2014

Applying Machine Learning Techniques In Diagnosing Bacterial Vaginosis

Yolanda Sherene Baker
North Carolina Agricultural and Technical State University

Applying Machine Learning Techniques in Diagnosing Bacterial Vaginosis

Yolanda Sherene Baker

North Carolina A&T State University

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Computer Systems Technology

Major: Information Technology

Major Professor: Dr. Rajeev Agrawal

Greensboro, North Carolina

2014

The Graduate School
North Carolina Agricultural and Technical State University
This is to certify that the Master's Thesis

Yolanda Sherene Baker

has met the thesis requirements of
North Carolina Agricultural and Technical State University

Greensboro, North Carolina
2014

Approved by:

Dr. Rajeev Agrawal Major Professor	Dr. Gerry Dozier Committee Member
Dr. Clay Gloster, Jr. Committee Member	Dr. Evelyn Sowells Committee Member
Dr. Clay Gloster, Jr. Department Chair	Dr. Sanjiv Sarin Dean, The Graduate School

Biographical Sketch

Yolanda Sherene Baker earned her Bachelor of Social Work degree from North Carolina Agricultural and Technical State University (A&T) in 1995. She went on to pursue an Associate in Applied Science in Business Administration with a concentration in Human Resources and completed the degree requirements in 2007. In 2012, she entered the Master of Science in Information Technology program at A&T.

Yolanda was inducted into the Honor Society of Phi Kappa Phi and became the Student Vice-President for the 2013-2014 school year. She also received the Wadaran Latamore Kennedy 4.0 GPA Scholar award and the 2nd Place Oral Graduate Presentation award at the 11th Annual Ronald E. McNair National Research Symposium.

While pursuing her degree, Yolanda worked as a teacher assistant for the department of Computer Systems Technology. She also worked as a graduate research assistant for the department of Computer Science.

Yolanda has presented her research at the 2013 IEEE Intelligence and Security Informatics (ISI) international conference and a paper has been accepted at the 52nd Annual ACM Southeast Conference. Additionally she has published her research in the Encyclopedia of Social Network Analysis and Mining.

Yolanda's thesis, Applying Machine Learning Techniques in Diagnosing Bacterial Vaginosis, was supervised by Dr. Rajeev Agrawal.

Dedication

First and foremost, I give all praise, glory and honor to my Lord and Savior Jesus Christ; without Him I am nothing. In Him I live, move and have my being.

I dedicate this thesis to my children Nia and Lewis and my mother Lenora Thorne. You are the wind beneath my wings and the main reason I try to reach for the stars. I love you with all of my heart. I want to thank you for the sacrifices of time and money you have made during my tenure in graduate school. Your support means everything and I always want to make you proud.

I would also like to thank the following people: Pastor Terry Warr, Caroline Jones, Jennifer King-Fairley and Jacqueline Thorne for all of your support and prayers, especially when I needed it most; Dr. Kelvin Bryant for sharing your knowledge and time during our impromptu tutorial sessions in addition serving as my "practice" committee; Patrice Bryant, Paula Daniel and Natalie Hollomon for making sure my thesis was grammatically correct and typo free and; last but certainly not least, my fellow grad school eagles, Elizabeth "Beth" Winchester, James Ashe and Shanell Frazer. I would not have made it without our study groups, phone calls, text messages, resource exchanges and that ever so important "You Got This" encouragement when I needed a push to finish the race.

Acknowledgements

First, I would like to express my deepest appreciation to my advisor Dr. Rajeev Agrawal for demonstrating the perfect balance between pushing me further than I would have gone on my own while, also knowing when to pull back. He understood my limitations, but also had the insight to see my hidden potential. His patience and kindness will be unforgettable. I have been able to accomplish more than I ever imagined as a graduate student. Equally, I would like express my sincere gratitude for Dr. Gerry Dozier. He was not only inspirational in my pursuit of graduate school, but invited me to be a part of his research team and funded my research. This thesis would not have been possible without him. Words cannot express how thankful I am to him for his belief in me. Third, I would like to thank Dr. James A. Foster and Daniel Beck for supplying the dataset for this research and making themselves available to answer our questions. Next, I would like to thank Dr. Clay Gloster, Jr. for having an open door, providing valuable advice and support and enthusiastically accepting the invitation to be on my thesis committee. Finally, I would like to extend my gratitude to Dr. Evelyn Sowells who never failed to encourage each time we passed each other in the hallway. It is an honor to have her as a member of my thesis committee.

This research was funded by the National Science Foundation (NSF), Science & Technology Center: Bio/Computational Evolution in Action Consortium (BEACON). The author would like to thank the NSF and BEACON for their support of this research.

Table of Contents

List of Figures

List of Tables

Abstract

Bacterial Vaginosis (BV) is the most common of vaginal infections diagnosed amongst women of child bearing years. Yet, there is very little insight as to how it occurs. There are a vast number of criteria that can be taken into consideration in determining the presence of BV. The purpose of this thesis is two-fold: first, to discover the most significant features necessary to diagnose the infection, and second, to apply various classification algorithms on the selected features. In order to fulfill our purpose, we conducted an array of experiments on the data. We tested the full set of raw data, removed the time series features, tested the medical and clinical features in isolation, cleaned the data and performed the same experiments on the clean full, clean clinical and clean medical datasets. We compared the accuracy, precision, recall and F-measure and time elapsed for each feature selection and classification grouping. It is observed that certain feature selection algorithms provided only a few features; however, the classification results were as good as using a large number of features. After comparing all of the experiments, the algorithms performed best on the raw full and clean full datasets. However, the raw full dataset returned better comprehensive results.

CHAPTER 1

Introduction

1.1 Introduction

Machine learning (ML) algorithms are centered around predictions based on generalizations created from previous examples. The provision of larger amounts of data allows for tackling larger problems (Domingos, 2012). ML transforms massive amounts of raw data into knowledge that becomes useful for the analyst. It is employed in business, academia, government, science and other industries. Its utilization runs the gamut and has been applied to many different types of data including fraud detection (Akoglu & Faloutsos, 2013), business negotiations (Jim, 1996), facial recognition (Joseph Shelton et al., 2011), email messages (Kiritchenko & Matwin, 2011) and many other applications. New ML algorithms are being developed and computers are becoming more powerful, which can lend itself to addressing complex problems with more accuracy and expeditiousness in a way that is practically impossible for humans.

The medical field is quickly embracing machine learning methodologies as these approaches have shown progress in their usefulness in prediction and classification. This implementation could prove useful in discovering ways to lower the cost of medication, improve clinical studies and help facilitate better assessments by physicians (Salama, Abdelhalim, & Zeid, 2012). ML can improve the healthcare process as data continues to increase and decrease the human effort that would traditionally be required. It has been used in the medical field to diagnose lung cancer (Kancherla & Mukkamala), breast cancer (Osareh & Shadgar, 2010), asthma (Prasad, Prasad, & Sagar, 2011), heart disease (Al-Shayea, 2011), dementia (Williams, Weakley, Cook, & Schmitter-Edgecombe, 2013) and other diseases and conditions. ML has

recently been used to compare the performance of a variety of classification algorithms in detecting breast cancer (Yau & Othman, 2007). The algorithms compared were Bayes Network, Radial Basis Function Networks (RBF), Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm.

There is a minimal amount of published research using supervised machine learning to diagnose BV. In the past few years and as recent as this year, Srinivasan et al. (2012), Ravel et al. (2011) and Beck & Foster (2014) have used both supervised and unsupervised machine learning techniques to classify BV related microbiota. However, we are expanding this research by conducting experiments using a different dataset.

In this thesis, we use a myriad of feature selection and classification algorithms to identify Bacterial Vaginosis (BV) in women. BV is a very common condition that is signified by changes in vaginal microbiota or microflora. The rest of this thesis is organized as follows. Section 1.2 discusses the motivation and challenges for this research. Chapter 2 features related work in the areas of Bacterial Vaginosis and machine learning. Chapter 3 provides details about machine learning, the feature selection, search method and classification algorithms used for this research and the Weka workbench used to process the data. Chapter 4 describes the dataset, experiment process and the metrics used to analyze the performance of the algorithms. Chapter 5 examines the experiments conducted and the results. Finally, Chapter 6 will present the conclusion and future work.

1.2 Motivation and Problem Statement

There are several diseases which arise because of changes in the microbial communities in the body. Scientists continue to conduct research in a quest to find the catalysts that provoke these changes in the naturally occurring microbiota. The human body can be very sensitive to

change and while the structure of the body is generally the same for everyone; each person has unique qualities and this can include the makeup of microbial communities. Are these microbial differences due to genetics, environment, behavior and/or a combination of the three?

Bacterial Vaginosis (BV) is a disease that fits the above criteria. BV afflicts approximately 29% of women in child bearing age. Typically women are asked a series of questions and are then tested via vaginal swab to confirm diagnosis, but the root causes continue to elude scientists. The challenge becomes finding a common set of attributes that can begin providing answers to the aforementioned question. The additional challenges include determining the optimal methods for diagnosis resulting in accuracy, efficiency and time and cost savings. Do we solely rely on the experience and competence of physicians or should we look to computer aided medical diagnosis? Machine learning has been used in many domains including medical diagnosis, but is it an effective tool for the diagnosis of BV? If so, which feature selection and classification algorithms are the best to use on a particular dataset or is it a "one size fits all" solution?

This research is targeted at finding a common set of attributes or features that are correlated with a BV positive diagnosis and will begin looking at which machine learning feature selection and classification algorithm combinations that will optimize diagnosis.

CHAPTER 2

Related Work

2.1 Bacterial Vaginosis

As highlighted in chapter 1, BV is often characterized by changes in the vaginal microbiota; unfortunately, the causes of those changes are not well understood. Fortunately, it is easily treatable with antibiotics such as metronidazole and clindamycin (Srinivasan et al., 2012). BV is most often diagnosed by testing the vaginal fluid via Gram stain and/or by an assessment based on Amsel's clinical criteria. The Gram stain produces a Nugent Score ranging from 1 – 10. A score of seven or greater yields a positive BV diagnosis. On the other hand, three of the following four Amsel's criteria must be present for a positive diagnosis: 1) presence of a fishy like odor, 2) presence of a white discharge, 3) a vaginal pH of > 4.5 and 4) a minimum of 20% "clue cells" detection (Brotman, 2011). However, Nugent's criterion has become the gold standard for diagnosis (Rangari Amit, Parmjit, & Sharma, 2013). In many instances, a diagnosis is made with Amsel's clinical criteria and confirmed with Gram stain. One of the problems women face is that they may be asymptomatic, however, BV positive (Sujatha et al., 2013). BV can cause unfavorable outcomes for women including an odorous discharge, pelvic inflammatory disease (PID), premature labor and cause them to be more susceptible to contracting HIV and other sexually transmitted diseases (STDs), (Fredricks, Fiedler, Thomas, Oakley, & Marrazzo, 2007). The rate at which BV reoccurs is very high and also not well understood.

Srinivasan et al. (2012) performed deep sequencing of the 16S rRNA gene in an attempt to uncover the variety and make-up of vaginal bacteria in BV positive women. They discovered that there were only two bacteria, Leptotrichia amnionii and Eggerthella sp. that were linked to all four of Amsel's criteria. They also uncovered the fact that there was a greater presence of

Lactobacillus crispatus or Lactobacillus iners in women without BV. Unsupervised machine learning (clustering) was one of the methodologies used to make taxonomic connections with BV. They concluded that vaginal bacteria biota in women with BV is varied and in greater quantities in addition, there was no single bacterium present in 100% of the women.

Beck & Foster (2014) applied genetic programming, random forests, and logistic regression machine learning techniques on two BV datasets from Srinivasan et al. (2012) and Ravel et al. (2011) to hopefully discover BV related microbial relationships. While the associated microbe clusters were different in the two datasets, they did discover that some of the clusters had overlapping microbes. They performed experiments on both the Nugent score and Amsel's criteria. Their experiments resulted in logistic regression and random forest outperforming genetic programming. On the Nugent score, logistic regression and random forest maintained accuracy between 90% and 95%. Amsel's criteria produced slightly lower accuracy. However, none of the three classification algorithms fell below 80% accuracy.

2.2 Machine Learning in Medical Diagnosis

In the world of medicine, machine learning (ML) has been used in the process of simplifying diagnoses and minimizing misdiagnoses. However, it must be noted that this technology is a tool and does not replace the role of the physician; instead, it should be used to aid in the overall diagnostic process and evaluation of patients. Computer scientists' use of ML techniques on medical data is continuing to rise as they look for patterns to assist with diagnoses and enhancement of patient care (Savage, 2012). As we see improvements and the generation of new ML algorithms, we will see a decrease in the time it takes to diagnose and an increase in precision, effectiveness and satisfied patients. ML algorithms have gained a much deserved reputation in research for use in assisting with the diagnoses of numerous diseases (Filippo,

Alberto, Eladia Maria, Petr, & Josef, 2013). We will explore a few current applications of machine learning in the realm of medical diagnosis research. As we will come to see, there is not a "one size fits all" machine learning solution for the vast variety of medical challenges.

We will begin our exploration with heart disease, as it is the primary cause of death worldwide. The World Health Organization (2014) published that almost 17 million lives were lost in 2011 worldwide due to cardiovascular diseases. This amounts to three in ten deaths. While these numbers seem alarming (and they are), they were the same in 2008. With all of the information, research and other resources available, why aren't these numbers decreasing? In 2010 alone, the financial consequence (direct and indirect) in the United States was an estimated cost of $315.4 billion (Go et al., 2013).

The process for diagnosing heart disease can be quite extensive requiring patients to take numerous tests and are many times dependent upon the experience and proficiency of the physician. Unfortunately, some of these tests do not lead to an accurate diagnosis and treatment of the disease. With the use of ML, there is the possibility of increasing accuracy and reducing the features in the prediction of heart disease. Anbarasi, Anupriya, & Iyengar (2010) used a dataset with 909 instances and 13 features. By using Correlation-Based Feature Subset Evaluation (CfsSubsetEval) combined with Genetic Search. They were able to achieve a reduced feature set containing only six key features. They chose three classification algorithms: Naïve Bayes, Decision Tree and Classification via Clustering. Experiments using the three classification algorithms were performed on both the original dataset and the dataset with reduced features to validate the accuracy. Decision Tree (99.2%) ranked first after applying feature selection, the accuracy of Naïve Bayes (96.5%) remained steady on both the original and

reduced datasets and Classification via Clustering (88.3%) produced low results compared to the other two algorithms.

Cancer is the second leading cause of death worldwide with lung cancer as the number one cause of cancer death. The American Cancer Society has predicted that in 2014 there will be 224,201 new cases of lung cancer and predicted to claim 159,260 lives in the Unites States alone (American Cancer Society, 2014). Lung cancer like most cancers, can begin to progressively spread to other organs if it goes untreated, and even then, there's a possibility that the treatment may not work. In order to increase the likelihood of eradicating the cancer and increasing the survival rate, early detection and treatment is quintessential. Unfortunately, some of the current testing methods such as Computed Tomography scan (CT scan), chest radiography and Sputum analysis either require an extensive amount of time, money and/or can only detect the cancer in its advanced stage, thus, lowering the chances of survival (Taher & Sammouda, 2011).

Researchers continue to experiment with machine learning algorithms employing mostly supervised learning as an alternate means for classifying cancer. Hosseinzadeh, KayvanJoo, Ebrahimi, & Goliaei (2013) compared Support Vector Machines (SVM), Naïve Bayes and Artificial Neural Networks (ANN) in their ability to accurately identify and predict the specific type of lung tumor based on a number of factors such as the structure of the tumor. They determined that the SVM algorithm at 88% accuracy was the top performer of the three and that classification and feature selection had great potential in simple applications.

Only second to lung cancer, breast cancer is very invasive and the most primary cause of cancer related death amid women (Salama et al., 2012). ML has often been applied in diagnosing and detecting breast cancer because it has the ability to identify patterns that may be otherwise difficult to detect, as well as learn from previous instances. While there has been much research,

it is still unknown what causes breast cancer. Due to this fact, early detection is imperative in reducing the death rate. Detection should be able to reliably and accurately differentiate between malignant and benign tumors. The customary technique for diagnosis is usually performed by human observation. However, the number of patients is increasing and therefore computers to aid and automate the diagnosis process have been developed in the past several years. The qualitative data is converted to a quantitative feature classification problem which has more objectivity (Osareh & Shadgar, 2010).

Yau & Othman (2007) compare the accuracy, time and error rate of five classification algorithms on breast cancer data that consisted of 699 rows and 9 columns of data. The algorithms compared were Bayes network classifier, radial basis function, decision tree and single conjunctive rule learner. They determined that Bayes network classifier was the best algorithm of those compared based on the accuracy or the percentage of correctly classified instances of 89.71%, model build time of 0.19 seconds and average error at 0.2140. In comparison, Aruna, Rajagopalan, & Nandakishore (2011) evaluated three sets of breast cancer data in the areas of accuracy, precision, specificity and sensitivity. The Wisconsin Diagnostic Breast Cancer Dataset contained 569 instances and 32 features, Wisconsin Breast Cancer Dataset consisted of 683 instances (444 benign, 239 malignant) and Breast Tissue Dataset was comprised of 106 instances and 9 features. Naïve Bayes, SVM Gaussian RBF kernel (SVM-RBF), RBF neural networks, decision trees, J48 and simple classification and regression trees (CART) were applied to evaluate the performance of each algorithm. They determined based on their experiments that SVM-RBF was the top ranked performer in all areas. Both sets of breast cancer machine learning experiments were conducted using the Weka workbench.

Asthma is a common lung disease that affects approximately 235 million people worldwide. That number is inclusive of the 25 million affected with asthma in the U.S. of which, 7 million are children. In similar fashion as some of the aforementioned illnesses, the exact cause of asthma is unknown and may be a combination of environment and genetics. People are afflicted with asthma in varying degrees ranging from mild to acute and can sometimes become fatal if not treated timely. Typical symptoms of asthma are chest tightening, shortness of breath, wheezing and coughing. Asthma diagnoses is usually attained by a physician performing a series of tests that may include: Spirometry (tests the lung function), chest x-ray, EKG and bronchoprovocation among others (National Heart, Lung and Blood Institute, 2014).

Physicians will often use a stethoscope as a non-invasive way to listen for sounds produced by the lungs; this method can be unreliable for many reasons including the physician's hearing ability and the frequency of the sounds. There is a recent study using computerized lung sound analysis that has the potential to help physicians make quicker and more accurate diagnoses. Emanet, Öz, Bayram, & Delen (2014) use an embedded real-time microprocessor system and an inexpensive microphone to transmit the sounds. After the data was retrieved, they applied Random Forest, AdaBoost combined with Random Forest and artificial neural networks (ANN) machine learning algorithms for classification. Random Forest and AdaBoost combined with Random Forest performed well reaching an accuracy of approximately 90%, while ANNs were at about 80%.

The final disease we will explore using ML techniques as a diagnostic tool is Alzheimer's disease (AD). The most common cause of dementia is Alzheimer's disease. It manifests as memory loss and the deterioration of other intellectual capabilities that impedes the normal way of living. While there has been and still is lots of research to find a cure for Alzheimer's, one has

not been discovered as of yet. The best available remedies merely have the ability to retard the dementia symptoms in order to increase the length of experiencing a good quality of life (Alzheimer's Association, 2014).

Changes in the structure and function of the brain have been assessed using diagnostic tools such as a magnetic resonance imaging (MRI), computer-aided diagnosis (CAD) and single-photon emission computed tomography (SPECT). However, Yasuo et al. (2013) found that there was not any current research using classification based on MRI images alone. They applied a support vector machine (SVM) and an artificial neural network (ANN) to a dataset containing four morphological and six functional images on 30 patients (15 with AD and 15 without AD). Unfortunately, their results were much lower than classification based on data retrieved from baseline principal component analysis (PCA) and SPECT images. The experiments yielded 0.660 on morphological and 0.903 on functional images using SVM for classification. They attributed the low rates to the meager dataset and lack of algorithm variety.

2.3 Machine Learning in Other Real World Applications

Machine learning (ML) not only finds its place in the field of medicine, but has also been very beneficial in other applications such as education, science, security, business and so on. Algorithm performance is often highlighted as a ML outcome, and should be, but there are others that should also be taken into consideration such as increase in quality of life, lives saved, interventions implemented and time, effort and money conserved to name a few. These additional outcomes can help connect ML to other real world problems. It’s not enough to simply run an algorithm on dataset, it should include determining the most relevant features, analyzing and interpreting the results and convincing others that this technique is worthwhile for large scale implementation (Wagstaff, 2012).

The field of biometrics has embraced machine learning to assist in the identification and authentication process. There are several modes of biometric identification including fingerprints, iris, signature, voice and face. Shelton et al. (2012) developed the Genetic and Evolutionary Feature Extraction – Machine Learning ($GEFE_{ML}$) algorithm for facial recognition in the area of Genetic & Evolutionary Biometrics (GEB). This algorithm works based on the principles of Darwinism's natural selection. They compared the performance of their $GEFE_{ML}$ with that of the traditional Local Binary Pattern (LBP) feature extraction technique. $GEFE_{ML}$ accuracy was comparable to LBP and reduced processing time by 45% (in terms of computational complexity).

CHAPTER 3

Exploring and Identifying Appropriate Feature Selection and Classification Algorithms

3.1 Machine Learning

Machine learning (ML) was birthed in the 1930s beginning with Ronald A. Fisher and in the 1950s with linear perceptron from Frank Osenblatt. From the 1960s until the late 1980s, ML experienced highs and lows, however, things began looking up with the entrance of neural networks. The 1990s brought about the resurrection of ML with support vector machines (SVM) (Alexander, 2013). Over the past decade, it has been a rapidly developing field and has often been applied successfully to complex and real world challenges.

Machine learning utilizes a variety of artificial intelligence and statistical tools to train on past data in order to create reasonable generalizations, discover patterns, classify previously unseen data or predict new directions (Hosseinzadeh et al., 2013). The primary objective of ML is to minimize classification errors on the training data. It has the ability to deliver precise or nearly perfect predictions (Anu, Agrawal, & Bhattacharya, in-press). ML works extremely well on massive datasets that may go beyond the bounds of human analyzation and interpretation.

The term machine learning is often mistakenly used interchangeably with data mining. While data mining can make use of ML algorithms, such as SimpleKMeans in a clustering application, the emphasis is different. Data mining attempts to search through data to gather information that can be converted into a structure that is comprehensible for further use. In other words, it is a knowledge discovery process. As previously mentioned, ML emphasizes generalization, prediction and representation. There are many machine learning packages on the market that each has a host of algorithms for exploration. Figure 1 displays a few of the open source packages.

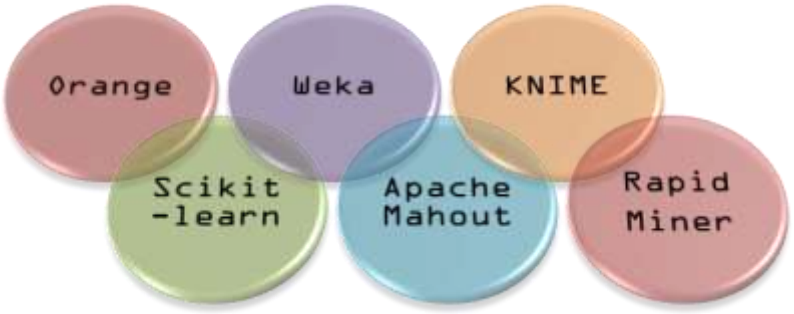

Figure 1. Open source machine learning software packages.

Machine learning algorithms are generally classified into two learning schemes: supervised and unsupervised learning. The type and availability of data and the anticipated outcome will determine which learning scheme will be employed (Dua, 2011). In supervised learning, the output values of the model are defined prior to creating it, such as BV positive or negative. In contrast, the model itself dictates its output for unsupervised learning (Kantardzic, 2003). The most familiar tasks of supervised learning are regression and classification. This research focuses on the task of classification learning, which will be defined later.

3.2 Feature Selection and Classification

3.2.1 Feature selection. Feature selection (FS) is the process of choosing the most significant features and forming a subgroup or subset that will be the most valuable for prediction and analysis. The goal is to discover a subset of features that perform as well (or better) than the original set. It is assumed that any given dataset contains data or features that are not relevant, duplicates and/or noisy data, thus necessitating feature selection (Hall, 1999). There are major benefits of applying this machine learning technique. FS reduces the amount of data that has to be analyzed in turn reducing storage and runtime. This pre-processing step may cost you time in the beginning, but will improve the outcome and efficiency in the end. This is especially true when dealing with enormous amounts of data. In addition, by executing FS we

can anticipate that algorithms will learn more quickly and accuracy will be improved because irrelevant features have been reduced or completely eliminated. Simply stated, feature selection should produce top performance with minimal processing energy.

Feature selection generally falls into one of two categories: minimum subset and feature ranking. Minimum subset algorithms produce exactly what the name suggests; it creates a subset of features with the least amount of relevant features that will yield maximum results. However, there is no distinction between the features in terms of ranking. On the other hand, feature ranking algorithms do not reduce the dataset, but instead it orders the features based on evaluation measures that have been specified (Kantardzic, 2003).

The two primary FS approaches fall within the two categories listed above: filter methods and wrapper methods. Filters create a subset before learning begins that is the most favorable. Based on overall characteristics, an autonomous evaluation is made. Because filters run much faster than wrappers, they may be the preferred method for large and highly dimensional datasets (Witten, Frank, & Hall, 2011). Wrappers assess the subset by "wrapping around" a classification algorithm that will be used for learning. They usually outperform filters in terms of accuracy; however, the computational cost is very high when used on large datasets. Feature selection algorithms are typically coupled with a search method such as genetic search, exhaustive search and best first (Rajarajeswari & Somasundaram, 2012). A given search method will roam through the features in order to locate good subsets.

3.2.2 Classification. Classification, as previously mentioned, is one of two supervised learning techniques. The objective of a classifier algorithm is to accurately group objects into a predefined set of classes. In other words, it predicts the class of each instance (Dua, 2011). This approach is mostly used in artificial intelligence (AI), machine learning and pattern recognition.

Just as with machine learning, classification has been used in a variety of applications such as medical diagnosis, biometrics, cybersecurity, risk analysis, manufacturing, etc. (Anbarasi et al., 2010). There are a range of major classification techniques that include: Neural networks, Bayesian classifiers, meta learners, decision trees, etc. (David, Saeb, & Al Rubeaan, 2013). Choosing the best classifier for a particular problem is extremely important, yet this task has not been given much research attention (Peng, Kou, Ergu, Wu, & Shi, 2012).

3.3 Weka

Weka (Waikato Environment for Knowledge Analysis) was created at the University of Waikato in New Zealand and has a compilation of data preprocessing tools and machine learning algorithms. Weka's interface allows users to easily use the tools and apply algorithms on a variety of datasets by accessing the "Explorer" through its graphical user interface (GUI) pictured in Figure 2. It was written in Java and runs on most operating systems such as Windows, Mac OS and Linux.

Figure 2. Weka GUI.

Figure 3 displays the following features in the Weka Explorer for working with data: Preprocessing, classification, clustering, association rule mining, attribute selection and visualization. Weka imports data files that are in the Attribute-Relation File Format (ARFF), comma-separated values file format (CSV) as well as a few others. Once you load your dataset,

there are many options to choose. These include classification and clustering options such as, cross-validation, percentage split and classes-to-cluster evaluation.

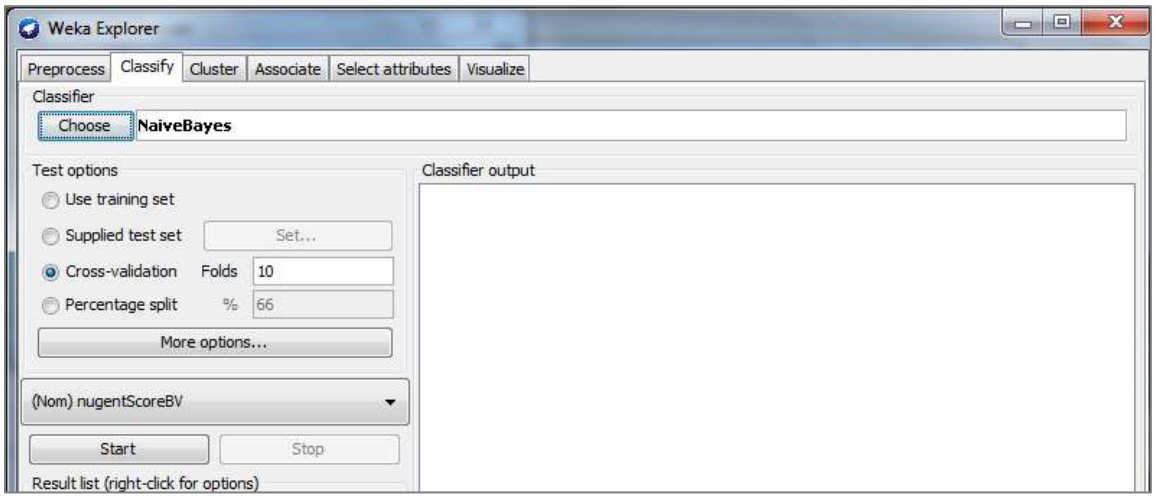

Figure 3. Weka Explorer.

The following section will list and describe all of the feature (attribute) selection, search method and classification algorithms we used for this research.

3.4 Algorithms Used in Weka

3.4.1 Feature selection algorithms.

- CfsSubsetEval: Produces subsets of features that have a low association with each other and greatly interrelated with the class (Witten et al., 2011).
- ClassifierSubsetEval: A wrapper method that uses a classifier to approximate the value of a set of attributes. The attribute subsets evaluated are derived from training or an isolated testing set of data (Witten et al., 2011).
- ConsistencySubsetEval: Feature sets are assessed based on the level of class consistency. It searches for high class consistency with minimal features; however the subset consistency

cannot fall below the consistency of the full feature set (Chue-Poh, Ka-Sing, & Weng-Kin, 2008).

- FilteredSubsetEval: A filter is applied to the training data prior to performing feature selection. An error message will be generated if the filter alters the original features number or ordering (Witten et al., 2011).
- WrapperSubsetEval: Similar to ClassifierSubsetEval, a classifier is used to determine the value of a subset of features. However, cross-validation is used to approximate the precision of the learning scheme for the feature subset (Witten et al., 2011).

3.4.2 Search method algorithms.

- BestFirst: Explores a random subset of features using greedy hill climbing and supplemented with backtracking. Backtracking is controlled by selecting the number of sequential non-improving nodes allowed. An empty set of features may be initially selected for a forward search, a full feature set for a backward search or begin midway and search both ways so that all possible distinct feature additions and deletions at any location can be examined (Sindhu, Geetha, & Kannan, 2012).
- RankSearch: An attribute evaluator is used to rank all of the features. A forward selection search produces a ranked list using the selected subset evaluator. After the list is produced, each subset increments in size and is then evaluated. It begins with the best feature then the best feature plus the next best feature until it produces the best subset which is reported (Witten et al., 2011).
- GeneticSearch: Based on the principles of evolution's survival of the fittest, the genetic search begins with an empty feature set along with rules generated randomly for the initial population. Afterwards, new populations and offspring are formed from the rules of the

current population. Crossover and mutation are administered to create offspring. This process repeats until every rule in final population fulfills the fitness threshold (Anbarasi et al., 2010).

- LinearForwardSelection: Is an extension of BestFirst. The user selects *m* number of features that should not be exceeded in each step. Runtime is reduced because the number of evaluations has been decreased. LinearForwardSelection uses one of two methods; fixed set or fixed width. Both rank the features using a subset evaluator. Fixed set uses only the *m* best features in the succeeding forward selection while fixed width increases *k* in each successive step (Gutlein, Frank, Hall, & Karwath, 2009).
- SubsetSizeForwardSelection: Is an extension of LinearForwardSelection. The search executes *k*-folds cross validation that can be specified by the user. The prime subset-size is then chosen by executing a LinearForwardSelection on every fold. Lastly, the whole data set is used to execute a LinearForwardSelection up to the prime subset-size (Gutlein et al., 2009).
- GreedyStepwise: Executes a greedy backward or forward search through the area of feature subsets. It might start from a random place in the area or may start with all or none of the features. The search ends when adding or deleting any of the residual features causes a decrease in the evaluation. GreedyStepwise also has the ability to yield a ranked list of features by crisscrossing through the area and recording the order of the selected features (Witten et al., 2011).

3.4.3 Classification algorithms.

- Bagging: Uses a random classifier and combines or aggregates copies of that classifier to improve performance. Bagging for classification takes a majority vote for a predicted class by a sequence of classifiers (Breiman, 1996).
- RandomForest: A collection or ensemble of decision trees. It uses the outcomes of the trees that are individually "weak" classifiers to make one strong classifier. This is done by way of each tree voting on the most common class (Breiman, 2001).
- LogitBoost: A boosting algorithm that uses logistic regression. Boosting increases the performance of classification by joining weak classifiers. It handles noisy data very well (Cai, Feng, Lu, & Chou, 2006).
- KStar (K*): A nearest neighbor instance based learner. Instance based means that it compares pre-classified examples to classify an instance. Nearest neighbor finds the instance that is most similar in the training set (Cleary & Trigg, 1995).
- FT: Has the capacity to process both nominal and numeric features, binary and multi-class variables and missing values. The leaves have linear functions and angled splits. . (Witten et al., 2011).
- J48: Written in Java and is derived from the C4.5 Revision 8 algorithm for use in the Weka workbench. This classifier is decision tree based. This means that it is configured like a tree. Tests are performed on one or more features creating non-leaf nodes (i.e. root node) and classification results are represented by leaf nodes. Classification takes place by beginning at the root of the tree, testing the node specific feature and generating a branch for each value. The method is repeated on each of the nodes on the branch and terminates when a leaf node is produced (Wang, Makedon, Ford, & Pearlman, 2005).

- AdaBoostM1: Uses a series of iterations during training to add weak learners thereby creating a strong learner. Each iteration adds a new weak learner to the collection and its weighting vector adjusts to concentrate on misclassified examples in previous cycles (Friedman, Hastie, & Tibshirani, 2000).
- NaïveBayes: A simple probabilistic classifier based on the supposition of class conditional independence of features and that the prediction is not biased by any concealed features (John & Langley, 1995).
- RBFNetwork: Comprised of three layers: input, hidden and output. It is similar to the k-means algorithm in that the expected target value will most likely have similar values of those that are nearby. The name radial basis function derived its name because it uses radius distance (Sherrod, 2014).
- OneR: one feature is tested by a set of rules and creates a one level decision tree (Witten et al., 2011).

CHAPTER 4

Our Approach to Predict the Presence of Disease with Microbiome Community Etiology

4.1 Dataset

In this chapter, we provide our experiment process using the machine learning techniques defined in chapter three. We also define the evaluation metrics used to calculate the accuracy, precision, recall and F-measure in predicting the presence of the disease. The dataset used in our experiment is comprised of 25 women studied over a 10 week period. This data is a subset of a larger dataset of 400 women (Ravel et al., 2011). Dr. James A. Foster and Daniel Beck form the University of Idaho provided us with the de-identified data in a .csv file. The study was arranged so that samples and information were retrieved from the women every day during the 10 week period, however, some women missed days. There were also a few weeks that void of any data in the spreadsheet. There are a total of 1601 instances and 418 features. The BV data consists of three sub-categories of features: time series, clinical and medical data.

4.1.1 Time series data. The time series data documents day and week numbers for features "DIA_DAY" and "DIA_WEEK" respectively (day 1, week 9), day of the week for feature "DAYOFWK" (Monday = 2), day number of the study for feature "TIME" (1, 2…70) and patient id number for feature P_ID (1, 2…25). The same data is repeated with the exception of "TIME" beginning with "P_ID.1" and so on. For our experiments, we used the data without the feature names. The numeric labels for this data range from features 1 – 11. A sample of the time series data with feature names are shown below in Table 1.

Table 1

Time Series Sample

TIME	P_ID	SITE2	DIA_DAY	DIA_WEEK	DAYOFWK
1	1	2	1	1	3

Table 1

Cont.

2	1	2	2	1	4
3	1	2	3	1	5
4	1	2	4	1	6
5	1	2	5	1	7
6	1	2	6	1	1
7	1	2	7	1	2
8	1	2	1	2	2
9	1	2	2	2	4
10	1	2	3	2	5

4.1.2 Clinical data. The clinical data is a combination of results from Amsel's criteria and a questionnaire. The questionnaire included questions such as sexual activity, contraceptive use, tobacco use, etc. The numeric labels for this data range from features 12 – 38. The clinical data feature names are displayed in Table 2.

Table 2

Clinical Data Features

Clinical Data Features				
Vag_Int	Sper_Use	Fem_Powd	Meds	Vag_Itch
Anal_Sex	Lubr_Use	Menstrua	Swabs	Vag_Burn
Oral_Sex	Partner	Tampon	Slide	Vag_Dis
Fing_Pen	Thong	Pad	Ph_Glove	
Sexy_Toy	Douching	Stress	Vag_Odor	
Cond_Use	Fem_Spra	Tob_Use	Vag_Irr	

4.1.3 Medical data. The medical data was obtained from vaginal swabs which were used to perform 454 sequencing of the V12 region of the 16S gene. The numeric labels for this data range from features 39 – 418. Table 3 exhibits the medical data feature names.

Table 3

Medical Data Features

Medical Data Features			
Acholeplasma	Bosea	Helcobacillus	Propioniferax
Achromobacter	Brachybacterium	Helcococcus	Propionimicrobium
Acidaminococcus	Bradyrhizobiaceae.1	Herbaspirillum	Propionivibrio
Acidimicrobiaceae.1	Bradyrhizobium	Hydrogenophaga	Proteobacteria.10
Acidimicrobiaceae.2	Brevibacillus	Hydrogenophilus	Proteobacteria.11
Acidimicrobiales.1	Brevibacterium	Hyphomicrobiaceae.1	Proteobacteria.12
Acidovorax	Brevundimonas	Ignatzschineria	Proteobacteria.14
Acinetobacter	Burkholderia	Incertae_Sedis_XII.1	Proteobacteria.15
Actinobacteria.1	Burkholderiales.1	Incertae_Sedis_XII.2	Proteobacteria.16
Actinobacteria.2	Caenimonas	Janibacter	Proteobacteria.17
Actinobaculum	Caldicellulosiruptor	Janthinobacterium	Proteobacteria.5
Actinomycetales.1	Campylobacter	Jeotgalicoccus	Proteobacteria.6
Actinomycetales.10	Capnocytophaga	Jonquetella	Proteobacteria.7
Actinomycetales.11	Carboxydocella	Kingella	Proteobacteria.8
Actinomycetales.12	Carboxydothermus	Klebsiella	Proteobacteria.9
Actinomycetales.13	Carnobacteriaceae.1	Kocuria	Pseudomonas
Actinomycetales.14	Carnobacteriaceae.2	L.iners	Pseudonocardiaceae.1
Actinomycetales.15	Catenibacterium	L.mucosae	Pseudoxanthomonas
Actinomycetales.16	Chromatiales.1	L.otu1	Psychrobacter
Actinomycetales.17	Chryseobacterium	L.otu6	Ralstonia
Actinomycetales.2	Cloacibacterium	L.reuteri	Raoultella
Actinomycetales.3	Clostridiales.1	Lachnospiraceae.10	Rheinheimera
Actinomycetales.4	Clostridiales.10	Lachnospiraceae.2	Rhizobiales.1
Actinomycetales.5	Clostridiales.11	Lachnospiraceae.3	Rhizobiales.2
Actinomycetales.6	Clostridiales.12	Lachnospiraceae.4	Rhizobium
Actinomycetales.7	Clostridiales.13	Lachnospiraceae.5	Rhodanobacter
Actinomycetales.8	Clostridiales.14	Lachnospiraceae.6	Rhodococcus
Actinomycetales.9	Clostridiales.15	Lachnospiraceae.7	Roseburia
Aeromonadaceae.1	Clostridiales.16	Lachnospiraceae.8	Roseomonas
Aeromonas	Clostridiales.17	Lactobacillales.1	Rothia

Table 3

Cont.

Akkermansia	Clostridiales.18	Lactobacillales.2	Rubrobacter
Alistipes	Clostridiales.19	Lactobacillales.3	Ruminococcaceae.1
Alloscardovia	Clostridiales.2	Lactobacillales.4	Ruminococcaceae.2
Alphaproteobacteria.1	Clostridiales.20	Lactobacillales.5	Ruminococcaceae.3
Alphaproteobacteria.2	Clostridiales.21	Lactococcus	Ruminococcaceae.5
Amaricoccus	Clostridiales.3	Leptotrichia	Ruminococcaceae.7
Anoxybacillus	Clostridiales.4	Leptotrichiaceae.1	Ruminococcaceae.8
Aquabacterium	Clostridiales.6	Leptotrichiaceae.2	Ruminococcus
Aquincola	Clostridiales.7	Leuconostoc	Rummeliibacillus
Archaea.1	Clostridiales.8	Marinobacter	Saprospiraceae.1
Archaea.2	Clostridiales.9	Marinomonas	Sarcina
Archaea.3	Clostridium	Marmoricola	Schlegelella
Archaea.4	Collinsella	Massilia	Sedimentibacter
Archaea.5	Comamonadaceae.1	Megamonas	Selenomonas
Archaea.6	Comamonas	Meiothermus	Shewanella
Archaea.7	Coprobacillus	Mesorhizobium	Silanimonas
Archaea.8	Coprococcus	Methylobacillus	Skermanella
Archaea.9	Cupriavidus	Methylobacterium	Slackia
Arthrobacter	Cytophagaceae.1	Methyloversatilis	Solibacillus
Asaccharobacter	Dechloromonas	Microbacterium	Solobacterium
Aspromonas	Deinococcus	Mitsuokella	Sphingobium
Asticcacaulis	Delftia	Modestobacter	Sphingomonas
Atopobacter	Dermabacter	Mogibacterium	Sphingopyxis
Aurantimonas	Dermacoccus	Moryella	Sporacetigenium
Azonexus	Desulfobacterium	Mucilaginibacter	Sporomusa
Azospira	Devosia	Mycobacterium	Stenotrophomonas
Bacillaceae.1	Diaphorobacter	Mycoplasma	Streptomyces
Bacillaceae.2	Dietzia	Neisseria	Streptophyta
Bacillales.1	Dolosigranulum	Neisseriaceae.1	Subdoligranulum
Bacillariophyta	Dorea	Nesterenkonia	Succinispira
Bacilli.1	Dyella	Nitratireductor	Sutterella
Bacilli.2	Dysgonomonas	Nitrobacter	Symbiobacterium
Bacilli.3	Enhydrobacter	Nocardioides	Syntrophomonas
Bacillus	Enterobacter	Nosocomiicoccus	TM7_genera _incertae_sedis
Bacteria.1	Enterobacteriaceae.1	Novosphingobium	Tepidimonas

Table 3

Cont.

Bacteria.10	Enterobacteriaceae.2	Ochrobactrum	Thermanaeromonas
Bacteria.11	Enterococcaceae.1	Odoribacter	Thermicanus
Bacteria.12	Eremococcus	Oligella	Thermobacillus
Bacteria.13	Erysipelothrix	Oribacterium	Thermolithobacter
Bacteria.14	Erythrobacter	Oscillibacter	Thermomicrobia.1
Bacteria.15	Escherichia.Shigella	Paenibacillus	Thermomonas
Bacteria.16	Eubacterium	Paludibacter	Thermus
Bacteria.17	Exiguobacterium	Pantoea	Treponema
Bacteria.18	Facklamia	Parabacteroides	Trichococcus
Bacteria.19	Fangia	Paracoccus	Turicibacter
Bacteria.3	Fastidiosipila	Parasutterella	Ureaplasma
Bacteria.4	Firmicutes.1	Pasteurella	Varibaculum
Bacteria.5	Firmicutes.2	Patulibacter	Variovorax
Bacteria.6	Flavisolibacter	Pediococcus	Veillonella
Bacteria.7	Flavobacteriaceae.1	Pedobacter	Veillonellaceae.1
Bacteria.8	Flavobacteriaceae.2	Pelagibacter	Veillonellaceae.2
Bacteroidales.1	Flavobacteriaceae.3	Pelomonas	Vibrio
Bacteroidales.2	Flavobacterium	Peptococcus	Vogesella
Bacteroidales.3	Fusobacterium	Petrimonas	Weeksella
Bacteroides	Gallicola	Petrobacter	Weissella
Bacteroidetes.1	Gammaproteobacteria.1	Phascolarctobacterium	Zimmermannella
Bacteroidetes.2	Geobacillus	Phenylobacterium	Zoogloea
Bacteroidetes.4	Geothrix	Planifilum	corGroup1
Bacteroidetes.5	Globicatella	Planococcus	corGroup2
Barnesiella	Gp10	Planomicrobium	corGroup3
Bavariicoccus	Gp6	Plesiomonas	corGroup4
Bifidobacterium	Granulicatella	Prevotellaceae.2	corGroup5
Blastococcus	Gulosibacter	Propionibacteriaceae.1	corGroup6
Blastomonas	Haematobacter	Propionibacteriaceae.2	corGroup7
Blautia	Haemophilus	Propionibacterium	corGroup8

4.2 Experiment Process

For all of our experiments, we used the Weka workbench explorer feature. The experiment process shown in Figure 4 will be outlined in this section in detail. We began by

converting the data into a format that was acceptable for use in Weka. The first column was an anonymous identifier for each woman; therefore, it was deleted because it was not part of the feature set. The Nugent score results were contained in the second column. If the Nugent score was ≥ 7 and ≤ 10, it was given a score of "1" indicating BV positive otherwise it was given a score of "0" indicating BV negative. We converted all numeric values to nominal values: 1's to "yes" and 0's to "no" in the nugentScoreBV column to grant access to a greater number of Weka's algorithms. The data in the nugentScoreBV column was moved to the last column of the dataset as required by Weka for supervised learning.

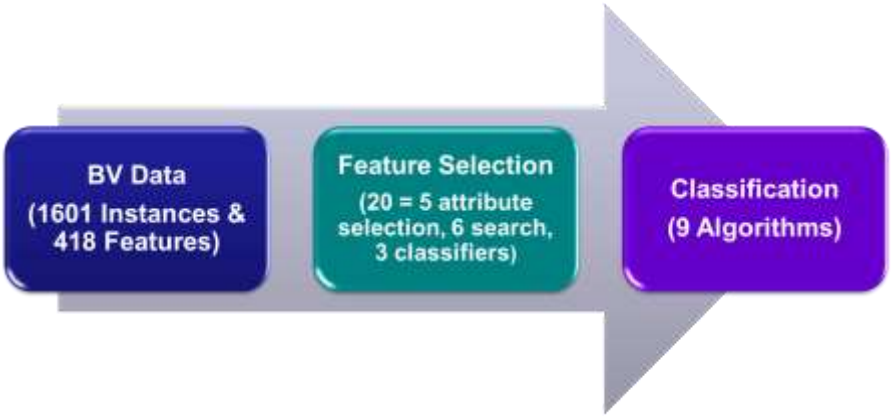

Figure 4. Experiment process.

We used several of Weka's attribute selection algorithms for our feature selection process and classifier algorithms to test the accuracy of diagnosing the presence of Bacterial Vaginosis (BV). Weka has two groups of feature selection algorithms: Attribute subset evaluators and single-attribute evaluators. The single-attribute evaluator algorithms reduce the feature set in theory in that it uses the ranker method and ranks all of the features in descending order. It requires the user to manually set a threshold to discard the lower ranking features or specify the number of features to preserve. We used five of the six attribute subset evaluator algorithms (each defined in Chapter 3) on the full training set because they automated the feature

selection/reduction process (CfsSubsetEval, ClassifierSubsetEval, ConsistencySubsetEval, FilteredSubsetEval and WrapperSubsetEval). Note that the CostSensitiveSubsetEval algorithm could not to be used because it did not produce any output for this dataset.

Weka has two types of feature selectors: Search method and ranking method (11 total methods). We discarded the sole ranking method because it is only used with the single-attribute evaluator algorithms mentioned above. We then eliminated four of the remaining ten algorithms for one of two reasons: Processing time exceeded eight hours or the method failed to produce any output. The discarded search methods were ExhaustiveSearch, RaceSearch, RandomSearch, and ScatterSearchV1. During our feature selection process, we decided to use only two very popular algorithms (NaïveBayes and Bagging) for our WrapperSubsetEval algorithm. The default for ClassifierSubsetEval was the ZeroR algorithm, however, it only yielded one feature. Therefore, we decided to choose a very similar algorithm (OneR) to use for this feature selector.

Weka has a total of 93 classification algorithms, 27 of which were not available for the type of dataset used, leaving 66 for classification. We quickly realized that it would be a formidable task to run experiments using all of the available algorithms. So to begin the classification process, we initially selected six popular algorithms (Bagging, ConjunctiveRule, J48, NaiveBayes, OneR and RandomForest). After administering the experiments, we discovered that OneR and ConjuctiveRule had the same output for all of the feature sets. In addition, they had lower accuracy than most of the other algorithms. Therefore, we chose to eliminate the OneR and ConjuctiveRule algorithms and replace them with RBFNetwork and AdaBoostM1 then added an additional three algorithms (LogitBoost, KStar and FT) to increase the variety for our experiments. In Weka, deterministic algorithms such as RandomForest will yield repeatable

experiments because Weka uses a “seed value” to remove the randomization. In order to randomize the results, a manual manipulation of the seed value will be necessary.

Table 4 features the combinations of the five feature selection, six search methods and three classifier algorithms (used for wrapper methods) assembled to create 20 distinct feature selection sets.

Table 4

Feature Selection Sets

Code	**Attribute Evaluator**	**Search Method**	**Classifier Algorithms**
FS1	CfsSubsetEval	BestFirst	
FS2	CfsSubsetEval	GeneticSearch	
FS3	CfsSubsetEval	RankSearch	
FS4	ClassifierSubsetEval	GeneticSearch	OneR
FS5	ConsistencySubsetEval	BestFirst	
FS6	ConsistencySubsetEval	GeneticSearch	
FS7	ConsistencySubsetEval	LinearForwardSelection	
FS8	ConsistencySubsetEval	RankSearch	
FS9	ConsistencySubsetEval	SubsetSizeForwardSelection	
FS10	FilteredSubsetEval	BestFirst	
FS11	FilteredSubsetEval	GeneticSearch	
FS12	FilteredSubsetEval	RankSearch	
FS13	WrapperSubsetEval	BestFirst	NaiveBayes
FS14	WrapperSubsetEval	BestFirst	Bagging
FS15	WrapperSubsetEval	GeneticSearch	Bagging
FS16	WrapperSubsetEval	GreedyStepwise	NaiveBayes
FS17	WrapperSubsetEval	LinearForwardSelection	NaiveBayes
FS18	WrapperSubsetEval	RankSearch	NaiveBayes
FS19	WrapperSubsetEval	SubsetSizeForwardSelection	NaiveBayes
FS20	WrapperSubsetEval	SubsetSizeForwardSelection	Bagging

The nine classification algorithms used for our experiments are shown in Table 5. The default settings were maintained for all feature selection, search method and classification algorithms except as noted above where search methods were required for the feature selection process and classifiers for the wrapper methods. We saved the feature selection reports produced by Weka in Notepad++. The actual reduced feature sets were saved as *.arff files that would then be imported back into Weka for the classification process.

Table 5

Classification Algorithms

Code	**Algorithm**
A1	Bagging
A2	RBFNetwork
A3	J48
A4	NaiveBayes
A5	AdaBoostM1
A6	RandomForest
A7	LogitBoost
A8	KStar (K*)
A9	FT (Functional Trees)

We used 10-fold cross-validation for testing and training. Holdout methods are commonly used when there is a limited quantity of data. In 10-fold cross-validation, the data is divided into 10 approximately equal parts or folds. The first fold is used for testing and folds two through ten are used for training. Each successive fold is used for testing and the remainders for training until all ten iterations are complete. The error rate is calculated for each of the 10 folds and averaged to produce comprehensive error estimation. According to Witten et al. (2011), theoretical substantiation and large-scale testing have shown that 10-folds seem to produce prime error estimations.

4.2.1 Raw Full Data Experiment Process. We conducted our first set of experiments on the full set of raw data that included all three data subsets: Time series, clinical and medical data. This data was untouched with the exception of making the adjustments required by Weka mentioned in section 4.1. This dataset included the cells and sometimes rows of missing data. We calculated the time taken for each feature selection and classification algorithm to produce output. We then created an elapsed time table. We additionally created a feature set and metrics table that will be presented in chapter 5.

4.2.2 Time Series Removed Experiment Process. We removed the columns containing the first eleven time series features detailed in section 4.1, leaving only the clinical and medical features for experiments. We applied feature selection and classification to this already reduced dataset. We calculated the time taken for each feature selection and classification algorithm to produce output. We then created an elapsed time table. We additionally created a feature set and metrics table that will be featured in chapter 5.

4.2.3 Clinical Experiment Process. We retained only the columns containing the clinical data (features 12 – 38) which includes the questionnaire results and Amsel's clinical criteria as mentioned previously. Feature selection and classification algorithms were applied to both giving us information on time elapsed and metric results. Tables were created from this output.

4.2.4 Medical Experiment Process. We retained only the columns containing the medical data (features 39 – 418) which was derived from the data obtained via the 454 sequencing of the V12 region of the 16S gene. We calculated the time taken for each feature selection and classification algorithm to produce output. We then created an elapsed time table and additionally created feature set and metrics tables.

4.2.5 Clean Full Experiment Process. We decided to address the issue of missing data for this set of experiments. We used a simple yet laborious process of "cleaning" the data. We examined the data in intervals of one week (days 1-7) at a time for each woman over the 10 week period. If data was missing for an entire week, we simply eliminated those rows. In instances where less than seven days of results were shown, we inserted rows to make a complete week. We then calculated the mean (M) for each feature (column of data) within the week that was being examined:

$$M = \frac{\sum x_i}{n}$$

Where M is the mean, $\sum x_i$ is the sum of data for the week being examined and n is the number of cells with data for the week being examined. We then inserted M into all of the cells with missing data in the column for the week being examined. This was repeated until the entire table was void of missing data. We then applied feature selection and classification algorithms giving time elapsed and metric results. We created tables using this output.

4.2.6 Clean Clinical Experiment Process. We retained only the columns containing the now clean clinical data (features 12 – 38) which includes the questionnaire results and Amsel's clinical criteria as mentioned previously. As with all previous experiments, feature selection and classification algorithms were applied to both giving us information on time elapsed and metric results. Tables were created from this output.

4.2.7 Clean Medical Experiment Process. We retained only the columns containing the clean medical data (features 39 – 418) that included the added M data in addition to data obtained via the 454 sequencing of the V12 region of the 16S gene. We calculated the time taken for each feature selection and classification algorithm to produce output. We then created an elapsed time table and additionally created feature set and metrics tables.

4.3 Metrics Defined

In classification where there are solely two classes such as with our data where yes = BV positive and no = BV negative, there are only four possible outcomes shown in Figure 5.

PREDICTED CLASS			
POSITIVE (Yes)	NEGATIVE (No)		
TP	FN	POSITIVE (Yes)	ACTUAL CLASS
FP	TN	NEGATIVE (No)	

Figure 5. Confusion Matrix.

In the framework of our research, the confusion matrix components have the following descriptions:

- True positive (TP) is the number of correctly classified positive cases of BV,
- False negative (FN) is the number of positive cases of BV incorrectly classified as negative,
- False positive (FP) is the number of negative cases of BV incorrectly classified as positive, and
- True negative (TN) is the number of correctly classified negative cases of BV.

The overall accuracy (AC) is the percentage of correctly classified cases of BV. It is calculated using the number of correctly classified instances, TP and TN divided by the total number of classified BV cases:

$$AC = \frac{TP + TN}{TP + FN + FP + TN}$$

The precision (PR) is the percentage of positive predictions retrieved that were actually positive cases of BV. It is the number of true positives divided by the number of all retrieved positive results:

$$PR = \frac{TP}{TP + FP}$$

The recall (RC) is the percentage of positive predictions retrieved from all positive cases of BV. It is the number of true positives divided by the number of all positive cases of BV:

$$RC = \frac{TP}{TP + FN}$$

The F-measure (FM) is the harmonic mean of precision and recall. The harmonic mean is usually used when determining the average of rates:

$$FM = 2 * \frac{PR * RC}{PR + RC}$$

CHAPTER 5

Experiments and Results

In this chapter we present the results from our research that includes a variety of experiments performed on the raw full, time series removed, clinical only, medical only, cleaned full, cleaned clinical only and cleaned medical only datasets.

5.1 Raw Full Dataset

Table 6 displays the results of the feature selection process on the raw full dataset. Based solely on feature reduction, FS19 was the top ranking performer by reducing the features down to a feature set of 6. FS6 was the lowest ranking performer by only reducing the features down to a feature set of 183.

Table 6

Raw Full Feature Set

Raw Full Feature Set					
Set	**Attribute Evaluator**	**Search Method**	**Classifier**	**# of Feat.**	**Selected Feature List**
FS1	Cfs Subset Eval	Best First		15	34, 105, 211, 262, 285, 320, 322, 338, 357, 404, 411, 414, 416, 417, 418
FS2	Cfs Subset Eval	Genetic Search		143	1, 2, 3, 6, 9, 14, 16, 18, 19, 23, 25, 26, 30, 34, 35, 39, 49, 54, 63, 73, 82, 86, 88, 91, 92, 93, 96, 99, 105, 122, 129, 131, 136, 161, 167, 169, 173, 175, 178, 187, 188, 195, 202, 209, 219, 226, 229, 230, 231, 234, 236, 237, 238, 239, 240, 241, 243, 244, 245, 247, 249, 250, 252, 256, 257, 259, 261, 262, 263, 266, 269, 270, 271, 272, 274, 275, 278, 279, 280, 281, 282, 283, 284, 285, 289, 290, 293, 295, 299, 305, 306, 308, 309, 310, 314, 320, 324, 330, 331, 333, 337, 338, 339, 340, 345, 346, 348, 349, 350, 352, 355, 357, 359, 360, 362, 364, 366, 368, 370, 371, 374, 376, 377, 382, 384, 385, 386, 389, 390, 394, 395, 399, 403, 404, 405, 407, 409, 410, 411, 413, 414, 416, 417

Table 6

Cont.

FS3	Cfs SubsetEval	RankSearch		16	34, 53, 105, 112, 211, 256, 262, 285, 320, 327, 338, 404, 411, 414, 416, 417
FS4	Classifier SubsetEval	Genetic Search	OneR	99	3, 4, 6, 7, 8, 9, 11, 26, 28, 35, 38, 39, 41, 48, 61, 62, 63, 64, 68, 72, 76, 77, 80, 84, 92, 107, 111, 127, 129, 130, 131, 137, 140, 165, 171, 172, 181, 184, 186, 192, 199, 205, 206, 214, 216, 217, 220, 227, 228, 230, 232, 235, 243, 244, 250, 251, 258, 260, 262, 263, 266, 275, 285, 286, 289, 290, 291, 298, 301, 302, 303, 304, 306, 313, 315, 323, 336, 337, 338, 341, 343, 346, 348, 349, 351, 352, 354, 359, 361, 363, 373, 378, 381, 382, 397, 400, 409, 411, 415
FS5	Consistency SubsetEval	BestFirst		9	2, 28, 33, 285, 411, 412, 414, 416, 418
FS6	Consistency SubsetEval	Genetic Search		183	6, 8, 10, 11, 12, 14, 19, 22, 24, 27, 28, 29, 30, 33, 36, 38, 39, 40, 42, 49, 51, 55, 57, 60, 64, 66, 71, 76, 79, 81, 86, 89, 92, 94, 95, 99, 100, 101, 105, 106, 108, 110, 112, 113, 115, 116, 117, 118, 122, 123, 125, 126, 127, 128, 131, 136, 146, 147, 149, 150, 151, 152, 159, 162, 163, 165, 167, 168, 169, 170, 177, 178, 179, 182, 184, 186, 187, 188, 195, 196, 197, 198, 201, 205, 211, 217, 218, 219, 221, 225, 226, 227, 231, 234, 236, 239, 240, 243, 244, 245, 248, 249, 250, 251, 258, 260, 261, 263, 266, 267, 273, 275, 278, 279, 281, 282, 284, 285, 288, 289, 293, 294, 295, 301, 303, 309, 315, 316, 321, 324, 325, 327, 329, 330, 332, 334, 339, 340, 341, 342, 343, 345, 348, 349, 350, 352, 354, 358, 360, 366, 367, 368, 369, 371, 374, 376, 379, 380, 382, 384, 386, 387, 388, 390, 392, 393, 394, 395, 397, 398, 404, 405, 406, 407, 409, 410, 411, 412, 413, 414, 416, 417, 418

Table 6

Cont.

FS7	Consistency SubsetEval	Linear Forward Selection		13	7, 34, 49, 105, 122, 285, 320, 338, 357, 411, 414, 416, 418
FS8	Consistency SubsetEval	RankSearch		91	2, 7, 16, 20, 23, 25, 29, 30, 33, 34, 40, 43, 49, 53, 63, 64, 87, 88, 91, 104, 105, 112, 120, 122, 123, 124, 126, 129, 135, 139, 145, 148, 157, 159, 162, 163, 164, 171, 172, 191, 207, 209, 211, 218, 226, 229, 230, 245, 246, 248, 249, 250, 251, 254, 256, 259, 262, 275, 279, 282, 285, 296, 297, 298, 311, 318, 320, 321, 322, 324, 325, 327, 334, 337, 338, 339, 355, 357, 359, 375, 394, 400, 404, 409, 411, 412, 413, 414, 416, 417, 418
FS9	Consistency SubsetEval	Subset SizeForwardSelection		11	7, 34, 49, 105, 122, 285, 357, 411, 414, 416, 418
FS10	Filtered SubsetEval	BestFirst		11	105, 285, 320, 338, 357, 404, 411, 414, 416, 417, 418
FS11	FilteredSubsetEval	Genetic Search		143	1, 2, 3, 6, 9, 14, 16, 18, 19, 23, 25, 26, 30, 34, 35, 39, 49, 54, 63, 73, 82, 86, 88, 91, 92, 93, 96, 99, 105, 122, 129, 131, 136, 161, 167, 169, 173, 175, 178, 187, 188, 195, 202, 209, 219, 226, 229, 230, 231, 234, 236, 237, 238, 239, 240, 241, 243, 244, 245, 247, 249, 250, 252, 256, 257, 259, 261, 262, 263, 266, 269, 270, 271, 272, 274, 275, 278, 279, 280, 281, 282, 283, 284, 285, 289, 290, 293, 295, 299, 305, 306, 308, 309, 310, 314, 320, 324, 330, 331, 333, 337, 338, 339, 340, 345, 346, 348, 349, 350, 352, 355, 357, 359, 360, 362, 364, 366, 368, 370, 371, 374, 376, 377, 382, 384, 385, 386, 389, 390, 394, 395, 399, 403, 404, 405, 407, 409, 410, 411, 413, 414, 416, 417

Table 6

Cont.

FS12	Filtered SubsetEval	RankSearch		8	256, 320, 327, 338, 404, 411, 414, 417
FS13	Wrapper SubsetEval	BestFirst	Naïve Bayes	14	2, 7, 29, 33, 34, 104, 130, 228, 241, 285, 295, 411, 416, 417
FS14	Wrapper SubsetEval	BestFirst	Bagging	7	1, 2, 3, 12, 201, 250, 411
FS15	Wrapper SubsetEval	Genetic Search	Bagging	156	1, 2, 3, 9, 12, 13, 18, 20, 21, 24, 25, 28, 29, 30, 31, 32, 33, 34, 35, 36, 40, 42, 43, 45, 46, 47, 49, 52, 53, 54, 56, 59, 62, 63, 64, 65, 67, 69, 72, 77, 78, 85, 89, 92, 93, 98, 100, 101, 103, 109, 111, 114, 123, 127, 129, 130, 131, 132, 136, 137, 142, 151, 153, 154, 163, 165, 171, 172, 178, 181, 184, 186, 187, 192, 199, 201, 205, 206, 210, 214, 216, 217, 218, 220, 221, 222, 229, 232, 235, 243, 244, 246, 250, 251, 258, 260, 262, 263, 266, 267, 270, 271, 273, 275, 280, 285, 286, 288, 289, 290, 291, 293, 294, 298, 301, 302, 306, 310, 313, 323, 332, 338, 339, 342, 343, 347, 348, 349, 351, 352, 354, 359, 361, 363, 364, 371, 373, 378, 379, 383, 386, 389, 393, 394, 397, 402, 403, 404, 405, 409, 410, 412, 413, 416, 417, 418
FS16	Wrapper SubsetEval	Greedy Stepwise	Naïve Bayes	14	2, 7, 29, 33, 34, 104, 130, 228, 241, 285, 295, 411, 416, 417
FS17	Wrapper SubsetEval	Linear Forward Selection	Naïve Bayes	9	34, 105, 285, 320, 334, 337, 338, 411, 417
FS18	Wrapper SubsetEval	RankSearch	Naïve Bayes	10	91, 104, 256, 320, 327, 338, 404, 411, 414, 417
FS19	Wrapper SubsetEval	Subset SizeForwardSelection	Naïve Bayes	6	105, 112, 285, 320, 411, 417
FS20	Wrapper SubsetEval	Subset SizeForwardSelection	Bagging	7	2, 91, 285, 296, 334, 404, 411

Table 7

Raw Full: Precision, Recall and F-Measure Rates

Features		**Classifiers**								
Name	**Metric**	A1	A2	A3	A4	A5	A6	A7	A8	A9
FS1	AC	96.3148%	95.6902%	96.2523%	96.0025%	96.1274%	96.7520%	96.5646%	96.6271%	97.0643%
	PR	0.896	0.836	0.886	0.869	0.885	0.886	0.908	0.941	0.914
	RC	0.863	0.9	0.871	0.876	0.863	0.908	0.867	0.835	0.896
	FM	0.879	0.867	0.879	0.872	0.874	0.897	0.887	0.885	0.905
FS2	AC	96.8145%	92.7545%	95.94%	94.6908%	96.4397%	97.0019%	97.0019%	ERROR	96.6896%
	PR	0.909	0.753	0.862	0.795	0.887	0.91	0.904		0.88
	RC	0.884	0.795	0.88	0.888	0.884	0.896	0.904		0.912
	FM	0.896	0.773	0.871	0.839	0.885	0.903	0.904		0.895
FS3	AC	96.065%	95.94%	95.7527%	96.5646%	96.0025%	97.1268%	96.5646%	96.065%	96.8145%
	PR	0.888	0.862	0.882	0.901	0.884	0.898	0.901	0.927	0.913
	RC	0.855	0.88	0.839	0.876	0.855	0.92	0.876	0.811	0.88
	FM	0.871	0.871	0.86	0.888	0.869	0.909	0.888	0.865	0.896
FS4	AC	96.752%	91.5678%	96.3773%	93.8164%	95.8776%	95.94%	96.2523%	ERROR	95.7527%
	PR	0.909	0.75	0.89	0.793	0.865	0.862	0.883		0.847
	RC	0.88	0.687	0.876	0.815	0.871	0.88	0.876		0.888
	FM	0.894	0.717	0.883	0.804	0.868	0.871	0.879		0.867
FS5	AC	96.5022%	95.8776%	96.1899%	95.3154%	95.8776%	96.9394%	96.877%	96.752%	97.1268%
	PR	0.891	0.893	0.892	0.882	0.867	0.894	0.909	0.923	0.914
	RC	0.884	0.835	0.859	0.807	0.867	0.912	0.888	0.863	0.9
	FM	0.887	0.863	0.875	0.843	0.867	0.903	0.898	0.892	0.907
FS6	AC	96.5022%	91.5678%	96.1274%	95.0031%	95.8151%	96.6271%	96.9394%	ERROR	95.8151%
	PR	0.897	0.721	0.87	0.812	0.87	0.888	0.92		0.853
	RC	0.876	0.747	0.884	0.884	0.859	0.896	0.88		0.884
	FM	0.886	0.734	0.876	0.846	0.865	0.892	0.899		0.868

Table 7

Cont.

FS7	AC	96.1274%	96.3148%	96.1899%	96.3148%	96.1274%	95.7527%	96.752%	97.1268%	96.8145%
	PR	0.885	0.871	0.873	0.88	0.888	0.866	0.902	0.947	0.899
	RC	0.863	0.896	0.884	0.884	0.859	0.859	0.888	0.863	0.896
	FM	0.874	0.883	0.878	0.882	0.873	0.863	0.895	0.903	0.897
FS8	AC	96.752%	93.8788%	95.8151%	94.0037%	95.94%	97.1893%	96.877%	95.3779%	96.0025%
	PR	0.912	0.796	0.855	0.772	0.877	0.902	0.916	0.903	0.86
	RC	0.876	0.815	0.88	0.871	0.859	0.92	0.88	0.787	0.888
	FM	0.893	0.806	0.867	0.819	0.868	0.911	0.898	0.841	0.874
FS9	AC	96.1899%	95.6277%	96.1899%	95.5653%	96.1274%	97.1893%	96.752%	97.1268%	95.94%
	PR	0.885	0.851	0.87	0.859	0.888	0.898	0.902	0.943	0.883
	RC	0.867	0.871	0.888	0.855	0.859	0.924	0.888	0.867	0.851
	FM	0.876	0.861	0.879	0.857	0.873	0.911	0.895	0.904	0.867
FS10	AC	96.1274%	96.3148%	95.5653%	96.5022%	96.0025%	97.1893%	96.4397%	96.5646%	96.9394%
	PR	0.888	0.886	0.871	0.897	0.878	0.905	0.897	0.937	0.91
	RC	0.859	0.876	0.839	0.876	0.863	0.916	0.871	0.835	0.892
	FM	0.873	0.881	0.855	0.886	0.87	0.91	0.884	0.883	0.901
FS11	AC	96.8145%	92.7545%	95.94%	94.6908%	96.4397%	97.0019%	97.0019%	ERROR	96.6896%
	PR	0.909	0.753	0.862	0.795	0.887	0.91	0.904		0.88
	RC	0.884	0.795	0.88	0.888	0.884	0.896	0.904		0.912
	FM	0.896	0.773	0.871	0.839	0.885	0.903	0.904		0.895
FS12	AC	96.065%	95.94%	95.6902%	96.1899%	95.8776%	95.253%	95.7527%	95.5653%	96.4397%
	PR	0.904	0.883	0.913	0.912	0.865	0.847	0.866	0.964	0.94
	RC	0.835	0.851	0.799	0.835	0.871	0.847	0.859	0.743	0.823
	FM	0.868	0.867	0.852	0.872	0.868	0.847	0.863	0.839	0.878
FS13	AC	96.9394%	96.5022%	96.3773%	97.3766%	96.1899%	97.1268%	96.9394%	97.3142%	97.3766%
	PR	0.913	0.881	0.875	0.916	0.885	0.901	0.907	0.936	0.912
	RC	0.888	0.896	0.896	0.916	0.867	0.916	0.896	0.888	0.92
	FM	0.9	0.888	0.885	0.916	0.876	0.908	0.901	0.911	0.916

Table 7

Cont.

FS14	AC	96.6896%	95.6902%	96.5022%	95.4403%	95.8776%	96.9394%	96.0025%	96.877%	95.7527%
	PR	0.908	0.866	0.9	0.846	0.865	0.907	0.871	0.934	0.876
	RC	0.876	0.855	0.871	0.863	0.871	0.896	0.871	0.859	0.847
	FM	0.892	0.861	0.886	0.855	0.868	0.901	0.871	0.895	0.861
FS15	AC	96.752%	91.8176%	95.3779%	93.6914%	96.6896%	97.5016%	96.5022%	ERROR	95.8151%
	PR	0.896	0.746	0.849	0.803	0.938	0.927	0.921		0.855
	RC	0.896	0.719	0.855	0.787	0.843	0.912	0.847		0.88
	FM	0.896	0.732	0.852	0.795	0.888	0.919	0.883		0.867
FS16	AC	96.9394%	96.5022%	96.3773%	97.3766%	96.1899%	97.1268%	96.9394%	97.3142%	97.3766%
	PR	0.913	0.881	0.875	0.916	0.885	0.901	0.907	0.936	0.912
	RC	0.888	0.896	0.896	0.916	0.867	0.916	0.896	0.888	0.92
	FM	0.9	0.888	0.885	0.916	0.876	0.908	0.901	0.911	0.916
FS17	AC	96.4397%	96.1274%	96.3773%	96.9394%	95.8776%	95.6902%	95.8151%	94.6908%	96.5022%
	PR	0.925	0.858	0.896	0.917	0.865	0.844	0.873	0.941	0.897
	RC	0.839	0.9	0.867	0.884	0.871	0.888	0.855	0.703	0.876
	FM	0.88	0.878	0.882	0.9	0.868	0.865	0.864	0.805	0.886
FS18	AC	96.0650%	95.94%	95.6902%	96.3148%	95.8776%	95.5028%	95.7527%	95.4403%	96.4397%
	PR	0.904	0.883	0.913	0.924	0.865	0.858	0.866	0.963	0.94
	RC	0.835	0.851	0.799	0.831	0.871	0.851	0.859	0.735	0.823
	FM	0.868	0.867	0.852	0.875	0.868	0.855	0.863	0.834	0.878
FS19	AC	96.2523%	96.3148%	95.6277%	96.9394%	95.94%	95.7527%	96.1274%	94.6284%	96.3148%
	PR	0.913	0.871	0.891	0.917	0.865	0.847	0.879	0.95	0.906
	RC	0.839	0.896	0.819	0.884	0.876	0.888	0.871	0.691	0.851
	FM	0.874	0.883	0.854	0.9	0.87	0.867	0.875	0.8	0.878
FS20	AC	96.6896%	95.8776%	96.5022%	95.0656%	95.8776%	96.2523%	96.065%	96.1274%	96.5022%
	PR	0.912	0.907	0.9	0.866	0.865	0.874	0.872	0.935	0.907
	RC	0.871	0.819	0.871	0.807	0.871	0.888	0.876	0.807	0.863
	FM	0.891	0.861	0.886	0.836	0.868	0.88	0.874	0.866	0.885

Table 7 exhibits the results of the classification process on the raw full dataset. We calculated the accuracy (AC), precision (PR), recall (RC) and F-measure (FM). FS15 A6 (97.5016%) ranked highest in accuracy, FS12 A8 (0.964) in precision FS9 A6 (0.924) for recall and FS15 A6 for F-Measure (0.919). Based solely on AC and FM, FS15 A6 is the top ranking performer.

The raw full time elapsed table featured in Table 8 shows that based on the sum of the time elapsed for feature selection and classification, FS9 (0:00:21) had the best performance and FS15(1:17:02) had the poorest performance.

Table 8

Raw Full Time Elapsed

Features		**Classifiers (Time Elapsed)**								
Name	**Time**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	0:00:01	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:15	0:00:03
FS2	0:00:10	0:00:06	0:00:04	0:00:04	0:00:02	0:00:09	0:00:04	0:00:20	0:01:33	0:00:12
FS3	0:00:11	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:02	0:00:12	0:00:02
FS4	0:00:11	0:00:03	0:00:02	0:00:03	0:00:01	0:00:07	0:00:04	0:00:14	0:00:49	0:00:12
FS5	0:00:05	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:12	0:00:01
FS6	0:00:02	0:00:07	0:00:04	0:00:05	0:00:01	0:00:13	0:00:04	0:00:24	0:01:43	0:00:16
FS7	0:00:02	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:00	0:00:02	0:00:13	0:00:02
FS8	0:00:03	0:00:03	0:00:02	0:00:03	0:00:01	0:00:06	0:00:02	0:00:13	0:00:47	0:00:08
FS9	0:00:02	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:13	0:00:02
FS10	0:00:01	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:02	0:00:02	0:00:13	0:00:01
FS11	0:00:10	0:00:05	0:00:04	0:00:04	0:00:01	0:00:10	0:00:03	0:00:19	0:01:31	0:00:11
FS12	0:00:11	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:07	0:00:01
FS13	0:17:07	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:11	0:00:02
FS14	0:29:54	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:06	0:00:02
FS15	1:14:25	0:00:06	0:00:04	0:00:06	0:00:01	0:00:11	0:00:04	0:00:21	0:01:31	0:00:13
FS16	0:11:37	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:11	0:00:02
FS17	0:00:54	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:02	0:00:01	0:00:05	0:00:01
FS18	0:17:05	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:02	0:00:02	0:00:07	0:00:01
FS19	0:00:24	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:05	0:00:02
FS20	0:02:25	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:05	0:00:02

When considering runtime, reduction in features and recall in Figure 6 below, we have determined that FS16 A9 is the better algorithm to use for this dataset.

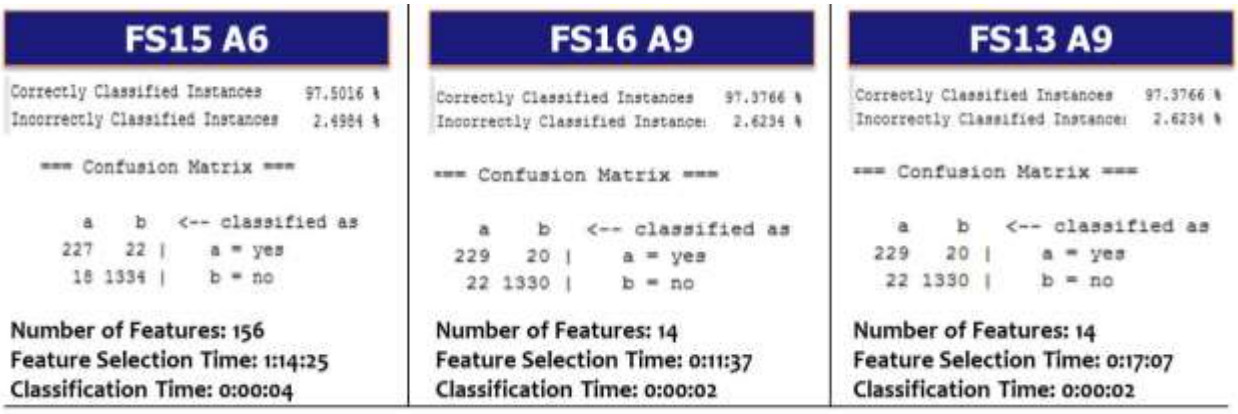

Figure 6. Top Three Raw Full Set.

5.2 Time Series Removed Dataset

Table 9 displays the results of the feature selection process on time series dataset. Based solely on feature reduction, FS14, FS19 and FS20 were the top ranking performers by reducing the features down to a feature set of 6. FS15 was the lowest ranking performer by only reducing the features down to a feature set of 219.

Table 9

Time Series Removed Feature Set

Time Series Removed Dataset					
Set	**Attribute Evaluator**	**Search Method**	**CL**	**# of Feat.**	**Selected Feature List**
FS1	Cfs Subset Eval	BestFirst		15	34, 105, 211, 262, 285, 320, 322, 338, 357, 404, 411, 414, 416, 417, 418

Table 9

Cont.

FS2	Cfs Subset Eval	Genetic Search		141	14, 16, 18, 20, 21, 26, 28, 29, 34, 35, 37, 38, 42, 45, 49, 51, 52, 55, 56, 60, 74, 75, 79, 81, 83, 87, 91, 104, 109, 112, 114, 115, 117, 121, 122, 125, 126, 127, 129, 132, 138, 141, 145, 148, 157, 160, 161, 162, 163, 167, 172, 179, 183, 188, 198, 199, 200, 204, 210, 214, 221, 222, 230, 233, 235, 238, 247, 253, 256, 257, 259, 262, 268, 269, 270, 275, 276, 278, 279, 281, 282, 283, 284, 285, 287, 288, 294, 298, 299, 300, 301, 307, 308, 309, 310, 312, 314, 318, 319, 320, 321, 324, 325, 327, 328, 329, 334, 339, 340, 342, 345, 349, 353, 355, 357, 358, 359, 362, 363, 364, 365, 371, 373, 383, 395, 397, 401, 402, 404, 405, 406, 407, 408, 409, 411, 413, 414, 415, 416, 417, 418
FS3	Cfs SubsetEval	RankSearch		16	34, 53, 105, 112, 211, 256, 262, 285, 320, 327, 338, 404, 411, 414, 416, 417
FS4	Classifier SubsetEval	Genetic Search	OneR	116	13, 15, 17, 29, 33, 34, 38, 39, 43, 44, 49, 51, 52, 57, 61, 66, 68, 69, 81, 88, 91, 96, 99, 101, 102, 104, 108, 112, 113, 121, 124, 125, 134, 135, 138, 141, 155, 157, 161, 164, 174, 175, 182, 183, 184, 188, 191, 196, 202, 206, 211, 212, 214, 216, 217, 218, 220, 221, 222, 233, 234, 238, 242, 243, 247, 251, 253, 254, 262, 264, 270, 274, 279, 282, 284, 286, 287, 290, 291, 292, 297, 299, 301, 305, 309, 310, 311, 314, 315, 318, 320, 324, 325, 329, 336, 339, 342, 343, 346, 349, 358, 363, 365, 368, 371, 373, 387, 393, 397, 400, 402, 404, 405, 411, 416, 418
FS5	Consistency SubsetEval	BestFirst		9	28, 33, 285, 411, 413, 414, 416, 417, 418

Table 9

Cont.

FS6	Consistency SubsetEval	Genetic Search		182	12, 14, 16, 18, 20, 23, 25, 26, 27, 28, 29, 30, 31, 33, 34, 36, 37, 45, 49, 51, 52, 53, 55, 56, 61, 62, 69, 71, 72, 76, 78, 80, 83, 85, 86, 89, 96, 98, 101, 105, 107, 108, 110, 111, 114, 115, 117, 120, 121, 123, 125, 126, 127, 131, 132, 134, 137, 139, 144, 145, 146, 149, 151, 153, 154, 156, 157, 158, 160, 162, 164, 166, 168, 170, 173, 174, 175, 176, 183, 186, 188, 193, 195, 204, 207, 209, 210, 212, 213, 216, 218, 221, 223, 226, 228, 231, 232, 235, 236, 238, 240, 241, 247, 248, 250, 253, 255, 256, 258, 259, 261, 263, 265, 267, 270, 271, 272, 274, 276, 277, 279, 282, 285, 286, 288, 292, 293, 299, 301, 304, 305, 307, 312, 315, 318, 321, 324, 325, 328, 329, 332, 333, 336, 338, 341, 342, 343, 344, 347, 349, 351, 352, 355, 357, 358, 359, 360, 361, 362, 363, 364, 365, 368, 369, 370, 371, 374, 381, 384, 388, 392, 394, 402, 404, 405, 407, 408, 409, 411, 413, 416, 418
FS7	Consistency SubsetEval	Linear Forward Selection		13	34, 49, 105, 122, 285, 320, 338, 357, 411, 414, 416, 417, 418
FS8	Consistency SubsetEval	RankSearch		89	16, 20, 23, 25, 29, 30, 33, 34, 40, 43, 49, 53, 63, 64, 87, 88, 91, 104, 105, 112, 120, 122, 123, 124, 126, 129, 135, 139, 145, 148, 157, 159, 162, 163, 164, 171, 172, 191, 207, 209, 211, 218, 226, 229, 230, 245, 246, 248, 249, 250, 251, 254, 256, 259, 262, 275, 279, 282, 285, 296, 297, 298, 311, 318, 320, 321, 322, 324, 325, 327, 334, 337, 338, 339, 355, 357, 359, 375, 394, 400, 404, 409, 411, 412, 413, 414, 416, 417, 418

Table 9

Cont.

FS9	Consistency SubsetEval	Subset SizeForward Selection		11	34, 49, 105, 122, 285, 357, 411, 414, 416, 417, 418
FS10	Filtered SubsetEval	BestFirst		11	105, 285, 320, 338, 357, 404, 411, 414, 416, 417, 418
FS11	FilteredS ubsetEval	Genetic Search		141	14, 16, 18, 20, 21, 26, 28, 29, 34, 35, 37, 38, 42, 45, 49, 51, 52, 55, 56, 60, 74, 75, 79, 81, 83, 87, 91, 104, 109, 112, 114, 115, 117, 121, 122, 125, 126, 127, 129, 132, 138, 141, 145, 148, 157, 160, 161, 162, 163, 167, 172, 179, 183, 188, 198, 199, 200, 204, 210, 214, 221, 222, 230, 233, 235, 238, 247, 253, 256, 257, 259, 262, 268, 269, 270, 275, 276, 278, 279, 281, 282, 283, 284, 285, 287, 288, 294, 298, 299, 300, 301, 307, 308, 309, 310, 312, 314, 318, 319, 320, 321, 324, 325, 327, 328, 329, 334, 339, 340, 342, 345, 349, 353, 355, 357, 358, 359, 362, 363, 364, 365, 371, 373, 383, 395, 397, 401, 402, 404, 405, 406, 407, 408, 409, 411, 413, 414, 415, 416, 417, 418
FS12	Filtered SubsetEval	RankSearch		8	256, 320, 327, 338, 404, 411, 414, 417
FS13	Wrapper SubsetEval	BestFirst	Naïve Bayes	12	34, 41, 130, 131, 167, 228, 241, 285, 295, 411, 416, 417
FS14	Wrapper SubsetEval	BestFirst	Bagging	6	126, 137, 139, 402, 411, 413

Table 9

Cont.

FS15	Wrapper SubsetEval	Genetic Search	Bagging	219	12, 14, 15, 16, 18, 19, 20, 23, 25, 29, 30, 31, 32, 33, 34, 36, 37, 39, 40, 42, 44, 45, 46, 50, 53, 55, 56, 58, 59, 60, 62, 63, 65, 66, 68, 71, 72, 75, 76, 80, 81, 83, 88, 89, 96, 98, 99, 103, 106, 107, 108, 109, 111, 113, 114, 115, 116, 117, 119, 120, 122, 124, 125, 128, 130, 131, 132, 133, 139, 140, 145, 147, 148, 154, 155, 156, 157, 159, 160, 165, 167, 168, 170, 171, 172, 174, 176, 177, 178, 180, 183, 185, 186, 187, 188, 190, 191, 194, 195, 197, 199, 201, 202, 203, 204, 206, 207, 208, 209, 210, 214, 217, 220, 221, 222, 223, 224, 225, 226, 228, 230, 231, 232, 233, 234, 235, 236, 237, 240, 242, 243, 244, 248, 250, 251, 253, 254, 255, 257, 258, 262, 263, 265, 266, 268, 270, 275, 277, 278, 279, 283, 284, 285, 286, 293, 294, 296, 298, 300, 302, 304, 306, 309, 310, 312, 316, 321, 322, 325, 327, 328, 329, 330, 332, 333, 334, 336, 338, 339, 340, 344, 347, 348, 350, 353, 354, 355, 357, 358, 360, 361, 362, 363, 364, 365, 367, 369, 371, 374, 376, 384, 386, 390, 391, 392, 394, 396, 398, 401, 402, 404, 407, 408, 410, 411, 413, 415, 416, 417
FS16	Wrapper SubsetEval	Greedy Stepwise	Naïve Bayes	12	34, 41, 130, 131, 167, 228, 241, 285, 295, 411, 416, 417
FS17	Wrapper SubsetEval	Linear Forward Selection	Naïve Bayes	9	34, 105, 285, 320, 334, 337, 338, 411, 417
FS18	Wrapper SubsetEval	Rank Search	Naïve Bayes	10	91, 104, 256, 320, 327, 338, 404, 411, 414, 417
FS19	Wrapper SubsetEval	Subset SizeForwardSelection	Naïve Bayes	6	105, 112, 285, 320, 411, 417
FS20	Wrapper SubsetEval	Subset SizeForwardSelection	Bagging	6	285, 357, 409, 411, 416, 417

Table 10

Time Series Removed: Precision, Recall and F-Measure Rates

Features		**Classifiers**								
Name	**Metric**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	AC	96.3148%	95.6902%	96.2523%	96.0025%	96.1274%	96.7520%	96.5646%	96.6271%	97.0643%
	PR	0.896	0.836	0.886	0.869	0.885	0.886	0.908	0.941	0.914
	RC	0.863	0.9	0.871	0.876	0.863	0.908	0.867	0.835	0.896
	FM	0.879	0.867	0.879	0.872	0.874	0.897	0.887	0.885	0.905
FS2	AC	97.0643%	92.5671%	96.065%	94.8157%	96.0025%	97.1268%	96.6271%	ERROR	96.5022%
	PR	0.924	0.769	0.875	0.824	0.878	0.904	0.905		0.875
	RC	0.884	0.747	0.871	0.847	0.863	0.912	0.876		0.904
	FM	0.903	0.758	0.873	0.836	0.87	0.908	0.89		0.889
FS3	AC	96.065%	95.94%	95.7527%	96.5646%	96.0025%	97.1268%	96.5646%	96.065%	96.8145%
	PR	0.888	0.862	0.882	0.901	0.884	0.898	0.901	0.927	0.913
	RC	0.855	0.88	0.839	0.876	0.855	0.92	0.876	0.811	0.88
	FM	0.871	0.871	0.86	0.888	0.869	0.909	0.888	0.865	0.896
FS4	AC	97.0019%	92.2548%	96.5022%	94.3161%	96.0025%	96.752%	96.877%	ERROR	96.2523%
	PR	0.914	0.78	0.9	0.816	0.874	0.886	0.92		0.851
	RC	0.892	0.699	0.871	0.819	0.867	0.908	0.876		0.92
	FM	0.902	0.737	0.886	0.818	0.871	0.897	0.897		0.884
FS5	AC	96.6896%	96.1274%	96.3148%	96.3773%	96.065%	97.1268%	96.8145%	96.6271%	95.94%
	PR	0.905	0.888	0.896	0.893	0.878	0.889	0.919	0.908	0.88
	RC	0.88	0.859	0.863	0.871	0.867	0.932	0.871	0.871	0.855
	FM	0.892	0.873	0.879	0.882	0.873	0.91	0.895	0.889	0.868
FS6	AC	97.1268%	85.8214%	96.1899%	94.5659%	96.0025%	96.6271%	96.9394%	ERROR	95.6277%
	PR	0.911	0.615	0.879	0.793	0.874	0.888	0.917		0.871
	RC	0.904	0.237	0.876	0.88	0.867	0.896	0.884		0.843
	FM	0.907	0.342	0.877	0.834	0.871	0.892	0.9		0.857

Table 10

Cont.

FS7	AC	96.4397%	96.0025%	96.1899%	96.4397%	96.1274%	96.9394%	96.5646%	96.9394%	96.3148%
	PR	0.897	0.849	0.882	0.884	0.885	0.894	0.908	0.946	0.889
	RC	0.871	0.904	0.871	0.888	0.863	0.912	0.867	0.851	0.871
	FM	0.884	0.875	0.877	0.886	0.874	0.903	0.887	0.896	0.88
FS8	AC	96.5022%	93.8788%	95.6902%	94.0037%	95.94%	97.1893%	96.9394%	95.0031%	96.0025%
	PR	0.9	0.794	0.854	0.772	0.877	0.908	0.92	0.904	0.866
	RC	0.871	0.819	0.871	0.871	0.859	0.912	0.88	0.759	0.88
	FM	0.886	0.806	0.863	0.819	0.868	0.91	0.899	0.825	0.873
FS9	AC	96.4397%	95.6902%	95.94%	96.065%	96.1274%	97.1268%	96.5646%	96.8145%	96.1899%
	PR	0.897	0.833	0.874	0.866	0.885	0.901	0.908	0.942	0.87
	RC	0.871	0.904	0.863	0.884	0.863	0.916	0.867	0.847	0.888
	FM	0.884	0.867	0.869	0.875	0.874	0.908	0.887	0.892	0.879
FS10	AC	96.1274%	96.3148%	95.5653%	96.5022%	96.0025%	97.1893%	96.4397%	96.5646%	96.9394%
	PR	0.888	0.886	0.871	0.897	0.878	0.905	0.897	0.937	0.91
	RC	0.859	0.876	0.839	0.876	0.863	0.916	0.871	0.835	0.892
	FM	0.873	0.881	0.855	0.886	0.87	0.91	0.884	0.883	0.901
FS11	AC	97.0643%	92.5671%	96.065%	94.8157%	96.0025%	97.1268%	96.6271%	ERROR	96.5022%
	PR	0.924	0.769	0.875	0.824	0.878	0.904	0.905		0.875
	RC	0.884	0.747	0.871	0.847	0.863	0.912	0.876		0.904
	FM	0.903	0.758	0.873	0.836	0.87	0.908	0.89		0.889
FS12	AC	96.065%	95.94%	95.6902%	96.1899%	95.8776%	95.253%	95.7527%	95.5653%	96.4397%
	PR	0.904	0.883	0.913	0.912	0.865	0.847	0.866	0.964	0.94
	RC	0.835	0.851	0.799	0.835	0.871	0.847	0.859	0.743	0.823
	FM	0.868	0.867	0.852	0.872	0.868	0.847	0.863	0.839	0.878
FS13	AC	96.4397%	96.1274%	96.1274%	97.1893%	95.94%	97.1268%	96.6896%	96.3773%	96.8145%
	PR	0.9	0.858	0.858	0.895	0.88	0.904	0.905	0.948	0.906
	RC	0.867	0.9	0.9	0.928	0.855	0.912	0.88	0.811	0.888
	FM	0.883	0.878	0.878	0.911	0.868	0.908	0.892	0.874	0.897

Table 10

Cont.

FS14	AC	96.6271%	93.8164%	96.4397%	94.8157%	95.8776%	96.5646%	96.4397%	93.5665%	96.1899%
	PR	0.911	0.841	0.907	0.84	0.865	0.891	0.897	0.94	0.92
	RC	0.867	0.743	0.859	0.823	0.871	0.888	0.871	0.627	0.827
	FM	0.889	0.789	0.882	0.832	0.868	0.889	0.884	0.752	0.871
FS15	AC	97.0019%	85.3217%	96.1274%	93.7539%	96.065%	97.1268%	96.8145%	ERROR	95.3779%
	PR	0.91	0.576	0.867	0.773	0.878	0.908	0.913		0.83
	RC	0.896	0.213	0.888	0.847	0.867	0.908	0.88		0.884
	FM	0.903	0.311	0.877	0.808	0.873	0.908	0.896		0.856
FS16	AC	96.4397%	96.1274%	96.1274%	97.1893%	95.94%	97.1268%	96.6896%	96.3773%	96.8145%
	PR	0.9	0.858	0.858	0.895	0.88	0.904	0.905	0.948	0.906
	RC	0.867	0.9	0.9	0.928	0.855	0.912	0.88	0.811	0.888
	FM	0.883	0.878	0.878	0.911	0.868	0.908	0.892	0.874	0.897
FS17	AC	96.4397%	96.1274%	96.3773%	96.9394%	95.8776%	95.6902%	95.8151%	94.6908%	96.5022%
	PR	0.925	0.858	0.896	0.917	0.865	0.844	0.873	0.941	0.897
	RC	0.839	0.9	0.867	0.884	0.871	0.888	0.855	0.703	0.876
	FM	0.88	0.878	0.882	0.9	0.868	0.865	0.864	0.805	0.886
FS18	AC	96.0650%	95.94%	95.6902%	96.3148%	95.8776%	95.5028%	95.7527%	95.4403%	96.4397%
	PR	0.904	0.883	0.913	0.924	0.865	0.858	0.866	0.963	0.94
	RC	0.835	0.851	0.799	0.831	0.871	0.851	0.859	0.735	0.823
	FM	0.868	0.867	0.852	0.875	0.868	0.855	0.863	0.834	0.878
FS19	AC	96.2523%	96.3148%	95.6277%	96.9394%	95.94%	95.7527%	96.1274%	94.6284%	96.3148%
	PR	0.913	0.871	0.891	0.917	0.865	0.847	0.879	0.95	0.906
	RC	0.839	0.896	0.819	0.884	0.876	0.888	0.871	0.691	0.851
	FM	0.874	0.883	0.854	0.9	0.87	0.867	0.875	0.8	0.878
FS20	AC	96.3773%	95.8151%	95.94%	95.8151%	95.8151%	96.5022%	96.4397%	95.8151%	96.4397%
	PR	0.893	0.87	0.88	0.87	0.873	0.878	0.9	0.938	0.921
	RC	0.871	0.859	0.855	0.859	0.855	0.9	0.867	0.783	0.843
	FM	0.882	0.865	0.868	0.865	0.864	0.889	0.883	0.853	0.881

Table 10 exhibits the results of the classification process on the time series removed dataset. We calculated the accuracy (AC), precision (PR), recall (RC) and F-measure (FM). FS10 A6, FS13 A4 and FS16 A4 (97.1893%) ranked highest in accuracy, FS12 A8 (0.964) in precision FS5 A6 (0.932) for recall and FS13 A4 and FS16 A4 for F-Measure (0.911). Based solely on AC and FM, FS13 A4 and FS16 A4 are the top ranking performers.

The raw full time elapsed table featured in Table 11 shows that based on the sum of the time elapsed for feature selection and classification, FS9, FS10 and FS12 (0:00:23) had the best performance and FS15 (1:18:57) had the poorest performance.

Table 11

Time Series Removed Time Elapsed

FEATURES		**CLASSIFIERS (TIME ELAPSED)**								
Name	**Time**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	0:00:07	0:00:00	0:00:01	0:00:00	0:00:00	0:00:02	0:00:01	0:00:03	0:00:15	0:00:02
FS2	0:00:11	0:00:04	0:00:03	0:00:07	0:00:02	0:00:00	0:00:05	0:00:21	0:01:39	0:00:16
FS3	0:00:10	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:03	0:00:13	0:00:02
FS4	0:00:43	0:00:04	0:00:03	0:00:04	0:00:01	0:00:09	0:00:04	0:00:17	0:01:13	0:00:10
FS5	0:00:05	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:16	0:00:01
FS6	0:00:03	0:00:06	0:00:04	0:00:05	0:00:02	0:00:14	0:00:05	0:00:27	0:01:44	0:00:18
FS7	0:00:02	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:02	0:00:02	0:00:14	0:00:02
FS8	0:00:03	0:00:03	0:00:03	0:00:03	0:00:01	0:00:06	0:00:03	0:00:13	0:00:48	0:00:08
FS9	0:00:02	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:13	0:00:01
FS10	0:00:02	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:02	0:00:02	0:00:14	0:00:01
FS11	0:00:10	0:00:09	0:00:03	0:00:07	0:00:02	0:00:12	0:00:06	0:00:23	0:01:39	0:00:24
FS12	0:00:10	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:08	0:00:02
FS13	0:12:17	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:10	0:00:01
FS14	0:24:05	0:00:01	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:08	0:00:01
FS15	1:15:16	0:00:07	0:00:05	0:00:09	0:00:02	0:00:17	0:00:06	0:00:31	0:01:59	0:00:25
FS16	0:07:48	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:12	0:00:01
FS17	0:01:02	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:02	0:00:01	0:00:06	0:00:02
FS18	0:12:49	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:08	0:00:01
FS19	0:00:24	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:05	0:00:01
FS20	0:01:57	0:00:00	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:08	0:00:01

When considering runtime, reduction in features and recall in Figure 7 below, we have determined that FS16 A4 is the better algorithm to use for this dataset.

Figure 7. Top Three Time Series Removed.

5.3 Clinical Dataset

Table 12 displays the results of the feature selection process on the clinical dataset. Based solely on feature reduction, FS13, FS16, FS17 and FS19 were the top ranking performers by reducing the features down to a feature set of 2. FS15 was the lowest ranking performer by only reducing the features down to a feature set of 19.

Table 12

Clinical Feature Set

Clinical Feature Set					
Set	**Attribute Evaluator**	**Search Method**	**Classifier**	**# of Feat.**	**Selected Feature List**
FS1	Cfs Subset Eval	BestFirst		6	20, 21, 25, 29, 33, 34
FS2	Cfs Subset Eval	Genetic Search		6	20, 21, 25, 29, 33, 34
FS3	Cfs SubsetEval	RankSearch		7	20, 21, 23, 25, 29, 33, 34
FS4	Classifier SubsetEval	Genetic Search	OneR	5	13, 14, 24, 27, 34
FS5	Consistency SubsetEval	BestFirst		10	20, 21, 23, 25, 28, 29, 30, 33, 34, 38

Table 12

Cont.

FS6	Consistency SubsetEval	Genetic Search		13	18, 20, 21, 23, 25, 28, 29, 30, 31, 33, 34, 37, 38
FS7	Consistency SubsetEval	Linear Forward Selection		10	20, 21, 23, 25, 28, 29, 30, 33, 34, 38
FS8	Consistency SubsetEval	RankSearch		11	16, 20, 21, 23, 25, 28, 29, 30, 33, 34, 38
FS9	Consistency SubsetEval	Subset SizeForward Selection		10	20, 21, 23, 25, 28, 29, 30, 33, 34, 38
FS10	Filtered SubsetEval	BestFirst		3	29, 33, 34
FS11	FilteredS ubsetEval	Genetic Search		3	29, 33, 34
FS12	Filtered SubsetEval	RankSearch		5	23, 25, 29, 33, 34
FS13	Wrapper SubsetEval	BestFirst	Naïve Bayes	2	20, 34
FS14	Wrapper SubsetEval	BestFirst	Bagging	10	12, 13, 20, 25, 29, 30, 32, 33, 34, 37
FS15	Wrapper SubsetEval	Genetic Search	Bagging	19	12, 13, 16, 19, 20, 23, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38
FS16	Wrapper SubsetEval	Greedy Stepwise	Naïve Bayes	2	20, 34
FS17	Wrapper SubsetEval	Linear Forward Selection	Naïve Bayes	2	20, 34
FS18	Wrapper SubsetEval	RankSearch	Naïve Bayes	6	16, 23, 25, 29, 33, 34
FS19	Wrapper SubsetEval	Subset SizeForward Selection	Naïve Bayes	2	20, 34
FS20	Wrapper SubsetEval	Subset SizeForward Selection	Bagging	10	12, 13, 20, 25, 29, 30, 32, 33, 34, 37

Table 13

Clinical: Precision, Recall and F-Measure Rates

Features		**Classifiers**								
Name	**Metric**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	AC	88.0075%	86.1961%	87.6327%	84.5721%	86.3835%	88.1949%	87.0081%	88.3823%	88.1949%
	PR	0.699	0.646	0.701	0.506	0.746	0.727	0.652	0.899	0.727
	RC	0.402	0.249	0.357	0.325	0.189	0.386	0.353	0.285	0.386
	FM	0.51	0.359	0.473	0.396	0.301	0.504	0.458	0.433	0.504
FS2	AC	88.0075%	86.1961%	87.6327%	84.5721%	86.3835%	88.1949%	87.0081%	88.3823%	88.1949%
	PR	0.699	0.646	0.701	0.506	0.746	0.727	0.652	0.899	0.727
	RC	0.402	0.249	0.357	0.325	0.189	0.386	0.353	0.285	0.386
	FM	0.51	0.359	0.473	0.396	0.301	0.504	0.458	0.433	0.504
FS3	AC	88.0075%	86.8207%	87.5703%	84.6971%	86.3835%	88.4447%	86.8207%	88.3198%	87.8826%
	PR	0.697	0.707	0.695	0.512	0.746	0.746	0.664	0.897	0.72
	RC	0.406	0.261	0.357	0.333	0.189	0.39	0.309	0.281	0.361
	FM	0.513	0.381	0.472	0.404	0.301	0.512	0.422	0.428	0.481
FS4	AC	86.0712%	86.0712%	85.8838%	86.1337%	86.1337%	86.0087%	86.1337%	84.4472%	86.0087%
	PR	0.75	0.75	0.745	0.737	0.737	0.766	0.737	0	0.727
	RC	0.157	0.157	0.141	0.169	0.169	0.145	0.169	0	0.161
	FM	0.259	0.259	0.236	0.275	0.275	0.243	0.275	0	0.263
FS5	AC	88.9444%	86.5084%	87.7577%	82.6359%	86.1337%	88.8819%	87.4453%	88.6321%	87.5703%
	PR	0.737	0.657	0.669	0.435	0.574	0.698	0.64	0.838	0.667
	RC	0.45	0.277	0.422	0.39	0.422	0.502	0.442	0.333	0.402
	FM	0.559	0.39	0.517	0.411	0.486	0.584	0.523	0.477	0.501
FS6	AC	88.9444%	86.321%	88.1324%	83.0106%	86.1337%	88.6946%	87.4453%	88.9444%	87.8201%
	PR	0.747	0.653	0.681	0.447	0.574	0.683	0.64	0.867	0.669
	RC	0.438	0.257	0.446	0.386	0.422	0.51	0.442	0.341	0.43
	FM	0.552	0.369	0.539	0.414	0.486	0.584	0.523	0.49	0.523

Table 13

Cont.

FS7	AC	88.9444%	86.5084%	87.7577%	82.6359%	86.1337%	88.8819%	87.4453%	88.6321%	87.5703%
	PR	0.737	0.657	0.669	0.435	0.574	0.698	0.64	0.838	0.667
	RC	0.45	0.277	0.422	0.39	0.422	0.502	0.442	0.333	0.402
	FM	0.559	0.39	0.517	0.411	0.486	0.584	0.523	0.477	0.501
FS8	AC	88.8819%	86.446%	87.6952%	82.6983%	86.1337%	88.6946%	87.4453%	88.5696%	87.5703%
	PR	0.735	0.7	0.667	0.437	0.574	0.687	0.64	0.83	0.667
	RC	0.446	0.225	0.418	0.39	0.422	0.502	0.442	0.333	0.402
	FM	0.555	0.34	0.514	0.412	0.486	0.58	0.523	0.476	0.501
FS9	AC	88.9444%	86.5084%	87.7577%	82.6359%	86.1337%	88.8819%	87.4453%	88.6321%	87.5703%
	PR	0.737	0.657	0.669	0.435	0.574	0.698	0.64	0.838	0.667
	RC	0.45	0.277	0.422	0.39	0.422	0.502	0.442	0.333	0.402
	FM	0.559	0.39	0.517	0.411	0.486	0.584	0.523	0.477	0.501
FS10	AC	86.9457%	86.0712%	86.5084%	86.0712%	86.1961%	87.258%	86.8832%	86.321%	86.9457%
	PR	0.778	0.724	0.732	0.623	0.769	0.836	0.76	0.826	0.763
	RC	0.225	0.169	0.209	0.265	0.161	0.225	0.229	0.153	0.233
	FM	0.349	0.274	0.325	0.372	0.266	0.354	0.352	0.258	0.357
FS11	AC	86.9457%	86.0712%	86.5084%	86.0712%	86.1961%	87.258%	86.8832%	86.321%	86.9457%
	PR	0.778	0.724	0.732	0.623	0.769	0.836	0.76	0.826	0.763
	RC	0.225	0.169	0.209	0.265	0.161	0.225	0.229	0.153	0.233
	FM	0.349	0.274	0.325	0.372	0.266	0.354	0.352	0.258	0.357
FS12	AC	87.6327%	86.3835%	87.3204%	86.0712%	86.321%	87.5703%	87.3204%	86.6958%	87.5703%
	PR	0.758	0.763	0.788	0.616	0.768	0.731	0.74	0.821	0.75
	RC	0.301	0.181	0.253	0.277	0.173	0.317	0.285	0.185	0.301
	FM	0.431	0.292	0.383	0.382	0.282	0.443	0.412	0.302	0.43
FS13	AC	86.321%	86.0712%	86.5084%	86.5084%	86.1337%	86.5084%	86.5084%	86.321%	86.321%
	PR	0.778	0.741	0.824	0.824	0.737	0.824	0.824	0.969	0.778
	RC	0.169	0.161	0.169	0.169	0.169	0.169	0.169	0.124	0.169
	FM	0.277	0.264	0.28	0.28	0.275	0.28	0.28	0.221	0.277

Table 13

Cont.

FS14	AC	88.4447%	86.5084%	88.1949%	82.5734%	84.6346%	89.1318%	87.945%	88.757%	89.0693%
	PR	0.703	0.677	0.679	0.426	0.507	0.714	0.657	0.879	0.734
	RC	0.446	0.253	0.458	0.345	0.45	0.502	0.47	0.321	0.466
	FM	0.545	0.368	0.547	0.381	0.477	0.59	0.548	0.471	0.57
FS15	AC	89.6315%	85.0718%	88.6946%	83.3854%	86.0712%	88.2573%	87.0706%	88.757%	89.0069%
	PR	0.771	0.586	0.736	0.455	0.676	0.669	0.714	0.863	0.735
	RC	0.474	0.137	0.426	0.349	0.201	0.486	0.281	0.329	0.458
	FM	0.587	0.221	0.539	0.395	0.31	0.563	0.403	0.477	0.564
FS16	AC	86.321%	86.0712%	86.5084%	86.5084%	86.1337%	86.5084%	86.5084%	86.321%	86.321%
	PR	0.778	0.741	0.824	0.824	0.737	0.824	0.824	0.969	0.778
	RC	0.169	0.161	0.169	0.169	0.169	0.169	0.169	0.124	0.169
	FM	0.277	0.264	0.28	0.28	0.275	0.28	0.28	0.221	0.277
FS17	AC	86.321%	86.0712%	86.5084%	86.5084%	86.1337%	86.5084%	86.5084%	86.321%	86.321%
	PR	0.778	0.741	0.824	0.824	0.737	0.824	0.824	0.969	0.778
	RC	0.169	0.161	0.169	0.169	0.169	0.169	0.169	0.124	0.169
	FM	0.277	0.264	0.28	0.28	0.275	0.28	0.28	0.221	0.277
FS18	AC	87.5078%	86.6334%	87.4453%	86.0712%	86.321%	87.3204%	87.133%	86.6334%	87.3204%
	PR	0.753	0.716	0.786	0.616	0.768	0.721	0.736	0.807	0.725
	RC	0.293	0.233	0.265	0.277	0.173	0.301	0.269	0.185	0.297
	FM	0.422	0.352	0.396	0.382	0.282	0.425	0.394	0.301	0.422
FS19	AC	86.321%	86.0712%	86.5084%	86.5084%	86.1337%	86.5084%	86.5084%	86.321%	86.321%
	PR	0.778	0.741	0.824	0.824	0.737	0.824	0.824	0.969	0.778
	RC	0.169	0.161	0.169	0.169	0.169	0.169	0.169	0.124	0.169
	FM	0.277	0.264	0.28	0.28	0.275	0.28	0.28	0.221	0.277
FS20	AC	88.4447%	86.5084%	88.1949%	82.5734%	84.6346%	89.1318%	87.945%	88.757%	89.0693%
	PR	0.703	0.677	0.679	0.426	0.507	0.714	0.657	0.879	0.734
	RC	0.446	0.253	0.458	0.345	0.45	0.502	0.47	0.321	0.466
	FM	0.545	0.368	0.547	0.381	0.477	0.59	0.548	0.471	0.57

Table 13 exhibits the results of the classification process on the clinical dataset. We calculated the accuracy (AC), precision (PR), recall (RC) and F-measure (FM). FS15 A1 (89.6315%) ranked highest in accuracy, FS13 A8, FS16 A8, FS17 A8 and FS19 A8 (0.969) in precision FS6 A6 (0.51) for recall and FS14 A6 and FS20 A6 for F-Measure (0.59).

The clinical time elapsed table featured in Table 14 shows that based on the sum of the time elapsed for feature selection and classification, FS11 and FS16 (0:00:04) had the best performance and FS15 (0:07:37) had the poorest performance.

Table 14

Clinical Time Elapsed

Features		**Classifiers (Time Elapsed)**								
Name	**Time**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	0:00:00	0:00:01	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:03	0:00:02
FS2	0:00:00	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:02	0:00:02
FS3	0:00:00	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:02	0:00:03	0:00:02
FS4	0:00:01	0:00:01	0:00:00	0:00:00	0:00:00	0:00:00	0:00:02	0:00:01	0:00:01	0:00:01
FS5	0:00:00	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:02	0:00:01	0:00:04	0:00:03
FS6	0:00:00	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:02	0:00:01	0:00:05	0:00:03
FS7	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:02	0:00:01	0:00:04	0:00:03
FS8	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:02	0:00:02	0:00:04	0:00:03
FS9	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:00	0:00:01	0:00:02	0:00:04	0:00:03
FS10	0:00:00	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:01
FS11	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:01
FS12	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:00	0:00:02	0:00:01	0:00:02	0:00:02
FS13	0:00:05	0:00:01	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00
FS14	0:02:30	0:00:01	0:00:00	0:00:00	0:00:00	0:00:00	0:00:02	0:00:02	0:00:04	0:00:03
FS15	0:07:19	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:03	0:00:03	0:00:06	0:00:04
FS16	0:00:02	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:00
FS17	0:00:05	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:00	0:00:00
FS18	0:00:06	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:01
FS19	0:00:02	0:00:00	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:00	0:00:01	0:00:00
FS20	0:01:34	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:02	0:00:01	0:00:04	0:00:03

When considering runtime, reduction in features and recall in Figure 8 below, we have determined that FS20 A6 is the better algorithm to use for this dataset.

Figure 8. Top Three Clinical.

5.4 Medical Dataset

Table 15 displays the results of the feature selection process on the medical dataset. Based solely on feature reduction, FS20 was the top ranking performer by reducing the features down to a feature set of 7. FS6 was the lowest ranking performer by only reducing the features down to a feature set of 203.

Table 15

Medical Feature Set

Clinical Feature Set					
Set	**Attribute Evaluator**	**Search Method**	**Classifier**	**# of Feat.**	**Selected Feature List**
FS1	Cfs Subset Eval	BestFirst		18	49, 104, 105, 112, 123, 129, 245, 248, 256, 262, 282, 285, 320, 327, 338, 355, 357, 404

Table 15

Cont.

FS2	Cfs Subset Eval	Genetic Search		109	40, 42, 49, 57, 63, 71, 75, 91, 104, 105, 106, 107, 112, 120, 122, 126, 127, 129, 131, 137, 139, 140, 144, 149, 156, 157, 159, 163, 173, 175, 177, 178, 183, 186, 192, 193, 199, 201, 209, 219, 222, 226, 228, 231, 233, 235, 237, 245, 248, 252, 254, 255, 256, 259, 260, 262, 263, 268, 269, 274, 275, 276, 282, 283, 289, 291, 293, 297, 299, 300, 306, 307, 308, 310, 314, 317, 319, 320, 321, 323, 327, 331, 332, 333, 334, 336, 337, 338, 340, 352, 355, 356, 357, 360, 362, 363, 375, 376, 381, 387, 389, 390, 392, 397, 399, 401, 404, 409, 410
FS3	Cfs SubsetEval	RankSearch		20	49, 63, 91, 104, 105, 112, 123, 129, 245, 248, 256, 262, 285, 320, 327, 338, 355, 357, 359, 404
FS4	Classifier SubsetEval	Genetic Search	OneR	106	40, 41, 48, 49, 52, 54, 56, 57, 61, 63, 67, 70, 71, 75, 88, 89, 93, 95, 106, 107, 108, 109, 111, 113, 114, 115, 126, 127, 129, 131, 134, 137, 141, 148, 151, 152, 156, 157, 171, 172, 173, 183, 190, 193, 194, 204, 207, 208, 211, 212, 215, 216, 221, 224, 227, 229, 232, 239, 241, 250, 251, 252, 253, 255, 259, 263, 264, 265, 268, 269, 276, 284, 285, 290, 294, 296, 298, 299, 300, 308, 309, 312, 337, 341, 342, 344, 346, 349, 350, 359, 363, 366, 371, 376, 379, 380, 386, 390, 391, 392, 395, 399, 403, 404, 406, 408
FS5	Consistency SubsetEval	BestFirst		28	49, 64, 88, 91, 104, 105, 122, 123, 129, 139, 145, 157, 159, 218, 230, 245, 250, 262, 282, 285, 311, 320, 327, 338, 339, 357, 399, 404

Table 15

Cont.

FS6	Consistency SubsetEval	Genetic Search		203	39, 43, 47, 48, 49, 52, 53, 56, 57, 58, 61, 64, 66, 68, 70, 74, 76, 80, 81, 83, 85, 87, 88, 90, 91, 95, 97, 98, 100 ,101, 104, 105, 110, 111, 115, 116, 117, 118, 120, 121, 123, 127, 128, 132, 133, 135, 137, 138, 139, 140, 141, 146, 147, 148, 149, 150, 152, 153, 154, 155, 157, 158, 159, 160, 161, 162, 163, 167, 168, 170, 171, 172, 173, 174, 175, 176, 178, 180, 183, 184, 185, 186, 187, 188, 190, 194, 197, 198, 202, 203, 205, 208, 209, 210, 211, 212, 217, 218, 219, 220, 222, 223, 225, 226, 228, 234, 235, 237, 238, 239, 240, 241, 242, 243, 244, 245, 248, 249, 250, 252, 253, 254, 255, 258, 259, 261, 262, 267, 269, 270, 272, 273, 274, 275, 277, 278, 279, 281, 282, 283, 285, 287, 288, 289, 291, 293, 294, 296, 297, 303, 307, 310 ,311, 312, 314, 315, 317, 319, 320, 322, 324, 326 ,327, 329, 330, 331, 334, 336, 337, 338, 339, 343, 344, 347, 350, 351, 353, 354, 356, 357, 364, 366, 368, 370, 371, 373, 374, 375, 379, 385, 387, 393, 394, 395, 396, 397, 398, 399, 401, 402, 404, 405, 406
FS7	Consistency SubsetEval	Linear Forward Selection		23	49, 91, 104, 105, 112, 120, 122, 129, 139, 157, 256, 259, 262, 282, 285, 320, 325, 327, 338, 355, 357, 404, 409
FS8	Consistency SubsetEval	Rank Search		78	40, 43, 49, 53, 63, 64, 87, 88, 91, 104, 105, 112, 120, 122, 123, 124, 126, 129, 135, 139, 145, 148, 157, 159, 162, 163, 164, 171, 172, 175, 191, 207, 209, 211, 218, 222, 226, 229, 230, 245, 246, 247, 248, 249, 250, 251, 254, 256, 259, 262, 275, 279, 282, 285, 296 297, 298, 311, 318, 320, 321, 322, 324, 325, 327, 334, 337, 338, 339, 355, 357, 359, 375, 394, 399, 400, 404, 409
FS9	Consistency SubsetEval	Subset Size Forward Selection		20	91, 104, 105, 112, 120, 129, 139, 157, 256, 259, 262, 282, 285, 320, 325, 327, 338, 355, 357, 404
FS10	Filtered SubsetEval	BestFirst		14	49, 104, 105, 112, 129, 256, 262, 282, 285, 320, 327, 338, 357, 404

Table 15

Cont.

FS11	Filtered SubsetEval	Genetic Search		109	40, 42, 49, 57, 63, 71, 75, 91, 104, 105, 106, 107, 112, 120, 122, 126, 127, 129, 131, 137, 139, 140, 144, 149, 156, 157, 159, 163, 173, 175, 177, 178, 183, 186, 192, 193, 199, 201, 209, 219, 222, 226, 228, 231, 233, 235, 237, 245, 248, 252, 254, 255, 256, 259, 260, 262, 263, 268, 269, 274, 275, 276, 282, 283, 289, 291, 293, 297, 299, 300, 306, 307, 308, 310, 314, 317, 319, 320, 321, 323, 327, 331, 332, 333, 334, 336, 337, 338, 340, 352, 355, 356, 357, 360, 362 363, 375, 376, 381, 387, 389, 390, 392, 397, 399, 401, 404, 409, 410
FS12	Filtered SubsetEval	Rank Search		13	63, 91, 104, 112, 129, 256, 285, 320, 327, 338, 357, 359, 404
FS13	Wrapper SubsetEval	BestFirst	Naïve Bayes	16	41, 56, 63, 104, 105, 123, 145, 153, 222, 248, 249, 256, 285, 320, 338, 404
FS14	Wrapper SubsetEval	BestFirst	Bagging	14	43, 49, 50, 54, 61, 256, 261, 285, 293, 298, 320, 327, 338, 404
FS15	Wrapper SubsetEval	Genetic Search	Bagging	192	39, 40, 43, 47, 50, 53, 55, 57, 60, 62, 63, 64, 69, 70, 72, 73, 75, 77, 78, 80, 81, 82, 83, 85, 89, 90, 91, 92, 100, 102, 103, 104, 108, 109, 110, 112, 113, 115, 118, 119, 121, 122, 126, 127, 129, 130, 131, 133, 135, 136, 140, 142, 148, 152, 153, 155, 156, 157, 158, 160, 163, 165, 168, 169, 172, 173, 174, 175, 176, 177, 178, 181, 185, 188, 189, 192, 193, 196, 198, 199, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 212, 216, 218, 221, 223, 226, 227, 229, 231, 234, 235, 236, 237, 240, 242, 244, 246, 248, 249, 253, 254, 255, 258, 262, 264, 265, 269, 271, 272, 274, 277, 278, 279, 280, 281, 282, 284, 285, 286, 287, 288, 289, 292, 294, 296, 297, 298, 300, 302, 303, 304, 305, 306, 309, 310, 314, 317, 318, 320, 321, 322, 323, 325, 326, 327, 328, 335, 336, 338, 342, 345, 346, 347, 353, 357, 360, 361, 362, 365, 366, 367, 370, 371, 373, 375, 376, 377, 382, 383, 386, 387, 388, 390, 392, 393, 395, 397, 401, 403, 404, 408, 410
FS16	Wrapper SubsetEval	Greedy Stepwise	Naïve Bayes	16	41, 56, 63, 104, 105, 123, 145, 153, 222, 248, 249, 256, 285, 320, 338, 404

Table 15

Cont.

FS17	Wrapper SubsetEval	Linear Forward Selection	Naïve Bayes	14	64, 104, 105, 112, 129, 157, 256, 285, 320, 327, 338, 357, 404, 409
FS18	Wrapper SubsetEval	RankSearch	Naïve Bayes	11	91, 104, 112, 129, 256, 285, 320, 327, 338, 359, 404
FS19	Wrapper SubsetEval	Subset SizeForwardSelection	Naïve Bayes	14	64, 104, 105, 112, 129, 157, 256, 285, 320, 327, 338, 357, 404, 409
FS20	Wrapper SubsetEval	Subset SizeForwardSelection	Bagging	7	49, 256, 285, 320, 327, 338, 404

Table 16 exhibits the results of the classification process on the medical dataset. We calculated the accuracy (AC), precision (PR), recall (RC) and F-measure (FM). FS14 A1 (94.7533%) ranked highest in accuracy, FS12 A8 (0.989) in precision FS3 A6 (0.876) for recall and FS17 A2 and FS17 A2 for F-Measure (0.861).

The medical elapsed table featured in Table 17 shows that based on the sum of the time elapsed for feature selection and classification, FS10 (0:00:20) had the best performance and FS15 (2:59:48) had the poorest performance.

Table 16

Medical: Precision, Recall and F-Measure Rates

Features		**Classifiers**								
Name	**Metric**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	AC	94.5659%	95.1905%	93.6914%	95.3154%	90.2561%	95.3779%	94.3785%	90.5059%	95.253%
	PR	0.84	0.877	0.811	0.878	0.872	0.846	0.912	0.98	0.847
	RC	0.803	0.803	0.775	0.811	0.438	0.859	0.707	0.398	0.847
	FM	0.821	0.839	0.793	0.843	0.583	0.853	0.796	0.566	0.847
FS2	AC	94.6284%	92.817%	93.6914%	93.3791%	89.7564%	94.6284%	93.5041%	ERROR	94.8782%
	PR	0.856	0.775	0.846	0.783	0.851	0.841	0.892		0.849
	RC	0.787	0.759	0.727	0.795	0.414	0.807	0.663		0.815
	FM	0.82	0.767	0.782	0.789	0.557	0.824	0.76		0.832
FS3	AC	94.5659%	95.128%	93.3167%	95.3779%	90.381%	95.253%	94.441%	90.9432%	95.0656%
	PR	0.84	0.901	0.809	0.882	0.874	0.829	0.921	0.981	0.84
	RC	0.803	0.771	0.747	0.811	0.446	0.876	0.703	0.426	0.843
	FM	0.821	0.831	0.777	0.845	0.59	0.852	0.797	0.594	0.842
FS4	AC	92.0675%	88.9444%	91.0681%	90.0687%	88.6321%	91.4428%	90.9432%	ERROR	92.4422%
	PR	0.777	0.698	0.765	0.718	0.813	0.728	0.806		0.776
	RC	0.687	0.51	0.614	0.594	0.349	0.719	0.55		0.723
	FM	0.729	0.589	0.682	0.651	0.489	0.723	0.654		0.748
FS5	AC	94.5034%	94.5659%	94.2536%	94.5659%	90.2561%	95.6277%	94.0037%	90.0687%	94.8782%
	PR	0.845	0.852	0.843	0.84	0.872	0.857	0.87	0.925	0.849
	RC	0.791	0.787	0.775	0.803	0.438	0.863	0.723	0.394	0.815
	FM	0.817	0.818	0.808	0.821	0.583	0.86	0.789	0.552	0.832
FS6	AC	94.5659%	88.5072%	93.7539%	92.5047%	90.2561%	95.4403%	94.0037%	ERROR	94.3785%
	PR	0.846	0.637	0.828	0.738	0.872	0.879	0.856		0.809
	RC	0.795	0.606	0.755	0.803	0.438	0.819	0.739		0.835
	FM	0.82	0.621	0.79	0.769	0.583	0.848	0.793		0.822

Table 16

Cont.

FS7	AC	94.8157%	94.5034%	94.0662%	94.7533%	89.2567%	95.128%	94.1287%	89.7564%	95.5028%
	PR	0.855	0.858	0.85	0.842	0.789	0.846	0.886	0.967	0.864
	RC	0.803	0.775	0.751	0.815	0.422	0.839	0.715	0.353	0.843
	FM	0.828	0.814	0.797	0.829	0.55	0.843	0.791	0.518	0.854
FS8	AC	94.8782%	92.3798%	94.441%	93.1918%	90.2561%	95.253%	94.1287%	89.1943%	94.6284%
	PR	0.855	0.766	0.854	0.759	0.872	0.842	0.871	0.888	0.825
	RC	0.807	0.735	0.775	0.823	0.438	0.855	0.731	0.349	0.831
	FM	0.831	0.75	0.813	0.79	0.583	0.849	0.795	0.501	0.828
FS9	AC	95.0656%	94.3161%	94.8157%	94.8782%	89.3816%	94.8157%	94.1911%	89.8813%	95.3779%
	PR	0.863	0.819	0.864	0.861	0.793	0.835	0.868	0.958	0.866
	RC	0.811	0.815	0.791	0.799	0.43	0.831	0.739	0.365	0.831
	FM	0.836	0.817	0.826	0.829	0.557	0.833	0.798	0.529	0.848
FS10	AC	95.0031%	94.7533%	94.1911%	95.0656%	89.2567%	95.1905%	94.3785%	89.8813%	95.3779%
	PR	0.854	0.834	0.831	0.873	0.789	0.85	0.896	0.968	0.866
	RC	0.819	0.827	0.787	0.799	0.422	0.839	0.723	0.361	0.831
	FM	0.836	0.831	0.808	0.834	0.55	0.844	0.8	0.526	0.848
FS11	AC	94.6284%	92.817%	93.6914%	93.3791%	89.7564%	94.6284%	93.5041%	ERROR	94.8782%
	PR	0.856	0.775	0.846	0.783	0.851	0.841	0.892		0.849
	RC	0.787	0.759	0.727	0.795	0.414	0.807	0.663		0.815
	FM	0.82	0.767	0.782	0.789	0.557	0.824	0.76		0.832
FS12	AC	95.0031%	95.128%	94.3161%	95.0031%	89.6939%	95.0656%	94.5034%	90.0687%	95.3779%
	PR	0.897	0.917	0.88	0.916	0.796	0.897	0.935	0.989	0.915
	RC	0.767	0.755	0.735	0.747	0.454	0.771	0.695	0.365	0.775
	FM	0.827	0.828	0.801	0.823	0.578	0.829	0.797	0.534	0.839
FS13	AC	94.5659%	95.0031%	93.7539%	95.5653%	89.2567%	94.6908%	93.8164%	89.0693%	95.253%
	PR	0.832	0.879	0.82	0.894	0.881	0.839	0.899	0.987	0.868
	RC	0.815	0.787	0.767	0.811	0.357	0.815	0.679	0.301	0.819
	FM	0.824	0.831	0.793	0.851	0.509	0.827	0.773	0.462	0.843

Table 16

Cont.

FS14	AC	94.7533%	93.2542%	93.3791%	92.817%	89.1318%	94.8157%	93.3791%	88.1949%	94.6908%
	PR	0.857	0.831	0.839	0.807	0.879	0.855	0.891	0.955	0.866
	RC	0.795	0.711	0.711	0.707	0.349	0.803	0.655	0.253	0.779
	FM	0.825	0.766	0.77	0.754	0.5	0.828	0.755	0.4	0.82
FS15	AC	94.8782%	85.4466%	94.6908%	93.0668%	89.3816%	94.9407%	94.0037%	ERROR	94.9407%
	PR	0.841	0.608	0.833	0.778	0.788	0.844	0.901		0.831
	RC	0.827	0.181	0.823	0.775	0.434	0.827	0.691		0.847
	FM	0.834	0.279	0.828	0.777	0.56	0.836	0.782		0.839
FS16	AC	94.5659%	95.0031%	93.7539%	95.5653%	89.2567%	94.6908%	93.8164%	89.0693%	95.2530%
	PR	0.832	0.879	0.82	0.894	0.881	0.839	0.899	0.987	0.868
	RC	0.815	0.787	0.767	0.811	0.357	0.815	0.679	0.301	0.819
	FM	0.824	0.831	0.793	0.851	0.509	0.827	0.773	0.462	0.843
FS17	AC	94.8782%	95.7527%	94.1911%	95.5028%	89.3816%	95.0031%	94.3161%	90.1936%	95.3779%
	PR	0.861	0.876	0.855	0.9	0.793	0.848	0.873	0.979	0.876
	RC	0.799	0.847	0.755	0.799	0.43	0.827	0.743	0.378	0.819
	FM	0.829	0.861	0.802	0.847	0.557	0.837	0.803	0.545	0.846
FS18	AC	94.441%	95.0656%	93.2542%	95.0031%	89.569%	95.3154%	94.5659%	89.3192%	95.253%
	PR	0.857	0.925	0.89	0.933	0.887	0.91	0.955	0.988	0.918
	RC	0.771	0.743	0.647	0.731	0.378	0.775	0.683	0.317	0.763
	FM	0.812	0.824	0.749	0.82	0.53	0.837	0.796	0.48	0.833
FS19	AC	94.8782%	95.7527%	94.1911%	95.5028%	89.3816%	95.0031%	94.3161%	90.1936%	95.3779%
	PR	0.861	0.876	0.855	0.9	0.793	0.848	0.873	0.979	0.876
	RC	0.799	0.847	0.755	0.799	0.43	0.827	0.743	0.378	0.819
	FM	0.829	0.861	0.802	0.847	0.557	0.837	0.803	0.545	0.846
FS20	AC	94.8157%	94.3785%	93.6914%	94.1287%	89.1318%	94.6284%	93.4416%	88.2573%	94.7533%
	PR	0.861	0.866	0.856	0.886	0.879	0.853	0.891	0.969	0.863
	RC	0.795	0.755	0.715	0.715	0.349	0.791	0.659	0.253	0.787
	FM	0.827	0.807	0.779	0.791	0.5	0.821	0.758	0.401	0.824

Table 17

Medical Time Elapsed

Features		Classifiers (Time Elapsed)								
Name	**Time**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	0:00:02	0:00:01	0:00:00	0:00:01	0:00:00	0:00:02	0:00:04	0:00:03	0:00:10	0:00:02
FS2	0:00:08	0:00:07	0:00:03	0:00:07	0:00:01	0:00:08	0:00:09	0:00:16	0:00:52	0:00:16
FS3	0:00:08	0:00:25	0:00:01	0:00:02	0:00:00	0:00:04	0:00:15	0:00:05	0:00:32	0:00:03
FS4	0:00:20	0:00:08	0:00:02	0:00:07	0:00:01	0:00:08	0:00:19	0:00:14	0:00:59	0:00:21
FS5	0:00:14	0:00:02	0:00:01	0:00:02	0:00:00	0:00:02	0:00:04	0:00:04	0:00:16	0:00:03
FS6	0:00:02	0:00:14	0:00:04	0:00:15	0:00:02	0:00:14	0:00:11	0:00:27	0:01:39	0:00:37
FS7	0:00:02	0:00:11	0:00:01	0:00:02	0:00:00	0:00:01	0:00:08	0:00:03	0:00:09	0:00:03
FS8	0:00:03	0:00:05	0:00:02	0:00:05	0:00:01	0:00:06	0:00:08	0:00:11	0:00:32	0:00:09
FS9	0:00:02	0:00:02	0:00:01	0:00:01	0:00:00	0:00:02	0:00:07	0:00:03	0:00:09	0:00:03
FS10	0:00:02	0:00:01	0:00:01	0:00:01	0:00:00	0:00:02	0:00:04	0:00:02	0:00:05	0:00:02
FS11	0:00:08	0:00:06	0:00:02	0:00:08	0:00:01	0:00:07	0:00:08	0:00:15	0:00:50	0:00:15
FS12	0:00:07	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:04	0:00:01	0:00:05	0:00:02
FS13	0:13:53	0:00:02	0:00:01	0:00:01	0:00:00	0:00:01	0:00:05	0:00:02	0:00:07	0:00:02
FS14	1:19:37	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:04	0:00:02	0:00:05	0:00:02
FS15	2:56:07	0:00:17	0:00:04	0:00:14	0:00:02	0:00:13	0:00:24	0:00:26	0:01:18	0:00:43
FS16	0:09:58	0:00:02	0:00:01	0:00:01	0:00:00	0:00:01	0:00:05	0:00:02	0:00:07	0:00:02
FS17	0:01:32	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:04	0:00:02	0:00:06	0:00:01
FS18	0:12:51	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:04	0:00:01	0:00:04	0:00:01
FS19	0:01:07	0:00:01	0:00:00	0:00:01	0:00:00	0:00:01	0:00:05	0:00:02	0:00:06	0:00:01
FS20	0:02:13	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:02	0:00:01	0:00:03	0:00:01

When considering runtime, reduction in features and recall in Figure 9 below, we have determined that FS19 A2 is the better algorithm to use for this dataset.

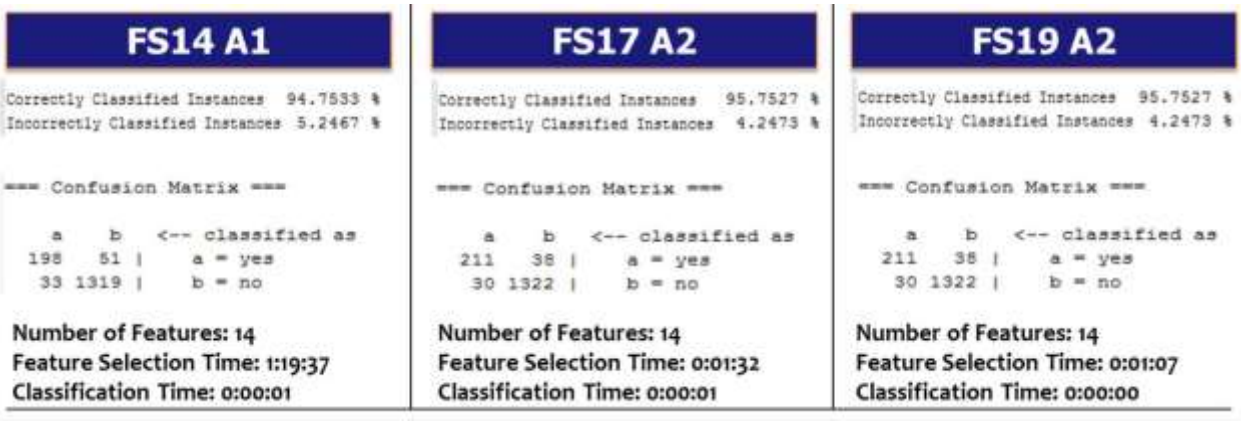

Figure 9. Top Three Medical.

5.5 Clean Full Dataset

Table 18 displays the results of the feature selection process on the clean full dataset. Based solely on feature reduction, FS20 was the top ranking performer by reducing the features down to a feature set of 4. FS15 was the lowest ranking performer by only reducing the features down to a feature set of 214.

Table 18

Clean Full Feature Set

Clinical Feature Set					
Set	**Attribute Evaluator**	**Search Method**	**Classifier**	**# of Feat.**	**Selected Feature List**
FS1	Cfs Subset Eval	BestFirst		13	34,105,248,285,320,322,338,404,411,414,416,417,418
FS2	Cfs Subset Eval	Genetic Search		134	3,13,19,27,29,31,34,35,54,64,67,73,75,82,86,87,88,96,97,104,112,122,128,129,131,145,147,148,155,164,167,169,173,175,176,181,187,188,194,196,209,218,222,228,229,232,236,239,243,244,248,251,252,256,258,260,261,263,266,268,271,274,276,280,282,287,289,291,299,301,302,305,306,307,308,310,311,314,316,317,318,320,324,327,328,332,334,335,337,338,339,340,348,350,352,353,354,356,357,359,360,361,362,363,364,366,370,375,376,379,381,382,383,384,385,387,390,392,394,395,396,399,403,404,405,407,410,411,413,414,415,416,417,418
FS3	Cfs SubsetEval	Rank Search		16	34,91,104,105,248,256,320,327,338,394,404,411,414,416,417,418
FS4	Classifier SubsetEval	Genetic Search	OneR	103	2,3,6,10,14,19,30,40,41,43,45,47,53,58,62,68,69,70,74,78,89,90,91,92,93,97,100,101,104,108,110,111,127,129,130,131,137,140,165,171,172,178,181,184,186,192,199,205,206,214,216,217,220,222,227,228,232,235,243,244,246,250,251,258,260,262,263,266,275,285,286,290,291,298,301,302,304,306,313,315,323,336,338,343,346,348,349,351,352,354,359,361,363,373,378,381,382,397,400,409,411,415,418

Table 18

Cont.

FS5	Consistency SubsetEval	BestFirst		9	2,28,33,285,411,412,414,416,418
FS6	Consistency SubsetEval	Genetic Search		170	1,2,4,5,9,12,13,17,19,22,23,25,27,28,30,31,32,33,34,35,38,40,41,42,43,45,46,47,50,52,53,54,57,59,60,61,62,63,65,66,67,69,70,71,77,79,81,82,84,91,94,99,102,104,106,109,111,115,117,118,119,123,129,131,132,142,144,145,151,153,154,155,159,161,165,171,178,181,184,186,192,193,194,195,197,199,201,203,205,206,210,213,214,215,217,220,222,223,229,232,235,243,244,245,251,258,259,260,262,263,267,272,274,275,276,277,281,285,286,289,290,295,298,301,302,303,304,305,313,314,325,328,331,333,336,338,341,343,344,345,346,348,349,350,352,361,363,366,371,372,383,386,389,392,393,394,395,396,403,404,405,409,410,411,412,413,414,415,416,417
FS7	Consistency SubsetEval	Linear Forward Selection		12	2,34,49,105,285,320,338,357,411,414,416,418
FS8	Consistency SubsetEval	Rank Search		96	2,7,16,20,23,25,29,30,33,34,43,49,53,54,63,64,69,87,88,91,104,105,112,120,122,123,124,126,129,131,135,139,141,145,148,157,159,162,163,164,171,172,191,207,209,211,218,226,229,230,245,246,248,249,250,251,254,256,259,262,275,279,282,285,296,297,298,311,318,320,321,322,324,325,327,334,335,337,338,339,355,357,359,375,384,394,400,404,409,411,412,413,414,416,417,418
FS9	Consistency SubsetEval	SubsetSize Forward Selection		10	2,34,49,105,285,357,411,414,416,418
FS10	Filtered SubsetEval	BestFirst		10	105,285,320,338,404,411,414,416,417,418

Table 18

Cont.

FS11	Filtered SubsetEval	Genetic Search		134	3,13,19,27,29,31,34,35,54,64,67,73,75,82,86,87,88,96,97,104,112,122,128,129,131,145,147,148,155,164,167,169,173,175,176,181,187,188,194,196,209,218,222,228,229,232,236,239,243,244,248,251,252,256,258,260,261,263,266,268,271,274,276,280,282,287,289,291,299,301,302,305,306,307,308,310,311,314,316,317,318,320,324,327,328,332,334,335,337,338,339,340,348,350,352,353,354,356,357,359,360,361,362,363,364,366,370,375,376,379,381,382,383,384,385,387,390,392,394,395,396,399,403,404,405,407,410,411,413,414,415,416,417,418
FS12	Filtered SubsetEval	Rank Search		10	1,104,256,320,327,338,404,411,414,417
FS13	Wrapper SubsetEval	BestFirst	Naïve Bayes	11	1,14,21,43,75,130,145,185,213,285,411
FS14	Wrapper SubsetEval	BestFirst	Bagging	15	2,5,20,65,207,256,269,305,320,334,348,373,399,402,411
FS15	Wrapper SubsetEval	Genetic Search	Bagging	214	1,2,7,9,10,14,15,17,19,21,23,27,28,31,32,33,34,35,36,39,41,43,44,47,48,49,54,60,61,62,63,64,66,67,68,70,71,74,75,79,80,85,92,94,96,99,100,102,103,107,112,113,114,119,121,123,125,126,127,128,133,134,136,139,142,143,145,149,150,157,158,160,161,163,165,169,171,172,178,180,181,183,184,185,186,188,189,190,193,194,196,200,201,204,206,210,212,213,215,216,217,218,219,220,222,223,224,225,227,228,229,230,231,232,233,234,235,237,239,240,243,244,245,246,247,248,249,251,254,255,259,263,265,266,267,271,273,274,276,278,279,281,283,284,286,291,292,293,294,295,296,299,301,302,304,306,310,313,314,317,318,319,320,321,322,323,326,327,330,331,332,333,334,336,338,339,341,344,352,353,354,355,356,357,358,359,366,368,369,370,371,372,377,379,380,381,382,383,386,387,388,389,392,393,395,396,398,404,405,407,409,414,416,417

Table 18

Cont.

FS16	Wrapper SubsetEval	Greedy Stepwise	Naïve Bayes	7	14,21,130,145,185,285,411
FS17	Wrapper SubsetEval	Linear Forward Selection	Naïve Bayes	9	104,105,124,285,327,337,338,411,417
FS18	Wrapper SubsetEval	Rank Search	Naïve Bayes	10	91,104,256,320,327,338,404,411,414,417
FS19	Wrapper SubsetEval	SubsetSize Forward Selection	Naïve Bayes	5	104,105,285,411,417
FS20	Wrapper SubsetEval	SubsetSize Forward Selection	Bagging	4	7,104,411,418

Table 19 exhibits the results of the classification process on the clean full dataset. We calculated the accuracy (AC), precision (PR), recall (RC) and F-measure (FM). FS8 A6 (97.5015%) ranked highest in accuracy, FS20 A8 (0.977) in precision, FS5 A1 (0.981) for recall and FS8 A6 for F-Measure (0.919). Based solely on AC and FM, FS8 A6 is the top ranking performer.

The clean full time elapsed table featured in Table 20 shows that based on the sum of the time elapsed for feature selection and classification, FS7 (0:00:23) had the best performance and FS14 (1:59:59) had the poorest performance.

Table 19

Clean: Precision, Recall and F-Measure Rates

Features		**Classifiers**								
Name	**Metric**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	AC	96.1333%	96.4902%	96.1333%	96.4902%	95.8358%	97.0851%	96.7281%	96.6686%	96.7876%
	PR	0.887	0.878	0.893	0.87	0.87	0.935	0.916	0.941	0.895
	RC	0.864	0.902	0.856	0.913	0.864	0.875	0.871	0.841	0.902
	FM	0.875	0.89	0.874	0.891	0.867	0.904	0.893	0.888	0.898
FS2	AC	96.7281%	93.0399%	97.0851%	94.5866%	96.4307%	97.1446%	97.0256%	ERROR	96.0143%
	PR	0.91	0.793	0.918	0.817	0.889	0.958	0.918		0.838
	RC	0.879	0.754	0.894	0.845	0.883	0.856	0.89		0.924
	FM	0.894	0.773	0.906	0.831	0.886	0.904	0.904		0.879
FS3	AC	96.0738%	96.1927%	96.0143%	96.6092%	95.8953%	96.4902%	96.7281%	96.0738%	96.7281%
	PR	0.884	0.888	0.899	0.894	0.871	0.915	0.91	0.93	0.92
	RC	0.864	0.867	0.841	0.89	0.867	0.856	0.879	0.811	0.867
	FM	0.874	0.877	0.869	0.892	0.869	0.885	0.894	0.866	0.893
FS4	AC	96.8471%	92.9209%	96.7876%	94.884%	95.8358%	96.5497%	96.3712%	ERROR	96.0143%
	PR	0.907	0.817	0.907	0.83	0.862	0.926	0.889		0.856
	RC	0.89	0.708	0.886	0.848	0.875	0.848	0.879		0.898
	FM	0.899	0.759	0.897	0.839	0.868	0.885	0.884		0.876
FS5	AC	96.4902%	96.0738%	96.0738%	95.5979%	96.0143%	97.204%	96.3712%	97.0851%	95.7168%
	PR	0.977	0.899	0.893	0.889	0.872	0.939	0.898	0.928	0.869
	RC	0.981	0.845	0.852	0.822	0.875	0.879	0.867	0.883	0.856
	FM	0.979	0.871	0.872	0.854	0.873	0.908	0.882	0.905	0.863
FS6	AC	96.9661%	85.9607%	96.4307%	95.4194%	95.9548%	96.8471%	96.9066%	ERROR	97.1446%
	PR	0.918	0.566	0.892	0.828	0.868	0.945	0.914		0.903
	RC	0.886	0.455	0.879	0.894	0.875	0.848	0.886		0.917
	FM	0.902	0.504	0.885	0.86	0.872	0.894	0.9		0.91

Table 19

Cont.

FS7	AC	96.6092%	96.4902%	96.1927%	96.1333%	96.0143%	96.6092%	96.5497%	97.204%	96.7876%
	PR	0.903	0.881	0.897	0.849	0.877	0.919	0.896	0.943	0.904
	RC	0.879	0.898	0.856	0.917	0.867	0.86	0.883	0.875	0.89
	FM	0.891	0.889	0.876	0.882	0.872	0.888	0.889	0.908	0.897
FS8	AC	96.7281%	94.0512%	96.4307%	93.6347%	95.9548%	97.5015%	96.8471%	ERROR	96.6092%
	PR	0.91	0.801	0.902	0.756	0.868	0.934	0.917		0.9
	RC	0.879	0.826	0.867	0.879	0.875	0.905	0.879		0.883
	FM	0.894	0.813	0.884	0.813	0.872	0.919	0.897		0.891
FS9	AC	96.6092%	95.7763%	95.9548%	95.4789%	96.0143%	96.4902%	96.5497%	97.0851%	96.3117%
	PR	0.9	0.853	0.883	0.829	0.877	0.912	0.896	0.942	0.898
	RC	0.883	0.883	0.856	0.898	0.867	0.86	0.883	0.867	0.864
	FM	0.891	0.868	0.869	0.862	0.872	0.885	0.889	0.903	0.88
FS10	AC	96.3117%	96.3712%	96.1927%	96.2522%	95.9548%	96.9661%	96.4902%	96.1927%	96.4307%
	PR	0.898	0.886	0.888	0.879	0.877	0.931	0.908	0.924	0.892
	RC	0.864	0.883	0.867	0.883	0.864	0.871	0.864	0.826	0.879
	FM	0.88	0.884	0.877	0.881	0.87	0.9	0.885	0.872	0.885
FS11	AC	96.7281%	93.0399%	97.0851%	94.5866%	96.4307%	97.1446%	97.0256%	ERROR	96.0143%
	PR	0.91	0.793	0.918	0.817	0.889	0.958	0.918		0.838
	RC	0.879	0.754	0.894	0.845	0.883	0.856	0.89		0.924
	FM	0.894	0.773	0.906	0.831	0.886	0.904	0.904		0.879
FS12	AC	95.8953%	95.8953%	95.7763%	96.1333%	95.8358%	95.4194%	95.9548%	95.122%	96.3117%
	PR	0.892	0.885	0.918	0.913	0.862	0.87	0.871	0.96	0.935
	RC	0.841	0.848	0.803	0.833	0.875	0.833	0.871	0.72	0.822
	FM	0.865	0.867	0.857	0.871	0.868	0.851	0.871	0.823	0.875
FS13	AC	96.7281%	96.3712%	96.5497%	96.7281%	95.8358%	96.5497%	96.3117%	94.2296%	96.6092%
	PR	0.897	0.88	0.879	0.886	0.862	0.906	0.888	0.947	0.888
	RC	0.894	0.89	0.905	0.909	0.875	0.871	0.875	0.67	0.898
	FM	0.896	0.885	0.892	0.897	0.868	0.888	0.882	0.785	0.893

Table 19

Cont.

FS14	AC	96.7281%	94.1701%	96.8471%	94.7055%	95.8358%	97.204%	96.0738%	96.6092%	96.1927%
	PR	0.91	0.855	0.907	0.86	0.862	0.919	0.875	0.948	0.894
	RC	0.879	0.758	0.89	0.792	0.875	0.902	0.875	0.83	0.86
	FM	0.894	0.803	0.899	0.824	0.868	0.91	0.875	0.885	0.876
FS15	AC	97.204%	91.1362%	96.3117%	94.2891%	95.8358%	96.9661%	96.4902%	ERROR	96.1927%
	PR	0.936	0.719	0.888	0.804	0.898	0.949	0.932		0.855
	RC	0.883	0.716	0.875	0.841	0.83	0.852	0.837		0.913
	FM	0.908	0.717	0.882	0.822	0.862	0.898	0.882		0.883
FS16	AC	96.3117%	96.3712%	96.1333%	96.7281%	95.8358%	95.1814%	96.3117%	93.8727%	96.4902%
	PR	0.888	0.872	0.859	0.886	0.862	0.851	0.895	0.945	0.896
	RC	0.875	0.902	0.902	0.909	0.875	0.841	0.867	0.648	0.879
	FM	0.882	0.886	0.88	0.897	0.868	0.846	0.881	0.769	0.887
FS17	AC	96.1333%	96.3117%	96.0738%	97.0256%	95.8358%	96.1333%	95.8358%	94.2891%	96.6092%
	PR	0.9	0.856	0.878	0.925	0.862	0.887	0.865	0.947	0.888
	RC	0.848	0.92	0.871	0.883	0.875	0.864	0.871	0.674	0.898
	FM	0.873	0.887	0.875	0.903	0.868	0.875	0.868	0.788	0.893
FS18	AC	95.8953%	95.8953%	95.7763%	96.1333%	95.8358%	95.4194%	95.9548%	95.122%	96.3117%
	PR	0.892	0.885	0.918	0.913	0.862	0.87	0.871	0.96	0.935
	RC	0.841	0.848	0.803	0.833	0.875	0.833	0.871	0.72	0.822
	FM	0.865	0.867	0.857	0.871	0.868	0.851	0.871	0.823	0.875
FS19	AC	96.1927%	96.4902%	96.0738%	96.8471%	95.8358%	96.1927%	95.8358%	94.5271%	96.4307%
	PR	0.903	0.873	0.878	0.901	0.862	0.882	0.865	0.943	0.889
	RC	0.848	0.909	0.871	0.898	0.875	0.875	0.871	0.693	0.883
	FM	0.875	0.891	0.875	0.899	0.868	0.878	0.868	0.799	0.886
FS20	AC	96.4307%	96.0143%	96.7281%	95.6573%	95.8358%	96.6686%	96.2522%	96.3712%	96.6092%
	PR	0.911	0.896	0.913	0.88	0.862	0.916	0.888	0.977	0.937
	RC	0.856	0.845	0.875	0.837	0.875	0.867	0.871	0.788	0.841
	FM	0.883	0.869	0.894	0.858	0.868	0.891	0.88	0.872	0.886

Table 20

Clean Time Elapsed

Features		Classifiers (Time Elapsed)								
Name	**Time**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	0:00:02	0:00:01	0:00:02	0:00:01	0:00:01	0:00:01	0:00:01	0:00:02	0:00:15	0:00:02
FS2	0:00:10	0:00:04	0:00:02	0:00:04	0:00:01	0:00:10	0:00:04	0:00:20	0:01:39	0:00:14
FS3	0:00:09	0:00:01	0:00:01	0:00:00	0:00:00	0:00:02	0:00:02	0:00:03	0:00:16	0:00:02
FS4	0:00:10	0:00:04	0:00:03	0:00:03	0:00:01	0:00:08	0:00:04	0:00:16	0:01:21	0:00:12
FS5	0:00:05	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:14	0:00:02
FS6	0:00:02	0:00:05	0:00:05	0:00:05	0:00:02	0:00:13	0:00:04	0:00:26	0:01:52	0:00:14
FS7	0:00:01	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:14	0:00:02
FS8	0:00:04	0:00:04	0:00:03	0:00:03	0:00:02	0:00:07	0:00:04	0:00:14	0:00:59	0:00:10
FS9	0:00:02	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:02	0:00:01	0:00:14	0:00:01
FS10	0:00:01	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:14	0:00:23
FS11	0:00:10	0:00:05	0:00:03	0:00:05	0:00:01	0:00:10	0:00:05	0:00:21	0:01:36	0:00:14
FS12	0:00:10	0:00:01	0:00:02	0:00:01	0:00:01	0:00:01	0:00:01	0:00:02	0:00:10	0:00:02
FS13	0:11:11	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:02	0:00:02	0:00:09	0:00:03
FS14	1:59:40	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:10	0:00:03
FS15	1:19:33	0:00:14	0:00:07	0:00:08	0:00:01	0:00:17	0:00:05	0:00:33	0:02:07	0:00:22
FS16	0:03:24	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:06	0:00:02
FS17	0:01:15	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:06	0:00:01
FS18	0:26:58	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:08	0:00:01
FS19	0:00:19	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:05	0:00:01
FS20	0:01:36	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:07	0:00:01

When considering runtime, reduction in features and recall in Figure 10 below, we have determined that FS8 A6 is the better algorithm to use for this dataset.

Figure 10. Top Three Clean Full.

5.6 Clean Clinical Dataset

Table 21 displays the results of the feature selection process on the clean clinical dataset. Based solely on feature reduction, FS13, FS16 and FS19 were the top ranking performers by reducing the features down to a feature set of 2. FS15 was the lowest ranking performer by only reducing the features down to a feature set of 16.

Table 21

Clean Clinical Feature Set

Clinical Feature Set					
Set	**Attribute Evaluator**	**Search Method**	**Classifier**	**# of Feat.**	**Selected Feature List**
FS1	Cfs Subset Eval	BestFirst		6	20, 21, 25, 29, 33, 34
FS2	Cfs Subset Eval	Genetic Search		6	20, 21, 25, 29, 33, 34
FS3	Cfs SubsetEval	RankSearch		7	20, 21, 23, 25, 29, 33, 34
FS4	Classifier SubsetEval	Genetic Search	OneR	5	13, 14, 24, 27, 34
FS5	Consistency SubsetEval	BestFirst		10	20, 21, 23, 25, 28, 29, 30, 33, 34, 38
FS6	Consistency SubsetEval	Genetic Search		14	13, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 33, 34, 38
FS7	Consistency SubsetEval	Linear Forward Selection		10	20, 21, 23, 25, 28, 29, 30, 33, 34, 38
FS8	Consistency SubsetEval	RankSearch		11	16, 20, 21, 23, 25, 28, 29, 30, 33, 34, 38
FS9	Consistency SubsetEval	Subset SizeForward Selection		10	20, 21, 23, 25, 28, 29, 30, 33, 34, 38
FS10	Filtered SubsetEval	BestFirst		3	29, 33, 34
FS11	FilteredS ubsetEval	Genetic Search		3	29, 33, 34
FS12	Filtered SubsetEval	RankSearch		4	23, 29, 33, 34
FS13	Wrapper SubsetEval	BestFirst	Naïve Bayes	2	20, 34

Table 21

Cont.

FS14	Wrapper SubsetEval	BestFirst	Bagging	11	12, 13, 14, 17, 20, 29, 30, 33, 34, 37, 38
FS15	Wrapper SubsetEval	Genetic Search	Bagging	16	12, 13, 16, 17, 18, 19, 20, 21, 22, 25, 28, 29, 33, 35, 36, 37
FS16	Wrapper SubsetEval	Greedy Stepwise	Naïve Bayes	2	20, 34
FS17	Wrapper SubsetEval	Linear Forward Selection	Naïve Bayes	8	20, 23, 27, 28, 29, 33, 34, 38
FS18	Wrapper SubsetEval	RankSearch	Naïve Bayes	4	23, 29, 33, 34
FS19	Wrapper SubsetEval	Subset SizeForward Selection	Naïve Bayes	2	20, 34
FS20	Wrapper SubsetEval	Subset SizeForward Selection	Bagging	11	12, 13, 14, 17, 20, 29, 30, 33, 34, 37, 38

Table 22 exhibits the results of the classification process on the clean clinical dataset. We calculated the accuracy (AC), precision (PR), recall (RC) and F-measure (FM). FS14 A1 and FS20 A1 (89.8275%) ranked highest in accuracy, FS13 A8, FS16 A8 and FS19 A8 (0.974) in precision FS14 A6 and FS20 A6 (0.515) for recall and FS14 A6 and FS20 A6 for F-Measure (0.613).

The clean clinical elapsed table featured in Table 23 shows that based on the sum of the time elapsed for feature selection and classification, FS10 and FS11 (0:00:05) had the best performance and FS15 (0:06:29) had the poorest performance.

Table 22

Clean Clinical: Precision, Recall and F-Measure Rates

Features		**Classifiers**								
Name	**Metric**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	AC	88.3403%	86.972%	87.6264%	86.4366%	86.2582%	88.5187%	87.091%	88.6377%	88.9352%
	PR	0.73	0.703	0.697	0.65	0.709	0.748	0.664	0.91	0.787
	RC	0.409	0.295	0.375	0.295	0.212	0.405	0.36	0.307	0.405
	FM	0.524	0.416	0.488	0.406	0.327	0.526	0.467	0.459	0.535
FS2	AC	88.3403%	86.972%	87.6264%	86.4366%	86.2582%	88.5187%	87.091%	88.6377%	88.9352%
	PR	0.73	0.703	0.697	0.65	0.709	0.748	0.664	0.91	0.787
	RC	0.409	0.295	0.375	0.295	0.212	0.405	0.36	0.307	0.405
	FM	0.524	0.416	0.488	0.406	0.327	0.526	0.467	0.459	0.535
FS3	AC	88.2808%	87.4479%	87.6264%	86.4366%	86.2582%	88.5187%	86.7936%	88.6377%	88.2808%
	PR	0.731	0.748	0.697	0.65	0.709	0.748	0.667	0.91	0.731
	RC	0.402	0.303	0.375	0.295	0.212	0.405	0.318	0.307	0.402
	FM	0.518	0.431	0.488	0.406	0.327	0.526	0.431	0.459	0.518
FS4	AC	86.3772%	86.2582%	86.3772%	86.3772%	86.3772%	86.4961%	86.3772%	84.2951%	86.1392%
	PR	0.778	0.789	0.818	0.761	0.761	0.894	0.761	0	0.772
	RC	0.186	0.17	0.17	0.193	0.193	0.159	0.193	0	0.167
	FM	0.3	0.28	0.282	0.308	0.308	0.27	0.308	0	0.274
FS5	AC	88.7567%	87.4479%	87.8049%	86.9126%	87.1505%	89.4111%	87.8049%	89.0541%	88.8162%
	PR	0.733	0.73	0.686	0.647	0.669	0.75	0.675	0.877	0.732
	RC	0.447	0.318	0.413	0.367	0.36	0.489	0.432	0.352	0.455
	FM	0.555	0.443	0.515	0.469	0.468	0.592	0.527	0.503	0.561
FS6	AC	89.3516%	87.2695%	87.9833%	86.7936%	87.1505%	89.0541%	87.8049%	88.9352%	89.2326%
	PR	0.771	0.716	0.691	0.642	0.669	0.725	0.675	0.868	0.771
	RC	0.458	0.314	0.424	0.36	0.36	0.489	0.432	0.348	0.447
	FM	0.575	0.437	0.526	0.461	0.468	0.584	0.527	0.497	0.566

Table 22

Cont.

FS7	AC	88.7567%	87.4479%	87.8049%	86.9126%	87.1505%	89.4111%	87.8049%	89.0541%	88.8162%
	PR	0.733	0.73	0.686	0.647	0.669	0.75	0.675	0.877	0.732
	RC	0.447	0.318	0.413	0.367	0.36	0.489	0.432	0.352	0.455
	FM	0.555	0.443	0.515	0.469	0.468	0.592	0.527	0.503	0.561
FS8	AC	88.9352%	86.3772%	87.8049%	86.9126%	87.1505%	89.2326%	87.8049%	88.9946%	88.5187%
	PR	0.738	0.684	0.686	0.647	0.669	0.729	0.675	0.869	0.713
	RC	0.458	0.246	0.413	0.367	0.36	0.5	0.432	0.352	0.451
	FM	0.565	0.362	0.515	0.469	0.468	0.593	0.527	0.501	0.552
FS9	AC	88.7567%	87.4479%	87.8049%	86.9126%	87.1505%	89.4111%	87.8049%	89.0541%	88.8162%
	PR	0.733	0.73	0.686	0.647	0.669	0.75	0.675	0.877	0.732
	RC	0.447	0.318	0.413	0.367	0.36	0.489	0.432	0.352	0.455
	FM	0.555	0.443	0.515	0.469	0.468	0.592	0.527	0.503	0.561
FS10	AC	86.9126%	85.8418%	86.3177%	86.6151%	86.1392%	86.8531%	87.1505%	86.7936%	86.7936%
	PR	0.775	0.641	0.693	0.741	0.731	0.753	0.786	0.862	0.744
	RC	0.235	0.223	0.231	0.227	0.186	0.242	0.25	0.189	0.242
	FM	0.36	0.331	0.347	0.348	0.296	0.367	0.379	0.311	0.366
FS11	AC	86.9126%	85.8418%	86.3177%	86.6151%	86.1392%	86.8531%	87.1505%	86.7936%	86.7936%
	PR	0.775	0.641	0.693	0.741	0.731	0.753	0.786	0.862	0.744
	RC	0.235	0.223	0.231	0.227	0.186	0.242	0.25	0.189	0.242
	FM	0.36	0.331	0.347	0.348	0.296	0.367	0.379	0.311	0.366
FS12	AC	86.9126%	86.1987%	86.4366%	86.8531%	86.1392%	86.972%	87.091%	86.4961%	87.0910%
	PR	0.775	0.67	0.7	0.753	0.731	0.765	0.764	0.863	0.742
	RC	0.235	0.239	0.239	0.242	0.186	0.246	0.258	0.167	0.273
	FM	0.36	0.352	0.356	0.367	0.296	0.372	0.385	0.279	0.399
FS13	AC	86.7341%	86.6151%	86.7936%	86.3772%	86.3772%	86.7341%	86.7936%	86.4961%	86.6746%
	PR	0.847	0.81	0.85	0.761	0.761	0.847	0.85	0.974	0.845
	RC	0.189	0.193	0.193	0.193	0.193	0.189	0.193	0.144	0.186
	FM	0.31	0.312	0.315	0.308	0.308	0.31	0.315	0.251	0.304

Table 22

Cont.

FS14	AC	89.8275%	87.4479%	89.4706%	85.5443%	85.7823%	89.768%	88.2213%	88.9352%	89.3516%
	PR	0.782	0.785	0.754	0.571	0.575	0.756	0.672	0.842	0.746
	RC	0.489	0.277	0.489	0.318	0.364	0.515	0.489	0.364	0.489
	FM	0.601	0.409	0.593	0.409	0.445	0.613	0.566	0.508	0.59
FS15	AC	89.649%	84.5925%	88.6972%	83.7002%	86.9126%	89.2326%	87.4479%	88.5782%	89.0541%
	PR	0.788	0.619	0.715	0.473	0.69	0.755	0.651	0.853	0.753
	RC	0.466	0.049	0.466	0.326	0.303	0.466	0.432	0.33	0.451
	FM	0.586	0.091	0.564	0.386	0.421	0.576	0.519	0.475	0.564
FS16	AC	86.7341%	86.6151%	86.7936%	86.3772%	86.3772%	86.7341%	86.7936%	86.4961%	86.6746%
	PR	0.847	0.81	0.85	0.761	0.761	0.847	0.85	0.974	0.845
	RC	0.189	0.193	0.193	0.193	0.193	0.189	0.193	0.144	0.186
	FM	0.31	0.312	0.315	0.308	0.308	0.31	0.315	0.251	0.304
FS17	AC	87.9833%	86.6151%	88.5187%	87.091%	86.6746%	88.4593%	86.972%	88.5782%	88.3403%
	PR	0.691	0.714	0.713	0.753	0.756	0.719	0.669	0.909	0.687
	RC	0.424	0.246	0.451	0.265	0.223	0.436	0.337	0.303	0.473
	FM	0.526	0.366	0.552	0.392	0.345	0.542	0.448	0.455	0.561
FS18	AC	86.9126%	86.1987%	86.4366%	86.8531%	86.1392%	86.972%	87.091%	86.4961%	87.091%
	PR	0.775	0.67	0.7	0.753	0.731	0.765	0.764	0.863	0.742
	RC	0.235	0.239	0.239	0.242	0.186	0.246	0.258	0.167	0.273
	FM	0.36	0.352	0.356	0.367	0.296	0.372	0.385	0.279	0.399
FS19	AC	86.7341%	86.6151%	86.7936%	86.3772%	86.3772%	86.7341%	86.7936%	86.4961%	86.6746%
	PR	0.847	0.81	0.85	0.761	0.761	0.847	0.85	0.974	0.845
	RC	0.189	0.193	0.193	0.193	0.193	0.189	0.193	0.144	0.186
	FM	0.31	0.31	0.315	0.308	0.308	0.31	0.315	0.251	0.304
FS20	AC	89.8275%	87.4479%	89.4706%	85.5443%	85.7823%	89.768%	88.2213%	88.9352%	89.3516%
	PR	0.782	0.785	0.754	0.571	0.575	0.756	0.672	0.842	0.746
	RC	0.489	0.277	0.489	0.318	0.364	0.515	0.489	0.364	0.489
	FM	0.601	0.409	0.593	0.409	0.445	0.613	0.566	0.508	0.59

Table 23

Clean Clinical Time Elapsed

Features		Classifiers (Time Elapsed)								
Name	**Time**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	0:00:00	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:02	0:00:03	0:00:02
FS2	0:00:00	0:00:01	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:02	0:00:02
FS3	0:00:00	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:03	0:00:02
FS4	0:00:01	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:02	0:00:01
FS5	0:00:00	0:00:01	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:05	0:00:03
FS6	0:00:00	0:00:02	0:00:01	0:00:01	0:00:00	0:00:01	0:00:02	0:00:02	0:00:06	0:00:04
FS7	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:02	0:00:02	0:00:04	0:00:03
FS8	0:00:00	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:05	0:00:03
FS9	0:00:00	0:00:01	0:00:01	0:00:01	0:00:01	0:00:01	0:00:02	0:00:01	0:00:04	0:00:03
FS10	0:00:00	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:01
FS11	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:00	0:00:02	0:00:01
FS12	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:02	0:00:01
FS13	0:00:06	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:01
FS14	0:03:40	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:02	0:00:02	0:00:04	0:00:03
FS15	0:06:10	0:00:02	0:00:01	0:00:00	0:00:00	0:00:02	0:00:02	0:00:02	0:00:06	0:00:04
FS16	0:00:02	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01
FS17	0:00:20	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:04	0:00:02
FS18	0:00:06	0:00:00	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:00	0:00:02	0:00:01
FS19	0:00:02	0:00:00	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01
FS20	0:02:37	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:02	0:00:01	0:00:05	0:00:03

When considering runtime, reduction in features and recall in Figure 11 below, we have determined that FS20 A6 is the better algorithm to use for this dataset.

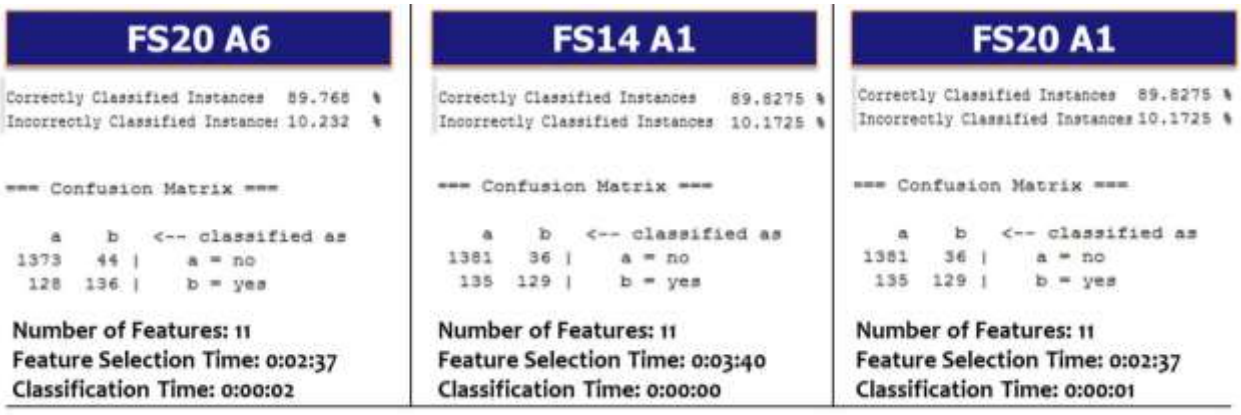

Figure 11. Top Three Clean Clinical.

5.7 Clean Medical Dataset

Table 24 displays the results of the feature selection process on the clean medical dataset. Based solely on feature reduction, FS16 was the top ranking performer by reducing the features down to a feature set of 4. FS6 was the lowest ranking performer by only reducing the features down to a feature set of 170.

Table 24

Clean Medical Feature Set

Clinical Feature Set					
Set	**Attribute Evaluator**	**Search Method**	**Classifier**	**# of Feat.**	**Selected Feature List**
FS1	Cfs Subset Eval	BestFirst		13	43, 105, 248, 285, 320, 322, 338, 404, 411, 414, 416, 417, 418
FS2	Cfs Subset Eval	Genetic Search		84	39, 53, 54, 64, 70, 73, 76, 81, 86, 89, 90, 97, 104, 105, 108, 113, 120, 123, 124, 125, 130, 139, 149, 151, 152, 157, 159, 164, 174, 177, 181, 187, 188, 195, 196, 199, 201, 206, 208, 209, 213, 224, 226, 227, 234, 240, 244, 247, 250, 256, 257, 263, 264, 272, 280, 282, 295, 296, 298, 299, 300, 301, 311, 318, 320, 332, 338, 354, 355, 356, 375, 381, 397, 398, 402, 404, 405, 406, 411, 412, 413, 414, 417, 418
FS3	Cfs SubsetEval	RankSearch		16	43, 91, 104, 105, 248, 256, 320, 327, 338, 394, 404, 411, 414, 416, 417, 418
FS4	Classifier SubsetEval	Genetic Search	OneR	39	39, 53, 64, 69, 70, 73, 86, 90, 95, 97, 105, 113, 123, 130, 152, 164, 195, 199, 206, 208, 224, 226, 234, 253, 257, 272, 279, 290, 298, 301, 313, 356, 379, 381, 406, 409, 411, 417
FS5	Consistency SubsetEval	BestFirst		11	105, 141, 245, 248, 411, 412, 413, 414, 416, 417, 418

Table 24

Cont.

FS6	Consistency SubsetEval	Genetic Search		170	39, 40, 42, 46, 53, 54, 56, 57, 58, 62, 63, 66, 67, 70, 71, 73, 77, 78, 79, 81, 82, 83, 84, 86, 88, 91, 93, 94, 97, 99, 101, 103, 105, 106, 111, 112, 113, 115, 117, 120, 121, 122, 123, 124, 125, 126, 128, 129, 132, 133, 138, 139, 141, 142, 143, 145, 146, 150, 151, 152, 153, 156, 160, 167, 169, 170, 171, 173, 176, 180, 182, 184, 185, 186, 187, 188, 193, 200, 203, 205, 207, 208, 212, 213, 214, 215, 216, 220, 223, 224, 230, 231, 233, 239, 240, 242, 243, 245, 247, 248, 251, 252, 254, 255, 256, 258, 259, 260, 262, 264, 272, 276, 277, 282, 284, 285, 299, 302, 305, 309, 310, 312, 321, 325, 333, 334, 335, 336, 337, 338, 344, 347, 350, 351, 356, 357, 362, 363, 364, 365, 366, 369, 370, 371, 372, 373, 374, 375, 382, 383, 384, 385, 387, 390, 391, 392, 395, 397, 398, 399, 403, 405, 406, 410, 411, 412, 413, 415, 416, 418
FS7	Consistency SubsetEval	Linear Forward Selection		11	49, 105, 285, 320, 338, 357, 411, 414, 416, 417, 418
FS8	Consistency SubsetEval	RankSearch		82	43, 49, 53, 54, 63, 64, 69, 88, 91, 104, 105, 112, 120, 122, 123, 124, 126, 129, 131, 135, 139, 141, 145, 148, 157, 159, 162, 163, 164, 171, 172, 191, 207, 209, 218, 226, 229, 230, 245, 248, 249, 250, 251, 254, 256, 262, 275, 279, 282, 285, 296, 297, 298, 311, 318, 320, 321, 322, 324, 325, 327, 334, 335, 337, 338, 339, 355, 357, 359, 375, 384, 394, 400, 404, 409, 411, 412, 413, 414, 416, 417, 418
FS9	Consistency SubsetEval	Subset SizeForward Selection		9	49, 105, 285, 357, 411, 414, 416, 417, 418

Table 24

Cont.

FS10	Filtered SubsetEval	BestFirst		10	105, 285, 320, 338, 404, 411, 414, 416, 417, 418
FS11	Filtered SubsetEval	Genetic Search		84	39, 53, 54, 64, 70, 73, 76, 81, 86, 89, 90, 97, 104, 105, 108, 113, 120, 123, 124, 125, 130, 139, 149, 151, 152, 157, 159, 164, 174, 177, 181, 187, 188, 195, 196, 199, 201, 206, 208, 209, 213, 224, 226, 227, 234, 240, 244, 247, 250, 256, 257, 263, 264, 272, 280, 282, 295, 296, 298, 299, 300, 301, 311, 318, 320 ,332, 338, 354, 355, 356, 375, 381, 397, 398, 402, 404, 405, 406, 411, 412, 413, 414, 417, 418
FS12	Filtered SubsetEval	Rank Search		10	91, 104, 256, 320, 327, 338, 404, 411, 414, 417
FS13	Wrapper SubsetEval	BestFirst	Naïve Bayes	9	41, 45, 145, 172, 228, 285, 334, 411, 416
FS14	Wrapper SubsetEval	BestFirst	Bagging	8	46, 179, 244, 381, 402, 409, 411, 415
FS15	Wrapper SubsetEval	Genetic Search	Bagging	169	39, 42, 43, 45, 49, 50, 51, 55, 58, 61, 63, 65, 66, 67, 68, 70, 71, 73, 77, 78, 79, 80, 82, 86, 88, 89, 91, 94, 97, 99, 100, 101, 102, 105, 109, 110, 111, 112, 113, 115, 116, 119, 121, 127, 131, 132, 137, 139, 141, 142, 145, 147, 149, 150, 152, 153, 155, 156, 157, 163, 164, 166, 167, 170, 173, 174, 176, 177, 181, 183, 185, 186, 187, 188, 189, 197, 198, 200, 202, 203, 204, 205, 220, 223, 224, 226, 228, 231, 233, 236, 237, 238, 242, 247, 251, 252, 253, 255, 256, 257, 260, 262, 264, 265, 267, 270, 272, 273, 274 277, 278, 279, 281, 282, 284, 285, 286, 288, 289, 290, 294, 298, 301, 305, 312, 313, 316, 321, 324, 33,3 339, 340, 341, 345, 346, 349, 353, 357, 360, 361, 363, 364, 365, 366, 367, 371, 375, 380, 382, 383, 384, 387, 388, 390, 391, 392, 395, 398, 399 403, 404, 405, 406, 409, 411, 413, 415, 416 417

Table 24

Cont.

FS16	Wrapper SubsetEval	Greedy Stepwise	Naïve Bayes	4	130, 145, 285, 411
FS17	Wrapper SubsetEval	Linear Forward Selection	Naïve Bayes	9	104, 105, 124, 285, 327, 337, 338, 411, 417
FS18	Wrapper SubsetEval	RankSearch	Naïve Bayes	10	91, 104, 256, 320, 327, 338, 404, 411, 414, 417
FS19	Wrapper SubsetEval	Subset SizeForward Selection	Naïve Bayes	5	104, 105, 285, 411, 417
FS20	Wrapper SubsetEval	Subset SizeForward Selection	Bagging	9	120, 276, 282, 285, 320, 404, 411, 416, 417

Table 25 exhibits the results of the classification process on the clean medical dataset. We calculated the accuracy (AC), precision (PR), recall (RC) and F-measure (FM). FS1 A6, FS5 A3 and FS17 A4 (97.0256%) were equal in accuracy, FS4 A8 (0.969) in precision FS17 A2 (0.92) for recall and FS5 A3 and FS17 A4 for F-Measure (0.903). Based solely on AC and FM, FS17 A4 is the top ranking performer.

The clean medical time elapsed table featured in Table 26 shows that based on the sum of the time elapsed for feature selection and classification, FS9 (0:00:22) had the best performance and FS15 (0:59:26) had the poorest performance.

Table 25

Clean Medical: Precision, Recall and F-Measure Rates

Features		**Classifiers**								
Name	**Metric**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	AC	96.2522%	96.6092%	96.2522%	96.4307%	95.9548%	97.0256%	96.4902%	96.3712%	96.2522%
	PR	0.894	0.9	0.885	0.889	0.877	0.931	0.908	0.925	0.888
	RC	0.864	0.883	0.875	0.883	0.864	0.875	0.864	0.837	0.871
	FM	0.879	0.891	0.88	0.886	0.87	0.902	0.885	0.879	0.88
FS2	AC	96.3712%	93.2183%	96.0738%	94.5271%	95.8358%	96.9661%	96.0143%	ERROR	96.2522%
	PR	0.944	0.762	0.899	0.795	0.862	0.946	0.883		0.897
	RC	0.86	0.826	0.845	0.879	0.875	0.856	0.86		0.86
	FM	0.882	0.793	0.871	0.835	0.868	0.899	0.871		0.878
FS3	AC	96.0738%	95.8358%	96.4902%	96.2522%	95.8953%	96.6092%	96.4307%	95.8358%	96.4307%
	PR	0.893	0.873	0.893	0.9	0.871	0.916	0.902	0.918	0.905
	RC	0.852	0.86	0.883	0.856	0.867	0.864	0.867	0.807	0.864
	FM	0.872	0.866	0.888	0.878	0.869	0.889	0.884	0.859	0.884
FS4	AC	96.1927%	93.4563%	96.2522%	95.3004%	95.8358%	95.6573%	95.8953%	95.122%	95.8953%
	PR	0.913	0.806	0.921	0.849	0.862	0.863	0.874	0.969	0.901
	RC	0.837	0.769	0.833	0.852	0.875	0.86	0.864	0.712	0.83
	FM	0.874	0.787	0.875	0.851	0.868	0.861	0.869	0.821	0.864
FS5	AC	96.3712%	96.6092%	97.0256%	96.4902%	95.8358%	96.8471%	96.5497%	96.2522%	96.3117%
	PR	0.901	0.916	0.921	0.905	0.862	0.924	0.909	0.943	0.891
	RC	0.864	0.864	0.886	0.867	0.875	0.871	0.867	0.811	0.871
	FM	0.882	0.889	0.903	0.886	0.868	0.897	0.888	0.872	0.881
FS6	AC	96.6686%	86.5556%	96.6092%	94.4676%	96.2522%	96.9661%	96.8471%	ERROR	95.8953%
	PR	0.903	0.711	0.9	0.798	0.885	0.928	0.92		0.871
	RC	0.883	0.242	0.883	0.867	0.875	0.875	0.875		0.867
	FM	0.893	0.362	0.891	0.831	0.88	0.901	0.897		0.869

Table 25

Cont.

FS7	AC	96.4307%	96.2522%	95.7763%	96.1333%	95.9548%	96.6092%	96.4902%	96.6092%	96.1333%
	PR	0.905	0.879	0.864	0.867	0.877	0.919	0.908	0.94	0.884
	RC	0.864	0.883	0.867	0.89	0.864	0.86	0.864	0.837	0.867
	FM	0.884	0.881	0.866	0.879	0.87	0.888	0.885	0.886	0.876
FS8	AC	96.6686%	93.6942%	96.6686%	93.5158%	95.8358%	96.9066%	96.5497%	60.2023%	96.3117%
	PR	0.913	0.797	0.919	0.751	0.862	0.938	0.909	0.904	0.885
	RC	0.871	0.803	0.864	0.879	0.875	0.86	0.867	0.688	0.879
	FM	0.891	0.8	0.891	0.81	0.868	0.897	0.888	0.782	0.882
FS9	AC	96.4307%	96.0143%	96.1927%	96.2522%	95.9548%	96.9661%	96.4902%	96.6092%	96.1927%
	PR	0.902	0.866	0.885	0.865	0.877	0.921	0.908	0.94	0.882
	RC	0.867	0.883	0.871	0.902	0.864	0.883	0.864	0.837	0.875
	FM	0.884	0.874	0.878	0.883	0.87	0.901	0.885	0.886	0.878
FS10	AC	96.3117%	96.3712%	96.1927%	96.2522%	95.9548%	96.9661%	96.4902%	96.1927%	96.4307%
	PR	0.898	0.886	0.888	0.879	0.877	0.931	0.908	0.924	0.892
	RC	0.864	0.883	0.867	0.883	0.864	0.871	0.864	0.826	0.879
	FM	0.88	0.884	0.877	0.881	0.87	0.9	0.885	0.872	0.885
FS11	AC	96.3712%	93.2183%	96.0738%	94.5271%	95.8358%	96.9661%	96.0143%	ERROR	96.2522%
	PR	0.904	0.762	0.899	0.795	0.862	0.946	0.883		0.897
	RC	0.86	0.826	0.845	0.879	0.875	0.856	0.86		0.86
	FM	0.882	0.793	0.871	0.835	0.868	0.899	0.871		0.878
FS12	AC	95.8953%	95.8953%	95.7763%	96.1333%	95.8358%	95.4194%	95.9548%	95.122%	96.3117%
	PR	0.892	0.885	0.918	0.913	0.862	0.87	0.871	0.96	0.935
	RC	0.841	0.848	0.803	0.833	0.875	0.833	0.871	0.72	0.822
	FM	0.865	0.867	0.857	0.871	0.868	0.851	0.871	0.823	0.875
FS13	AC	96.3712%	96.7876%	96.4902%	96.9066%	95.8953%	96.4307%	96.3712%	96.4307%	96.4902%
	PR	0.889	0.917	0.89	0.902	0.876	0.905	0.898	0.915	0.89
	RC	0.879	0.875	0.886	0.902	0.86	0.864	0.867	0.852	0.886
	FM	0.884	0.895	0.888	0.902	0.868	0.884	0.882	0.882	0.888

Table 25

Cont.

FS14	AC	96.4902%	94.3486%	96.3712%	94.765%	95.8358%	96.1333%	96.3117%	95.4194%	96.2522%
	PR	0.912	0.851	0.911	0.841	0.862	0.896	0.895	0.952	0.907
	RC	0.86	0.777	0.852	0.822	0.875	0.852	0.867	0.746	0.848
	FM	0.885	0.812	0.881	0.831	0.868	0.874	0.881	0.837	0.877
FS15	AC	96.5497%	89.1136%	96.9066%	95.003%	96.2522%	96.7281%	96.7876%	ERROR	96.3117%
	PR	0.893	0.694	0.921	0.808	0.885	0.92	0.92		0.888
	RC	0.886	0.549	0.879	0.894	0.875	0.867	0.871		0.875
	FM	0.89	0.613	0.899	0.849	0.88	0.893	0.895		0.882
FS16	AC	96.3712%	96.2522%	96.2522%	96.6686%	95.8358%	95.0625%	96.3117%	93.7537%	96.6092%
	PR	0.889	0.894	0.868	0.882	0.862	0.849	0.888	0.954	0.897
	RC	0.879	0.864	0.898	0.909	0.875	0.833	0.875	0.633	0.886
	FM	0.884	0.879	0.883	0.896	0.868	0.841	0.882	0.761	0.891
FS17	AC	96.1333%	96.3117%	96.0738%	97.0256%	95.8358%	96.1333%	95.8358%	94.2891%	96.6092%
	PR	0.9	0.856	0.878	0.925	0.862	0.887	0.865	0.947	0.888
	RC	0.848	0.92	0.871	0.883	0.875	0.864	0.871	0.674	0.898
	FM	0.873	0.887	0.875	0.903	0.868	0.875	0.868	0.788	0.893
FS18	AC	95.8953%	95.8953%	95.7763%	96.1333%	95.8358%	95.4194%	95.9548%	95.122%	96.3117%
	PR	0.892	0.885	0.918	0.913	0.862	0.87	0.871	0.96	0.935
	RC	0.841	0.848	0.803	0.833	0.875	0.833	0.871	0.72	0.822
	FM	0.865	0.867	0.857	0.871	0.868	0.851	0.871	0.823	0.875
FS19	AC	96.1927%	96.4902%	96.0738%	96.8471%	95.8358%	96.1927%	95.8358%	94.5271%	96.4307%
	PR	0.903	0.873	0.878	0.901	0.862	0.882	0.865	0.943	0.889
	RC	0.848	0.909	0.871	0.898	0.875	0.875	0.871	0.693	0.883
	FM	0.875	0.891	0.875	0.899	0.868	0.878	0.868	0.799	0.886
FS20	AC	96.6092%	96.0738%	96.1333%	96.3712%	95.8953%	96.6092%	96.3117%	96.3712%	96.0738%
	PR	0.9	0.884	0.89	0.901	0.876	0.912	0.901	0.951	0.899
	RC	0.883	0.864	0.86	0.864	0.86	0.867	0.86	0.811	0.845
	FM	0.891	0.874	0.875	0.882	0.868	0.889	0.88	0.875	0.871

Table 26

Clean Medical Time Elapsed

Features		Classifiers (Time Elapsed)								
Name	**Time**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
FS1	0:00:00	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:16	0:00:02
FS2	0:00:08	0:00:03	0:00:03	0:00:03	0:00:01	0:00:07	0:00:05	0:00:13	0:00:43	0:00:15
FS3	0:00:09	0:00:01	0:00:02	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:17	0:00:02
FS4	0:00:15	0:00:02	0:00:01	0:00:01	0:00:01	0:00:03	0:00:05	0:00:06	0:00:16	0:00:03
FS5	0:00:05	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:02	0:00:22	0:00:02
FS6	0:00:03	0:00:06	0:00:04	0:00:08	0:00:01	0:00:13	0:00:05	0:00:26	0:01:58	0:00:23
FS7	0:00:02	0:00:00	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:15	0:00:01
FS8	0:00:03	0:00:03	0:00:02	0:00:03	0:00:01	0:00:07	0:00:03	0:00:13	0:00:50	0:00:09
FS9	0:00:02	0:00:00	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:14	0:00:02
FS10	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:01	0:00:01	0:00:02	0:00:15	0:00:01
FS11	0:00:08	0:00:03	0:00:03	0:00:04	0:00:01	0:00:07	0:00:05	0:00:13	0:00:42	0:00:11
FS12	0:00:08	0:00:01	0:00:00	0:00:01	0:00:00	0:00:01	0:00:01	0:00:02	0:00:09	0:00:01
FS13	0:07:55	0:00:01	0:00:00	0:00:00	0:00:00	0:00:01	0:00:02	0:00:02	0:00:10	0:00:01
FS14	0:27:31	0:00:01	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:09	0:00:02
FS15	0:56:05	0:00:06	0:00:06	0:00:08	0:00:01	0:00:13	0:00:06	0:00:25	0:01:53	0:00:23
FS16	0:01:28	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:04	0:00:01
FS17	0:01:19	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:02	0:00:02	0:00:06	0:00:01
FS18	0:16:45	0:00:00	0:00:00	0:00:00	0:00:00	0:00:01	0:00:00	0:00:02	0:00:00	0:00:02
FS19	0:00:20	0:00:01	0:00:01	0:00:00	0:00:00	0:00:01	0:00:01	0:00:01	0:00:05	0:00:01
FS20	0:04:28	0:00:01	0:00:01	0:00:01	0:00:00	0:00:01	0:00:01	0:00:01	0:00:10	0:00:02

When considering runtime, reduction in features and recall in Figure 12 below, we have determined that FS5 A3 is the better algorithm to use for this dataset.

Figure 12. Top Three Clean Medical.

There is a vast amount of results from our experiment; however, after analyzing all of the experiments performed, and considering overall accuracy, runtime, reduction in features and recall, we have determined that FS16 A9 of the raw full dataset is the better algorithm to use for this problem. We gave more weight to recall than precision because if BV goes undiagnosed and therefore untreated, it can cause very harmful effects for women. On the other hand, if a woman is diagnosed as BV positive but in reality is negative (false positive); the consequence will merely be taking an inexpensive anti-biotic which will cause little to no harm for women. While the difference between the false negative outcomes for this data seems minimal, approximately 1 million pregnant women are diagnosed with BV yearly. This fact highlights the significance of the results. Untreated BV in women increases the chance of pre-term labor and pelvic inflammatory disease (PID).

The final feature set is constructed of features 2, 7, 29, 33, 34, 104, 130, 228, 241, 285, 295, 411, 416 and 417. The feature names are displayed in Table 27.

Table 27

Final Feature Names

P_ID	VAG_ODOR	Jonquetella	corGroup6
P_ID.1	Bacteria.10	Mycoplasma	corGroup7
TOB_USE	Bifidobacterium	Odoribacter	
PH_GLOVE	Haemophilus	corGroup1	

CHAPTER 6

Conclusion and Future Research

In this thesis we conducted experiments using twenty different feature selection algorithms and analyzed the time taken by each of them. We used nine classification algorithms using the selected features in the previous step and studied the precision and recall of BV disease. Six additional datasets were created by conducting experiments on the data subsets and cleaning the data. We compared the accuracy, precision, recall, F-measure and runtime for each feature selection and classification combination. Some of the features were not present in many women and had minimal effect on the overall outcome. On the other hand, the features which were present in all the women have significant effect on the classification results.

As the medical community continues to embrace machine learning approaches in order to discover ways to advance clinical studies, lower the cost of medication and aid physicians with expedited and more accurate diagnoses, research in this area must continue. Our future work will be dedicated towards conducting experiments with additional feature selection and classification algorithms, adjust the seed values of the deterministic algorithms to force randomization and manipulate the default settings on top performing algorithms. All of this will be done in an effort to find the optimal algorithm combinations to increase reduce features and increase accuracy for our BV dataset.

References

Akoglu, L., & Faloutsos, C. (2013). *Anomaly, event, and fraud detection in large network datasets.* Paper presented at the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy.

Al-Shayea, Q. K. (2011). Artificial Neural Networks in Medical Diagnosis. *International Journal of Computer Science Issues (IJCSI), 8*(2), 150-150. doi: 10.1605/01.301-0022274893.2013

Alexander, F. J. (2013). Machine Learning. *Computing in Science & Engineering, 15*(5), 9-11. doi: 10.1109/MCSE.2013.107

Alzheimer's Association (2014). What is Alzheimer's? Retrieved March, 2014, from http://www.alz.org/alzheimers_disease_what_is_alzheimers.asp

Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *International Journal of Engineering Science and Technology, 2*(10), 5370-5376.

Anu, J., Agrawal, R., & Bhattacharya, S. (in-press). *Ranking Tourist Attractions using Time Series GPS Data of Cabs*. Paper presented at the IEEE SoutheastCon 2014, Lexington, KY.

Aruna, S., Rajagopalan, D. S., & Nandakishore, L. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology (CS&IT), 2*, 37-45.

Beck, D., & Foster, J. A. (2014). Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PloS one, 9*(2), e87830.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Brotman, R. M. (2011). Vaginal microbiome and sexually transmitted infections: an epidemiologic perspective. *The Journal of clinical investigation, 121*(12), 4610-4617.

Cai, Y.-D., Feng, K.-Y., Lu, W.-C., & Chou, K.-C. (2006). Using LogitBoost classifier to predict protein structural classes. *Journal of Theoretical Biology, 238*(1), 172-176.

Chue-Poh, T. A. N., Ka-Sing, L. I. M., & Weng-Kin, L. A. I. (2008). Multi-Dimensional Features Reduction of Consistency Subset Evaluator on Unsupervised Expectation Maximization Classifier for Imaging Surveillance Application. *International Journal of Image Processing, 2*(1), 18-26.

Cleary, J. G., & Trigg, L. E. (1995). *K*: An Instance-based Learner Using an Entropic Distance Measure*. Paper presented at the 12th International Conference on Machine Learning, Lake Tahoe, CA..

David, S. K., Saeb, A. T., & Al Rubeaan, K. (2013). Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics. *Computer Engineering and Intelligent Systems, 4*(13), 28-38.

Domingos, P. (2012). A few useful things to know about machine learning (Vol. 55, pp. 78-87). New York: ACM.

Dua, S. (2011). *Data Mining and Machine Learning in Cybersecurity*. Boca Raton: CRC Press.

Emanet, N., Öz, H. R., Bayram, N., & Delen, D. (2014). A comparative analysis of machine learning methods for classification type decision problems in healthcare. *Decision Analytics, 1*(1), 6.

Filippo, A., Alberto, L., Eladia Maria, P.-M., Petr, V., & Josef, H. (2013). Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine, 11*(2), 47-58.

Fredricks, D. N., Fiedler, T. L., Thomas, K. K., Oakley, B. B., & Marrazzo, J. M. (2007). Targeted PCR for detection of vaginal bacteria associated with bacterial vaginosis. *Journal of clinical microbiology, 45*(10), 3270-3276.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics, 28*(2), 337-407.

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Michael J. Blaha, . . . Turner, M. B. (2013). Heart Disease and Stroke Statistics--2014 Update: A Report From the American Heart Association. *Circulation, 123*(3), e28-e292. doi: 10.1161/01.cir.0000441139.02102.80

Gutlein, M., Frank, E., Hall, M., & Karwath, A. (2009). *Large-scale attribute selection using wrappers.* Paper presented at the IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. (Doctoral dissertation), The University of Waikato.

Hosseinzadeh, F., KayvanJoo, A. H., Ebrahimi, M., & Goliaei, B. (2013). Prediction of lung tumor types based on protein attributes by machine learning algorithms. *SpringerPlus, 2*(1), 1-14. doi: 10.1186/2193-1801-2-238

Jim, R. O. (1996). A machine-learning approach to automated negotiation and prospects for electronic commerce. *Journal of Management Information Systems, 13*(3), 83.

John, G. H., & Langley, P. (1995). *Estimating continuous distributions in Bayesian classifiers*. Paper presented at the Eleventh conference on Uncertainty in artificial intelligence, Montreal, Quebec.

Kancherla, K., & Mukkamala, S. (2013). *Early lung cancer detection using nucleus segementation based features*.

Kantardzic, M. (2003). *Data mining: concepts, models, methods, and algorithms*. Piscataway, NJ: IEEE Press.

Kiritchenko, S., & Matwin, S. (2011). *Email classification with co-training*. Paper presented at the 2011 Conference of the Center for Advanced Studies on Collaborative Research, Toronto, Ontario, Canada.

National Heart, Lung and Blood Institute (2014). Asthma. from http://www.nhlbi.nih.gov/health/health-topics/topics/asthma/

Osareh, A., & Shadgar, B. (2010). *Machine learning techniques to diagnose breast cancer.* Paper presented at the HIBIT 2010 : 5th International Symposium on Health Informatics and Bioinformatics, Belek, Antalya, Turkey.

Peng, Y., Kou, G., Ergu, D., Wu, W., & Shi, Y. (2012). An Integrated Feature Selection and Classification Scheme. *Studies in Informatics and Control, 21*(3), 241-248.

Prasad, B. D. C. N., Prasad, P. E. S. N. K., & Sagar, Y. (2011). A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma) (Vol. 131, pp. 570-576). Berlin, Heidelberg: Springer Berlin Heidelberg.

Rajarajeswari, S., & Somasundaram, K. (2012). An Empirical Study on feature selection for Data Classification. *International Journal of Advanced Computer Research, 2*(5), 111-115.

Rangari Amit, A., Parmjit, S., & Sharma, V. (2013). Comparison of the amsel's composite clinical criteria and nugent's criteria for diagnosis of bacterial vaginosis:-a step towards

preventing mis-diagnosis. *Journal of Advance Researches in Biological Sciences, 5*(1), 37-44.

Ravel, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L., . . . Russell, J. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences of the United States of America, 108 Suppl 1*(11), 4680-4687.

Salama, G. I., Abdelhalim, M., & Zeid, M. A.-e. (2012). Breast Cancer Diagnosis on Three Different Datasets using Multi-classifiers. *International Journal of Computer and Information Technology, 1*(1), 36-43.

Savage, N. (2012). Better medicine through machine learning (Vol. 55, pp. 17-19). New York: ACM.

Shelton, J., Alford, A., Small, L., Leflore, D., Williams, J., Adams, J., . . . Ricanek, K. (2012). *Genetic & evolutionary biometrics: feature extraction from a machine learning perspective.* Paper presented at the Southeastcon, 2012 Proceedings of IEEE.

Shelton, J., Dozier, G., Bryant, K., Adams, J., Popplewell, K., Abegaz, T., . . . Ricanek, K. (2011). *Genetic based LBP feature extraction and selection for facial recognition.* Paper presented at the ACM SE 2011.

Sherrod, P. (2014). DTREG. Retrieved February, 2014, from http://www.dtreg.com/rbf.htm

Sindhu, S. S., Geetha, S., & Kannan, A. (2012). Decision tree based light weight intrusion detection using a wrapper approach. *Expert Systems With Applications, 39*(1), 129-141.

Society, A. C. (2014). What are the key statistics about lung cancer? Retrieved February 13, 2014, from http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-key-statistics

Srinivasan, S., Marrazzo, J. M., Fredricks, D. N., Hoffman, N. G., Morgan, M. T., Matsen, F. A., . . . Bumgarner, R. (2012). Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PloS one, 7*(6), e37818.

Sujatha, S., Martin, T. M., Congzhou, L., Frederick, A. M., Noah, G. H., Tina, L. F., . . . David, N. F. (2013). More Than Meets the Eye: Associations of Vaginal Bacteria with Gram Stain Morphotypes Using Molecular Phylogenetic Analysis: e78633. *PloS one, 8*(10).

Taher, F., & Sammouda, R. (2011). *Lung cancer detection by using artificial neural network and fuzzy clustering methods.* Paper presented at the 2011 IEEE GCC Conference and Exhibition (GCC), Dubai.

Wagstaff, K. (2012). Machine learning that matters. *arXiv preprint arXiv:1206.4656*.

Wang, Y., Makedon, F. S., Ford, J. C., & Pearlman, J. (2005). HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics (Oxford, England), 21*(8), 1530-1537.

Williams, J. A., Weakley, A., Cook, D. J., & Schmitter-Edgecombe, M. (2013). *Machine Learning Techniques for Diagnostic Differentiation of Mild Cognitive Impairment and Dementia.* Paper presented at the Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, WA.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques, third edition*. Burlington, Mass: Morgan Kaufmann Publishers.

World Health Organization (2014). The top 10 causes of death. Retrieved February 15, 2014, from http://www.who.int/mediacentre/factsheets/fs310/en/index2.html

Yasuo, Y., Fukai, T., Hidetaka, A., Takashi, Y., Chiaki, T., Ohara, T., . . . Hiroshi, H. (2013). Computer-aided differential diagnosis system for Alzheimer's disease based on machine learning with functional and morphological image features in magnetic resonance imaging. *Journal of Biomedical Science and Engineering, 6*(11), 1090.

Yau, T. M. S., & Othman, M. F. b. (2006). *Comparison of Different Classification Techniques Using WEKA for Breast Cancer*. Paper presented at the 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006 Berlin, Heidelberg.