ($p<0.05$) between the classification accuracy of logistic regression and ZeroR. There is no statistically

significant difference ($p<0.05$) between the classification accuracy of the models constructed using the three noteworthy combinations of feature selection and machine learning algorithm found and either ZeroR or logistic regression. The model constructed using the best machine learning algorithm without feature selection is ZeroR! These two facts are strong indications that the models we are evaluating are not appropriate for this target.

The best attribute selection method over this dataset is gain ratio attribute selection. This is the feature selection algorithm where all of the machine learning algorithms have a consistent classification accuracy that is higher than the classification accuracy for the models constructed using the attributes selected by the other feature selection models.

The support vector machines with a kernel exponent of 0.9 and logistic regression with feature selection perform well over this dataset. The classification accuracy of logistic regression is slightly higher than the classification accuracy of support vector machines. The ROC curve for logistic regression also has a greater area under and a better trade off between true and false positives than the curve for support vector machines.

Our model selector is able is able to very slightly increase the classification accuracy by less than one percentage point by selecting between models constructed using a logistic regression with feature selection and support vector machines with a kernel exponent of 0.9 by using logistic regression as the level-1 meta model to predict which model will make the best prediction for a given instance. The classification accuracy of this model constructed using our model selector is statistically significantly ($p<0.05$) better than ZeroR. The classification accuracy is not no statistically significant different ($p<0.05$) from logistic regression.

4.7 Attributes Selected by Medical Expert

When we introduced the use of logistic regression and ZeroR as baseline algorithms, we mentioned that the ideal baseline comparison would be the doctor's predictions of expected survival for each patient. Unfortunately, the doctors we are working with would be biased as they worked closely with each of the patients within our dataset. We thus had to rely on logistic regression and ZeroR to provide a baseline.

A large component of this thesis is the investigation into how to select the best set of attributes. We can compare the attributes selected by the feature selection algorithms already presented and the attribute selection performed by doctors. We asked one of the UMass medical doctors we have been collaborating with to select the 30 most relevant attributes from the dataset with all attributes. The attributes selected by this doctor are shown in Figure 106.

DemECOG	CxLiver	Histology	ResTransfusion	NoResNoHandle
SxWtloss	LabAlb	PreOutlook	ResAttemptUn	NoResMagnitude
SxChola	CTSMA	TxResect	ResPODays	NoResCeliacInvolve
SxAbd	CTHepatic	TxChemo	ResPOLeak	NoResSMAInvolve
SxBack	CTPortal	TxPal	ResPOLiverTB	NoResCirrhosis
CxHF	EUSCyto	TxPalCeliac	ResPathM	NoResMetastatic

Figure 106: Top 30 Attributes Selected By Medical Expert

In the following subsections we will compare the accuracy of the models constructed using these 30 attributes with the accuracy of models constructed using 30 attributes selected by Gain Ratio, ReliefF, and Support Vector Machine Attribute Evaluators. We will not be using Principal Components as this feature selection approach transforms the input space making a direct comparison with selecting the top 30 attributes impossible.

4.7.1 Attribute Selection Over Dataset with Six and Twelve Month Splits

Figure 107 shows the accuracy of the models constructed by the machine learning algorithm in

the far left column using the attribute selection technique in the other columns to select 30 attributes. The feature selection approach with the highest accuracy for each machine learning algorithm is highlighted in the table. In every case, the attribute selection algorithms performed better than the attributes selected by the medical expert.

Machine Learning Algorithm	Medical Expert	Gain Ratio	ReliefF	SVM
ZeroR	33	33	33	33
Logistic Regression	35	49	47	22
SMO with Kernel of 0.9	37	46	46	27
SMO with Kernel of 1.0	37	46	47	25
ANN with 1 Hidden Unit	38	46	48	27
ANN with 2 Hidden Units	39	49	51	25
Naïve Bayes	35	47	46	35
J4.8	38	44	39	33
Bayesian Network: 1 Parent	34	48	42	29
Bayesian Network: 2 Parents	30	43	41	32

Figure 107: Accuracy of Several Approaches to Attribute Selection

The highest overall classification accuracy was a model constructed using artificial neural networks with two hidden units over 30 attributes selected by ReliefF. Over the dataset with this target, this exact combination was found to be the third highest combination of machine learning algorithm and feature selection. The highest accuracy over the attributes selected by the medical expert also is constructed using artificial neural networks with two hidden units. The accuracy of the model constructed using the attributes selected by ReliefF is a statistically significant improvement ($p<0.05$) over ZeroR. The accuracy of the model constructed using the attributes selected by the medical expert is not statistically significantly different ($p<0.05$) from ZeroR.

Figure 108 lists the 30 attributes selected by ReliefF over the dataset with six and twelve month splits. The highlighted attributes are the ones also selected by the medical expert. Note that there are only five attributes in common between the set selected by the medical expert and by ReliefF.

SurOncName	SxPru	MedOncName	TxRadia	ResTFFP
ResPODischStatus	PTCStent	SxBack	TxChemoGem	CTNodeOmit
CxDiab	PTCDx	DemHeight	EUSNoNode	EUSOtherNode
SHCigarette	TxChemo	EUSTumorSizeX	SxSatiety	LabAmylase
SxOT	ResPOPulmComp	TxResect	SHAlcohol	CxPriorCancerChemo
EUSVascOmit	Histology	TxPal	EUSTumorSizeY	ResPOCourse

Figure 108: Top 30 Attributes Selected by ReliefF – Target with Six and Twelve Month Split

The highest combination of machine learning algorithm and feature selection over this dataset was artificial neural networks with one hidden unit using ReliefF to select the top 70 attributes. These seventy attributes are shown in Figure 109. Note that there is only an overlap of ten attributes, highlighted in the figure, between the seventy selected by ReliefF and the 30 selected by the medical expert.

Over this dataset the attributes selected by ReliefF can be used to construct models that have an accuracy that is statistically significantly better than ZeroR while the models constructed using the attributes selected by the medical expert have accuracies that are not statistically significantly different from ZeroR. There is also not a large amount of overlap between the attributes selected using ReliefF and the medical expert's attributes. These taken together indicates that ReliefF selects a better set of attributes than the medical expert over the dataset with this target though no statistically significant difference ($p<0.05$) exists.

SurOncName	MedOncName	ResTFFP	CxPriorCancerS	ResPODays	CxDiabOnset
ResPODischStatus	SxBack	CTNodeOmit	RadOncName	LabALT	LabBili
CxDiab	DemHeight	EUSOtherNode	ResTransfusion	EUSCeliacNode	SxCCS
SHCigarette	EUSTumorSizeX	LabAmylase	ResBloodLoss	CTTumorSizeX	SxFati
SxOT	TxResect	CxPriorCancerChemo	CTOtherNode	DemWeight	TxPalRad
EUSVascOmit	TxPal	ResPOCourse	ResPONG	EUSStagingN	SxChola
SxPru	TxRadia	EUSDx	ResPOLeak	FamilyMotherDx	SxVom
PTCDx	TxChemoGem	SxWtloss	ResPathR	SxInd	EUSStagingT
PTCStent	EUSNoNode	ERCPDx	EUSSMV	CxIHD	FamilyOther1
TxChemo	SxSatiety	PreOutlook	CxDiabDiet	SxBC	EUSCeliacClass
ResPOPulmComp	SHAlcohol	TxChemoFlu	CxPriorCancerR	FamilyOther2	
Histology	EUSTumorSizeY	ResPOInfection	CxPriorCancer	CTPortalClass	

Figure 109: Top 70 Attributes Selected by ReliefF – Target with Six and Twelve Month Split

4.7.2 Attribute Selection Over Dataset with Nine Month Split

Figure 110 shows the accuracy of the models constructed by the machine learning algorithm in the far left column using the attribute selection technique in the other columns to select 30 attributes. The feature selection approach with the highest accuracy for each machine learning algorithm is highlighted in the table. In every case, the attribute selection algorithms performed better than the attributes selected by the medical expert.

Machine Learning Algorithm	Medical Expert	Gain Ratio	ReliefF	SVM
ZeroR	50	50	50	50
Logistic Regression	49	49	58	40
SMO with Kernel of 0.9	52	53	58	39
SMO with Kernel of 1.0	50	51	57	40
ANN with 1 Hidden Unit	62	54	60	38
ANN with 2 Hidden Units	62	53	60	37
Naïve Bayes	52	60	55	51
J4.8	51	50	56	49
Bayesian Network: 1 Parent	62	57	59	48
Bayesian Network: 2 Parents	60	51	56	40

Figure 110: Accuracy of Several Approaches to Attribute Selection

The highest overall classification accuracy was a model constructed using artificial neural networks with two hidden units with 30 attributes selected by the medical expert. The highest accuracy over the attributes selected a feature selection algorithm was constructed using artificial neural networks with one hidden unit constructed with 30 attributes selected by ReliefF. Neither of these accuracies are statistically significantly different from ZeroR.

Figure 111 lists the 30 attributes selected by ReliefF over the dataset with a nine month split. The highlighted attributes are the ones also selected by the medical expert. Note that there are only seven attributes in common between the set selected by the medical expert and by ReliefF.

TxLap	SurOncName	DemECOG	ResBloodLoss	TxRadia
SxSatiety	PTCStent	CxDiab	SxWtloss	TxPal
ResPODischStatus	PTCDx	CxPriorCancerSurgery	ResPOPulmComp	TxChemo
SxPru	ResPOInfection	SxFati	TxChemoGem	LabCEA
SxBack	GIMDName	DemWeight	TxChemoFlu	CxDiabOral
TxResect	Histology	RadOncName	EUSStagingT	PresumptiveDx

Figure 111: Top 30 Attributes Selected by ReliefF - Nine Month Split

The highest combination of machine learning algorithm and feature selection over this dataset was support vector machines with kernel exponent of 0.9 using ReliefF to select the top one hundred attributes. This combination of machine learning algorithm and feature selection has an accuracy that is statistically significantly better than ZeroR. The 100 attributes selected by support vector machines are shown in Figure 112. Note that there is only an overlap of twelve attributes, highlighted in the figure, between the seventy selected by ReliefF and the 30 selected by the medical expert.

PresumptiveDx	LabCA19-9	EUSSMV	SxPru	SxBC	CxDiab
CTHepatic	CTDx	CTTumorSizeX	CxHF	SxChola	CxDiabDiet
CTHepaticClass	CTCeliac	PTCDx	CxResp	SHDrugUse	CxDiabOral
CTCeliacClass	CTVascOmit	CTTumorSizeY	CxIHD	SHOth	CxPriorCancer
CTSMAClass	LabAST	CTCeliacNode	SxDyspha	SHExposure	CxPriorCancerRadiation
CTSMA	CXRDx	CTNodeOmit	SxOT	CxPriorCancerSurgery	CxPriorCancerChemo
CTSMVClass	LabAmylase	CTOtherNode	SxSatiety	SHAlcohol	CxBleed
CTPortalClass	EUSHepaticClass	EUSVascOmit	SxWtloss	SHCigarette	CxMal
CTPortal	EUSInferior	EUSCeliacClass	SxJaun	FamilyOther1Dx	CxLiver
CTInferior	EUSSMA	EUSCeliac	SxWtlossP	FamilyOther2Dx	ResPODays
CTSMV	EUSHepatic	PTCStent	DemECOG	FamilyOther2	ResPOInfection
CTInferiorClass	EUSSMAClass	EUSDx	DemWeight	FamilyFatherDx	ResAttempt
LabBili	EUSPortal	PTCStentType	DemHeight	FamilyOther1	ResPOCourse
LabALT	EUSCeliacNode	SxAbd	SxNau	FamilyMotherDx	ResAttemptUn
LabAlka	EUSPortalClass	SxBack	SxCCS	CxDiabOnset	ResPOPulmComp
LabCEA	EUSInferiorClass	SxFati	SxVom	CxHyper	
LabAlb	EUSSMVClass	SxInd	SxChole	CxRF	

Figure 112: Top 100 Attributes Selected by Support Vector Machines - Nine Month Split

Over this dataset the accuracy of the models constructed with only 30 attributes are not statistically significantly different from ZeroR. This is the case for both the attributes selected by the medical expert and by the feature selection algorithms. There is also not a large amount of overlap between the 100 attributes selected using support vector machines and the medical expert's 30

attributes. These taken together show that 30 attributes is not enough to predict the expected survival of a patient over the dataset with this target and that feature selection algorithms are able to do as good, if not better than, the medical expert's predictions of the best attributes.

4.7.3 Attribute Selection Over Dataset with Six Month Split

Figure 113 shows the accuracy of the models constructed by the machine learning algorithm in the far left column using the attribute selection technique in the other columns to select 30 attributes. The feature selection approach with the highest accuracy for each machine learning algorithm is highlighted in the table. In every case, the attribute selection algorithms performed better than the attributes selected by the medical expert.

Machine Learning Algorithm	Medical Expert	Gain Ratio	ReliefF	SVM
ZeroR	66.7	66.7	66.7	66.7
Logistic Regression	64.8	66.5	70.0	61.3
SMO with Kernel of 0.9	63.3	65.5	69.0	60.7
SMO with Kernel of 1.0	65.3	70.3	68.0	46.7
ANN with 1 Hidden Unit	62.3	69.0	67.8	52.3
ANN with 2 Hidden Units	35.2	62.7	60.7	50.0
Naïve Bayes	68.3	70.8	67.7	54.0
J4.8	66.2	69.5	65.7	62.2
Bayesian Network: 1 Parent	71.0	71.7	68.5	58.2
Bayesian Network: 2 Parents	64.7	67.0	65.7	59.0

Figure 113: Accuracy of Several Approaches to Attribute Selection

The highest overall classification accuracy was a model constructed using artificial neural networks with two hidden units with the 30 attributes selected by gain ratio. The highest accuracy over the attributes selected the medical expert was also constructed using artificial neural networks with one hidden unit constructed with 30 attributes. Neither of these model's accuracies are statistically significantly different ($p<0.5$) than ZeroR or than each other.

Figure 114 lists the 30 attributes selected by ReliefF over the dataset with a nine month split. The highlighted attributes are the ones also selected by the medical expert. Note that there are eight

attributes in common between the set selected by the medical expert and by ReliefF.

NoResMagnitude	TxPalBypass	SxBC	SurOncName	PreOutlook
PTCStent	NoResSMAInvolve	SHExposure	CxHF	TxResect
PTCDx	CTHepatic	TxPal	CxDiabDiet	EUSCeliacNode
TxPalStens	TxPalJejTube	ResPODischStatus	EUSSMV	TxChemoTax
NoResNoHandle	TxPalGasTube	EUSDx	TxChemoIri	CxIHD
CxPriorCancerChemo	CTSMA	SxInd	CXRDx	ResPOPulmComp

Figure 114: Top 30 Attributes Selected by Gain Ratio - Six Month Split

The highest combination of machine learning algorithm and feature selection over this dataset was support vector machines with kernel exponent of 0.9 using ReliefF to select the top one hundred attributes. This combination of machine learning algorithm and feature selection has an accuracy that is also not statistically significantly different ($p<0.05$) than ZeroR. For this reason we will not compare the set of attributes selected by support vector machines and the medical expert.

Over this dataset the accuracy of the models constructed with only 30 attributes are not statistically significantly different ($p<0.05$) from ZeroR. This is the case for both the attributes selected by the medical expert and by the feature selection algorithms. This shows that 30 attributes are not enough to predict the expected survival time of a patient over the dataset with this target.

4.7.4 Summary

In only the dataset with a six and a twelve month split for the target was there a statistically significant difference ($p<0.05$) between model constructed using attributes selected by feature selection and ZeroR. Over this dataset there was not a statistically significant difference ($p<0.05$) between the best model constructed using the thirty attributes selected by the medical expert and ZeroR. This is an indication that the attributes selected by feature selection may be slightly better than those selected by the medical expert but is not conclusive. For every other target there was no statistically significant difference between combinations of machine learning algorithm and selection of 30 attributes.

Over all three datasets there was no statistically significant difference ($p<0.05$) between the best model constructed by a combination of machine learning algorithm and the medical expert's attributes and a combination of a machine learning algorithm and the set of attributes selected the feature selection algorithms. This means that machine learning algorithms can do as good a job of predicting a subset of attributes as this medical expert.

5 Conclusions and Future Work

This thesis set out with a goal of constructing models to predict the expected survival time of a patient diagnosed with pancreatic cancer. For the construction of these models we have detailed records including 190 attributes related to 60 patients seen at the University of Massachusetts Medical School. We constructed six datasets from this record, one group with all attributes and the other group with only preoperative. Each group of datasets has the survival target discretized in three different ways. We expect that any models we construct will perform better than selecting the most frequent survival class. The medical community understands logistic regression for modeling these relationships so logistic regression provides a logical benchmark for models constructed using machine learning algorithms.

This thesis has investigated applying a variety of machine learning techniques to construct these models of survival time. We have focused on finding the best combinations of feature selection and machine learning algorithm for each of the six datasets. We have also investigated the use of meta learning approaches, including our model selector, to combine these models constructed using machine learning algorithms into algorithms with greater predictive accuracy.

This investigation showed that in four of the six datasets models could be constructed by combining machine learning algorithms with feature selection that predict expected survival better than arbitrarily choosing the most likely class. Logistic regression over all 190 attributes is not able to do better than arbitrarily choosing the most likely class.

The two out of the six datasets where no difference between our models and arbitrarily choosing most likely class both have target values of less than six months and greater than six month survival. Our machine learning algorithms seem not to be appropriate for modeling datasets with this target.

Artificial neural networks, Bayesian networks with two parents, and support vector machines were able to construct models that performed well when the right set of attributes are selected. Bayesian networks were observed to be resistant to large numbers of attributes while support vector machines and artificial neural networks can be adversely affected if too many attributes are present.

Gain ratio attribute selection and ReliefF attribute selected the best subsets of attributes though there were also cases when support vector machine attribute evaluator also performed well. These three feature selection algorithms consistently selected a better set of features than principal components.

The accuracy of logistic regression can be increased through the use of feature selection. In one of the models presented, logistic regression with feature selection performs better than arbitrarily selecting the most likely class.

We compared the feature selection algorithms used here to the features selected by a domain expert. The predictive power of attributes selected by the domain expert is no better than, and in some cases slightly worse than, that of the attributes selected by the feature selection algorithms.

Traditional meta-learning algorithms of stacking, bagging, and boosting are not useful over this dataset as they produce neither an increase in the classification accuracy nor a decrease in the standard deviation.

We designed and implemented a new meta-learning algorithm, which we call model selector. Our model selector meta-learning algorithm selects which of several model constructed by competing machine learning algorithm is most likely to correctly predict the correct target for an unseen instance. It then runs that unseen instance thought the selected model to generate a target prediction for that instance. Our model selector has the potential to have a classification accuracy greater than the

classification accuracy of the input models, though not consistently. In four of the six datasets there was an increase in the classification accuracy.

5.1 Future Work

Future work into investigation of quality of life should be carried out to extend the usefulness of the models constructed to predict survival time as a prognostic tool. The same systematic investigation of various machine learning and feature selection algorithms should be performed over the quality of life dataset as has been performed in this thesis for the quality of life dataset.

There are many national databases of patients. These have the disadvantage of having only a small amount of information about each patient but contain a much larger sample of patients. It would be interesting to evaluate how the machine learning algorithms perform over these datasets in comparison to their performance over the smaller dataset worked with in this thesis.

It is clear from this work that Bayesian networks with a maximum of two parents are among the best approaches for modeling survival time. Bayesian networks show a resistance to being adversely affected by too many irrelevant attributes which is important over such a highly dimensional dataset. The models constructed using Bayesian networks should be further investigated and evaluated in conjunction with clinical experts.

Finally, our model selector shows potential to be effective at increasing classification accuracy. This method should be further evaluated over other datasets and refined as needed.

6 Appendix A: List of Dataset Attributes

Follows is a list of all attributes in the pancreatic cancer dataset discussed throughout this thesis. Attributes that are not highlighted are the pre-operative attributes. Note that Figure 5 on page 22 has a description of each of these attributes.

Attribute	Category	Description
PresumptiveDx	Presentation	Presumptive Diagnosis (Pancreatic tumor, periampullary tumor, etc...)
DemECOG	Presentation	Demographics - ECOG Score (0-4)
DemHeight	Presentation	Demographics - Height in Inches of Patient
DemWeight	Presentation	Demographics - Weight in Pounds of Patient at Admission
SxWtloss	Presentation	Initial Symptoms - Weight Loss
SxWtlossP	Presentation	Initial Symptoms - Weight Loss - Pounds
AxJaun	Presentation	Initial Symptoms - Juandice
SxChole	Presentation	Initial Symptoms - Cholecystitis
SxChola	Presentation	Initial Symptoms - Cholangitis
SxBC	Presentation	Initial Symptoms - Biliary Colic
SxNau	Presentation	Initial Symptoms - Nausea
SxVom	Presentation	Initial Symptoms - Vomiting
SxCCS	Presentation	Initial Symptoms - Clay Colored Stool
SxFati	Presentation	Initial Symptoms - Fatigue
SxPru	Presentation	Initial Symptoms - Pruritis
SxInd	Presentation	Initial Symptoms - Indigestion
SxAbd	Presentation	Initial Symptoms - Abdominal Pain
SxBack	Presentation	Initial Symptoms - Back Pain
SxDyspha	Presentation	Initial Symptoms - Dysphagia
SxSatiey	Presentation	Initial Symptoms - Early Satiety
SxOT	Presentation	Initial Symptoms - Other
CxHF	History	Comorbidities - Heart Failure
CxIHD	History	Comorbidities - Ischemic Heart Disease
CxResp	History	Comorbidities - Respiratory
CxDiab	History	Comorbidities - Diabetes
CxDiabOral	History	Comorbidities - Diabetes - Insulin - Oral
CxDiabDiet	History	Comorbidities - Diabetes - Insulin - Diet Control
CxDiabOnset	History	Comorbidities - Diabetes - Onset (1=Less than six months, 2 =Greater tl
CxRF	History	Comorbidities - Renal Failure
CxHyper	History	Comorbidities - Hypertension
CxBleed	History	Comorbidities - Bleeding Disorder
CxLiver	History	Comorbidities - Liver Failure
CxMal	History	Comorbidities - Malnutrition
CxPriorCancer	History	Comorbidities - Prior Cancer Dx
CxPriorCancerChemo	History	Comorbidities - Prior Cancer Dx - Chemo
CxPriorCancerRadiation	History	Comorbidities - Prior Cancer Dx - Radiation
CxPriorCancerSurgery	History	Comorbidities - Prior Cancer Dx - Surgery
SHCigarette	History	Social History - Cigarettes (significant use)
SHAlcohol	History	Social History - Alcohol (significant use)
SHDrugUse	History	Social History - Drug Use
SHExposure	History	Social History - Environmental Exposure
SHOth	History	Social History - Other

Attribute	Category	Description
FamilyFatherDx	History	Family History - Father Dx
FamilyMotherDx	History	Family History - Mother Dx
FamilyOther1	History	Family History - Other1
FamilyOther1Dx	History	Family History - Other1 Dx
FamilyOther2	History	Family History - Other2
FamilyOther2Dx	History	Family History - Other2 Dx
LabCEA	Serum	Laboratory - CEA
LabCA19-9	Serum	Laboratory - CA19-9
LabAlb	Serum	Laboratory - Albumin
LabBili	Serum	Laboratory - Bilirubin
LabAlka	Serum	Laboratory - Alkaline phosphotase
LabALT	Serum	Laboratory - ALT
LabAST	Serum	Laboratory - AST
LabAmylase	Serum	Laboratory - Amylase
CXRDx	DiagImg	CXR - Diagnosis
CTDx	DiagImg	CT - Diagnosis
CTVascOmit	DiagImg	CT - Vascular Omission
CTCeliac	DiagImg	CT - Celiac Involvement
CTCeliacClass	DiagImg	CT - Celiac Involvement Class
CTSMA	DiagImg	CT - SMA Involvement
CTSMAClass	DiagImg	CT - SMA Involvement Class
CTHepatic	DiagImg	CT - Hepatic Involvement
CTHepaticClass	DiagImg	CT - Hepatic Involvement Class
CTInferior	DiagImg	CT - Inferior Vena Cava Involvement
CTInferiorClass	DiagImg	CT - Inferior Vena Cava Involvement Class
CTSMV	DiagImg	CT - SMV Involvement
CTSMVClass	DiagImg	CT - SMV Involvement Class
CTPortal	DiagImg	CT - Portal Vein Involvement
CTPortalClass	DiagImg	CT - Portal Vein Involvement Class
CTCeliacNode	DiagImg	CT - Celiac Nodal Disease
CTOtherNode	DiagImg	CT - Other Nodal Disease
CTNodeOmit	DiagImg	CT - Node Omission
CTTumorSizeX	DiagImg	CT - Tumor Size (cm) - Width
CTTumorSizeY	DiagImg	CT - Tumor Size (cm) - Height
PTCDx	DiagImg	PTC - Diagnosis
PTCStent	DiagImg	PTC - Stent
PTCStentType	DiagImg	PTC - Stent Type
EUSDx	Endoscopy	EUS - Diagnosis
EUSVascOmit	Endoscopy	EUS - Omission
EUSCeliac	Endoscopy	EUS - Celiac Involvement
EUSCeliacClass	Endoscopy	EUS - Celiac Involvement Class
EUSSMA	Endoscopy	EUS - SMA Involvement
EUSSMAClass	Endoscopy	EUS - SMA Involvement Class
EUSHepatic	Endoscopy	EUS - Hepatic Involvement
EUSHepaticClass	Endoscopy	EUS - Hepatic Involvement Class
EUSInferior	Endoscopy	EUS - Inferior Vena Cava Involvement
EUSInferiorClass	Endoscopy	EUS - Inferior Vena Cava Involvement Class
EUSSMV	Endoscopy	EUS - SMV Involvement
EUSSMVClass	Endoscopy	EUS - SMV Involvement Class

Attribute	Category	Description
EUSPortal	Endoscopy	EUS - Portal Vein Involvement
EUSPortalClass	Endoscopy	EUS - Portal Vein Involvement Class
EUSCeliacNode	Endoscopy	EUS - Celiac Node Disease
EUSOtherNode	Endoscopy	EUS - Other Nodal Disease
EUSNoNode	Endoscopy	EUS - No Nodes Mentioned
EUSTumorSizeX	Endoscopy	EUS - Tumor Size (cm) - Width
EUSTumorSizeY	Endoscopy	EUS - Tumor Size (cm) - Height
EUSStagingT	Endoscopy	EUS - Staging - T
EUSStagingN	Endoscopy	EUS - Staging - N
EUSCyto	Endoscopy	EUS - FNA Cytology
ERCPDx	Endoscopy	ERCP - Diagnosis
ERCPStent	Endoscopy	ERCP - Stent
ERCPStentType	Endoscopy	ERCP - Stent Type
Histology	Path	Histology
PreOutlook	Prelim	Pre-Surgical Tumor Outlook (Potentially Resectable, Locally Advanced/U
TxResect	Treatment	Treatment - Resection
TxLab	Treatment	Treatment - Laparoscopy
TxRadia	Treatment	Treatment - Radiation
TxRadiaAdju	Treatment	Treatment - Radiation - Adjuvancy
TxChemo	Treatment	Treatment - Chemo
TxChemoAdju	Treatment	Treatment - Chemo - Adjuvancy
TxChemoAVA	Treatment	Treatment - Chemo - Avastin
TxChemoCap	Treatment	Treatment - Chemo - Capecitabine
TxChemoErb	Treatment	Treatment - Chemo - Erbitux
TxChemoFlu	Treatment	Treatment - Chemo - Fluorouracil (5-FU)
TxChemoFUDR	Treatment	Treatment - Chemo - FUDR
TxChemoGem	Treatment	Treatment - Chemo - Gemcitabine
TxChemoIri	Treatment	Treatment - Chemo - Irinotecan
TxChemoLeu	Treatment	Treatment - Chemo - Leukovorin
TxChemoLev	Treatment	Treatment - Chemo - Levamasole
TxChemoMit	Treatment	Treatment - Chemo - Mitomycin
TxChemoOxa	Treatment	Treatment - Chemo - Oxaliplatin
TxChemoTax	Treatment	Treatment - Chemo - Taxol
TxChemoOth	Treatment	Treatment - Chemo - Other
TxChemoOthSpecify	Treatment	Treatment - Chemo - Other - Specify
TxPal	Treatment	Treatment - Palliation
TxPalRes	Treatment	Treatment - Palliation - Pall. Resection
TxPalBypass	Treatment	Treatment - Palliation - Bypass
TxPalCeliac	Treatment	Treatment - Palliation - Celiac Block
TxPalPara	Treatment	Treatment - Palliation - Paracentesis
TxPalTho	Treatment	Treatment - Palliation - Thoracentesis
TxPalRad	Treatment	Treatment - Palliation - Pall. Radiation
TxPalTrans	Treatment	Treatment - Palliation - Transfusion
TxPalStens	Treatment	Treatment - Palliation - Pall. Stenting
TxPalPV	Treatment	Treatment - Palliation - PV Shunts
TxPalHAL	Treatment	Treatment - Palliation - HAL
TxPalGasTube	Treatment	Treatment - Palliation - Gastrostomy Tube
TxPalJejTube	Treatment	Treatment - Palliation - Jejunstomy Tube
TxPalOth	Treatment	Treatment - Palliation - Other

Attribute	Category	Description
TxExp	Treatment	Treatment - Experimental protocol (ie. vaccine)
TxGene	Treatment	Treatment - Gene Counseling
ResPxType	Res	Resection - Procedure Type (Whipple, total pancreatectomy, distal panc
ResORTime	Res	Resection - OR Time (hr.)
ResVenRes	Res	Resection - Venous Resection
ResVenRec	Res	Resection - Venous Reconstruction
ResArtRes	Res	Resection - Arterial Resection
ResArtRec	Res	Resection - Arterial Reconstruction
ResOrgans	Res	Resection - Other Organs Resection
ResBloodLoss	Res	Resection - Estimated Blood Loss (cc)
ResTransusion	Res	Resection - Tranfusion
ResTUnits	Res	Resection - Transfusion Units
ResTFFP	Res	Resection - Transfusion - FFP
ResTCells	Res	Resection - Transfusion - Cell
ResAttempt	Res	Resection - Resection Attempt
ResAttempUn	Res	Resection - Resection Unsuccessful Reason (Tumor involvement, Operat
ResPOCourse	Res	Resection - PO - Post-Op Care Path
ResPODays	Res	Resection - PO - Time in ICU (days)
ResPOInfection	Res	Resection - PO - Wound infection
ResPOLeak	Res	Resection - PO - Leak
ResPONG	Res	Resection - PO - NG/gastrotomy drainage
ResPOAbdominal	Res	Resection - PO - Abdominal Collection
ResPOPulmComp	Res	Resection - PO - Pulminary Complications
ResPOLiverInsuf	Res	Resection - PO - Liver Insufficiency
ResPOLiverTB	Res	Resection - PO - Liver Insufficiency - Total Bilirubin
ResPODischStatus	Res	Resection - Discharge Status
ResPathT	Path	Resection - Pathology Staging - T
ResPathN	Path	Resection - Pathology Staging - N
ResPathM	Path	Resection - Pathology Staging - M
RestPathR	Path	Resection - Pathology Staging - R
ResPathV	Path	Resection - Pathology Staging - V
ResPathSizeX	Path	Resection - Pathology Tumor Size (cm) - Width
NoResNoHandle	NoRes	No Resection - Couldn't Handle Proposed Treatment
NoResRefused	NoRes	No Resection - Refused Treatment
NoResMagnitude	NoRes	No Resection - Magnitude Not Worth Benefits
NoResCeliacInvolve	NoRes	No Resection - Celiac Trunk Involvement
NoResSMAInvolve	NoRes	No Resection - SMA Involvement
NoResHepaticInvolve	NoRes	No Resection - Hepatic Involvement
NoResIVCInvolve	NoRes	No Resection - Inferior Vena Cava Involvement
NoResSMVInvolve	NoRes	No Resection - SMV Involvement
NoResPVInvolve	NoRes	No Resection - Portal Vein Involvement
NoResCirrhosis	NoRes	No Resection - Cirrhosis
NoResMaetastatic	NoRes	No Resection - Metastatic
Age	Patient	Patient – Age
Gender	Patient	Patient – Gender
MedOncName	Patient	Patient – Medical Oncologist
SurgOncName	Patient	Patient – Surgical Oncologist
RadOncName	Patient	Patient – Radiation Oncologist
GIMDName	Patient	Patient – Gastroenterologist
Survival		From Admission date to death date.

7 Bibliographical References

[ACS07] “Cancer Facts & Figures 2007” American Cancer Society <http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf>.

[Ahm05] Farid E. Ahmen. "Artificial Neural Networks for Diagnosis and Survival Prediction in Colon Cancer." *Molecular Cancer* 4.29(2005) <http://www.molecular-cancer.com/content/4/1/29>.

[BCD05] R. Bittern, A. Cuschieri, S. D. Dolgobrodov, R. Marshall, P. Moore, and R. J. C. Steele. "An Artificial Neural Network for Analysing the Survival of Patients with Colorectal Cancer." *European Symposium on Artificial Neural Networks* 2005 proceedings 27-29 April 2005 <http://www.dice.ucl.ac.be/Proceedings/esann/esannpdf/es2005-162.pdf>.

[BGR97] Harry B. Burke, Phillip H. Goodman, David B. Rosen, Donald E. Henson, John N. Weinstein, Frank E. Harrell, Jr., Jeffery R. Marks, David P. Winchester,, and David G. Bostwick. "Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction." *Cancer* 79.4 15 Febuary 1997 857-862. 24 <http://brain.cs.unr.edu/publications/burke.goodman.ANNsurvival.Cancer97.pdf>.

[CS00] Nello Cristianini, and John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge: Cambridge University Press, 2000.

[DHR05] Vincent T. DeVita, Jr., Samuel Hellman, and Steven A. Rosenburg. Cancer Principles & Practice of Oncology. 7th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2005.

[DZ04] Saso Dzeroski and Bernard Zenko. “Is Combining Classifiers with Stacking Better than Selecting the Best One?”. *Machine Learning*, 54, 25-273, 2004.

[Faw03] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. HP Laboratories Palo Alto, 2003 <http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf>.

[Hay06] John Hayward. Mining Oncology Data: Knowledge Discovery in Clinical Performance of Cancer Patients. MS Thesis. Worcester, MA: Worcester Polytechnic Institute, 2006.

[Kon94] Igor Kononenko. "Estimating Attributes: Analysis and Extensions of RELIEF". *European Conference on Machine Learning*, 171-182, 1994.

[KRS97] Charles E. Kahn Jr, Linda M. Roerts, Katherine A. Shaffer, and others. Construction of a Bayesian Network for Mammographic Diagnosis of Breast Cancer. *Comput. Biol.* Med. Vol. 27 No 1. 1997.

[Mit97] Tom Mitchell. Machine Learning. Boston, MA: WCB McGraw-Hill, 1997.

[Qui86] Quinlan, J. R., "Induction of decision trees". *Machine Learning* Vol. 1 No 1. 1986.

[RK97] Marko Robnik-Sikonja and Ior Kononenko. An adaption of Relief for attribute estimation in regression. *Fourteenth International Conference of Machine Learning*, 296-304, 1997.

[RN03] Stuart J. Russell and Peter Norvig. Artificial Intelligence A Modern Approach. Upper Saddle River NJ: Prentice Hall, 2003.

[Smi02] Lindsay I. Smith. A tutorial on Principal Components Analysis. February, 2002 <http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf>.

[VD02] Richardo Vilalta and Youssef Drissi. "A Perspective View And Survey Of Meta-Learning". *Artificial Intelligence Review*, 18:2, 77-95, 2002.

[WF05] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.