University of the Philippines Manila
College of Arts and Sciences
Department of Physical Sciences and Mathematics

BOSOM Calculator: A Breast Cancer Outcome - Survival Online Measurement Calculator using Data Mining and Predictive Modeling on SEER data

A special problem in partial fulfillment
of the requirements for the degree of
Bachelor of Science in Computer Science

Submitted by:

GilTroy Paular Meren

April 2014

Permission is given for the following people to have access to this SP:

Available to the general public	Yes
Available only after consultation with author/SP adviser	No
Available only to those bound by confidentiality agreement	No

ACCEPTANCE SHEET

The Special Problem entitled "BOSOM Calculator:
A Breast Cancer Outcome - Survival Online Measurement Calculator using Data Mining and Predictive Modeling on SEER data" prepared and submitted by GilTroy Paular Meren in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science has been examined and is recommended for acceptance.

Vincent Peter C. Magboo, M.D., M.Sc.
Adviser

EXAMINERS:

	Approved	**Disapproved**
1. Gregorio B. Baes, Ph.D. (*candidate*)	_____	_____
2. Avegail D. Carpio, M.Sc.	_____	_____
3. Richard Bryann L. Chua, M.Sc.	_____	_____
4. Aldrich Colin K. Co, M.Sc. (*candidate*)	_____	_____
5. Ma. Sheila A. Magboo, M.Sc.	_____	_____
6. Geoffrey A. Solano, Ph.D. (*candidate*)	_____	_____
7. Bernie B. Terrado, M.Sc. (*candidate*)	_____	_____

Accepted and approved as partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science.

Ma. Sheila A. Magboo, M.Sc.
Unit Head
Mathematical and Computing Sciences Unit
Department of Physical Sciences and Mathematics

Marcelina B. Lirazan, Ph.D.
Chair
Department of Physical Sciences and Mathematics

Alex C. Gonzaga, Ph.D., Dr.Eng.
Dean
College of Arts and Sciences

Abstract

Illnesses of high mortality rate such as breast cancer elicit questions related to the patient's time left to live. The common methods used to arrive at an estimate include comparing the patient's health condition to previous medical records and treatments, referral to statistically-computed survival rates based from historical records, or consulting another breast cancer expert.

The application of data mining on medical records to create predictive models for cancer survivability has been proven to hold significant accuracy by numerous scientific and applied researches throughout the years. Agrawal et al.'s "Lung Cancer Outcome Calculator" provides a framework for developing a predicted survival calculator for different cancers based on a patient's health condition.

This research aims to develop the Breast Cancer Outcome - Survival Online Measurement Calculator (BOSOM Calculator), an online application that takes a patient's clinical cancer data to give a predicted cancer survival based on a dataset from the Surveillance, Epidemiology, and End Results Program (SEER).

Keywords: breast cancer, data mining, medical records, survivability, survival calculator, predictive modeling, online, SEER

Contents

List of Figures

List of Tables

List of Program Codes

I. Introduction

A. Background of the Study

Breast cancer is one of the leading cancer types worldwide. In 2008, the Global Cancer (GLOBOCAN) program reported that 1.38 million people or 22.9% of the total cancer cases were diagnosed with breast cancer. Around 458,503 death records were also noted [1]. From the comparison of the incidence and mortality rates of the top cancer types that affected people worlwide in 2008 indicated in Fig. 1, breast cancer ranked second and third highest in incidence and mortality rates respectively [1].

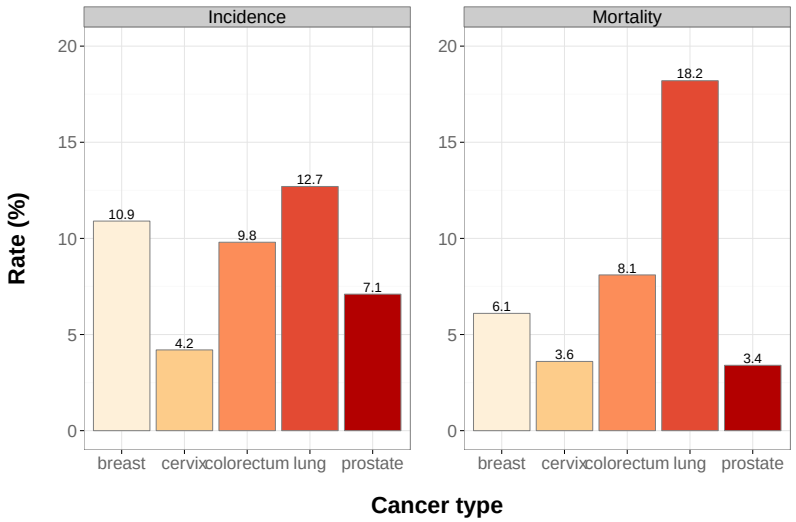


Figure 1: Graph of incidence and mortality rates of the top five cancer types worldwide in 2008 (GLOBOCAN)

Records from the Philippines show that breast cancer incidence was at 31.9% or 11,524 new cases and mortality of 11.9% or 4,085 deaths in 2008 [1]. Laudico et al. reported that it is the third cause of cancer deaths in both sexes (8%) and the first in women (18%). This shows the vulnerability of Filipinos to breast cancer especially the female population.

Patients who have breast cancer and other conditions known to have low survival rate often question the length of time they have left to live. In response, doctors present their prediction coming from research methods such as comparing the patient's health condition to past records to find similarities to patients with the condition, and through consultation with fellow doctors for their findings on the case being investigated [2]. The two parties (doctor and patient) may discuss possible medical readjustments on a number of factors including the estimated survival time [3–5].

As mentioned, patient medical records are significant sources of information in this field by aiding health care providers a basis of correlation of present to previous occurrences of a condition. Most cancer cases elicit numerous admissions to hospitals and consultations to doctors and these produce records of varied states of data quality and completeness but nevertheless valuable medical literature.

Data mining is about discovering patterns and finding relationships of information from existing records [6]. Past studies have demonstrated its potential by means of application to medical records and common examples are diagnosis and prognosis of a disease's recurrence using models created from medical records. Agrawal et al.'s online "Lung Cancer Outcome Calculator" (LCOC) was developed by creating predictive models from lung cancer cases provided by the Surveillance, Epidemiology, and End Results Program (SEER) using the Waikato Environment for Knowledge Analysis (WEKA) machine learning software. The study developed two sets of lung cancer predictive models — the first using more than 60 variables chosen from preprocessing and the second is a smaller subset of only 13 variables determined through their ability to predict in the context of the dataset. Both set of models were developed by training five classifiers namely alternating decision tree, J48 decision tree, random forest, `LogitBoost`, and random subspace. The calculator uses an ensemble voting method on the results of the aforementioned models to calculate the prediction. Tests showed that both performed with comparable high accuracy as seen in Fig. 2 [3].

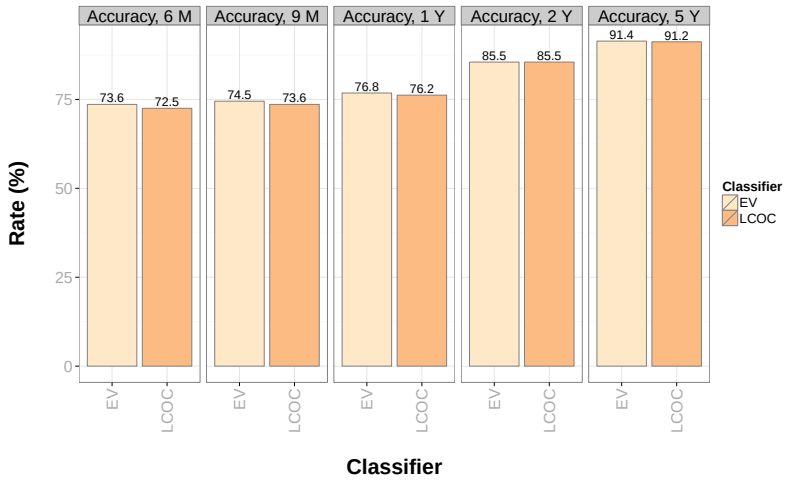


Figure 2: Graphs of the predictive accuracies per outcome variable of ensemble voting and the "Lung Cancer Outcome Calculator" from Agrawal et al.'s study

Several prognostic applications exist online for different cancer types. One of these is the LCOC that predicts survival for five time periods [3]; and the United Kingdom-based PREDICT Tool that uses ten breast cancer-specific variables in order to generate five and ten-year survival predictions [5]. Both of these systems have shown the application of data mining on prediction in medicine by using historical medical records.

B. Statement of the Problem

Cancer survival prediction for diagnosed breast cancer patients is commonly determined by their doctors from research and past experience. Unfortunately these require time and effort, and are only reliable to a clinical level [2].

In relation, general practice involves the following investigative methods:

1. Investigation of previous cases that exhibit the same nature as the current case to find similarities in conditions that might help in approximating a survival time [2];

2. Referencing statistically-computed cancer survival rates from a population. An example is a breast cancer survival rate table based on cancer stage as seen in Tables 1 and 2 [4, 7]; and/or

3. Consultation with another physician for their interpretation of the patient's clinical data. This serves as additional information on possible ways to relieve the condition [2].

There are currently two general cancer registries in the Philippines that provide native records for oncology studies and related research: the Philippine Cancer Society - Manila Cancer Registry (PCS-MCR) and Department of Health - Rizal Cancer Registry (DOH-RCR). The two organizations collect cases from the National Capital Region (NCR) [8]. Public access to the records is limited and these are currently paper-based documents. A 2009 study comparing the cancer survival of Filipinos (using data from PCS-MCR and DOH-RCR) to Filipino-Americans and Americans (SEER data) revealed that the survival rates of Filipinos are inferior to the other two. It was stated that the records collected by PCS-MCR and DOH-RCR are composed mostly late stages of cancer and they attributed this to cancer's negative social stigma [9], and that the residents in the United States of America have easier access to better cancer health care system [8].

A lung cancer survival study was able to develop an online calculator that uses clinical data as input and survival percentage as output [3]. Agrawal et al.'s study provides a design framework for a prediction calculator for other cancer types and their method takes advantage of algorithms (or "classifiers") to find connections between patient medical records and a survival outcome. Thus the aim of this research is to create the BOSOM Calculator — a breast cancer outcome-survival online measurement tool that uses data mining and predictive modeling on records in order to provide survival estimates based on predetermined variables such as age at diagnosis and cancer stage.

C. Objectives of the Study

General Objectives

The research intends to produce the BOSOM application (refers to both the website and calculator collectively) for breast cancer patients who wish to know a predicted survival within several time periods given several personal attributes, mainly breast cancer-related, that will be processed by trained predictive models. This online tool will provide patients and doctors an estimated prediction to help in discussing possible treatment options, medication changes or to consider palliative care or hospice care if applicable [3, 7, 9].

The first objective is to create two predictive models — one for the preprocessed data and another for the online calculator; second is to develop the online calculator that predicts breast cancer survivability using the latter predictive model; and last is to test the application in the server environment to validate its intended purpose.

Specific Objectives

The BOSOM application's users has the following capabilities:

1. To view the BOSOM Calculator website and navigate through the different pages available.
2. To view the BOSOM Calculator the user must click a link in the home page.
3. To enter the patient's clinical cancer data into the calculator's form for submission to the predictive model.
4. To view a survival prediction report to the user based on their clinical data. It will consist their entered data, a table and vertical bar graph of the survival rate versus time.
5. To have an export feature of the survival prediction report in the form of a portable document format (PDF) document for the user.

6. To view the PDF document/survival prediction report if the user wants to.

7. To view a general information page dedicated to breast cancer that contains links to local and international institutions that provide professional health care, additional information and research data for interested users.

D. Significance of the Study

Most cancer types have high mortality rates [1,8,9] and patients of such illnesses are more or less reluctant but interested in knowing the time they have left after diagnosis. Doctors consult past medical cases collected over time in their respective hospitals and medical centers, and ask another expert for validation and additional opinion to come up with an informed survival prediction [2].

It is hard for a patient to know that they have a limited time to live as predicted by a doctor. This would give rise to more questions from the patient and their family and some of which are: "Are there ways to prolong their time?", "Will chemotherapy help them live longer?", or "Should we stop with the radiation therapy since it might not help anymore?". People will not be contented on one finding thus they seek for alternatives (also known as "second opinion") by consulting another specialist and trying other available treatment methods.

Unfortunately not all options are generally applicable to everyone. Lin reported that patients who undergo cancer treatment methodologies, e.g. radiation therapy and chemotherapy, are most likely to die compared to the illness itself. Different cancer types and stages in turn respond to different treatments and medications [7]. Taking risks is part of the treatment but it establishes a thin line between healing and death.

A survival prediction calculator would prove to be beneficial to both the patient and their physicians through its convenience of use and availability to provide survival estimates in the form of an online application. Its importance is not

limited to: first, a predictive finding by a breast cancer doctor can be evaluated with the use of the calculator. It will serve as a reference to ease the patients' worries and as a form of second opinion. Next is that the result could give way to consider other methods of treatment - a high survival risk could prompt for more aggressive treatments such as radiation therapy or surgery if financial, medical and peer support are available in hopes of gradual healing and prolonging life; on the other hand, palliative care may be considered to make the patient's life more comfortable and relieve their symptoms [7]. Either way, learning about cancer survivability will enable patients to evaluate their current lifestyle, prepare themselves for possible events and inform their families and colleagues on what could happen in the future.

In the case of doctors and health care providers involved, they have the ability to modify the patient's current medication regime based on the predicted survival. Their decision-making process in giving medical treatment and medication for their patients could be reevaluated to avoid taking risks based on past medical records and opt for more successful, relevant, and proven methods [2, 7].

Furthermore, the application of data mining in medicine presents a new perspective in studying diseases. Historical data are available and modeling using algorithms such as neural networks and decision trees are proven to hold statistical accuracy and reliability for prognosis and diagnosis of diseases [4, 10]. The patterns and rules generated from these historical data are helpful in obtaining an actual representation of records i.e. similar to a doctor's findings from research and investigation.

A phone inquiry with the PCS-MCR revealed that their patient records are accessible only with permission from a government agency and a licensed physician among other legal requirements. These records are still paper-based and conversion to electronic format could take a large amount of time of this research given the limited timeframe. In addition, a 2009 study stated that they needed to in-

clude several significant data from the records such as survival status and cause of death. These where individually gathered from local registry civil offices and families of the person in the record [8]. This shows the incompatibility of the local records to this study due to current limited resources. It was also identified by Redaniel et al. that the records from the PCS-MCR and DOH-RCR are composed mostly of records with later cancer stages and this was correlated to a low survival rate from the predictive models they used. In comparison with Filipinos and Americans in the USA, their respective SEER datasets were proven to hold higher accuracy due to the diversity of cancer cases and the access to better health facilities [8].

Despite the fact that the dataset that was be used is based on predominatly American population of the SEER Program, there are studies that confirm applicabilty of the data to other geographical areas. The UK's PREDICT Tool was created using data from England and the tool performed well in a Canada-based validation study [5]. Wishart et al. concluded that geographic location is not a significant factor in predicting cancer survival using a model. Another study on the relationship of breast cancer incidence and geographical location in the United States resulted to no observable bearing to connect the two. The length of solar exposure and amount of synthesized vitamin D were also noted in the research [11].

The application will be made available for free use online as well as its source code. Programmers today use public code repositories such as GitHub and BitBucket to host source code. These websites handle version control, has a capability to support multiple contributors and a feature for a wiki or similar documentation framework. The structure of such online repositories is open for public maintenance and contribution. Popular regularly maintained ones include the jQuery JavaScript library, CodeIgniter PHP framework and the Git itself. The BOSOM Calculator will benefit from such arrangement as other programmers, researchers and statisticians could improve the capabilities and structure and even make derivatives for other cancer types and diseases.

The BOSOM Calculator will help in providing survival prediction to Filipino breast cancer patients and their families in order to guide them in reevaluating their lifestyle and medication process alongside a cancer specialist's prognosis.

E. Scope and Limitations

1. SEER data

   (a) Only the data from 1973 to December 2010 were used in the preprocessing. This is the registered information in the database named "Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2012 Sub (1973-2010 varying)" [12].

   (b) The SEER data is collected from 18 key locations (based on the indicated registry set - SEER 18) in the United States of America namely Atlanta (GA), Connecticut, Detroit (MI), Hawaii, Iowa, New Mexico, San Francisco-Oakland (CA), Seattle-Puget Sound (WA), Utah, Los Angeles (CA), San Jose-Monterey (CA), Rural Georgia, Alaska Native Tumor Registry, Greater California, Kentucky, Louisiana, New Jersey, and Greater Georgia (minus areas affected by Hurricane Katrina in 2005: parts of Louisiana, Georgia, New Jersey and Kentucky; and Hurricane Rita: most of Louisiana and western Kentucky) [12].

2. Data preprocessing and transformation

   (a) The statistical analysis software SEER*Stat was used to retrieve and preprocess the dataset either from the bundled database or the SEER server (given that Internet connection is available on the machine used).

   (b) The preprocessing and transformation of data were conducted in SEER*Stat and the statistical programming language R. The integrated development environment (IDE) RStudio was used for the R development.

   (c) The limit for SEER data was set from 1998 to 2003 as year of diagnosis.

Reasons include: a set of variables were introduced in 1998, the RX Summ series, which are not present in previous records [13], and only until 2003 in order to accommodate predictions up to ten years.

(d) Records with missing or "Blank(s)" values were eliminated. This decision is limited to a selected number of variables, enumerated in Table 13 in page 147.

(e) Records with irrelevant values for a number of variables were also eliminated from the dataset. An example is Behavior recode for analysis's values that were included are 'Benign', 'Borderline malignancy', 'In situ', and 'Malignant', while the rest were removed. Table 13 contains the other variables modified this way.

(f) The SEER*Stat variables selected from preprocessing phase were further filtered by means of different decisions to preserve non-redundancy and appropriateness. This process is discussed in Section B. of Chapter IV.

(g) The final dataset used in modeling was divided according to time of survival, with intervals of two years for uniformity. A priority for records with greater than or equal to eight years of survival months was applied in the inclusion to the dataset, as shown in Fig. 24.

3. Data mining and modeling

(a) The data modeling was implemented with WEKA API and not the GUI version. In contrast, the attribute selection used the WEKA GUI because no other specific methodology was required by the LCOC framework.

(b) The variables that were used in the modeling of the predictive survival depends on the output of the preprocessing phase.

(c) The smaller set of variables to be used in the calculator were determined

after the preprocessing phase. WEKA's correlation-based feature selection algorithm was used with ten-fold cross-validation. These variables were selected if their score was above 10% per cross-validation fold.

(d) The modeling was done twice: first with the complete dataset and the second only includes the variables that resulted from attribute selection.

(e) Algorithms

   i. The five algorithms alternating decision tree, J48 decision tree, random forest, `LogitBoost` and random subspace were used for the predictive modeling as they were proven to hold high predictive accuracy on the SEER data [3]. Their respective implementations in the WEKA software were specifically used for the modeling [3, 14].

   ii. The ensemble voting method indicated in the reference study (averaging) was be used to combine the algorithms' results for better accuracy. The average of the results from the five aforementioned models serves as the "predicted survival" of a given set of breast cancer data [3].

(f) Similar to LCOC, the modeling of data was accompanied by a ten times ten-fold cross-validation. It was applied to both modeling processes.

(g) The variables that were used in the modeling for the BOSOM Calculator depends on the output of the attribute selection on the complete dataset.

(h) The predicted survival are computed by the WEKA models and averaged to implement ensemble voting. These correspond to the probability of a record to have a class distribution of "1" or "alive".

(i) The recorded results of data modeling only include: "performance metrics" (e.g. accuracy, precision, etc.) per cross-validation run, a list of

the program's time executions, and a general summary of the modeling similar to the "result buffer" of WEKA GUI. The first only records cross-validation runs meaning only ten performance metrics per data modeling was obtained.

4. BOSOM application

   (a) The BOSOM Calculator's input fields directly depend on the results of attribute selection.

   (b) Only breast cancer survival rate is determined by the application. It will not give options for finding optimum treatment. The results of this calculator are only based on statistical and historical data and not to be used as a substitute for absolute medical diagnosis. Proper medical attention from a qualified physician is strongly advised for any questions regarding breast cancer and survival.

   (c) The models used to predict survival are not be updated with the data entered by users. However, it can be replaced with another model based on a newer set of SEER data as extracted from the WEKA software. By doing this, the calculator's prediction system and form module must be updated to match the new set of models' variables and values.

   (d) All data collected from the users are not be saved into a database for security and confidentiality reasons. Only PDF files containing the data provided and results are generated once the user clicks the "Save as PDF" or similar button.

   (e) The application does not correct inconsistent data among the input variables. Disproportional medical values are expected from users such as a declared low stage of cancer against an extreme spread of metastasis [7].

   (f) Only links to professional cancer sites in the country such as hospi-

tals and organizations or international cancer sites for information or other research purposes are included in the post-result viewing action. These are intended to help the user in finding more information or seek medication about their condition.

(g) Both the physical and web server where the application is located must be on in order for the BOSOM application to be viewable and usable to users.

(h) User interface (UI)

i. The UI is dependent of third-party style frameworks. Their limitations and incompatibilities to render specific elements to the browser, device and operating system are also a limitation of this application.

ii. In viewing the PDF file, there are three noted behaviors that may occur depending on the user's browser and device. These include: viewed directly into the browser's own PDF renderer, will not be viewed but a save prompt will appear, and it will be saved directly into the user's device.

F. Assumptions

Here are the assumptions for the users of the BOSOM application:

1. Users with a modern browser and Internet connection are allowed to use the application.

2. The website is viewable on all major devices and including in mobile and low resolution devices (will rely on CSS3 `media-queries` for rendering). Graceful degradation is expected for other devices such as older computers where features will revert to their basic equivalent. For example, gradients are

expected to be rendered as a solid color in Internet Explorer 9 and below as they are not supported [15].

3. The breast cancer data that will be provided by the users are proportional and realistic.
4. The users are aware of the purpose of the BOSOM Calculator and its results.
5. The users have a legitimate breast cancer data that will be entered into the BOSOM Calculator.
6. The user must have either a native PDF renderer in their browser or a PDF reader in their device in order to view the generated file.

II. Review of Related Literature

Cancer survival prediction using data mining on historical records is possible. The predictive models created from previous studies for predicting disease survivability were made from training data mining algorithms with medical records i.e. artificial neural networks, decision trees and Bayesian networks. Majority were reported to perform prediction with highly successful accuracies.

In 2005 Delen et al. compared three data mining and statistical algorithms namely the multi-layer perceptron artificial neural network (MLP ANN), C5 decision tree and logistic regression's ability to predict breast cancer survivability through predictive modeling on historical data. The 1973 to 2000 SEER dataset was subjected to preprocessing by modifying and removing records and variables based on the following criteria: if a record's time of survival did not exceed sixty months after diagnosis and was not fully updated in the course of the same length of time then it is removed; variables unrelated to breast cancer were removed; redundant and aggregated variables such as Morphology and "Extent of Disease" where split into new variables; only records between 1988 to 2000 are included in order to accommodate the new variables "Extent of Disease" and "AJCC Stage of Cancer"; and lastly records with illegal values, for example a tumor size greater than 200mm, were removed from the dataset. 202, 932 records and 17 variables were the result of the thorough preprocessing and these were used in creating three predictive models. Comparison was validated by testing for accuracy, sensitivity (percentage of true positives predicted), specificity (percentage of true negatives predicted) and k-fold cross validation of the model predictions. Microsoft Access database, IBM Statistical Package for Social Sciences (SPSS), STATISTICA software and the IBM Clementine data mining tool were used in analyzing the preprocessed dataset (specific versions not stated in the paper). The final mean results from the cross-validation show that the C5 decision tree performed best with a mean accuracy of 93.62% followed by MLP (91.2%) and logistic regression

(89.2%) [10].

Derivatives of their work also followed through the years, with an observable trend of SEER dataset and data mining software WEKA frequently used together. Bellaachia and Guven improved Delen et al.'s study by including the variables "Vital Status Recode" (VSR) and "Cause of Death" (COD) in the filtering of data. They identified a weakness in the latter's declared breast cancer prediction - the number of patients that are "not alive" did not match the VSR value "not alive" which was caused by not taking into account the variables VSR and COD during the preprocessing phase. The research used the 1973 to 2002 SEER data and WEKA. Their predictive models were created using Naïve Bayes, back-propagated ANN and C4.5 decision tree and tests showed an accuracy of 84.5%, 86.5% and 86.7% for each classifier. In relation to Delen et al.'s work, the discrepancy of the accuracy between the two researches was caused by different SEER datasets, preprocessing methods and data mining tools used [16].

A 2008 study compared seven classifiers on the 1992 to 1997 SEER dataset and found that logistic regression showed the highest accuracy and sensitivity of 85.8% and 97% respectively while the MLP ANN has the best specificity of 50.9% [17]. In addition, Rajesh and Anand compared ten classifiers where the C4.5 decision tree was observed to hold an accuracy of 92.2%. They trained the model with 1973 to 2008 SEER data on the WEKA software [18].

Adjuvant! is an early breast cancer in women decision making software that is capable of providing survival prediction for patients that will undergo "no additional therapy" and "endocrine therapy". These results are presented as graphs as seen in Fig. 3. The first prediction model was based from the SEER-9 Public Registries Files August 1998 dataset where the researchers identified that the information for the type of adjuvant therapy and relapse status were not included. According to its 2001 paper, Adjuvant! requires several breast cancer-related variables in order to provide a prediction and these are age, menopause status,

estrogen receptor (ER) status, tumor size and number of positive nodes to name a few [19].

Currently, Adjuvant! has both online and standalone software counterparts. The online site, declared as in its 8th version, is restricted to doctors and physicians who have an account and registration is available for prospective users [20]. Information pages for its usage, machine requirements and legal notice are provided for the public. A notice for an upcoming 9th version is posted in the login page hence the tool is updated with new breast cancer data.

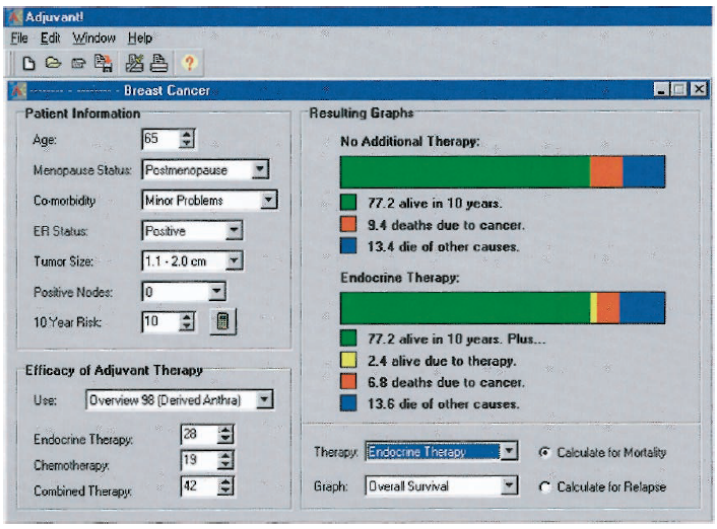


Figure 3: A screenshot of the main page of Adjuvant! taken from Ravdin et al.'s paper

The Lung Cancer Survivability Prediction Tool (LCSPT) was developed in 2009 for general lung cancer patients and is accessible in both mobile and desktop machines. Two prediction models based on histological data (age, gender, stage, cell type and tumor) and "additional" information (research authors mentioned "treatment options" and "smoking status") on an unidentified dataset were created. The variables' association to survivability was obtained using the Kaplan-Meier method and the predictive power was determined using a multivariate Cox proportional hazards model. They compared the two model's predicted and survival curves and stated that both were nearly equal thus suitable for the project (the

actual values were not mentioned in the paper). It was integrated on a proposed Survival Probability Prediction Architecture (SPPA) that enables integration of the prediction tool to an electronic health records (EHR) system, database and a model-view-controller web framework (which contains the prediction model) in a server in order to deliver content to devices [21].

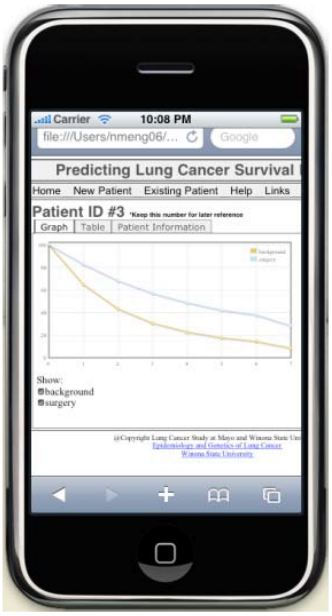


Figure 4: The Lung Cancer Survivability Prediction Tool as seen on an iPhone

PREDICT Tool, a breast cancer survival calculator, used a Cox proportional mortality model on 5,694 records collected from Eastern Cancer Registry and Information Center (ECRIC) in the United Kingdom from 1999 to 2003 (specifically cases in East Anglia, UK). The researchers considered the well-known SEER cancer dataset but opted for geographically more native dataset. In order to test the predictive accuracy of the model, they used 5,468 records from the West Midlands Cancer Intelligence Unit (WMCIU) breast cancer dataset from the same time period. As we can see in figure 5, it accepts ten variables that were specified in the ECRIC dataset [5]. Wishart et al. reported that a validation study on Canadian breast cancer patients showed that the PREDICT Tool still held a "good performance" thus the elimination of regional differences of the records as a hindrance

to survival prediction.

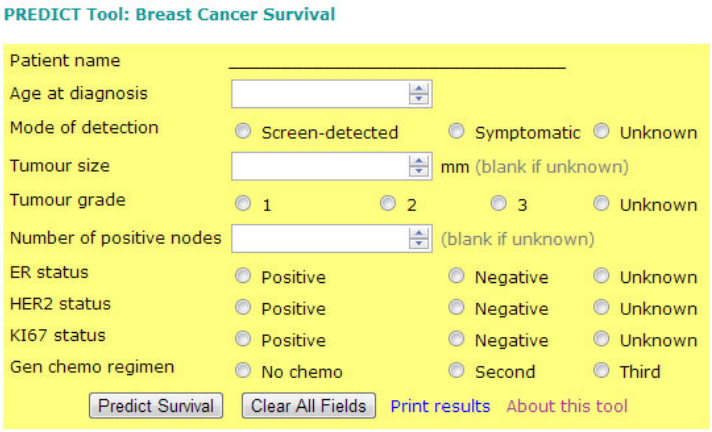


Figure 5: A screenshot of the PREDICT Tool's form

Agrawal et al. developed an online lung cancer outcome calculator using data mining and predictive modeling [22]. The study involved intensive preprocessing of 1973 to 2006 SEER data. The range of records were limited up to 2001 to accommodate the five-year prediction up to 2006 and records from below 1998 were omitted because variables such as "RX Summ-Surg Site 98-02" and "RX Summ-Scope Reg 98-02" were only introduced in that particular year. In connection, only cases that indicated the "Cause of Death" as lung cancer were considered and modifications to variables like creation of "number of regional lymph nodes that were removed and examined by the pathologist" and "number of malignant/in-situ tumors" derived from "Regional Nodes Examined" and "Sequence Number-Central" were done. Although new, these additional variables were found to have substantial predictive power [3].

A final dataset composed of 57,254 records and 64 variables were used in the modeling to determine the top five best performing classifiers from the WEKA software. J48 decision tree, alternating decision tree, `LogitBoost`, random subspace and random forest were selected and ensemble voting was used to combine their accuracies for single analysis. The predictive models were developed from

two different datasets based on the size of variables; where the first using the 64 variables and the latter exclusive to the 13 feature-selected based on predictive power. Figure 6 shows these 13 variables found the online calculator. Results show that both models have nearly similar predictive accuracies of 91.4% and 91.2% respectively separated only by 0.2% of discrepancy [3].

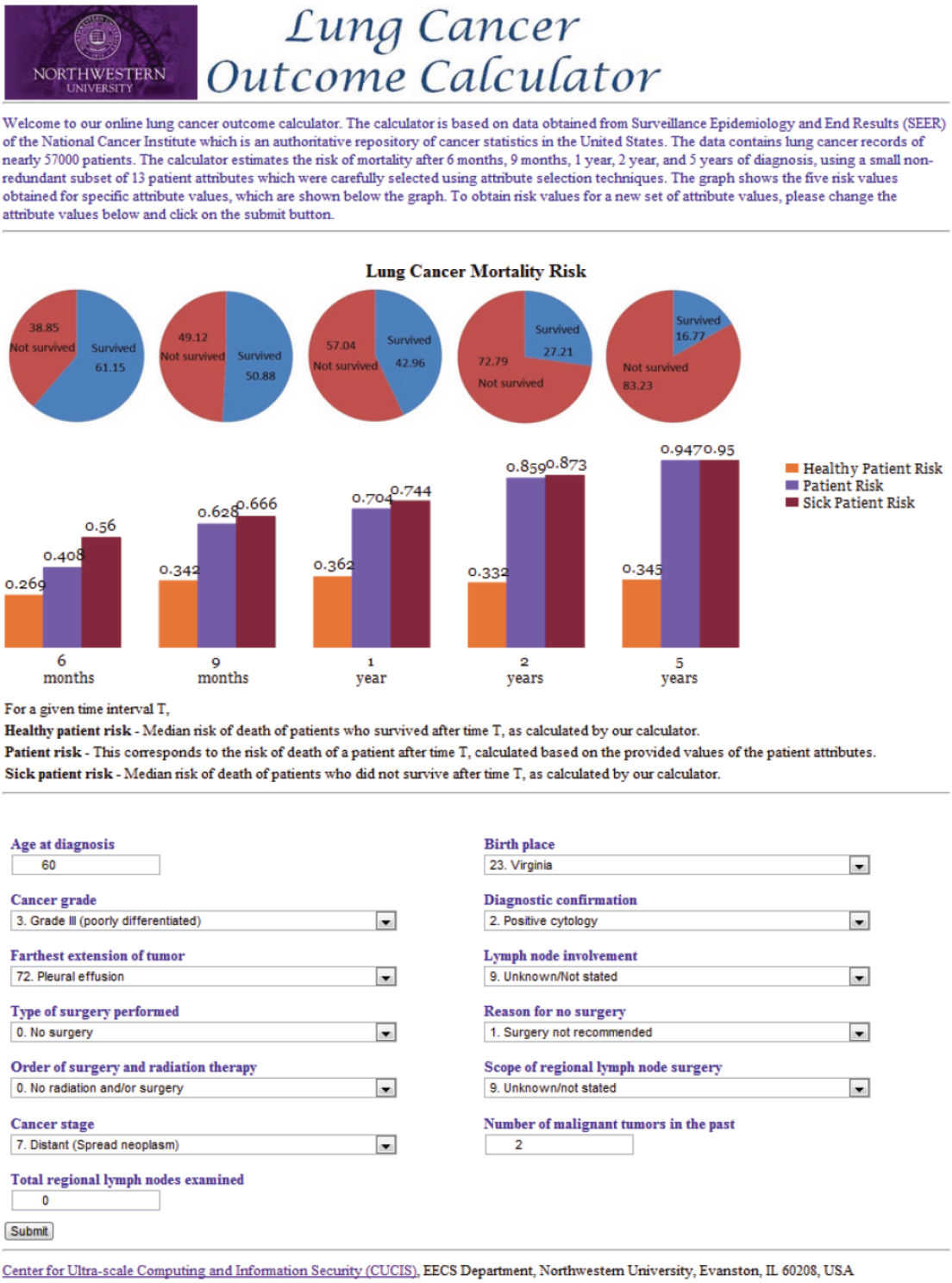


Figure 6: A screenshot of the input form and results page of the Lung Cancer Outcome Calculator

III. Theoretical Framework

A. Cancer and survival

Cancer occurs when cells undergo an abnormal growth process – instead of being repaired or replaced by the body [9], they continue to grow, develop and affect surrounding tissues. These cancer cells group together forming tumors that could spread to other organs (also known as metastasis) and lead to complications and other types of cancers [7].

1. Breast cancer

   Breast cancer starts from breast cells that turn into cancer cells caused by mutation defects. In addition, there exist risk factors that increase chances of a person developing the disease and the American Cancer Society reported three categories: unchangeable, lifestyle-related and the uncertain / controversial / unproven. In both sexes, unchangeable factors include aging, family history of breast cancer and gene mutations; lifestyle-related points to estrogen therapy (ET), alcoholism, obesity and physical inactivity [7, 9, 23]. This type of cancer is clinically and statistically known as more common in women than men due to the fact that exposure of the breast tissue to the estrogen hormone (biologically present in women) is proven to increase the risk [7, 9].

   There are different types of breast cancer and majority of these are a derivative or combination of non-invasive (or pre-cancer, in situ) and invasive cancers. Non-invasive cancers are benign tumors that do not penetrate beyond the confines of its starting location thus early detection may easily eliminate it through different kinds of breast surgery such as mastectomy. Meanwhile invasive-type cancers originate from the non-invasive types that were able to penetrate nearby tissues. These are more life-threatening because tumors and metastasis weakens the vital organs and lowers the mor-

tality of the patient [7,9].

2. Cancer prognosis

   The National Cancer Institute (NCI) defines prognosis as an "estimate of the likely course and outcome of a disease." It goes hand-in-hand with several investigative methods in order to present such information to the cancer patient and their families. The patient's condition, the current available treatment options (could be restricted by the hospital facilities, the patient's health, or financial limitations) and additional health problems are commonly noted in medical records and these are found in old records that contribute to the understanding of the condition [4]. Another are kinds of statistical estimates that often aid in prognosis such as: cancer-specific survival / disease-specific survival, relative survival, overall survivial and disease-free survival/recurrence-free survival / progression-free survival. Note that these statistical estimates have defined assumptions and where collected over a certain number of population [4].

   Doctors usually check previous records of patients with the same condition and compare the current case to find similarities or differences to aid in decision-making. Some opt to consult other doctors for their own opinions and findings to get alternative interpretations of the case [2].

3. Survival prediction

   The process of survival prediction deals with determining the time left for a patient to live, generally associated with diseases of of high mortality rate such as cancer. It is a part of a physician's prognostic investigation where a result takes the form of a numerical percentage of survival over a period of time that depends on a factor, for example cancer stage or time after diagnosis. Statisticians and researches contantly update several metrics for a period of time to serve as reference for various disciplines. As discussed before, one of their focus is on the field of medicine and survival estimates

are maintained for reference and as a glimpse to the overall status of how a given condition affects a population. Major types of cancer have their own computations from organizations like GLOBOCAN (as seen in Fig. 2) and in the case of breast cancer, stage was chosen as a focal point in Tables 1 and 2 that aid in providing supplemental population-based prediction [7].

In the presence of data mining in medicine, survival prediction is simulated by modeling or training a capable mathematical algorithm with medical records in order to obtain rules and patterns much like what a doctor's scientific analysis is conducted. Past studies have employed several algorithms such as artificial neural networks, logistic regression and decision trees to create a predictive model to estimate survival prediction of cancer patients.

Table 1: Breast cancer survival rates by stage in women (from NCDB 2001-2002)

Cancer stage	**5-year survival rate**
0	93%
I	88%
IIA	81%
IIB	74%
IIIA	67%
IIIB	41%
IIIC	49%
IV	15%

Table 2: Breast cancer survival rates by stage in men (from NCDB 2001-2002)

Cancer stage	**5-year survival rate**
0	100%
I	96%
II	84%
III	52%
IV	24%

B. Knowledge discovery in databases (KDD)

Fayyad et al. defines knowledge discovery in databases (KDD) as a process of finding practical and sensible interpretations from a large amount of data. It is composed of five consecutive steps namely: selection, preprocessing, transformation, data mining, and analysis, interpretation or evaluation [24].

1. Selection

   Selection, or data gathering deals with obtaining an appropriate dataset that will be used for the KDD study [24]. Traditional methods include surveys and interviews and recently data are available online for public use from repositories such as SEER for various cancer types and the University of California Irvine (UCI) that provides multiple kinds of datasets.

2. Data preprocessing

   Majority of data mining researches agree that cleaning the data before the actual mining process is directly related to the success (i.e. high prediction accuracy) of the desired output (classification and regression). It involves isolating only the necessary entries and variables required in the study. Records with missing or illegal entries might be removed entirely in order to preserve the quality of the data (but these can be left as other algorithms perform better with noise) [6].

   Witten et. al. attributed problems in datasets to human errors such as typographical errors and unknown, blank or illegal (i.e. a 9999 out-of-range value or -1 for positive real codes) values. It is the responsibility of the researchers to find discrepancies in the records in order to maximize the potential of the data. It is recommended to consult the experts responsible for creating or managing the desired dataset because they are valuable source of information for the identification of the variables and values not fully understood. It was also stated that most "dirty" datasets or containing

missing data and illegal values tend to impair the performance of classifiers such as decision trees and regression algorithms [6].

3. Transformation

   Data transformation refers to the modification of the preprocessed data to suite the needs of the research. This could adhere to the requirements of a software that will be used during the data mining process.

4. Data mining

   Data mining is the process of analysis and extraction of meaningful information from data using statistical and machine learning techniques wherein trends and patterns that occur in an observable frequency can be applied in practical applications [6]. It has a wide range of applications to fields that generate a large amount of information. In reality, there are numerous way to use data mining in solving problem. Customer behavior is an indispensable property of major companies such as supermarkets and social networking sites; data mining provides opportunities to improve strategies for profit and revenue by finding correlations between the customers to products and services offered. Medical, biological and chemical research teams can find possible relationships of entities such as proteins and viruses to diseases by using clinical and historical data from hospitals or academic institutions. Data mining on medical records will help define rules and patterns on how a patient's clinical data is related to their survival.

5. Analysis

   Interpretation and drawing conclusions based from the results of data mining comprises the analysis step. Performance metrics such as prediction accuracy, specificity and sensitivity are analyzed during this step. Visualization tools such as histograms, bar graphs and tree diagrams are also considered to aid in understanding the results better [24].

C. Predictive modeling

1. Classification

It is the process of interpreting a given dataset by the recognition of patterns, similarities or differences in order to achieve a practical decision or forecast [25]. Classification deals with nominal values, categorizing a given set of data to a predefined value i.e. breast cancer cases could be classified as either "malignant" or "benign". *Regression* is another method that solves a numeric outcome instead of a class.

The list below contains the top classifiers reported to have significant accuracy for an SEER-based lung cancer dataset study. These where chosen from around 30 other implementations based on execution time [3].

(a) Decision trees

A typical decision tree has internal nodes that indicate a variable which split into branches or the values. Further down the tree are the terminal nodes that denote the classification of a data. Decision trees are frequently used in data mining because their structure resemble the way real world rules are followed in order to make a decision depending on a set of instructions.

i. J48 decision tree

J48 is a Java programming language implementation of the C4.5 decision tree [6, 26]. It classifies a given dataset by splitting the variables individually to find the one with the highest information gain or the measurement of its contribution influence to the outcome [3, 27].

A J48 DT consists of two elements: a *leaf*, which indicates an outcome value or class and the *decision node* represents a test to

which path or leaf to proceed with the classification. The creation of a decision tree relies on the "gain criterion" in order to grow and split leaves. It is defined by Quinlan et. al. as "the information conveyed by a message depends on its probability and can be measured in bits as minus the log to base 2 of that probability".

Given a dataset T with n outcome classes, T is partitioned into $T_1, T_2, \ldots, T_n$ where each belongs to a class. There are also the set of cases S and a set of outcome classes C. For notation purposes $freq(x)$ refers to the number of occurrences x thus $freq(C_i, S)$ represents the number of cases in S that belongs to class C_i [26].

As seen in a subtree in Fig. 7, the decision from what path to follow between the two leaves requires gain criterion. The probability that a case from S belongs to a leaf C_i is represented by $\frac{freq(C_i,S)}{freq(S)}$ and in relation to gain, its information can be solved by the equation:

$$-log_2 \left[\frac{freq(C_i, S)}{freq(S)} \right] \tag{1}$$

The class membership of a case is computed by taking the summation of the products of the probability and its information:

$$info(S) = -\sum_{j=1}^{n} \frac{freq(C_i, S)}{freq(S)} \cdot log_2 \left[\frac{freq(C_i, S)}{freq(S)} \right] \tag{2}$$

In general, the formula for gain criterion is the difference between the information of the dataset's cases and nodes or $gain(x) = info(T) - info_x(T)$. This measures the information gained by partitions the dataset T to a set x [26].

$info(T)$ is the *entropy* of all cases or the "measure of the average

amount of information needed to identify a class of a case in T" while $info_x(T)$ is the entropy of the nodes or expected information requirement. This is computed using the formula:

$$info_x(T) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \cdot info(T_i) \tag{3}$$

The leaf with the higher information gain is selected as the class and the search will continue to the other leaf's subtree.

ii. Alternating decision trees (ADT)

ADTs solve problems with the help decision stumps (a smaller tree and usually binary). It is composed of decision (or splitter) nodes and prediction nodes. The more positive or negative a prediction, the more the classifier leans to a positive or negative classification.

Freund and Mason provided a formal structure of an alternating decision tree in Fig. 7 where the stumps are connected by decision nodes indicated by the dashed arrows. Decision nodes are associated with a boolean argument leading to the prediction nodes that contain a real number value [28]. Getting the sign of the sum of the prediction nodes from the path of root to the terminal leaf serves as the overall prediction of the tree [29].

Sok et al. reported that ADTs are capable of producing interpretable and easy-to-understand decision rules due to the alternating decision and prediction nodes; and that the sum of the paths from the root to the terminal leaf indicates the magnitude of the classification confidence.

iii. Random forest

A random forest is a collection of decision trees that decides for a final classification by majority voting [30]. A training set is divided into a number of approximately equally-sized subsets selected randomly with replacement (or using bootstrap) and each of these subsets form the individual decision trees [31]. Starting from the root, each decision tree is recursively built by finding the best split on a node's random set of variables and creating more trees on the left and right children until all variables are exhausted [32].

Random forests are naturally unpruned or not limited to a number of trees but Dittman said that the optimum number of trees is 100. A case is introduced into each tree and then the decision with the most number of instances (or other ensemble function) is the classification.

Leo Breiman noted that it is relatively robust to outliers and noise in the data; a faster procedure compared to classifiers boosting and bagging; and it is simple and easily parallelized. It was mentioned that its applicable in medical diagnosis as records contain multiple variables with only a small number of usable values [33].

(b) Boosting

Boosting is a classification algorithm that continuously splits on a dataset where in each iteration, weak classifications are reweighted in order to improve the splitting in the next iteration [29, 34, 35]. The continuous reweighting of the data increases the accuracy of the model. Common applications of boosting includes optical character recognition and medical diagnosis and prognosis where inaccurate data are

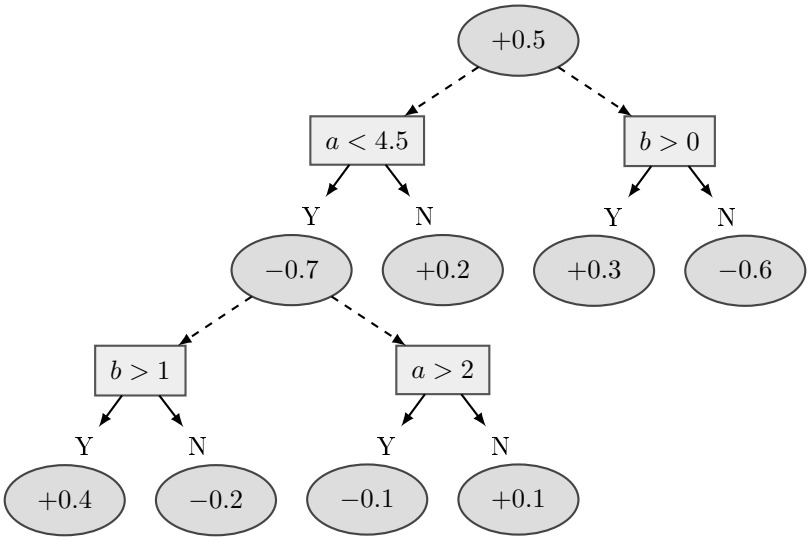


Figure 7: A diagram of Freund and Mason's general alternating decision tree example

frequently found [36].

The boosting algorithm `LogitBoost` starts by providing weights to a set of instances to be classified and these are all equal at the first iteration because the distribution of the training set is still unknown. At the end of the following iterations, the classifiers that were incorrectly classified are updated (most of the time increased) in order to be more prominent thus leaning towards to a better classification in the next [34, 35]. A weighted least square regression function is computed from the working responses and data points using the weights and the final classification is the sum of these regression values [35]. `LogitBoost` was reported to greatly reduce misclassification of instances due to its reweight method and this causes a better generalization of the data [30] but it takes a generous amount of time to train given its iterative nature [37].

(c) Random subspace

The random subspace method works by constructing a classifier (i.e. a decision tree, logistic regression) from a randomly selected subset

of the whole dataset [38, 39]. Voting is implemented to combine the results of these constructed classifiers [38]. Cai et al. expounded the classification phase where the dataset is divided into random subsets or subspaces that are repeated n number of times to create n subspace classifiers with individual results. It was introduced by Tim Kam Ho in 1998.

2. Ensemble learning

   Multiple models are going to be created from the aforementioned five classification algorithms; it will be beneficial if we combine their results to increase the predictive performance of the study [6]. Agrawal et al. were able to calculate their predictive accuracy by computing for the mean of the five algorithms.

   Current related literatures propose the use of ensemble techniques to combine multiple classifiers to improve their performance for survival prediction. Boosting was recommended (specifically `AdaBoostM1`) to complement individual six algorithms NB, MLP, Sequential Minimal Optimization (SMO), IBK, KStar and Bayes Net's respective accuracies. The last proved to be most effective in the diagnosis of colon cancer with 90.32% and 91.94% accuracies with and without boosting respectively [40]. Salama used WEKA's voting method in comparing five algorithms, their combinations in different groups applied to two breast cancer datasets from University of California Irvine (UCI) Machine Learning Repository. Varied results were gathered from different algorithm and feature selection combinations [41].

3. Attribute feature selection

   In order to create the BOSOM Calculator, a selection of the top variables in the breast cancer dataset must be done first to reduce the number of variables to be analyzed. A correlation-based feature subset selection was recommended by Agrawal et al. to determine the variables for the subset

dataset. mark Hall's `CfsSubSetEval` selects variables that has the closest relationship with the outcome and loose relationship with the other variables [42].

The success of this method is proven by the comparison of the predictive accuracy of the lung cancer outcome calculator versus the original set variables – the first earning 91.2% while the latter 91.4%, only a small portion of difference as seen in Fig. 2. Earlier studies have implemented feature selection in terms of sensitivity analysis [10] and information gain [16] and their results are seen in Table 3 [3, 10, 16].

Table 3: Selected breast cancer survival studies and their respective set of variables in decreasing predictive power

AGRAWAL (lung cancer)	DELEN (breast cancer)	BELLAACHIA (breast cancer)
RX Summ-Surg Prim Site	Grade	Extension of tumor (EOD)
Summary Stage 2000 (1998+)	No. of primaries	Stage of cancer
Regional Nodes Examined	Stage of cancer	Lymph node involv (EOD)
Reason for No Surgery	Radiation	Site Specific Surgery
RX Summ-Scope Reg 98-02	No. of lymph nodes	No. of positive nodes
EOD Lymph Node Involvement	Tumor size	Tumor size (EOD)
EOD Extension	Lymph node involvement	Histological type
Diagnostic Confirmation	Surgery	Age
Grade	No. of positive Nodes	Behavior code
Sequence Number-Central	Behavior	No. of nodes (EOD)
Birth Place	Marital status	Grade
RX Summ-Surg/Rad Sequence	Primary site code	Marital status
Age at Diagnosis	Age	Primary site
	Race	Radiation
	Histology	Race
	Extension of disease	No. of primaries

The researchers were able to create an online calculator using a smaller subset of variables feature selected from their preprocessed data. Specif-

ically, a correlation-based feature subset selection technique was used in order to find the variables with the highest predictive power, enumerated in Table 3. However, a past study used sensitivity analysis to correlate input to output variables. They tested this one the 10 data folds acquired from cross-validation and were able to rank 16 variables with "cancer grade" the highest at 70% [10]. Bellaachia et al. implemented information gain to achieve the same process and results showed that "Extension of Tumor (EOD)", "Stage of Cancer" and "Lymph node involv (EOD)" were the top three variables to predict breast cancer survivability.

4. Performance evaluation

In order to validate the predictive results of the classifiers and models, performance evaluation methods will be employed for quantitative measurement. These include testing for accuracy, sensitivity and specificity of the results and the k-fold cross-validation on the dataset.

(a) Accuracy, sensitivity, specificity, precision and ROC.

Majority of the literature in data mining and predictive survival states the use of the confusion matrix to evaluate the performance of algorithms and classifiers used. Table 4 shows a confusion matrix. Accuracy refers to the probability of predicting survival and death; recall or sensitivity is to correctly predicting survival; specificity is to death prediction; precision is defined by Sokolova et. al. as "class agreement of the data labels with the positive labels given by the classifier" and receiver operating characteristic curve (ROC) is the "ability of the model to avoid false classification" [17, 43] and their mathematical forms are provided in Equations 4. True positive (TP) refers to correctly predicted survival, true negative (TN) is to correctly predicted deaths and false positive (FP) and false negative (FN) are to incorrectly predicted survival and deaths respectively.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$recall/sensitivity = \frac{TN}{TN + FP} \tag{5}$$

$$sensitivity = \frac{TN}{TN + FP} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$ROC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{8}$$

Table 4: A confusion matrix for binary classification

	Survival	Death
Predicted survival	TP	FP
Predicted death	FN	TN

(b) k-fold cross-validation

Cross-validation works by dividing a dataset into k approximately equal parts wherein $\frac{k-1}{k}$ is trained for a purpose (in this study, prediction of survival) and then validated against the rest or $\frac{1}{k}$ of the dataset. This will run for k iterations in order to use all possibilities in the dataset. Majority of literature recommends the use of 10-fold cross-validation and this standard improves the results of an overall dataset prediction as it gives a chance for the parts to participate in the training and testing of the model [6, 10].

D. Surveillance, Epidemiology, and End Results Program (SEER)

1. Overview

   The Surveillance, Epidemiology, and End Results Program (SEER) of the USA National Cancer Institute is responsible for collecting cancer incidence and survival information in the United States of America. It currently encompasses 28% of the US population gathered from 18 cancer registries as enumerated in Chapter I. Section E.. Their cancer data are available for public use provided a "Data-Use Agreement" form that is signed by participating researchers. After confirmation from SEER, the data can be downloaded directly from the assigned web page or delivered via mail in a DVD format (due to the large file size).

2. SEER-18 data

   The SEER data used in this study is formally named as "`Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2012 Sub (1973-2010 varying)`" and was obtained on August 2013. It contains cancer records from January 1973 up to November 2010 with variables divided by Agrawal et al. as "demographic attributes (e.g. age, gender, location), diagnosis attributes (e.g. primary site, histology, grade, tumor size), treatment attributes (e.g. surgical procedure, radiation therapy), and outcome attributes (e.g. survival time, cause of death)" [3, 44]. The records can be accessed directly in the accompanying SEER*Stat software.

3. SEER*Stat

   The SEER*Stat statistics software, currently in version `8.1.2`, was used to obtain the SEER cancer data for various research purposes. It has the capability to do: frequency analysis, rate analysis (crude and age-adjusted),

survival analysis (observed, relative, cause-specific), limited-duration prevalence analysis, multiple primary - standardized incidence ratio analysis, left-truncated life tables analysis (beta version) and the case listing session.

The cancer database of the program is available from two locations: direct from SEER's server or from a local file. The former requires Internet connection and is constantly updated with new data, and the latter comes with the SEER*Stat and comprises the large bulk of file size.

All the variables used in the software are available for exploration using a Dictionary tool that can show the main categories (eg. "Site and Morphology", "Therapy"), their respective variables and a list of values for each variable. This data dictionary is helpful for quick reference to the SEER database's contents and modification of variables to suit one's needs. Merging, splitting and relabelling of variables and their values are possible here.

This study relied on the "Case Listing Session" feature. It is used to obtain the cancer records from the selected database. Preprocessing of records is required as filtering and selection of variables and values are a main feature of this. It composed of four main sections: "Data", "Selection" , "Table" and "Output".

The "Data" tab is where the database is selected, and as mentioned in the "SEER-18 data", the study used the default recommendation by the program.

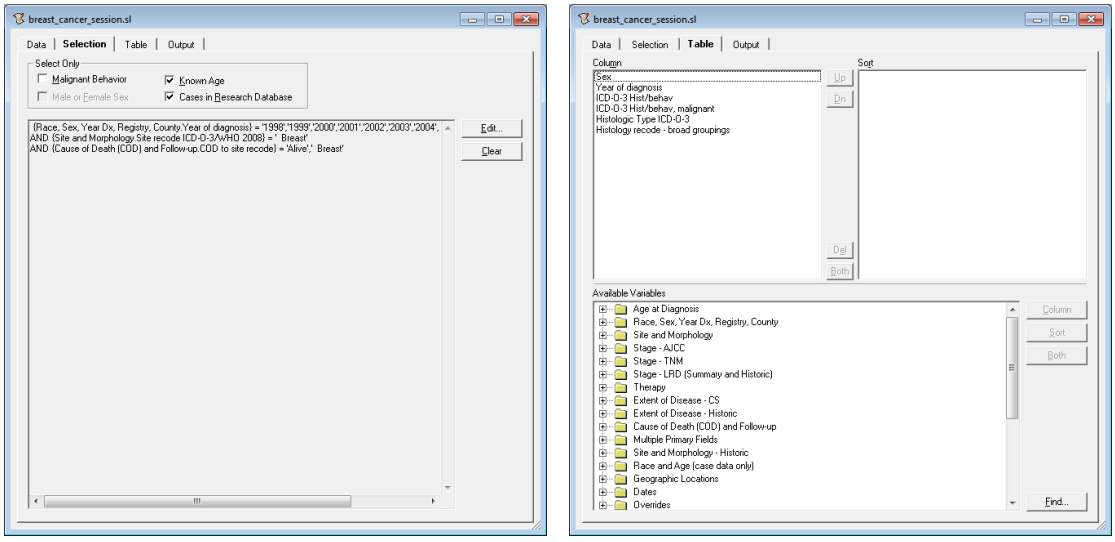
The "Selection" section two parts: the "Select Only" and a boolean query field. The first contains four items: "Malignant Behavior", "Known Age", "Male or Female Sex" and "Case in Research Database" wherein the third is disabled for clicking and the last is recommended to be checked in order to discard incomplete records gathered during Hurricane Katrina of 2005. The boolean query field is capable of firing filtering Standard Query Language-like (SQL) commands, set from a form. An image for this section is available

in Figure 8a.

Variables for inclusion in the datset are selected in "Table". There are around 331 available variables from 11 categories and it is advised to educate oneself first before selection. Different cancer staging and documentation standards govern the variables and these should be noted to prevent redundancies or misuse that could affect the results of the study.

The "Output" tab is only used for naming the data to serve as documentation.

Finally, result of the "Case Listing Session" is a "matrix", as it is called in the system. It is aan interactive spreadsheet of variable columns and record rows that can be copied into other applications such as Microsoft Excel or exported into a text or comma-separated values (CSV) file. This spreadsheet is saved in a "Case Listing Matrix" file format (or `.slm`).



(a) Selection tab **(b)** Table tab

Figure 8: Selected screenshots of some SEER*Stat Case Listing Session tabs taken from a Windows 7 system

E. Waikato Environment for Knowledge Analysis (WEKA)

The Waikato Environment for Knowledge Analysis or simply WEKA is an open-source data mining and machine learning software created by the Machine Learing Group of the Department of Computer Science from the University of Waikato, New Zealand. It is implemented in the Java programming language and capable of most common data mining-related tasks such as preprocessing, classification and clustering. The current releases are `3.6.10` (stable) for the general public and `3.7.10` for developers who want to extend and use the system's components [14, 45].

The WEKA GUI is composed of four main features: "Explorer", "Experimenter", "KnowledgeFlow" and "Simple CLI". "Explorer" is where the actual data mining method is applied; "Experimenter"; "KnowledgeFlow" is an interactive KDD and WEKA process diagram creator; and "Simple CLI" serves as a command line interface for executing commands. There are six major components of the Explorer application namely "Preprocess", "Classify", "Cluster", "Associate", "Select attributes" and "Visualize".

Explorer

To use the "Explorer" feature, the data must be loaded first in the "Preprocess" tab. The data can be uploaded in several ways namely: comma-separated values (CSV), Attribute-Relation File Format (ARFF) file, direct from a uniform resource locator (URL), from a local database through the Java Database Connectivity (JDBC) configuration, or from the available sample data generators. A filter for the data is available for application in the "Filter" section where more than 20 methods are implemented not limited to `NominalToBinary`, `NumericToNominal` and Synthetic Minority Over-sampling Technique (SMOTE). A simple tally of variables and records, in the form of a table and bar graph, are provided in the "Current relation" section and these can be analyzed in the "Selected attribute" section where a list of values is provided in tabular and graphical format. Addi-

tionally, thee variables can be removed in the "Attribute" section.

A status bar is located in the bottom of the window that shows summarized details of each task performed in the software. It also has a button named "Log" to see the entire status. The Java stack traces are also displayed in case of errors.

"Classify" is used in training the dataset with a large selection of around 90 classifying algorithms nested under the categories "bayes", "functions", "lazy", "meta", "mi", "misc", "rules" and "trees" which can be chosen in the "Classifier" section. The system also provides configuration and documentation of these algorithms. In the "Test options" section, several helper features are available: "Use training set", "Supplied test set", "Cross validation", and "Percentage split". The "More options" button shows a list of output statistics that can be viewed on the "Classifier output" once the training is started. A list of executed algorithms are indicated in the "Result list"; right-clicking one would show a dialog box that contains features for saving the algorithm's results as a model file (Java class file) and visualization (if applicable). The limitation of this feature comes from the classifiers themeselves; some of their implementations could be restricted to nominal and binary outcome variables, multiple outcome values or strictly numeric values.

"Select attributes" is used in determining the variables' ranking using attribute evaluators. Similar to the "Classify" tab's interface, there are 17 implemented evaluator algorithms and 11 search method algorithms that can be paired up to run simultaneously. The "Attribute Selection Mode" contains the option to use either "Use full-training set" or k-fold cross-validation. The results log are seen in the "Attribute selection output" section.

"Visualize" generates plot matrices of the data in a variable round robin scheme that can be customized in a wide variety of ways depending on the variables of interest.

The study extensively employed the WEKA API over the GUI version. The latter was only used for the attribute selection and confirmation of the trained

models. The API was chosen in order to gain more control with the data mining process and customize the data obtained from the software.

`Instance` and `Classifier`

All input data are represented as an `Instance` object and algorithms implemented as `Classifier`.

An `Instance` object transforms numeric, nominal, date and string variables into floating-point numbers. Numeric variables retain their values while the rest correspond to their array indices based on the order of declaration from the dataset. Manual creation of an `Instance` object is also possible in the API. All the variables' values are required to be explicitly declared as seen in Source Code 1 [46].

Source Code 1: Sample code for creating an `Instance` object

```
// Create empty instance with three attribute values
Instance inst = new DenseInstance(3);

// Set instance's values for the attributes "length", "weight", and
    "position"
inst.setValue(length, 5.3);
inst.setValue(weight, 300);
inst.setValue(position, "first");

// Set instance's dataset to be the dataset "race"
inst.setDataset(race);

// Print the instance
System.out.println("The instance: " + inst);
```

`Classifier` objects are implementations of classification and regression algorithms. They are capable of predicting an `Instance` object using the methods `classifyInstance` and `distributionForInstance`. An `Instance` object is passed to these methods in order to be predicted [46].

Attribute-Relation File Format

As explained before, WEKA is capable of handling files from different sources and format. One of these is the Attribution-Relation File Format which is a native file format to WEKA. It consists of three fields: `@relation`, `@attribute` and

`@data`. The first refers to the name of the entire dataset or file, the second is the variable used in the dataset and the last is the representation of a case or record from the dataset adhering to the order of `@attribute`s declared. The dataset variables are represented in the format "`@attribute variableName values`". The `values` dictate its type: continuous or numeric variables are set to `numeric` and categorical or nominal variables are declared explicitly separated by commas inside two curly braces i.e. `{value1, value2, valuen}`. A sample ARFF file can be seen in Source Code 6.

Result buffer

Both the "Classify" and "Select attributes" tabs have a large "result buffer" section where the details of each process done are exhausted. The result buffer from a classification can be divided into six major parts: WEKA data and classifier setup, classifier's training structure, predictions, performance metrics, performance metrics by class and the confusion matrix.

As seen in several examples in page 175, the first is formally called as "Run information" in the GUI and it contains valuable information regarding the training session's components: classifier used and its parameters, the dataset's number of records and the names of the variables, and lastly the testing mode [46].

"Classifier model (full training set)" shows the details of the trained classifier's details and structure and its content varies depending on the kind of classifier used. Decision tree-types would show a simple tree structure with weights and artificial neural networks show the weights of the layers for each iteration. In case k-fold cross-validation was selected as a test mode, it will also affect this part because each of the generated classifier structure per fold is also displayed, depending on the classifier and cross-validation interaction.

The next two parts consists of the performance metrics or measurement of the classifier's capability to predict. Accuracy, mean absolute error and root relative square error are some of the metrics in the first part. The second depends of the

outcome variable's values. Each of the eight metrics are provided with a individual values that corresponds to a class. "Summary" and "Detailed Accuracy By Class" are their names in the WEKA GUI. A confusion matrix of the training process is provided in the end.

The entire result buffer was recreated in the BOSOM Calculator's development, mirroring the Explorer's format, to get the results easier and quicker compared to the GUI version. The GUI version's limitation was the cross-validation and the number of ways necessary to save a results file of the trained model. This was automated in the modified version.

Prediction percentages

In WEKA, predictions refer to the probability that the data belongs to an outcome class [47]. It consists of the columns "Actual" and "Predicted" refer to the outcome class; and "Prediction" corresponds to the class distribution of the data depending on the selected class to output. In this study's implementation, both the values of outcome variable where recorded.

Model files

As discussed before, the trained classifiers in WEKA can be saved in the form of model files. These can be used again for re-training using new data to increase performance or to predict data to test its capability for prediction. The WEKA GUI and API both serializes a model file in order to read its content in order to be used [48]. Currently, WEKA has no feature to view the actual structure of the model file created from a classifier and data. It is only limited to: serialization and deserialization; visualization, if available; and source code extraction to selected classifiers only [48, 49]. After a model is serialized, it is casted as a `Classifier` object and this enables it to predict and be trained again by the WEKA API.

The closest way to see a model's structure is from the result buffer's "Classifier model (full training set)" part, although crude and needs further formatting in

order to be presentable.

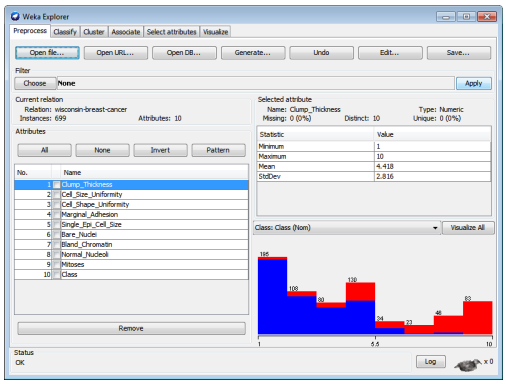


Figure 9: WEKA Explorer "Preprocess" tab with UCI Breast Cancer data taken from a Windows 7 system

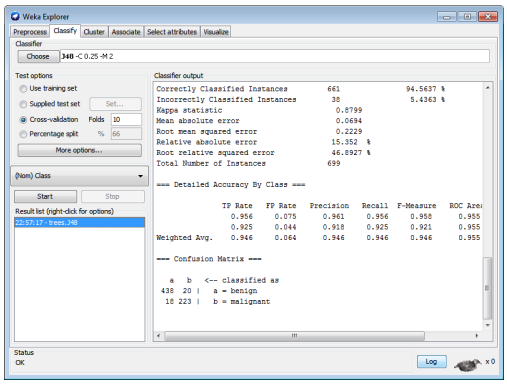


Figure 10: WEKA Explorer "Classify" tab with UCI Breast Cancer data taken from a Windows 7 system

F. Model-View-Contoller framework

In order to organize the structure of the website where the BOSOM Calculator will be integrated with, a model-view-controller (MVC) framework will be used to serve that purpose. The MVC is a design paradigm for separating content from the display in applications and websites which enables adjustments and synchronizations between components without causing much cascading effects. The model represents an object usually from reality and these contain inherent information on the object they represent that will be used throughout the application. Controllers are responsible for the application of processes to the information (whether in a model or entered by a user) such as implemented algorithms and business logic, connection to the database and serving the View. The view is the user interface where the results of the interaction between the model and controller are shown [50]. There can be multiple models which has its own controllers and views [51].

There are a number of open source MVC frameworks for the Java programming language available online where notable examples include Spring, Apache's Struts and Grails. Each has their own strengths and weaknesses but overall they enable native Java programs to be run seamlessly with web technologies.

Spring Framework

A Spring project or "web app" is created from a Java web project. Spring is introduced to the project by importing several dependent JAR files that must be noted.

A common Spring web app has the following entities interacting with each other: domain, controller, view, service and database access object (or simply "DAO"). The domain or model represents the actors and form data that are propagated throughout the application. In this study, the breast cancer data were represented by the class WekaData. A controller acts as a mediator between view and service; specifically, its responsibility (in relation to the aforementioned entities) is to serve

requested web pages and call higher level functions to perform tasks. It is also where the URL mappings are set using annotations. The views constitute the user interface; the framework recommends JavaServer Pages (JSP) as template. Service is where the business logic is implemented and communication with the DAO. Finally, as it name suggests, the DAO connects to the database either to deliver data to the views or vice versa.

Java web applications rely on servlets to deliver content to the server and a client's device. In Spring, the Dispatcher Servlet resolves requests through interaction with the controllers of the system to deliver content and perform tasks.

The application relies on the domain, service, controller and view layers in order to work. The DAO layer was not implemented because the prediction system does not rely on a database to manage data.

A web application is usually compressed into a web archive file (WAR) that is sent to a web server in order to run. This makes sure all the components and resources needed are intact.

IV. Design and Implementation

A. Use cases

1. Context diagram

   The BOSOM application both cater to all website visitors. The respective context diagrams of the website and BOSOM Calculator are provided in Figures 11 and 12.

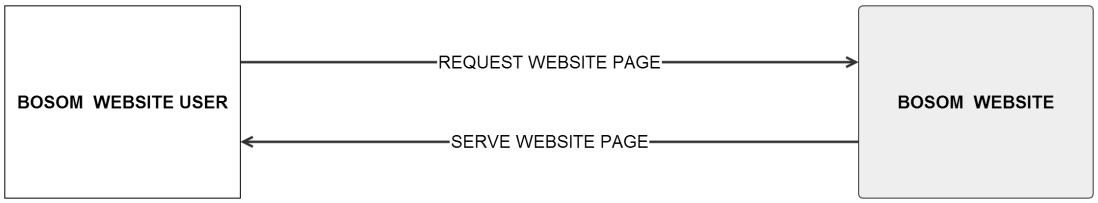


Figure 11: Context diagram of the BOSOM Calculator website

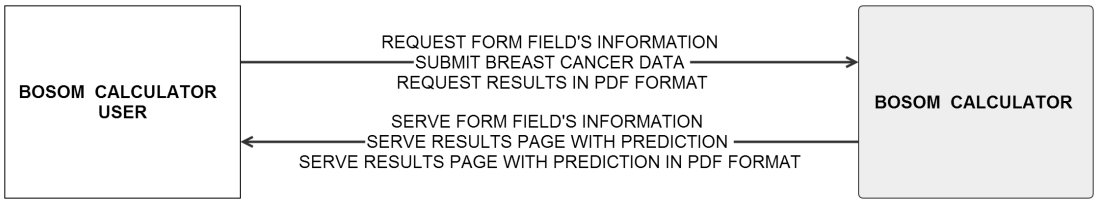


Figure 12: Context diagram of the BOSOM Calculator

2. Use case diagram

   The application caters to one type of user - the website visitor. All users are allowed to navigate throughout the site's pages including the BOSOM Calculator.

   The website and calculator are the only top-level entities in the application that the user interacts with. Users can visit the static pages (or that pages where minimal interaction is required i.e. the home page) and use the BOSOM Calculator to get their predicted survival. Figure 13 demonstrates this relationship of the three entities.

   As mentioned before, a user visits or "requests" for a page and the server

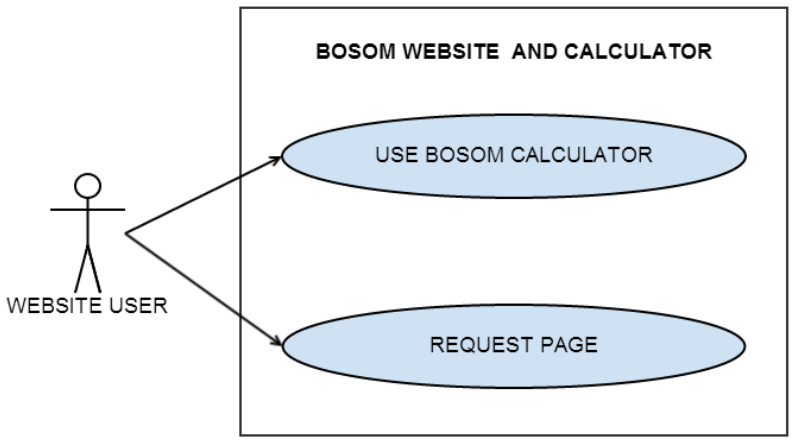


Figure 13: Top level use case diagram of the BOSOM website and Calculator

returns or "serves" the particular page as determined by its controller. This is shown in Fig. 14.

In the calculator's page, a user must answer a form and after submission they are redirected to a results page containing the survival. There is an option to view or save the results in PDF format provided. Figure 14 is the use case representation of how a user interacts with the BOSOM Calculator.

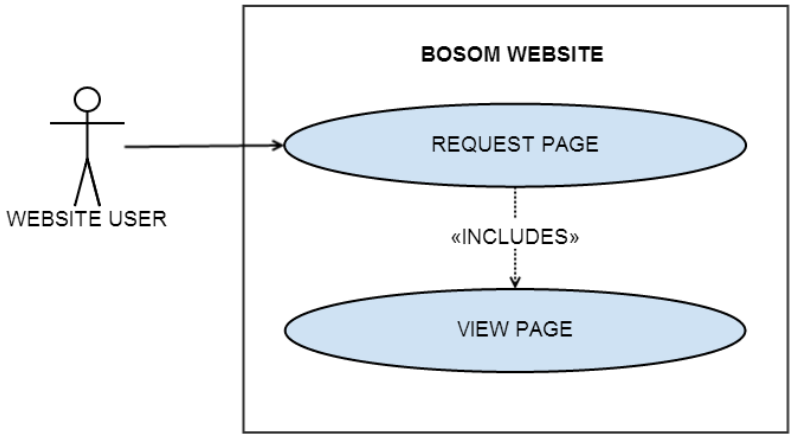


Figure 14: Use case diagram of the BOSOM website

3. Data flow diagram

The input data to the BOSOM Calculator will only be persistent from submission up to the result's PDF file creation or until a user leaves the page. There is no database setup that will store the breast cancer data from

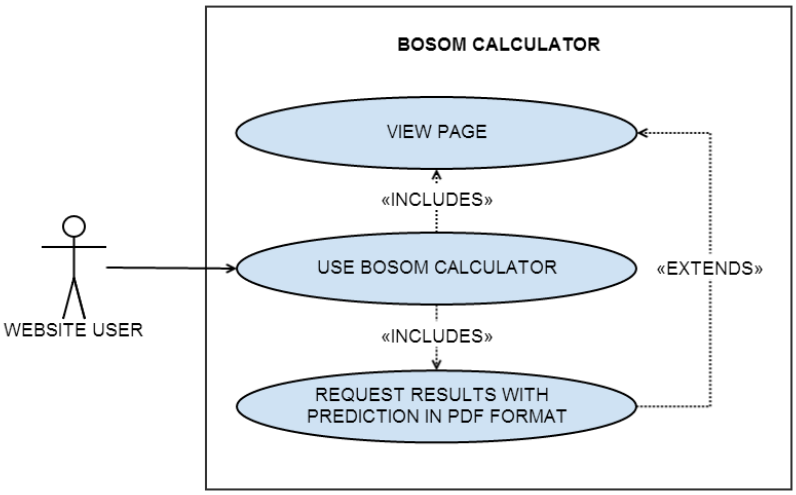


Figure 15: Use case diagram of the BOSOM Calculator

users.

The data is first sent to the main controller that's responsible for calling the following processes: "BOSOM Calculator Service" is where the models compute for the prediction; next is the PDF generator to add the computed predictions to the document; and last is the view where the user interface of the results page is assembled.

The other internal processes are laid out in Fig. 16 in page 50 along with their association with the entire system.

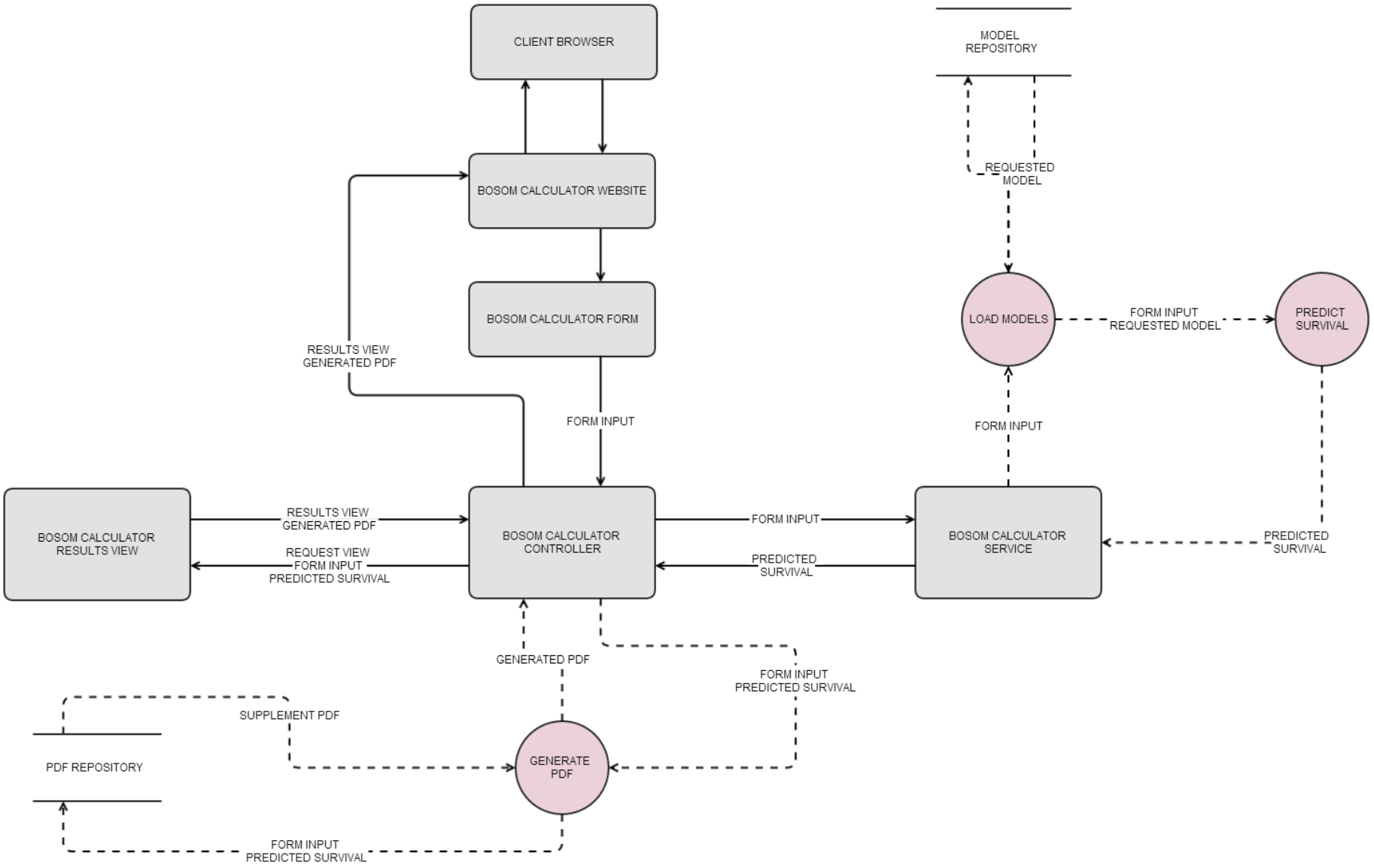


Figure 16: Data flow diagram of the BOSOM website and Calculator

B. Implementation

1. Data gathering

   The backbone of a data mining and knowledge analysis study is its data. Quality and completeness of the records are said to be directly proportional to the success of the data modeling in relation to its performance as determined by measurements such as accuracy, precision and specificity [6].

   The study's source of breast cancer data is the Surveillance, Epidemiology, and End Results Program (SEER), a division of USA National Cancer Institute. A "data-use agreement form" (seen in Fig. 52), was accomplished and submitted to request for access to the data. A confirmation e-mail was received later on containing instructions on how to download the data. SEER*Stat, a statistical software, is bundled with the cancer data's database. It was installed and ran on a capable machine. It was used to extract the breast cancer data that was used in this study.

2. Data preprocessing and transformation

   This section focuses on how the final dataset was achieved through systematic filtering and transformation of data. Both the SEER*Stat and R programming language were used to preprocess the data.

   Filtering of cancer data

   SEER*Stat's "Case Listing Session" feature, as explained in page 36, is has set of preprocessing modules for extraction and filtering of a desired cancer dataset.

   The first part was to pick a source database and the default labeled as "`Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2012 Sub (1973-2010 varying)`" was

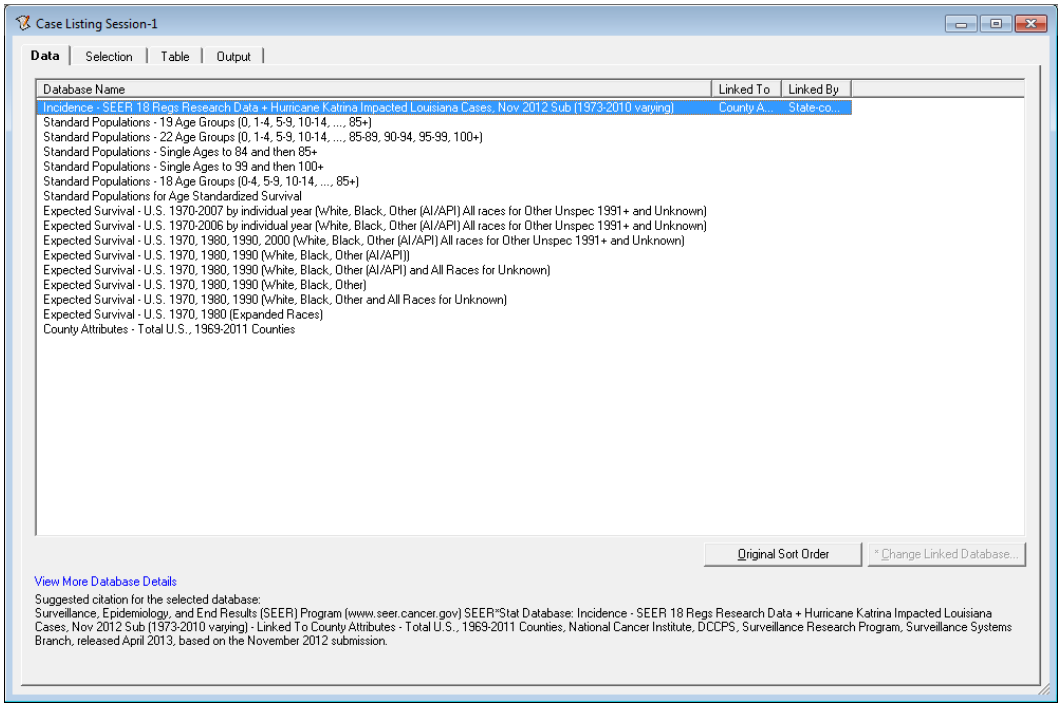


Figure 17: Screenshot of SEER*Stat "Data" tab showing the dataset used in the study

chosen because it was recommended by the system as well [12]. Fig. 17 is a screenshot of this activity from the SEER*Stat software.

The preliminary filtering was defined in the "Selection" tab; it works by marking form checkboxes and typing filter commands in the form provided. The checkboxes for "Malignant Behavior", "Known Age" and "Cases in Research Database" were unchecked, checked and checked respectively. The first's purpose was to include both malignant and non-malignant (or benign) records in the data, a classification of a cancer case; "Known Age" was checked to eliminate of possible outliers in the age variable; and the last was recommended by the system and the reason is to ensure consistency of records in terms of source location, because records from 2005's Hurricane Katrina were not included. The filter dialog box was supplied with breast cancer-specific commands including the ones to remove records with illegal or missing data and to limit inclusion of records based on a cancer staging

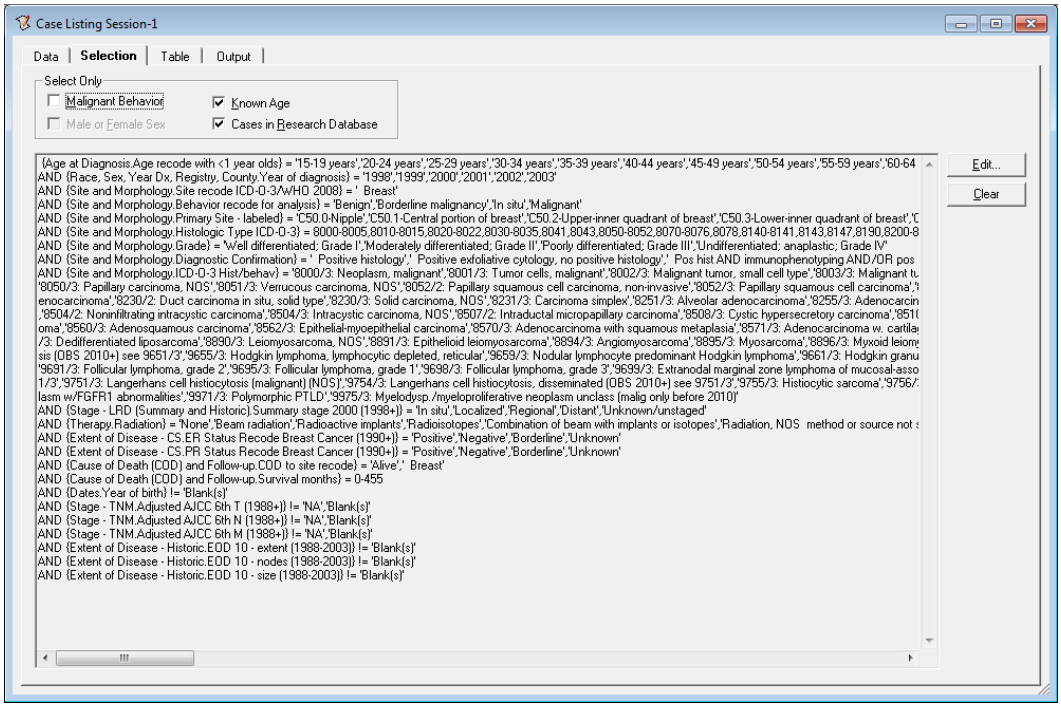


Figure 18: Screenshot of SEER*Stat "Selection" tab showing the options used in the study

system's edition. Table 13 contains these filter commands. Records with blanks and NA ("not applicable") values were purposely removed to increase the quality of data in terms of preserving completeness. This activity is seen in Fig. 18.

In connection with the SEER variable selection, there are several variables from Table 16 that were modified to separate their identified numeric and nominal values. `Regional nodes examined`, `Regional nodes positive`, and `Sequence number` were split to their appropriate types in Tables 5, 6, and 15. This step was implemented in the LCOC's as well in order to recognize the difference between the value's roles and possible predictive capability [3]. The SEER*Stat Data Dictionary showing the modified variables related to breast cancer is provided in Fig. 21 and the modification proper using the "Edit Merged Variable" dialog is provided in Fig. 22.

Lastly, the generated matrix was exported as a CSV file with the variable names set as header for later use. A dialog box for this activity is provided

in Fig. 19.

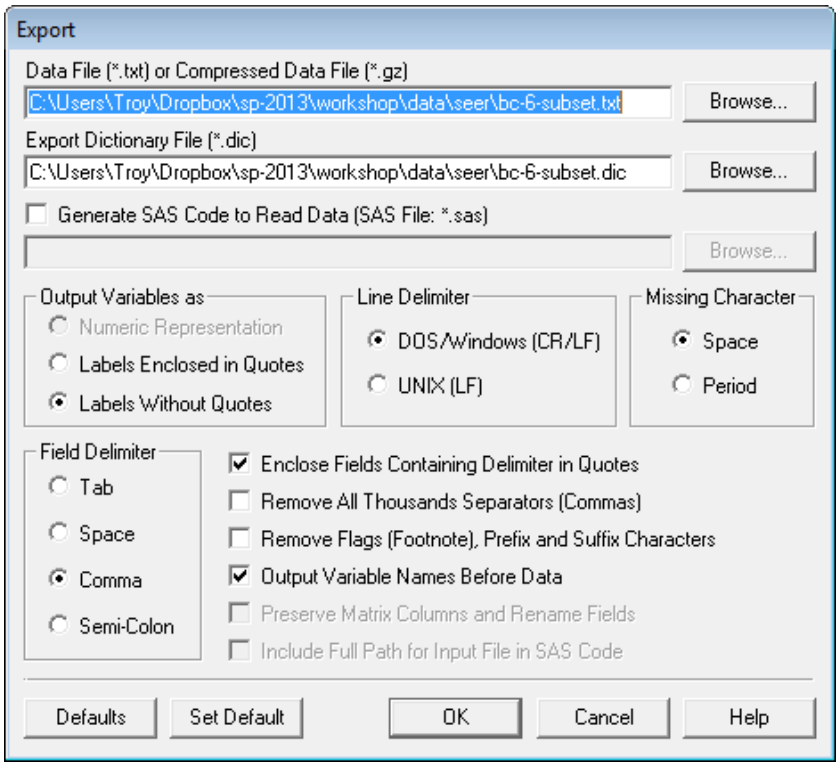


Figure 19: Screenshot of SEER*Stat Case Listing Matrix export feature

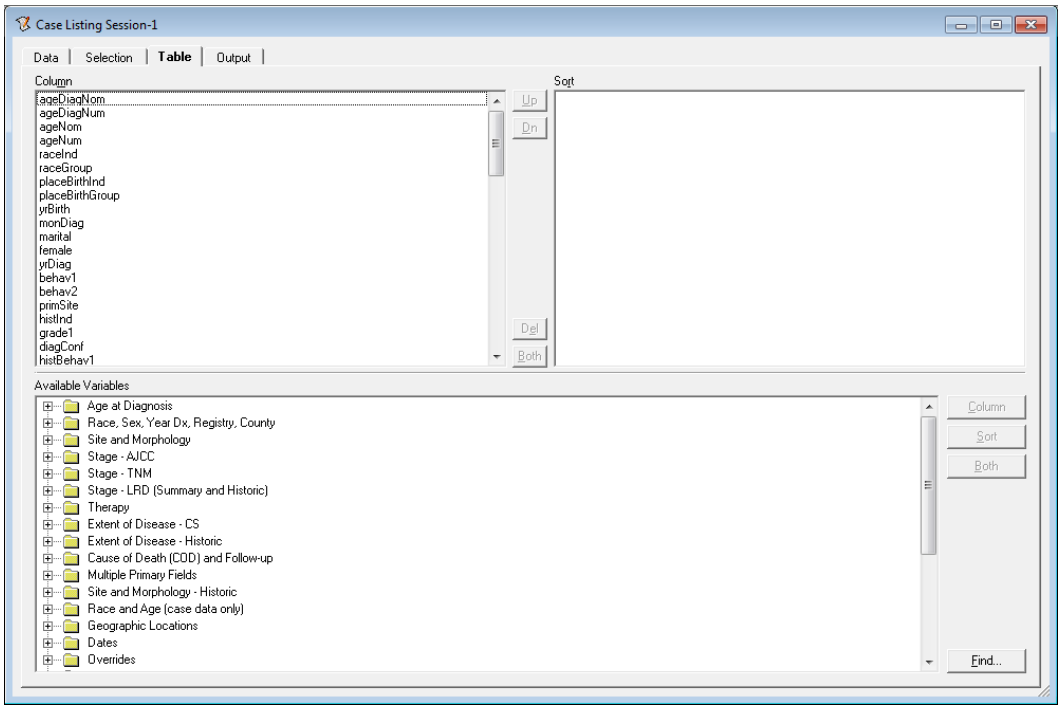


Figure 20: Screenshot of SEER*Stat "Table" tab showing the initial set of variables chosen for the study

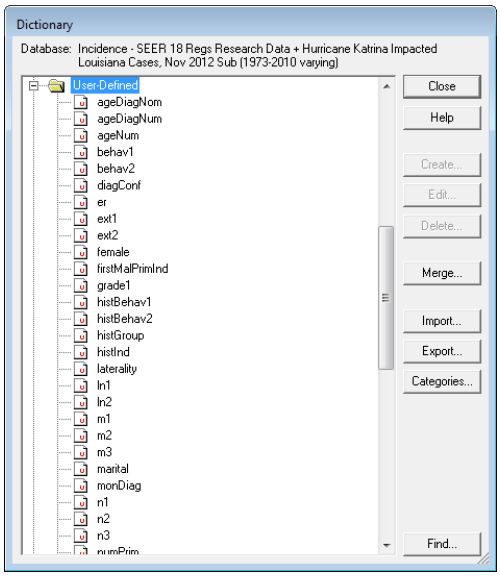


Figure 21: Screenshot of SEER*Stat's Data Dictionary showing the modified breast cancer-related variables

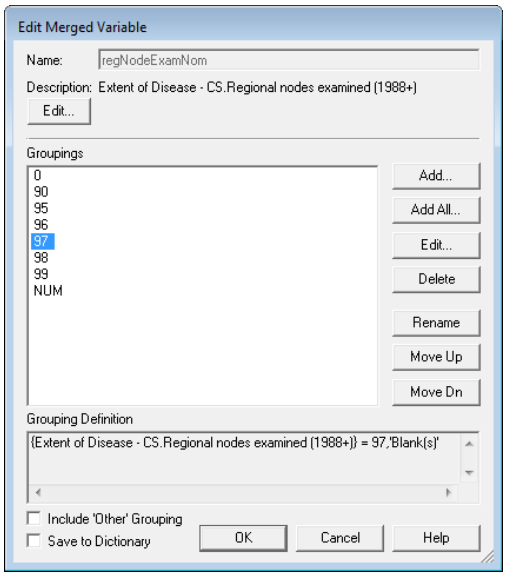


Figure 22: Screenshot of SEER*Stat's "Edit Merged Variable" feature for `Regional nodes examined (1988+)`

Table 5: Modification of SEER variable "Regional nodes examined"

Regional nodes examined - nominal	
Code	**Description**
0	No nodes examined
90	90 or more nodes examined
95	No regional nodes removed, but aspiration or core biopsy of regional nodes performed
96	Regional lymph node removal documented as sampling and number of nodes unknown/not stated
97	Regional lymph node removal documented as dissection and number of nodes unknown/not stated
98	Regional lymph nodes surgically removed but number of lymph nodes unknown/not stated and not documented as sampling or dissection; nodes examined, but number unknown
99	Unknown
Regional nodes examined - numeric	
Code	**Description**
1 - 89	Number of regional nodes examined

Table 6: Modification of SEER variable "Regional nodes positive"

Regional nodes positive - nominal	
Code	**Description**
0	All nodes examined negative.
90	90 or more nodes positive
95	Positive aspiration or core biopsy of lymph node(s)
97	Positive nodes - number unspecified
98	No nodes examined
99	Unknown if nodes are positive; not applicable
Regional nodes positive - numeric	
Code	**Description**
1 - 89	Number of regional nodes examined that are positive

Selection of breast cancer variables

The CSV file generated from SEER was imported to R. The variables whose type is nominal were explicitly indicated in the import function in order for R to recognize them as such; otherwise they are interpreted as numeric.

Six new variables were introduced in the data after import. The calculator provides survival prediction for a set of time intervals and these were taken from the SEER-provided survival months. The binary variables correspond to survival within less than two years, two, four, six, eight, and ten years. "Less than two years" was introduced to serve as supplementary data for the rest of the years and for the rest, an interval of two years was chosen for uniformity.

Source Code 2: Introduction of six binary time variables to the breast cancer dataset

```
dat$timeNot <- ifelse(dat$time < 24, 1, 0)
dat$time2 <- ifelse(dat$time >= 24, 1, 0)
dat$time4 <- ifelse(dat$time >= 48, 1, 0)
dat$time6 <- ifelse(dat$time >= 72, 1, 0)
dat$time8 <- ifelse(dat$time >= 96, 1, 0)
dat$time10 <- ifelse(dat$time >= 120, 1, 0)
```

Data preprocessing involves the recognition of relationship among variables in the data. SEER keeps track of staging standards and its time of implementation hence this must be noted when selecting variables of the similar group. Notable examples are the 6th and 7th editions of the American Joint Committee on Cancer (AJCC) applicable to 2004 and 2010 cases respectively. These include the three-part TNM cancer staging system composed of: primary tumor (T), regional lymph nodes (N), and distant metastasis (M) [52].

The related variables are grouped in Table 7 where the final chosen one is indicated. These seletions were based from research of the variables' relationships, cancer staging systems, and interview with an SEER represen-

tative. The rest were dropped for the final dataset. An interview with an SEER representative about merging these variables revealed that it is possible through cross-referencing of the related variables but it is not encouraged without training in cancer staging and coding. It was decided to not proceed with the variable merging and opt for choosing only one from the related variable groups.

After the non-redundant variables were identified from their groups, these rest were dropped from the data and the final set of variables used in predictive modeling are shown in Table 16. These 36 variables are highlighted. The Source Code 3 shows function that specifies the inclusion of these variables in the final dataset. This ensures that these are only variables that the algorithms in WEKA will recognize as predictive components in computing for a survival.

Source Code 3: Function to keep breast cancer variables specified for an R dataframe

```
filterDataframe <- function(dataset) {
  keeps <- c("ageDiagNum",
    "behav1","diagConf","er","ext2",
        "female","firstMalPrimInd","grade1",
    "histGroup","laterality", "m3",
    "n3","numMalTum","numPrim","pr",
    "primSite","raceGroup","rad","radSeqSurg","reasonNoCancerSurg",
    "regNodeExamNom","regNodeExamNum","regNodePosNom","regNodePosNum","stage3",
    "sumStage","surgPrimSite1","t3", "tumSizeNom2", "tumSizeNum2",
    "timeNot", "time2","time4","time6","time8","time10")
  return (dataset[,names(dataset) %in% keeps])
}
```

Preparation for output of data

In order to eliminate unnecessary values that are not legally a component of the variables, replacement of "" to SEER's "`Blank(s)`" and R's not applicable (`NA`) fields was applied to ensure uniform ity of representation of all missing values.

Table 7: Grouped SEER variables by relation

Group Name	SEER Name	Chosen Variable
Cancer stage	Derived AJCC Stage Group, 7th ed (2010+) Derived AJCC Stage Group, 6th ed (2004+) Adjusted AJCC 6th Stage (1988+) SEER modified AJCC stage 3rd (1988-2003)	Adjusted AJCC 6th Stage (1988+)
Primary tumor (T of TNM)	Derived AJCC T, 7th ed (2010+) Derived AJCC T, 6th ed (2004+) Adjusted AJCC 6th T (1988+)	Adjusted AJCC 6th T (1988+)
Regional lymph nodes (N of TNM)	Derived AJCC N, 7th ed (2010+) Derived AJCC N, 6th ed (2004+) Adjusted AJCC 6th N (1988+)	Adjusted AJCC 6th N (1988+)
Distant metastasis	Derived AJCC M, 7th ed (2010+) Derived AJCC M, 6th ed (2004+) Adjusted AJCC 6th M (1988+)	Adjusted AJCC 6th M (1988+)
Surgery of primary site	RX SummSurg Prim Site (1998+) Surgery of primary site (1998-2002)	RX SummSurg Prim Site (1998+)
Surgery of other sites / regions	RX SummSurg Oth Reg/Dis (2003+) Surgery of oth reg/dis sites (1998-2002)	RX SummSurg Oth Reg/Dis (2003+)
Tumor size	CS tumor size (2004+) EOD 10 - size (1988-2003)	EOD 10 - size (1988-2003)
Extension	CS Extension (2004+) EOD 10 - extent (1988-2003)	EOD 10 - extent (1988-2003)
Lymph nodes	CS lymph nodes (2004+) EOD 10 - nodes (1988-2003)	EOD 10 - nodes (1988-2003)

Source Code 4: Conversion of illegal SEER values from breast cancer dataset into spaces

```
dataset.raw <- as.matrix(filterDataframe(dataset.seer))
dataset.temp <- which(is.na(dataset.raw)==TRUE |
    dataset.raw=="Blank(s)")
dataset.raw[dataset.temp] <- ""
dataset.final <- as.data.frame(dataset.raw)
```

The final dataset is composed of subsets from each time period and exported in ARFF format using the RWeka package's `write.arff` function [53]. A selection of 100,000 records was decided and the division of subsets included are seen in Figure 24.

Results

The demographics of the SEER data starting from the filtering step up to the exportation are presented in this part.

Breast cancer data in SEER*Stat is classified with one of the variables named "`vital status recode`" (VSR). It states if a record is "`Alive`" or "`Dead`". The corresponding graph for both the complete set and subset are shown in Fig. 23. The number of records that survived in the datasets, the complete at 83.93% and the susbet at 87.34%, both dominated the composition in terms of vital status.

In order to create the subset dataset, survival time of the records were taken into account. Figure 24 shows the distribution breast cancer records based on time of survival. The large number of representatives relates to the many-to-many assignment of binary outcome variables to the records. For example, two-year and four-year survival can be applied to a record that has a survival time of 45 months. In the subset dataset, the emphasis on the eight and ten-year records caused additional population increase in the lower years'.

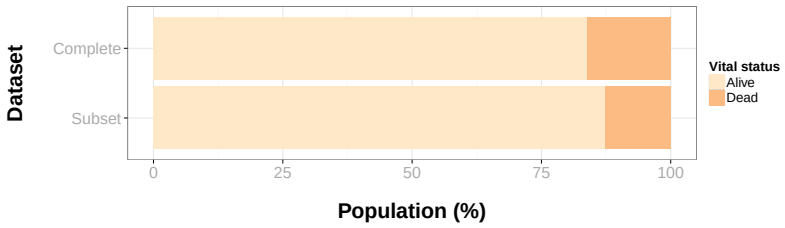


Figure 23: Graph of population distribution by vital status of the breast cancer datasets preprocessed from SEER

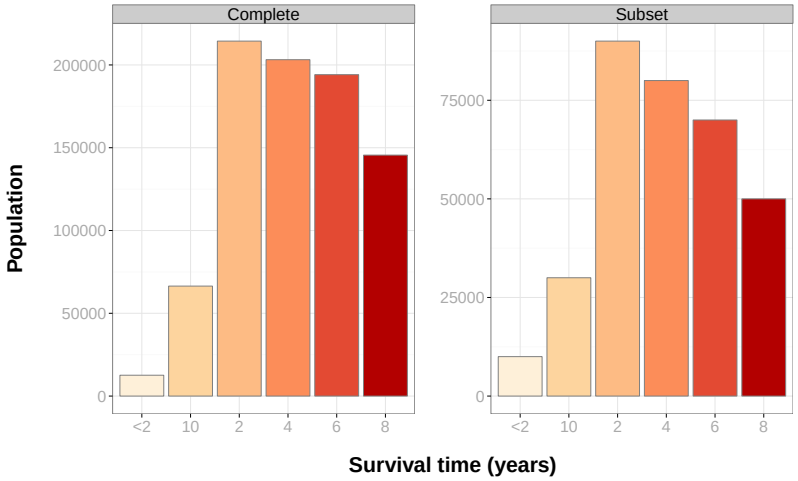


Figure 24: Graph of population distribution by survival time of the breast cancer datasets preprocessed from SEER

3. Data mining

The creation of predictive models were entirely done with the WEKA API. The breast cancer data were analyzed by classifiers in terms of relationship to the outcome variable through rules and computed weights. Please note that the terms "classifier" and "attribute" are used interchangeably with "algorithm" and "variable" respectively in this section; WEKA uses the first to refer to the latter.

A customized program was developed with the WEKA API to suit the needs of the study that were not implemented in the GUI version. The mechanism of this program is visualized in Fig. 25.

Creating the predictive models requires three major tasks: (1) to train the preprocessed breast cancer data using six classifiers; (2) to use an attribute selection algorithm on the data to get a subset of variables with comparable prediction ability to the complete set; and (3) to train the breast cancer data again only with the subset variables included. These tasks are discussed in parts in the succeeding sections, as illustrated in Fig. 25 in page 65.

From the preprocessing phase, the generated ARFF is read by the training program and saved as an `Instance` object.

Initialization

Before training, there are several parts are set manually before starting: the outcome time period and whether to use the full set or subset as determined by the set of attributes removed from the data. These are indicated in the filter code block by their index number in the data. Only one of the outcome variables two years, four years, six years, eight years and ten years were set as an outcome variable during training to make sure the other four does not affect the model's predictive performance. These are all set in the method `removeAttributes` defined in Source Code 5.

Source Code 5: The `removeAttributes` method of the `Training.java` class

```
private static Instances removeAttributes(Instances instance)
     throws Exception {
   // Indices of attributes to remove
   String[] optionsRemove = new String[] {
       "-R",
       "3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23,
           24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36" };
   Remove remove = new Remove();
   remove.setOptions(optionsRemove);
   remove.setInputFormat(instance);

   Instances instanceFiltered = Filter.useFilter(instance, remove);
   return instanceFiltered;
 }
```

Classifiers

The six classifiers (namely `ZeroR`, alternating decision tree, J48 decision tree, random forest, `LogitBoost` and random subspace) were set after the data is ready for training.

The study used the classifiers in their default parameters because the reference study did not specify any modifications. Table 14 contains the optional parameters for each of the five main classifiers used in the development. The content of these tables were taken from the online documentations (for columns "Code", "Name", "Value", "Author" and "Reference Journal") of the WEKA implementations and the WEKA GUI version 3.7.10 (for "Description") [46, 54–58].

Cross-validation

The cross-validation is run ten times to increase the representation of the records in the ten-fold setting during the classifier training. This process takes the longest to finish given the number of records, complexity of classifier's problem solving method and the number of times the cross validation process is repeated. As mentioned in Agrawal et al's paper, this method takes 3000 runs: five outcome variables, six classifiers, and ten runs

of ten-fold cross validation in total. The "performance metrics", i.e. accuracy, precision, specificity and others are saved incrementally per iteration and written to a CSV file at the end.

Results

In accordance to the output of the WEKA GUI's "Explorer" tab, a similar format was developed to mimic its "result buffer" that shows an exhaustive report of the recent run – classifier used and its parameters (all are default, indicated in tables); variables used from the dataset; a representation of creation model, if applicable, mostly weights or a decision tree; the performance metrics and a confusion matrix. This is saved in a generic text file.

After the ten runs of ten-fold cross-validation, a `.model` file is created and the next classifier is loaded until all six are used to train and create a model for prediction.

As indicated in the flowchart, the products of this part are: (1) a CSV of all classifiers' ten runs of cross validation, (2) a text file containing the amount of time it took for the training to finish and (3) the WEKA model files. Both the complete dataset and subset have these files generated for evaluation and use in the calculator (specifically the subset's models).

Attribute selection

Between the training of full set and subset datasets, WEKA's attribute selection was used to identify the attributes that will be included in the subset. (`CfsSubsetEval`) was selected as the "attribute evaluator" and a search algorithm called `BestFirst` was automatically set. Similarly to the training process, each outcome variable had a turn with ten-fold cross validation as "attribute selection mode". The output consists of the percentage of relevance of each variable to the prediction. The subset variables where chosen

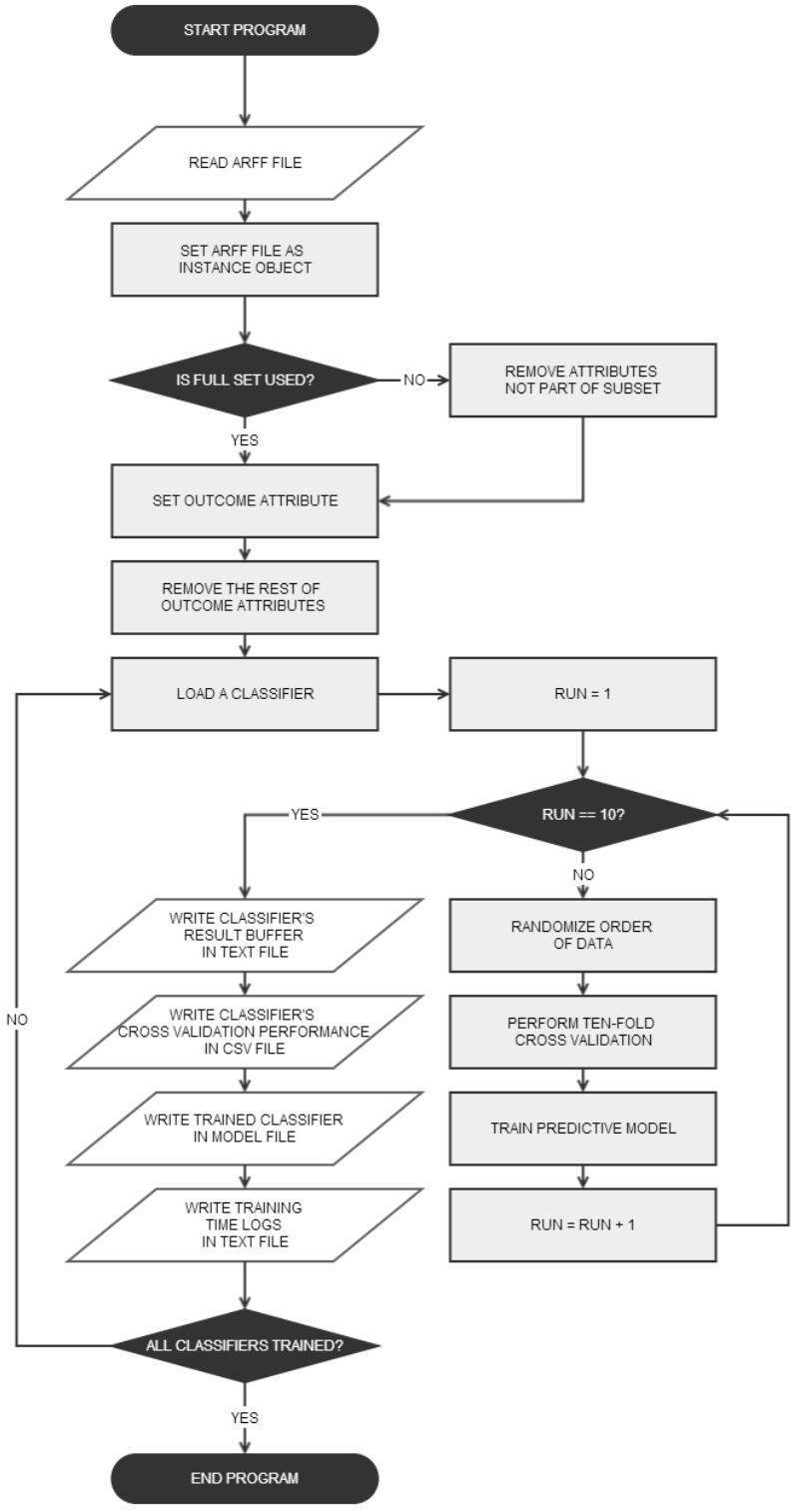


Figure 25: Flowchart of breast cancer predictive model creation and training using the WEKA API.

from all time period results if they registered a prediction relevance of at least 10%. These are recorded for the filtering in the subset training in the WEKA API.

4. Analysis

   The results of predictive modeling, mainly the performance metrics of the trained models, are dicussed in Chapter V..

C. Class diagrams

This part shows the relationship of classes used in the implementations of the predictive modeling and the BOSOM application's website and calculator systems.

NOTE: only the classes that the research made specifically have their attributes and operations shown in the class diagrams. The Java documentation for the rest are available for viewing online.

1. Predictive modeling

   The class diagram in Fig. 26 shows the entire class `Training.java` that is dependent on three categories of classes: (a) WEKA classifiers, (b) cross-validation, and (c) output generation. The first six classes beside `Training.java` are the classifiers used during the predictive modeling phase. This includes `Classifier.java`, `Instances.java`, and `Filter.java`. The seventh class from the second column is `Evaluation.java` which has methods for cross-validation. The ten runs of ten-fold cross-validation relied on this class. The last column has `AbstractOutput.java` and `CSV.java` which where used to generate the results of the training process.

2. BOSOM Calculator

   Fig. 27 shows a top-level view of how `CalcController.java` interacts with the prediction service layer and PDF creation controllers. It is also responsible for responding a view object to a user's request from the browser.

   `CalcController.java` passes data from the user to `CalcService.java` for prediction. The classes related to this service are shown in Fig. 28. The interfaces' implementations `CalcServiceImpl.java`, `CalcArffServiceImpl.java` and `CalcModelServiceImpl.java` work hand-in-hand to transform the breast cancer data into an `Instance` object and passed to a classifier for prediction.

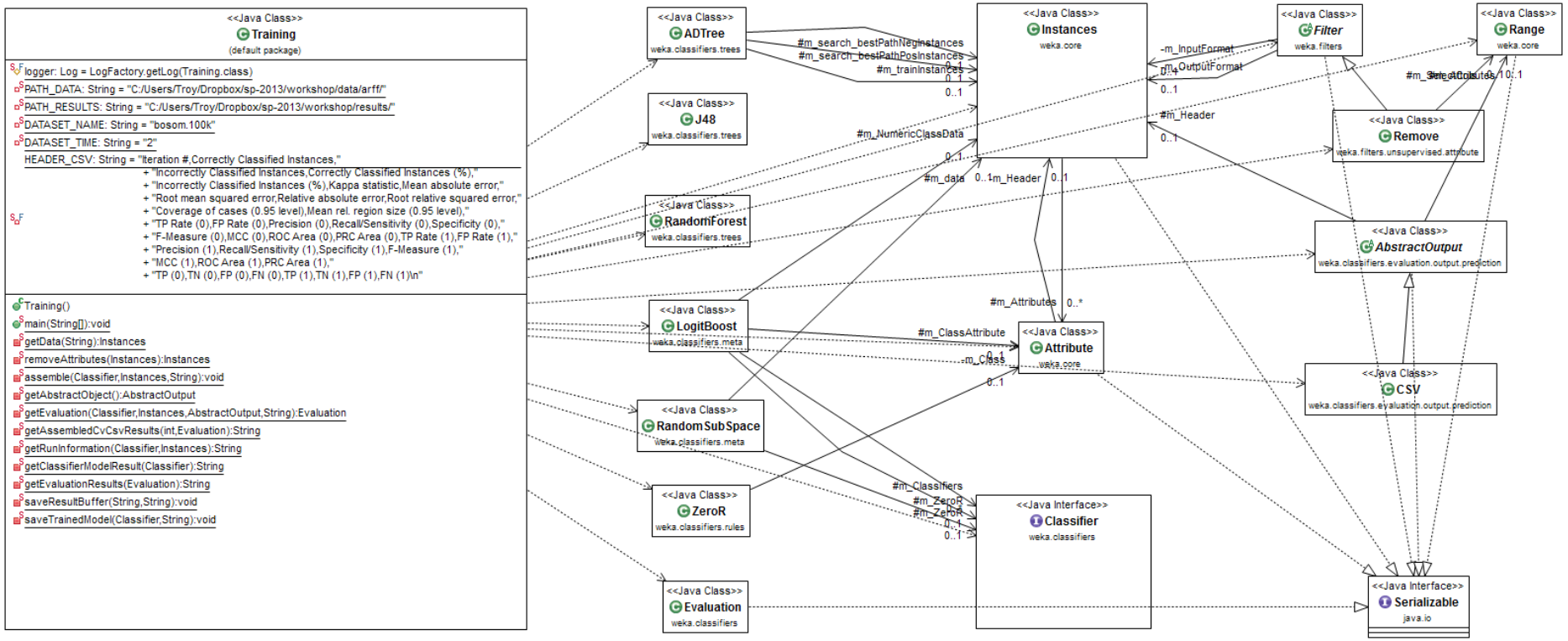


Figure 26: Class diagram of `Training.java`

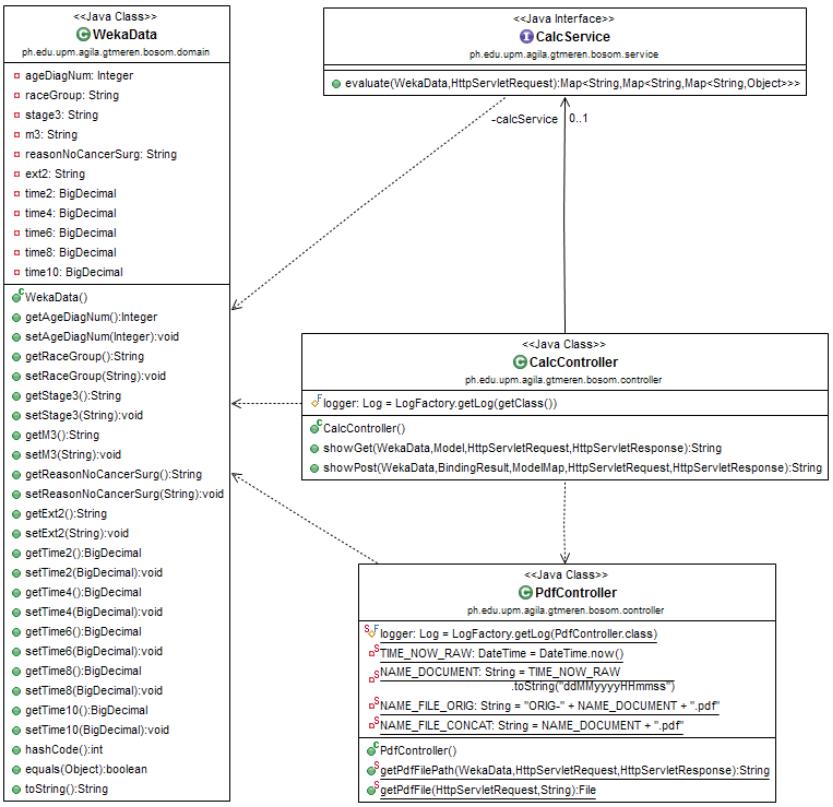


Figure 27: Top-level class diagram of `CalcController.java`

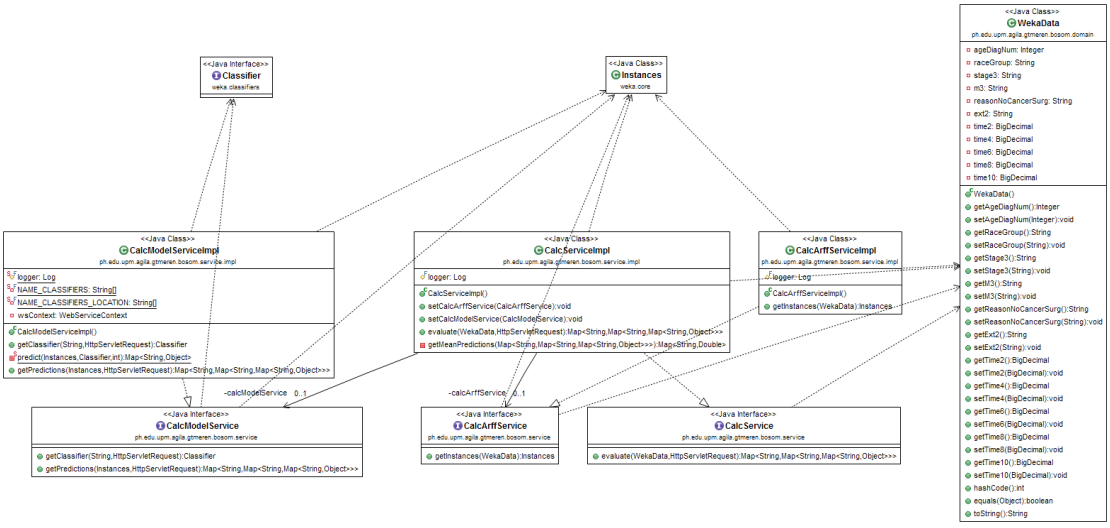


Figure 28: Class diagram of the BOSOM Calculator prediction module

The results of the BOSOM Calculator are available for PDF file export and several classes are employed for this task. Fig. 29 reveals the structure

of the PDF creation module where `CalcController.java` passes the breast cancer data and the output of the prediction module for PDF formatting.

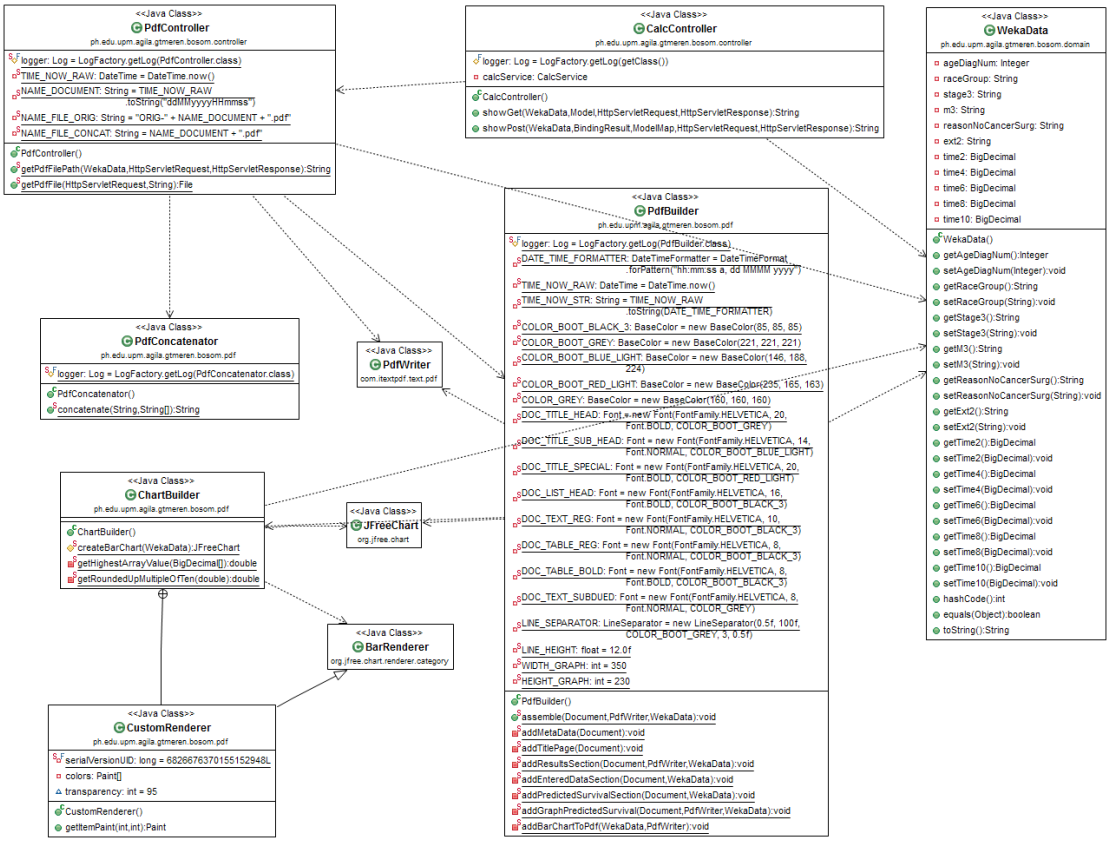


Figure 29: Class diagram of the BOSOM Calculator PDF creation module

The complete class diagram of the BOSOM application is provided in Fig. 30.

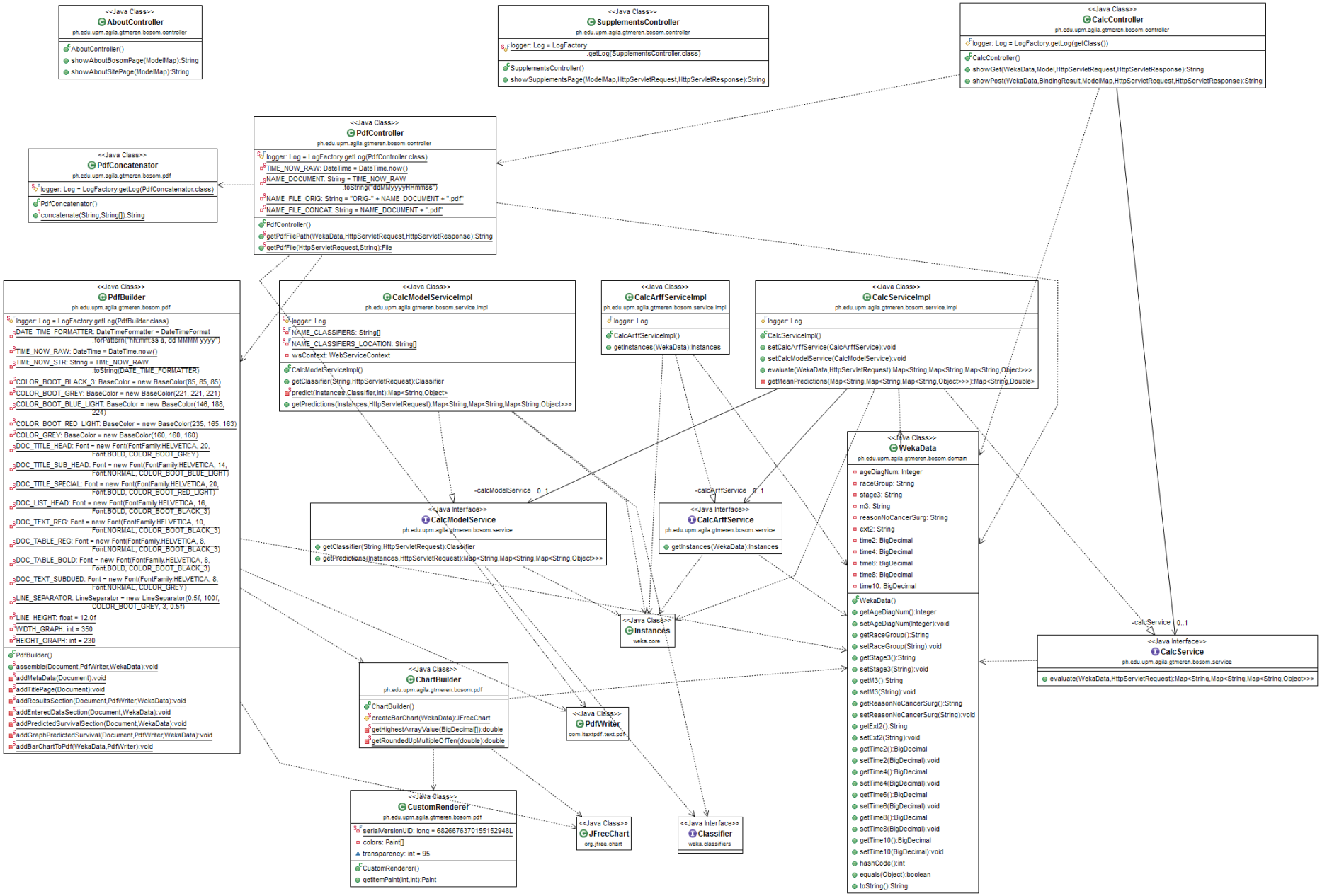


Figure 30: Class diagram of the whole BOSOM application

D. Architecture

D..1 System Architecture

The BOSOM application is made up of two major components namely the prediction system and the web application framework (WAP). A WAP enables the application to be built with web technology and served through browsers to users while the prediction system is responsible for predicting survival of a given breast cancer data. These two are thoroughly explained in the following paragraphs.

1. Web application framework

   Most modern web-based applications are developed with frameworks to increase productivity and hasten development [59]. Frameworks were create to lessen the programming effort a developer needs compared to traditional, non-framework websites and applications through the various fundamental features these implement such as session management, user interface templating and database connection [60].

   Java was chosen to be the main programming language of the BOSOM application because of the WEKA API's role. The latter is already written in Java thus a parallel implementation would ensure easier integration as opposed to a different programming language.

   Spring Framework is an open-source Java-based WAP based on the model-view-controller (MVC) software pattern that aims to separate the roles of each component. Spring is ideal for interacting with third-party technology because it is only referenced through a library and this setup gives way for other libraries (or "technologies") to be included in the application [61].

   The "entry point" of this application is the `Dispatcher Servlet` that calls for the index page to be viewed to a user. The rest of the views are served by a dedicated controller. `AboutController`, `CalcController`

and SupplementsController have URLs mapped to functions that return a view's name. These functions can call other functions from another controller or from the service layer.

The CalcController was developed to perform three main processes: deliver the form and results view, call the method to predict a WekaData object's survival, and call the method to create a PDF file, as indicated in the class diagram Fig. 27 and Fig. 29.

CalcController maps GET requests to the form view and POST to the results view, given that no errors occurred in the submission; else the first is returned to encourage correction of illegal input field values.

The models created from the first phase of this study were used in the prediction system. The domain object WekaData represents the data that was used to train the predictive models. WekaData has fields that correspond to the attributes of the model. When a user submits a successful form, it is read in the controller as a WekaData. This ensures the data's consistency inside the application.

Error handling is tied with the domain object. HibernateValidator, an annotation-based validation API, is implemented in WekaData, in the fields that are set with values from the form. How it is recognized by Spring is through XML configuration.

Next, a legal WekaData is passed to the service layer, to CalcServiceImpl, for prediction. This service calls CalcArffServiceImpl and CalcModelServiceImpl to construct an acceptable object for the model and to perform prediction using the model and object. These methods are explained further in the prediction system. After prediction, the WekaData is returned to the controller with predictions set to its fields.

Before the controller returns the results view, the PDF is constructed

first. `PdfController` manages the assembly of a PDF file with the help three classes. `WekaData` is passed to `PdfBuilder`, where the elements of the file are defined. The fields of the domain object are provided in these elements, i.e. a table. After constructing the page, `PdfBuilder` adds a bar chart of the predictions through `ChartBuilder`. `PdfController` passes the file location of the `PdfBuilder` file to `PdfConcatenator` to append a pre-made PDF file containing background information on the BOSOM Calculator. The concatenated PDF's file location is sent back to the controller to view in the results page. These classes are shown in Fig. 29.

In the results view, the `WekaData`'s fields are printed in a table; the predictions formatted in both table and graph; and the concatenated PDF file location set to a button as a link destination.

2. Prediction system

The backbone of the BOSOM application is the prediction system. It utilizes WEKA API to train and create the predictive models used by the calculator. It is integrated to the Spring application to reuse the models in performing prediction on new breast cancer data. The prediction system is based on the KDD process wherein each step has a similar function.

As mentioned before, the prediction system is implemented in the service layer with functions propagated among `CalcServiceImpl`, `CalcArffServiceImpl` and `CalcModelServiceImpl`.

Initializing the data

Once a `WekaData` object is passed to `CalcArffServiceImpl` through `CalcServiceImpl`, an `Instance` is constructed in order for the prediction models to recognize it. The content of an `Instance` object is provided in Source Code 6.

Normally the data WEKA accepts in the GUI is any of the formats ARFF, CSV and TXT but in the API it is not necessary to create such

files. `Instance` is used to store input data in WEKA. Numeric, nominal, date, string, and binary values are converted to floating point numbers where non-numeric variables set to their index in the dataset [46].

Source Code 6: A breast cancer data (`WekaData`) in WEKA's `Instance` format

```
@relation SeerBreastCancer

@attribute ageDiagNum numeric
@attribute raceGroup {Black,Other,Unknown,White}
@attribute stage3 {0,I,IIA,IIB,IIIA,IIIB,IIIC,IIINOS,IV,'UNK Stage'}
@attribute m3 {M0,M1,MX}
@attribute reasonNoCancerSurg {'Not performed, patient died prior to
    recommended surgery','Not recommended','Not recommended,
    contraindicated due to other conditions','Recommended but not
    performed, patient refused','Recommended but not performed,
    unknown reason','Recommended, unknown if performed','Surgery
    performed','Unknown; death certificate or autopsy only case'}
@attribute ext2 {00, 05, 10, 11, 13, 14, 15, 16, 17, 18, 20, 21, 23,
    24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 38, 40, 50, 60,
    70, 80, 85, 99}
@attribute time2 {0,1}
@attribute time4 {0,1}
@attribute time6 {0,1}
@attribute time8 {0,1}
@attribute time10 {0,1}

@data
50,Black,IIA,M0,'Surgery performed',11,?,?,?,?,?
```

An instance is created based on the model it will be tested on; hence great care was observed in this step. The model's attributes and values were explicitly added to the instance to act as a template – its purpose is to enable new instances to be recognized by the model. A `WekaData`'s field's values are assigned to the instance by array assignment to the corresponding attribute's index. Outcome attributes, however, are not yet set hence these are declared missing (or `Util.missingValue()` programmatically) and they are represented by question marks in the `@data` field. `CalcArffServiceImpl` returns the created instance object for prediction.

Loading the models

Similar to the instance, a model also has its own representation. The Classifier is an interface for all the classification and regression algorithms implemented in WEKA. In `CalcModelServiceImpl`, the trained models are read and casted as a Classifier object in order to be used.

Prediction proper

`CalcModelServiceImpl` handles instance prediction. An `Instance` object is passed by `CalcServiceImpl` and it is passed to a classifier as a parameter.

`classifyInstance` is the method used to predict an object's outcome attribute. It returns a double value that either represents the index of the classification or the predicted numeric value. In relation, `distributionForInstance` provides an instance's class membership; each nominal attribute value will have a respective value, numeric of course gets one prediction, while zero for unpredicted instance [46]. Source Code 7 shows how these were used in the program.

Source Code 7: WEKA Classifier methods to get an instance's class and class distribution

```
1 double outcomeValue = classifier.classifyInstance(instance);
2 double[] percentage = classifier.distributionForInstance(instance);
```

A linked hash map serves as the storage for the class and membership predictions. Its key – value access method is ideal for easier search for the values.

The five time periods' respective models are loaded in succession. Their filenames are declared in a string array and are mapped to the correct outcome variable (the time period).

Finally, the results collected in the linked hash map are extracted per time period. These are averaged (the ensemble voting technique) and set to the corresponding fields of the `WekaData` object.

The `WekaData` object is now predicted and sent back to `CalcController`.

D..2 Technical Architecture

1. Machine

   In the context of this study, "machine" refers to the computer used during software development. Its hardware and software specifications are provided below.

Table 8: Specifications of the machine used in the study

Component	Description
Operating system	Windows 7 Ultimate Service Pack 1
Machine	Acer© Aspire 4738ZG
Processor	Intel® Pentium® CPU P2600 @ 2.13GHz
Installed memory (RAM)	2.00 GB DDR3 memory
System type	64-bit operating system
Hard disk	500 GB SATA hard disk drive

2. Data preprocessing and transformation

   The data used in this study was provided by SEER. The cancer records where extracted from a database using SEER*Stat, a statistical analysis software. These records were exported to a CSV file using SEER*Stat's export tool. This format was chosen as it is accepted in majority of the data mining software that were used during the KDD process.

   In order to prepare the data for modeling, SEER*Stat and the open-source statistical programming language R were used. The integrated development environment (IDE) RStudio provided better usage and more convenient platform for using R because of features not limited to: project management, package management and integration with Git for version control.

   The preprocessed data were converted to ARFF through the `write.arff` R function from the RWeka package [53].

Table 9: Specifications of the data preprocessing and transformation step

Component	Description
SEER*Stat	Wersion 8.1.2 Built on 9 September 2013
R	Version 3.0.2 (2013-09-25) 64-bit
R IDE (RStudio)	Version 0.98.484

3. Data mining

Once the data is ready to use, the WEKA is used to create the models for the calculator. Due to the limitations of the GUI version regarding cross-validation, the open-source code (or "WEKA API") was used to build the desired setup. The following are the technologies required in this step.

Table 10: Specifications of the data mining step

Component	Description
Java Virtual Machine	`java version ï.7.0_25¨` Java(TM) SE Runtime Environment (build 1.7.0_25-b17) Java HotSpot(TM) 64-Bit Server VM (build 23.25-b01, mixed mode)
Java IDE (Eclipse)	Eclipse Java EE IDE for Web Developers Version: Luna Release
WEKA	Waikato Environment for Knowledge Analysis Version 3.7.10 (Developer version API)

4. Website framework

The website's backend program was developed with Spring Framework, an open-source Java-based web framework. The complete list of Java Archive (JAR) files necessary for the application is enumerated in Table 12.

The user interface is made from HTML, CSS, JavaScript and jQuery with Bootstrap and Flot.js for enhancing user experience.

The entire application is exported as a Web Application Archive (WAR) file that the server will process for online deployment.

5. Server

   The University of the Philippines Manila's Agila Computer Science Development Server currently hosts the application. It has the following specifications that the application fulfills in order to run successfully [62]:

   - Apache 2.2.22
   - Apache Tomcat 7.0.47
   - Java 1.6.0_26

6. End user device and browser

   All users are required to have consistent Internet connection in order to visit the website and use the application.

   The application requires any modern browser (Webkit and FireFox for best quality) to run with a capable device.

Table 11: Test environments for the BOSOM application

Component	Description
Device	General purpose laptops Samsung Galaxy 4S Samsung S2
Browser	Google Chrome 14.0 and newer (latest is 32.0) Mozilla Firefox 3.0 and newer (latest is 27.0) Safari 4.0 and newer (latest is 5.1) Opera 10.6 and newer (latest is 19.0) Internet Explorer 11 Internet Explorer 10 Internet Explorer 9 Internet Explorer 8 (Windows XP) Internet Explorer 7 (Windows XP)

All browsers are tested in a Windows 7 machine unless stated otherwise.

The generated PDF files need around 5 kilobytes of space if saved. Any basic PDF reader or capable program is required to view the contents of the file.

V. Results

The goal of this study is to develop a breast cancer prediction application. In order to assess its performance, the performance metrics for each model per cross-validation fold where recorded. A presentation of the developed application user interface follows.

A. Data mining

Sixty models were trained from two breast cancer datasets, both consisting of 100, 000 number of patients. The first dataset is the complete SEER dataset obtained from the preprocessing and transformation phase and the second only has a subset of variables included, selected using an attribute selection algorithm. Training involved ten-fold cross-validation repeated ten times to increase coverage of records used. Only the results of each iteration were recorded and not the entire 100 folds due to limitations of the API.

The term "outcome variable" refers to the survival time of a record and this could to be either `0` ("dead") or `1` ("alive"). This variable is formally the `Vital Status Recode` from the SEER*Stat database. Unfortunately, only the metrics for the class "dead" were recorded during the complete dataset training due to shortcomings in the training phase. The recorded performance metrics of each trained model were accuracy, precision, recall or sensitivity, specificity, and area under the receiver operating characteristic curve (ROC). These are abbreviated as ACC, PRE, REC, SPE and ROC respectively in Tables 22 and 24. The metrics were computed by the WEKA API, and the data presented are averages of the cross-validation iterations. Lastly, the classifiers `ZeroR`, alternating decision tree, J48 decision tree, random forest, `LogitBoost` and random subspace are abbreviated as ZR, ADT, J48, RF, LB and RS in the table respectively. The mean in the table represents the ensemble voting method, as required in the reference journal

by Agrawal et. al. that is used to combine the results of the classifiers.

A trend of declining accuracy in breast cancer survival prediction was observed in the complete dataset, in Table 22, as the survival time increases. This is related to the population representation for each survival time, where the earlier years had larger number of records as opposed to the later years, as seen in Fig. 24. This trend was also apparent in the other metrics except for the recall / sensitivity which had a sudden peak in the ten-year survival. In relation, it is sensible to conclude that deaths are more frequent to manifest in the later years, which could have led to the sudden rise in death prediction recall.

The complete breast cancer data was applied with an attribute selection algorithm, the correlation-based feature subset selection, to find a small number of variables with the most prediction ability. The WEKA GUI was used and `BestFirst` is set as the search algorithm. Ten-fold cross-validation was also applied during the process and the same five binary variables as outcome. In Table 23, variables that scored of at least 10% in any time period were included in the final selection and the six that fulfilled the criteria all registered scores of 100%. These are: "`Age at diagnosis`" (`ageDiagNum`), "`Race recode (White, Black, Other)`" (`raceGroup`), "`Adjusted AJCC 6th (1988+)`" (`stage3`), "`Adjusted AJCC 6th M (1988+)`" (`m3`), "`Reason no cancer-directed surgery`" (`reasonNoCancerSurg`), and "`EOD 10 - extent (1988 - 2003)`" (`ext2`). Based on the scores, it is obvious that `ageDiagNum`, `m3` and `reasonNoCancerSurg` hold significance in the prediction of all survival times. The subset models, and in turn the BOSOM Calculator, only used these variables for prediction. The complete list of these selected variable and their values are shown in Table 21.

The BOSOM Calculator models' performance is comparable to the original dataset in terms of the decreasing accuracy against increasing time of survival. In fact, the predictive accuracy decreased from 93% to a 75% within ten years. Again, the inferior representation of the later years in the dataset is one of the

possible causes of the decline. As for the rest of the performance metrics, the declining trend persisted. Agrawal et. al.'s ensemble voting and LCOC's metrics, however, are the inverse. The accuracies from six months to five-year survival where increasing, the same of the number of records with vital status recorded as "dead". It is observed that the accuracy of a model has a direct relationship with the number of dead records in the cancer dataset.

The accuracies of the ensemble voting and BOSOM Calculator, seen as "EV" and "BOSOM" in Fig. 31 respectively, are clearly near each other. For the best case, or predicting survival within two years, each performed with 94.0611% and 93.0816% with a difference of 0.9795%; in contrast, their worst were at 82.2363% 74.7720% with a fairly large difference of 7.4643%. In LCOC, the best was on the five-year survival at 91.4% and 91.2% and the worst, in six-month survival, is 73.6% and 72.5% [3]. This shows the better performance of the breast cancer predictive models, but it is important to note the differences of accuracies of the lung cancer models are less distant from each other. The LCOC's upper bound of difference is 1.1% for th six-month and a lower bound of 0 for the two-year survival. The BOSOM is less impressive with a large upper bound of 9.9852% for ten-year and lower bound of 0.9795% for two-year survival differences. Nevertheless, the BOSOM Calculator was able to perform with proximal accuracy to the models trained with the complete selection of 36 variables.

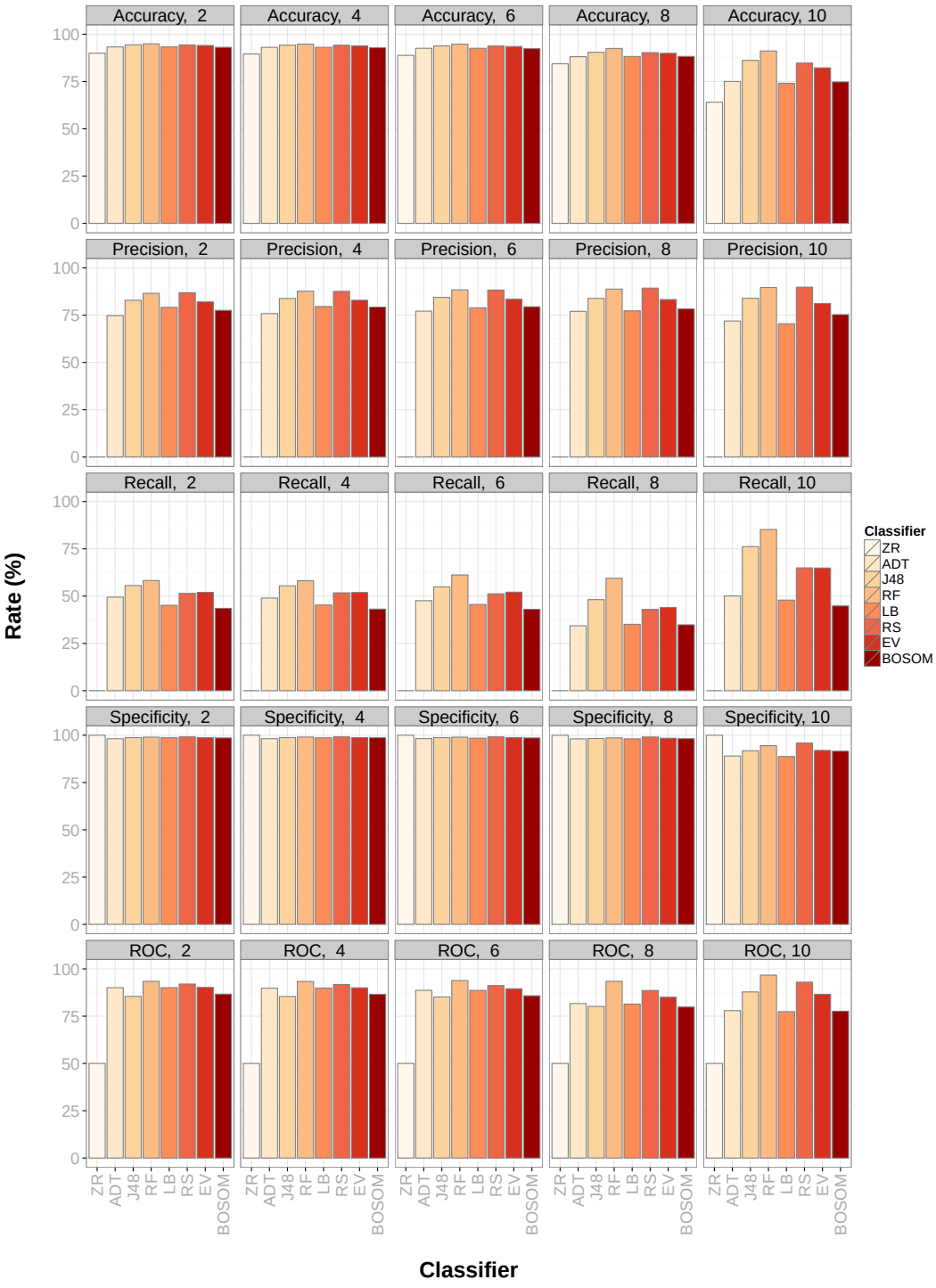


Figure 31: Graphs of the combined performance metrics per outcome variable of the baseline classifier, five predictive models, ensemble voting and BOSOM Calculator

B. Predictive modeling

The BOSOM Calculator uses 25 predictive models trained with a breast cancer dataset of 100,000 records. A custom program using the WEKA API was made to create and train the models. Its source code is provided in Source Code 10 and a flowchart in Figure 25. There were six classifiers namely `ZeroR`, alternating decision tree, J48 decision tree, random forest, `LogitBoost` and random subspace that were trained and each was set to predict five binary outcome variables: two, four, six, eight and ten years. Ten runs of ten-fold cross-validation was applied during the process to increase the probability of including all records in the dataset during the training.

As shown in Fig. 25, the classifiers were loaded one at a time until all models are created. The order of training was: `ZeroR`, random forest, `LogitBoost`, random subspace, J48 decision tree and alternating decision tree; these were arranged in increasing order of execution time. This ordering was determined during the complete dataset training phase and fortunately the hypothesis reflected on the smaller dataset's implementation. Tables 19 and tab:training-time-subset shows a breakdown of execution time for each model in the complete and subset respectively. The `ZeroR` performed the fastest since it only determines the dominant class attribute and alternating decision trees require additional tree traversal and weighing computations in order to arrive at a approximate solution to the classification problem.

Training the complete dataset took 21 hours to finish, with each outcome variable consuming around four hours. Figure 32 is a partial view of the console log during the training.

The subset dataset, as expected, took less time to finish. Seven hours of training time was spent on the five outcome variables in total. Figure 33 is a partial view of the console log during the training.

WEKA's "result buffer" is a valuable source of information about a generated model's description and performance. It is provided in the GUI version after training a model. There is no actual implementation of the result buffer generation in the API version but an equivalent was recreated in the program used in this study. Each classifier has its own representation of the trained model. Tables 17 and 18 provides the characteristics of each classifier.

The alternating decision tree models' number of number of nodes and predictor nodes were constant in both datasets. In spite of these, these trees have different structures as seen in Source Codes 43, 44, 45, 46 and 47. Next is the J48 decision tree whose number of leaves and the size of tree are proportional to the number of variables in the dataset. An average of 7991 leaves where found in the complete and only a mere 652 on the subset. The result buffers of J48 are often long due to the explicit structure of tree defined.

Random forests provide the out-of-bag (OOB) error estimates that correspond to the bootstrap cases that did not match the true case. It is obvious that the average OOB increased from 8.05% to 11.81% in between datasets. This could be attributed to the less number of variables used to determine the outocme variables.

Finally is the random subspace whose size of trees created where also proportional per dataset. In the complete, 7009 was the highest average size while 296 trees is to the subset.

`LogitBoost` was not included because of the lack of a generalized form of its iteration and weight results.

```
Feb 05, 2014 5:39:36 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//cv-results/bosom.100k.2.zr.folds.results.CSV]

Feb 05, 2014 5:39:40 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//main-results/bosom.100k.2.zr.result.buffer.TXT]

Feb 05, 2014 5:39:40 PM Training saveTrainedModel
INFO: Successfully created model [C:/Users/Troy/Dropbox/sp-2013/workshop/results//models/bosom.100k.2.zr.MODEL]

Feb 05, 2014 6:02:59 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//cv-results/bosom.100k.2.rf.folds.results.CSV]

Feb 05, 2014 6:02:59 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//main-results/bosom.100k.2.rf.result.buffer.TXT]

Feb 05, 2014 6:02:59 PM Training saveTrainedModel
INFO: Successfully created model [C:/Users/Troy/Dropbox/sp-2013/workshop/results//models/bosom.100k.2.rf.MODEL]

Feb 05, 2014 6:42:22 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//cv-results/bosom.100k.2.lb.folds.results.CSV]

Feb 05, 2014 6:42:22 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//main-results/bosom.100k.2.lb.result.buffer.TXT]

Feb 05, 2014 6:42:22 PM Training saveTrainedModel
INFO: Successfully created model [C:/Users/Troy/Dropbox/sp-2013/workshop/results//models/bosom.100k.2.lb.MODEL]
```

Figure 32: Partial console log of training the complete breast cancer dataset for predicting two-year survival

```
Feb 08, 2014 11:47:25 AM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//cv-results/bosom.100k.2.zr.folds.results.CSV]

Feb 08, 2014 11:47:32 AM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//main-results/bosom.100k.2.zr.result.buffer.TXT]

Feb 08, 2014 11:47:34 AM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//main-preds/bosom.100k.2.zr-preds.CSV]

Feb 08, 2014 11:47:35 AM Training saveTrainedModel
INFO: Successfully created model [C:/Users/Troy/Dropbox/sp-2013/workshop/results//models/bosom.100k.2.zr.MODEL]

Feb 08, 2014 12:10:40 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//cv-results/bosom.100k.2.rf.folds.results.CSV]

Feb 08, 2014 12:10:43 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//main-results/bosom.100k.2.rf.result.buffer.TXT]

Feb 08, 2014 12:10:50 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//main-preds/bosom.100k.2.rf-preds.CSV]

Feb 08, 2014 12:10:51 PM Training saveTrainedModel
INFO: Successfully created model [C:/Users/Troy/Dropbox/sp-2013/workshop/results//models/bosom.100k.2.rf.MODEL]

Feb 08, 2014 12:26:13 PM Training saveResultBuffer
INFO: Successfully created file [C:/Users/Troy/Dropbox/sp-2013/workshop/results//cv-results/bosom.100k.2.lb.folds.results.CSV]
```

Figure 33: Partial console log of training the subset breast cancer dataset for predicting two-year survival

C. BOSOM application

The Breast Cancer Outcome - Survival Online Measurement (BOSOM) Calculator is a web application that aims to provide a patient predicted survival estimates from two to ten years. It requires six fields from the user, in context to the patient who needs analysis: age at diagnosis, race, spread of metastasis, details of cancer-directed surgery, and extension of primary tumor, in order to provide a prediction. The other sections of the website are mostly geared towards informing the user on what BOSOM is about. The discussion of the sections of the BOSOM application are divided into the four main pages: "Home", "About", "BOSOM Calculator" and "Supplements".

NOTE: All the web pages presented in this part were rendered in a Google Chrome browser (version `33.0.1750.146 m`) unless stated.

1. Home page

   All the vistiors of the website are allowed to use the calculator. The home page is shown in Fig. 34. A welcome banner directly presents the application to the user, with a button to view the about page. This encourages the user to read first about the application before using the application.

   The header contains links to all four main pages, as well as an exhaustive counterpart in the footer. A simple description about the developer and the university where the application originated is also included.

2. About pages

   As seen in Figures 35 and 36, there are two pages under this section - "BOSOM Calculator" and "BOSOM Site". Both contain information about the application; the difference is the scope of the content.

   "BOSOM Calculator" serves to inform the user about the key princi-

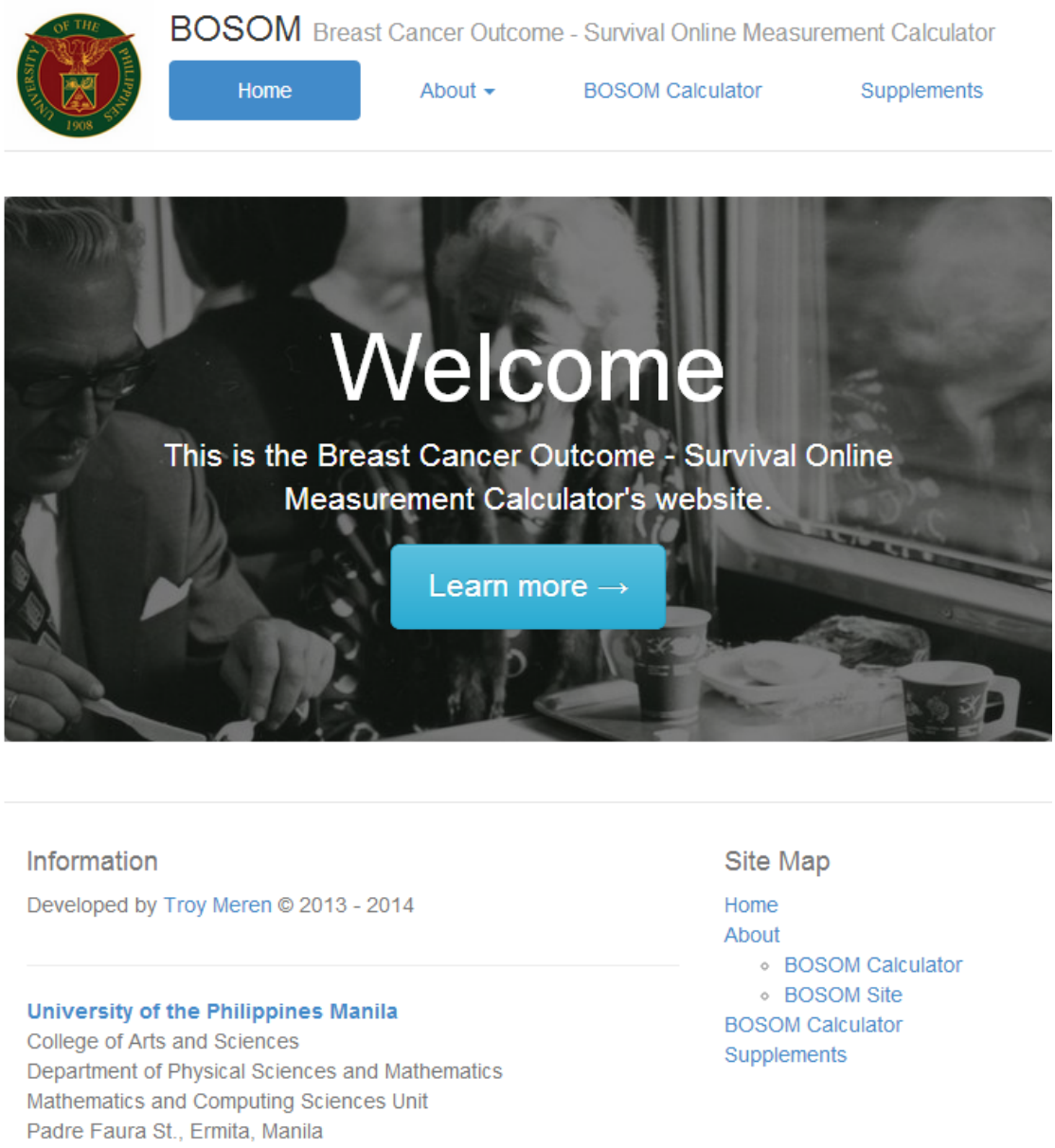


Figure 34: BOSOM application home page

ples that govern the application - breast cancer, data mining and predictive modeling. These were written in the most understandable way as possible, diverting from higher level terms to draw the user's attention into learning more about the subjects.

In relation, the "BOSOM Site" contains brief explanations about the technical components of the application and serves to attract users with familiarity to the technologies explain. Another feature of this section are the references to the open-source technologies used in the application, as a form of gratitude (by linking to their home pages or other required website) and recommendation for users who could be interested in learning more about them and hopefully apply to their own applications in the future.

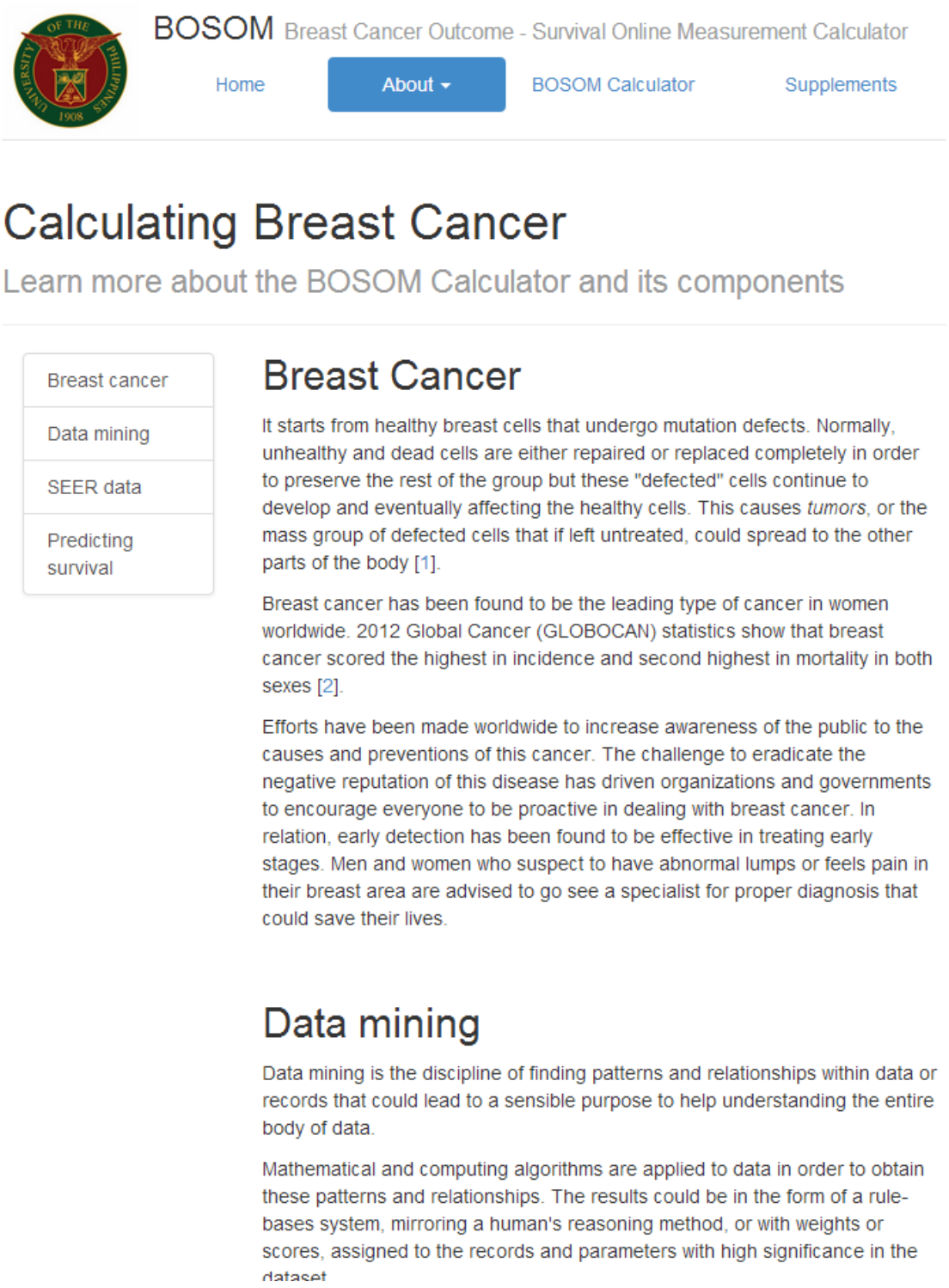


Figure 35: Partial view of the About BOSOM Calculator page

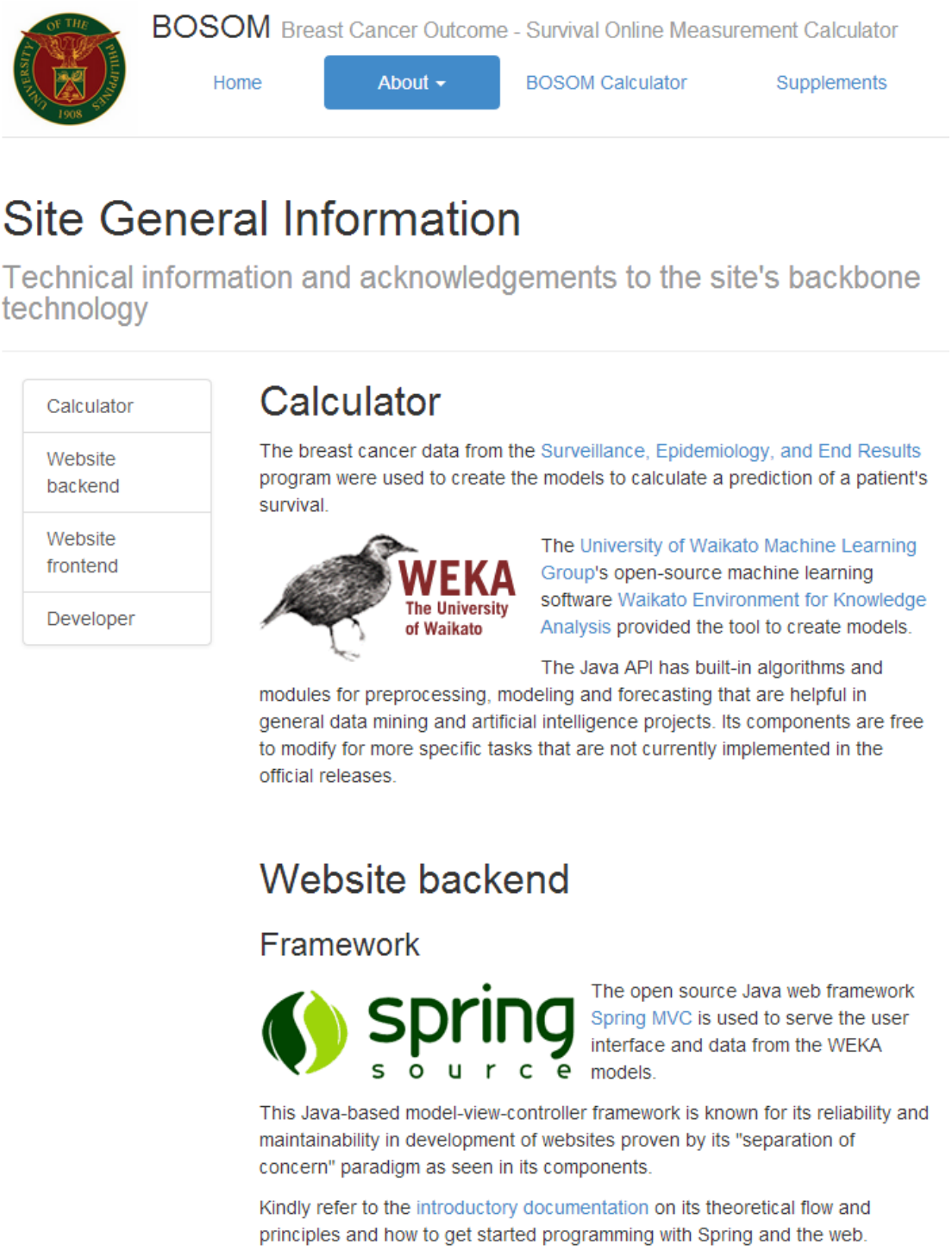


Figure 36: Partial view of the About BOSOM website page

3. BOSOM Calculator

The BOSOM Calculator has two parts - the form and the results page. When a user visits the page, the form page is revealed first, as seen in Fig. 37. The user has to input their breast cancer details to the fields provided. Some of the values are highly medical -intensive thus a "More information here" button serves to reveal a modal window consisting of the values' definition and reference. Figure 38 demonstrates the modal window for "Extension of primary tumor". To the left of the form is a simple reminder section whose purpose is to reiterate the instructions and a note that the application is not meant to replace a doctor's diagnosis of breast cancer.

There are six fields: five dropdown boxes and a text box. Depending on the user's browser, a submitted blank form will trigger either a small popup window near the empty field with an instruction, or red marks and notes as seen in Fig. 39. A large alert box will restate the need to accomplish the form without any errors. The first phenomenon is achieved through client-side validation or `HTML5` Validation and popup messages will show instructions about the illegal input. Figures 39, 40 and 41 show the empty fields and illegal input format popup events. The latter is server-side that uses `HibernateValidator`. The user can click the "Clear" button to remove all inputs on the fields. After submitting an error-free form (in the context that all forms are filled out correctly by the predetermined format), the user is directed to the results page.

A sample results page is composed of four sections namely: "Entered data", "Table for predicted survival", "Graph for predicted survival" and "Export results as PDF". The entered data is just a table of the form and the values the user answered as a reminder; the predicted survival are presented in both table and bar graph form to aid users who prefer either of the two forms of visual communication. These sections are presented in Figures 42,

43, 44 and 45 respectively. A PDF export feature is at the end, which is just a PDF-formatting of the results page, in case the user wants to save their BOSOM Calculator predictions.

During the models' computation process, the application prints logs of each classifier-model-prediction process for backtracking and investigation if ever errors occur. In fact, this is helpful in server deployment because these logs are saved into the server's own documentation of each process loaded in its premise and this can be used for inspection once errors occur. Figure 46 is a partial view of this logging process, as seen in an Eclipse version Luna Java IDE. This happens after a user submits a complete and error-free calculator form. The complete log is available in Appendix E..

Figure 47 is a sample BOSOM Calculator results report in PDF format. The PDF export has multiple possibilities of behavior depending on the user's browser and device. Ideally, the file will be displayed by the browser's native PDF reader and this was tested in Google Chrome, Mozilla Firefox and Internet Explorer on a general purpose computer or laptop. For mobile and tablet devices, the PDF file is not viewed immediately; instead, a save prompt will appear to ask the user if they want to save the file or the file is downloaded immediately. The BOSOM Calculator page's user interface in selected browsers and devices was shown in Table 11.

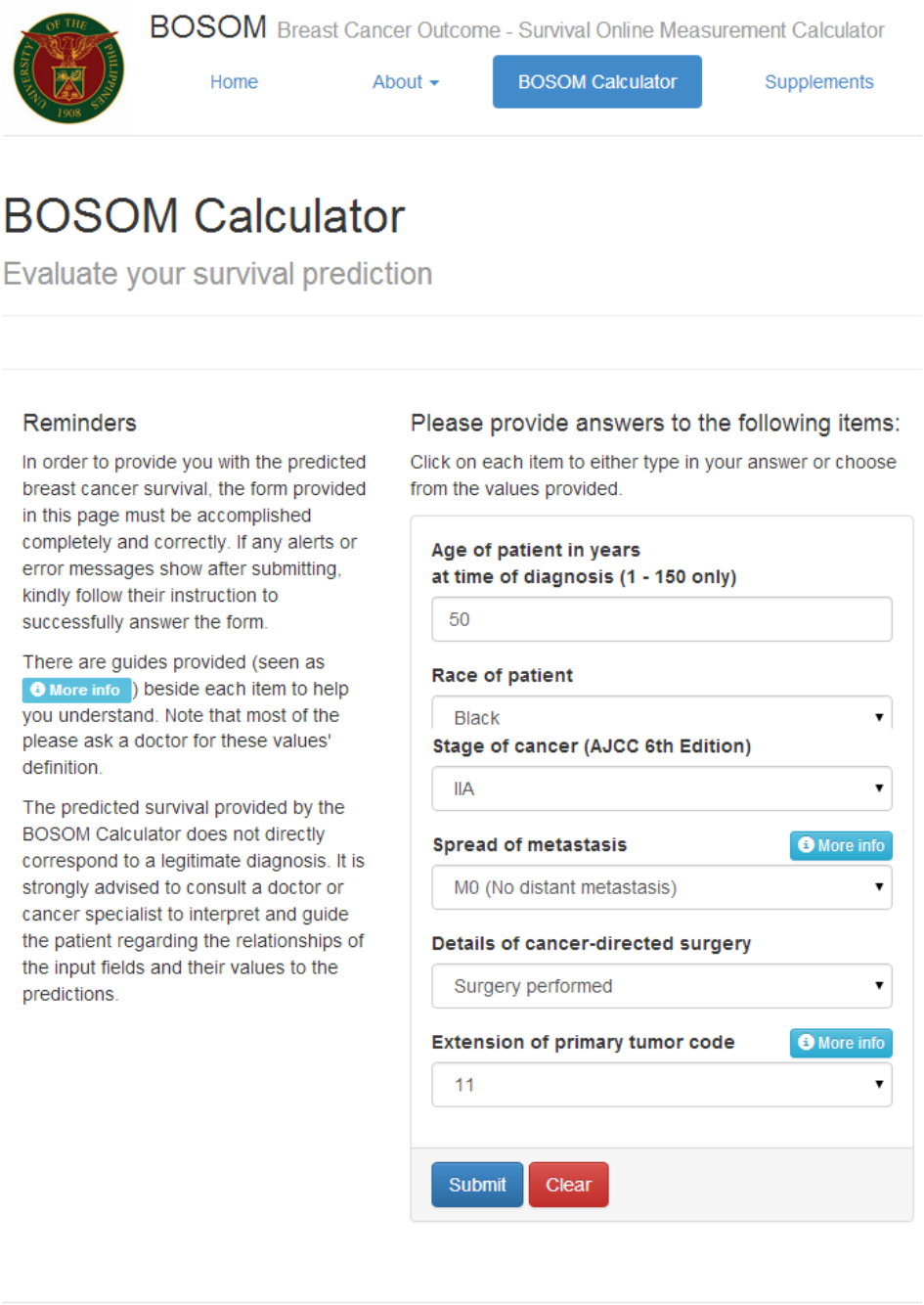


Figure 37: BOSOM Calculator form page with input data

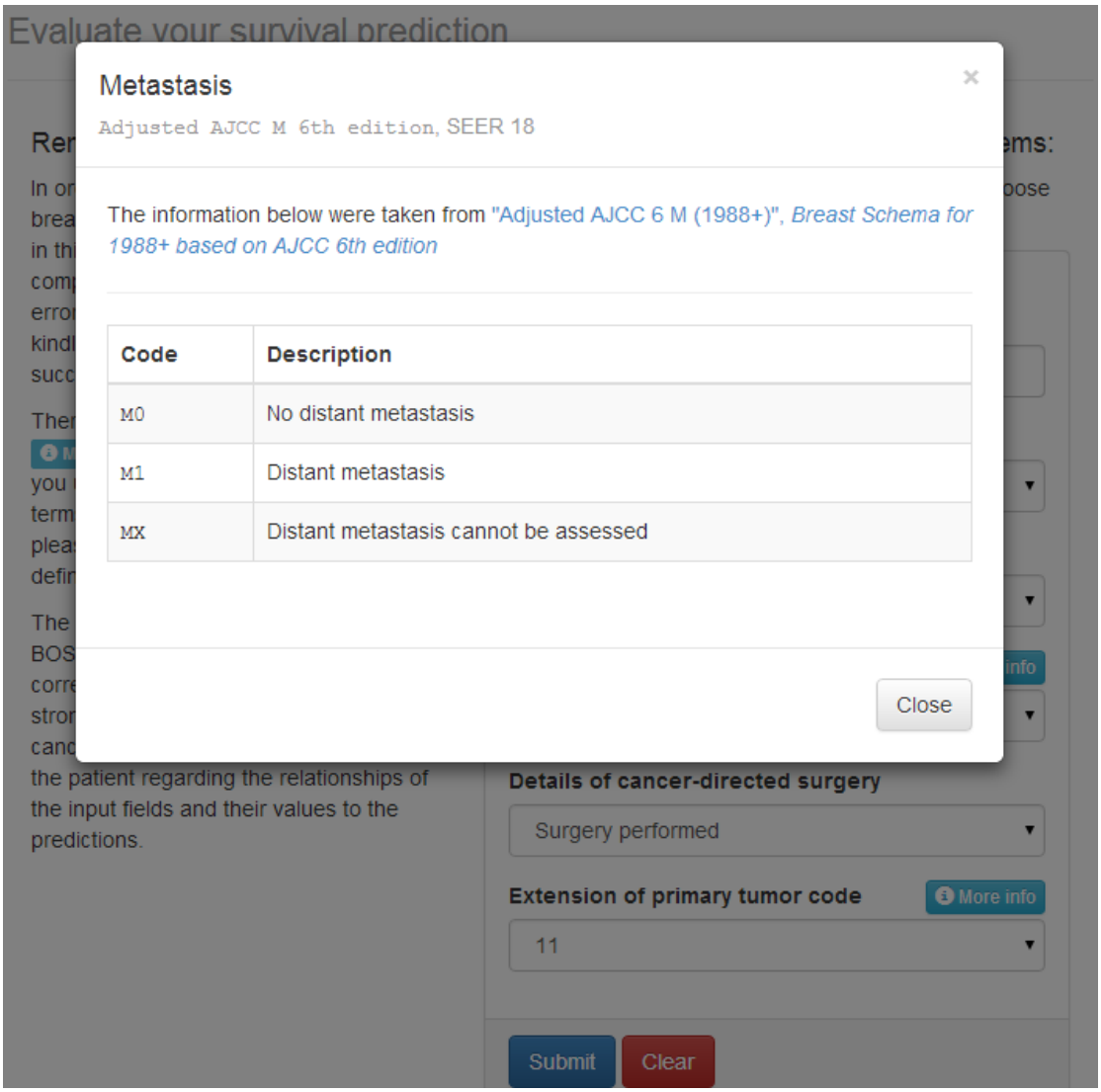


Figure 38: BOSOM Calculator modal window containing details for "spread of metastasis"

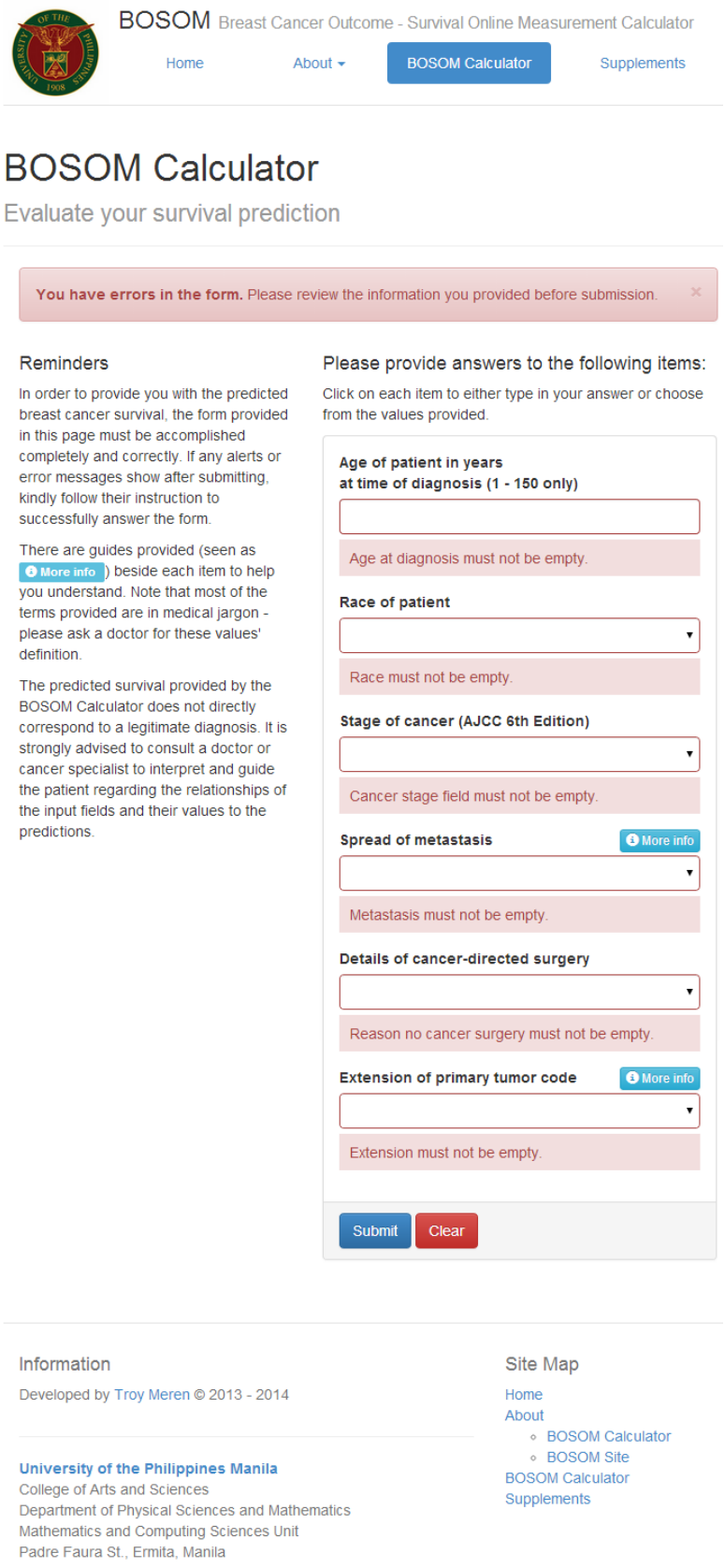


Figure 39: BOSOM Calculator server validation view

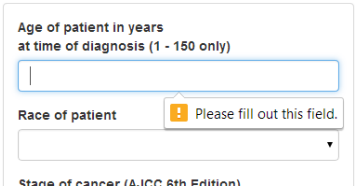


Figure 40: BOSOM Calculator client-side form validation for an empty field example

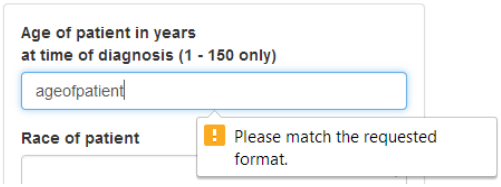


Figure 41: BOSOM Calculator client-side form validation for an illegal input format example

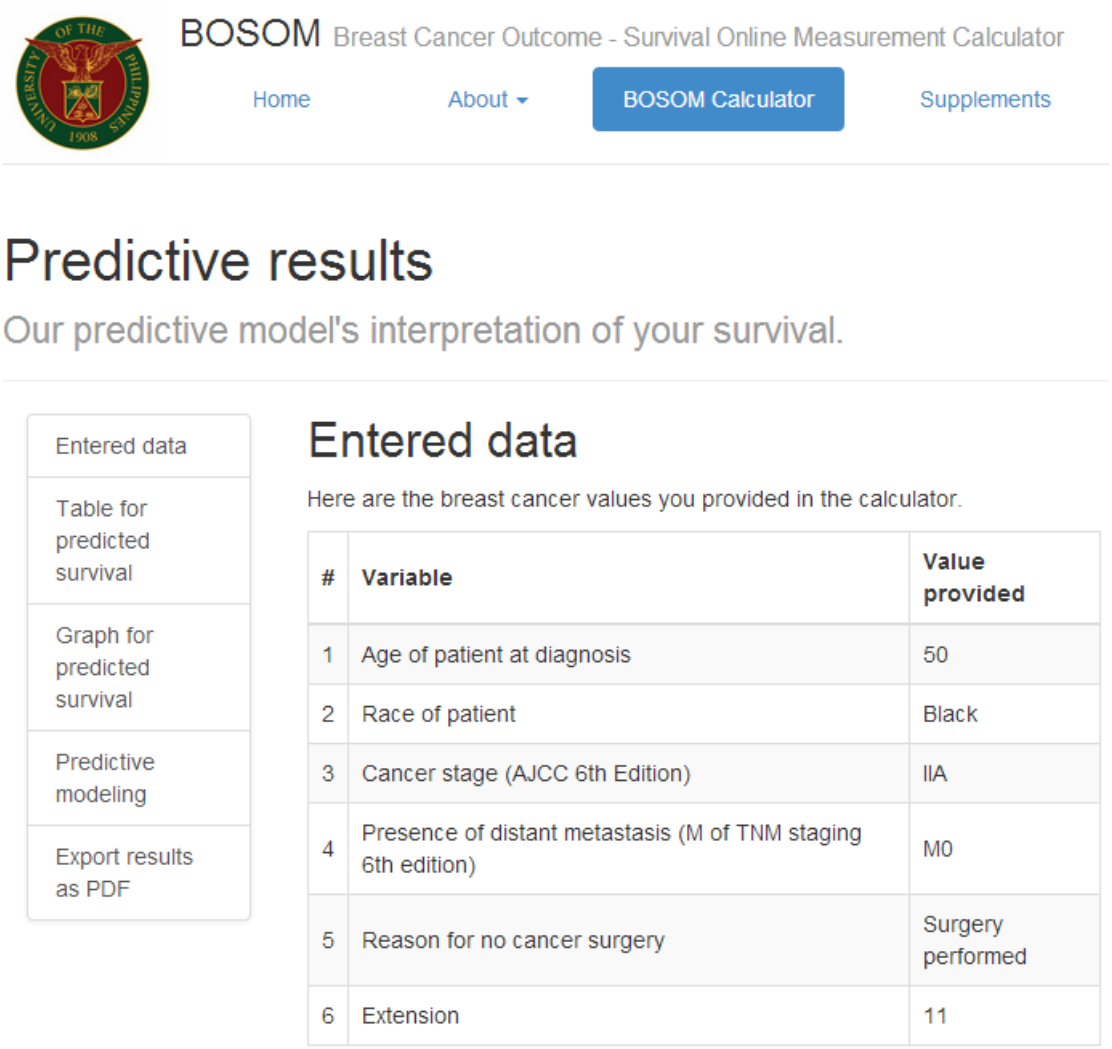


Figure 42: BOSOM Calculator results "Entered data" section

Table for predicted survival

Here are the predicted survivals as determined by our models based from past breast cancer patient records. These are from two to ten years, with two years of interval for uniformity.

Time period	Prediction model	Predicted survival (%)	Mean of predicted survivals (%)
2 years	ADT	71.99	87.82
	LB	92.50	
	J48	94.61	
	RF	86.29	
	RS	93.70	
4 years	ADT	72.64	88.01
	LB	93.49	
	J48	94.25	
	RF	86.29	
	RS	93.38	
6 years	ADT	79.65	83.12
	LB	86.39	
	J48	93.55	
	RF	63.31	

Figure 43: BOSOM Calculator results "Table for predicted survival" section

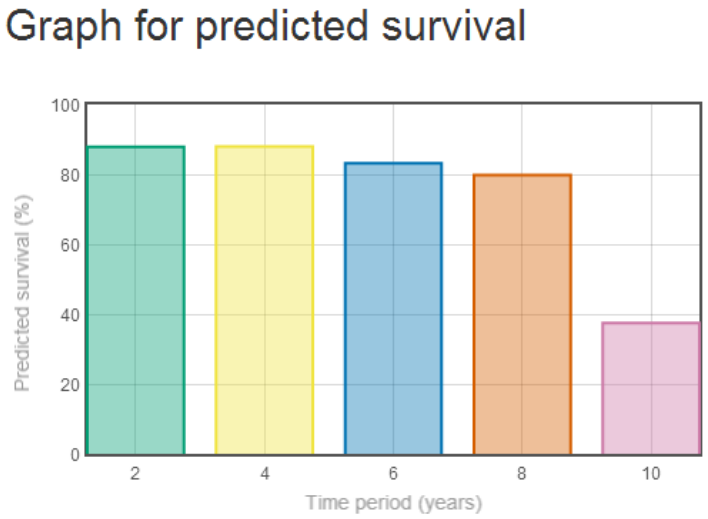


Figure 44: BOSOM Calculator results "Graph for predicted survival" section

Figure 45: BOSOM Calculator results "Export results as PDF" section

```
Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.controller.CalcController showF
INFO: Form data:  WekaData [ageDiagNum=50, raceGroup=Black, stage3=IIA, m3=M0, reasonN

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.impl.CalcArffServiceImp
INFO:
CalcArffServiceImpl: creating Instances data
@relation SeerBreastCancer

@attribute ageDiagNum numeric
@attribute raceGroup {Black,Other,Unknown,White}
@attribute stage3 {0,I,IIA,IIB,IIIA,IIIB,IIIC,IIINOS,IV,'UNK Stage'}
@attribute m3 {M0,M1,MX}
@attribute reasonNoCancerSurg {'Not performed, patient died prior to recommended surge
@attribute ext2 {00,05,10,11,13,14,15,16,17,18,20,21,23,24,25,26,27,28,30,31,33,34,35,
@attribute time2 {0,1}
@attribute time4 {0,1}
@attribute time6 {0,1}
@attribute time8 {0,1}
@attribute time10 {0,1}

@data
50,Black,IIA,M0,'Surgery performed',11,?,?,?,?,?

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.impl.CalcModelServiceIm
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time2/adt.MODEL

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.impl.CalcModelServiceIm
INFO: CalcModelServiceImpl: predicting class and its percentage distribution
Classifier: class weka.classifiers.trees.ADTree
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.28012585481457003
Percentage [1]: 0.71987414518543

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.impl.CalcModelServiceIm
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time2/lb.MODEL
```

Figure 46: Partial console log of the prediction process in BOSOM Calculator once a user submits a validated calculator form

BOSOM

Breast Cancer Outcome - Survival Online Measurement Calculator

Generated on: 12:54:53 AM, 12 March 2014

CALCULATOR RESULTS REPORT

Entered data

Here are the breast cancer values you provided in the calculator.

#	Variable	Value provided
1	Age at time of diagnosis	50
2	Race of patient	Black
3	Stage of cancer	IIA
4	Spread of etastasis	M0
5	Details of cancer-directed surgery	Surgery performed
6	Extension of primary tumor	M0

They are used to calculate the predicted survival rate given in the other sections.

Predicted survival

Here are the predicted survivals as determined by our models based from past breast cancer patient records. These are from two to ten years, with two years of interval for uniformity.

Time period	Survival
2 years	87.82%
4 years	88.01%
6 years	83.12%
8 years	79.80%
10 years	37.40%

Some of the values for each time period might not conform to the inverse relationship of survival prediction and time due to the data used.

Graph of predicted survival

Here is a chart representation of the predicted survival computed by our models.

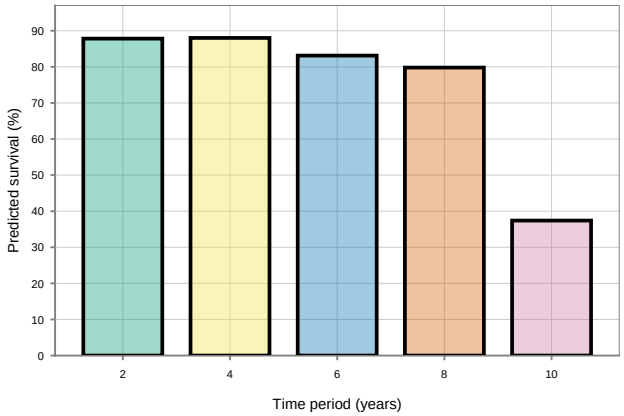


Figure 47: BOSOM Calculator results report in PDF format (page 1 of 2)

BOSOM | General Information

THE FORM

The BOSOM Calculator requires the following values for predicting breast cancer survival as determined by our study:

Item	Remarks	Values
Age of patient at time of diagnosis	No remarks	minimum = 1 maximum = 150
Race of patient	No remarks	▪ Black ▪ White ▪ Filipino or otherwise ▪ Unknown race
Stage of cancer	Based on Adjusted AJCC 6th Edition (1988+)	▪ 0 ▪ I ▪ IIA ▪ IIB ▪ IIIA ▪ IIIB ▪ IIINOS ▪ IV ▪ Unknown stage
Spread of metastasis	Based on Adjusted AJCC 6th Edition TNM (1988+)	▪ M0 (No distant metastasis) ▪ M1 (Distant metastasis) ▪ MX (Distant metastasis cannot be assessed)
Details of cancer-directed surgery	No remarks	▪ Not performed and patient died prior to recommended surgery ▪ Not recommended only ▪ Not recommended and contraindicated due to other conditions ▪ Recommended but not performed, patient refused ▪ Recommended but not performed for unknown reasons ▪ Recommended but unknown if performed ▪ Surgery performed ▪ Unknown OR death certificate or autopsy-only case
Extension of primary tumor	Based on EOD 10th Edition (1988 – 2003)	▪ 0 ▪ 5 ▪ 10 ▪ 11 ▪ 13 ▪ 14 ▪ 15 ▪ 16 ▪ 17 ▪ 18 ▪ 20 ▪ 21 ▪ 23 ▪ 24 ▪ 25 ▪ 26 ▪ 27 ▪ 28 ▪ 30 ▪ 31 ▪ 33 ▪ 34 ▪ 35 ▪ 36 ▪ 37 ▪ 38 ▪ 40 ▪ 50 ▪ 60 ▪ 70 ▪ 80 ▪ 85 ▪ 99

References

- "Adjusted AJCC 6th ed. T, N, M, and Stage". *Surveillance, Epidemiology, and End Results Program*. SEER, National Cancer Institute, NIH, DHHS, USA. Available: http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage/6th/breast.html
- Fritz, April and Ries, Lynn. "Extension". SEER Extent of Disease – 1988 Codes and Coding Instructions Third Edition. CSB, SP, DCCPS, NCI, DHHS, PHS, NIH, USA. 1998 January. Available: http://seer.cancer.gov/archive/manuals/EOD10Dig.pub.pdf

Figure 48: BOSOM Calculator results report in PDF format (page 2 of 2)

4. Supplements page

With respect to the flow of using the application, a supplements page, containing a list of breast cancer-catering facilities, was provided after getting the result of the calculator to direct the user to seek professional assistance if ever needed. The page is divided into local and international sectors in list form. All sectors' links to their websites are provided, with additions of other social networking sites if available.

The local breast cancer sector are mostly non-governmental organizations (NGOs). The list provides the major hospitals known to have a breast cancer facility and support groups aimed to educate and help Filipinos in overcoming the condition.

As shown in Fig. 49, the international-centered sectors was included for the purpose of providing alternative means of research outside the country. Majority are from the United States of America and some have an option to talk to their representative to entertain questions about the disease. SEER and NCI, for example, are forefronts of cancer research and development and they have public access articles for everyone to willing to learn more about cancer.

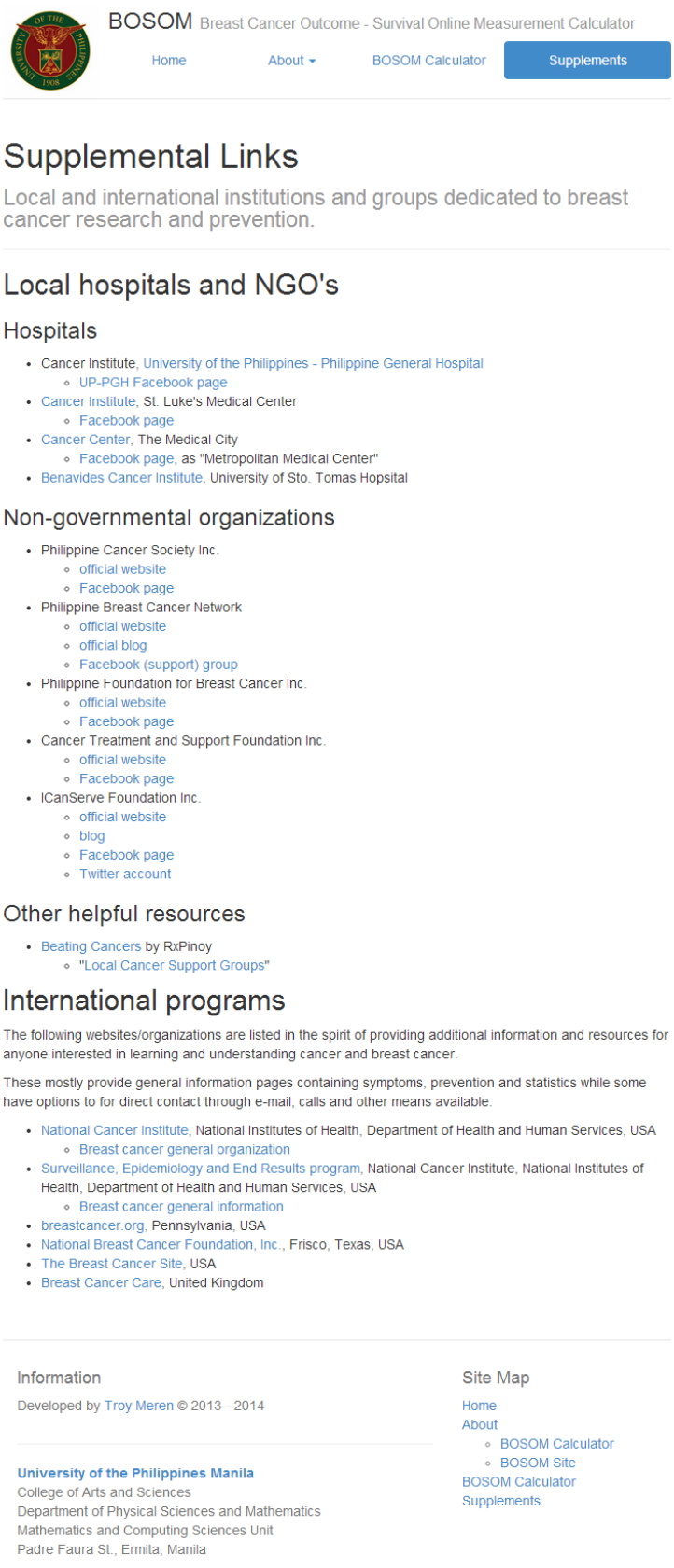


Figure 49: BOSOM application supplements page

5. Error pages

   There are several pages dedicated to informing the user whenever something amiss happened while navigating through the site.

   Illegal locations that are not included in the site map will be directed to the page shown in Fig. 50. This corresponds to an `HTTP 404` error or the page a user wants to visit does not exist. They are adivsed to return to the home page and refer to the links in both the header and footer for the available pages.

   Additional pages such as in Fig. 51 represents internal errors in the BOSOM application. It contains an apology for the user and they are encouraged to contact the developer to inform them of the mishap. This error page directly represents an umbrella catch for Java `Exception`s.

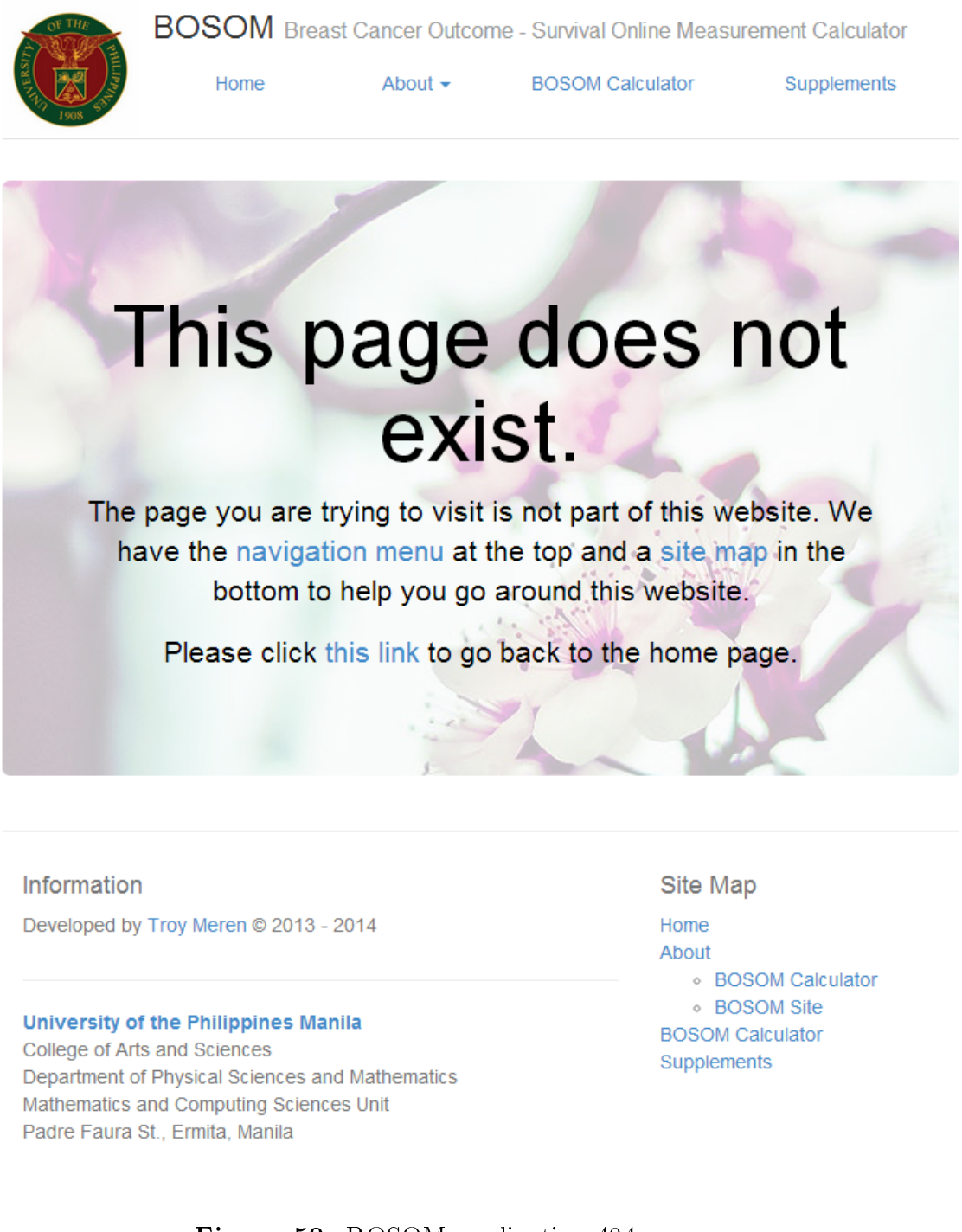


Figure 50: BOSOM application 404 error page

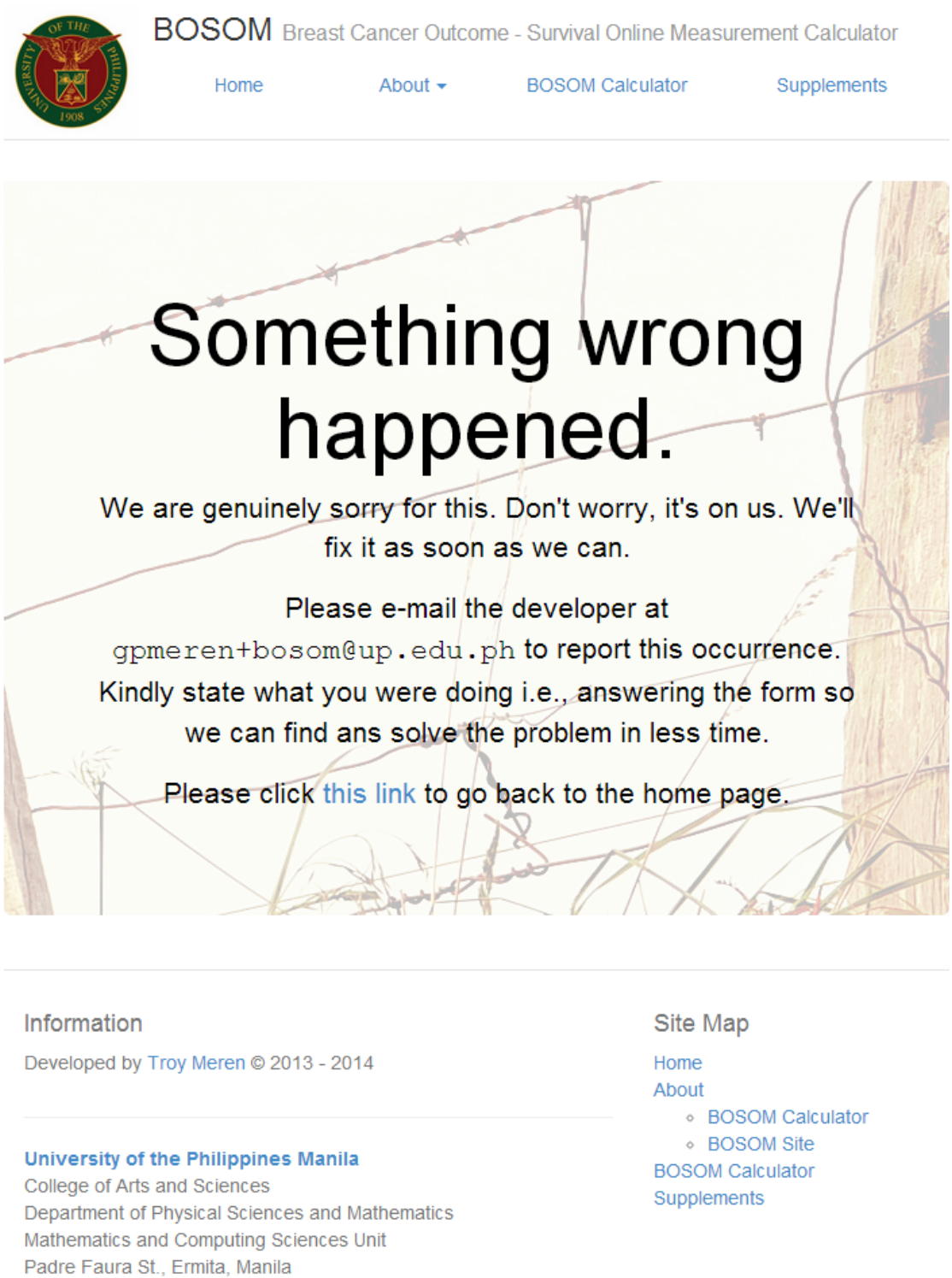


Figure 51: BOSOM application Java Exception error page

VI. Discussion

This study aimed to emulate the "Lung Cancer Outcome Calculator" by Agrawal et. al., an online application that provides predicted survival of lung cancer patients for six months, nine months, one year, two years and five years. It requires 13 variables that correspond to several medical data from the patient, and these where enumerated in Table 3 [3].

Breast cancer was chosen as the condition to apply the methodology used in LCOC because of its significant contribution to the incidence and mortality of cancer patients in the Philippines. Women in particular are more vulnerable to this type of cancer and the need to spread awareness about it has been an effort the government with the help of organizations and support groups nationwide.

The Breast Cancer Outcome - Survival Online Measurement (BOSOM) Calculator is a web application that provides a breast cancer patient's predicted survival within two to ten years, with two years interval. The calculator requires six variables: "age of patient at time of diagnosis", "race", "stage of cancer", "spread of metastis", "details of cancer-directed surgery" and "extension of primary tension" and submission will show a results page containing the predicted survival. All visitors of the BOSOM application are allowed to use to BOSOM Calculator as well.

The calculator uses 25 prediction models that correspond to five classifiers predicting one of the five outcome variables. The classifiers are namely alternating decision tree, J48 decision tree, random forest, `LogitBoost` and random subspace and these were trained with a preprocessed 100,000-record breast cancer dataset from the Surveillance, Epidemiology, and End Results Program (SEER). Both the training phase and BOSOM Calculator required the machine learning software Waikato Environment for Knowledge Analysis (WEKA) where the aforementioned classifiers are implemented. The impelementation of the majority of phases of this

study was based on the methodology applied in the development of LCOC, from the preprocessing to the data mining [3].

The training of models were applied to two datasets, only differing in the number of variables: the complete and the subset. The subset contains a small number of nonredundant variables determined using an attribute selection algorithm. The complete dataset required four to four and a half hours of training and the latter one to two hours.

The accuracy and other metrics of the models created were recorded to check their performance. Results show that the complete dataset and the BOSOM Calculator (subset) performed with comparable predictive accuracies of [94.0611%, 93.8669%, 93.5037%, 89.9190%, 82.2363%] and [93.0816%, 92.8752%, 92.3777%, 88.2827%, 74.7720%] per time period respectively. In relation to Agrawal et. al.'s study, theirs had [73.6%, 74.5%, 76.8%, 85.5%, 91.4%] and [72.5%, 73.6%, 76.2%, 85.5%, 91.2%]. The difference in the results are caused by different datasets, preprocessing results and machine used to train the models.

Any user can answer the form in the BOSOM Calculator. The results page will only be generated after a successful form validation of the answers the user provided. The results report consists of the breast cancer data from the user, the predicted survival in tabular and bar graph format, and an option to get the report in PDF format.

In addition, there are several pages in the website that can be visited by any user including a general information page containing explanations of the concepts that make up the application and a supplements page for external breast cancer help. The supplements page is made up of a list of hospitals, organizations and support groups that are currently active and all cater to breast cancer. Most of these have an active online presence (based form their latest social media activity) and they encourage people to contact them for questions and inquiries regarding the condition and their own organization.

VII. Conclusion

This study was able to create the Breast Cancer Outcome - Survival Online Measurement (BOSOM) Calculator - a web application that computes a person's breast cancer survival for two, four, six, eight and ten years. The data from the SEER Program of the USA was used and the training of the models for the calculator were done using WEKA wherein a custom program was developed to emulate the methodology used for the LCOC.

The development of the application involved three main steps: preprocessing, data mining and predictive modeling, and the web application. The first took the most time to finish because there were several repeats of filtering and variable selection in order to isolate a final dataset. Next was the data mining that took a while to finish (one hour to five hours) due to the number of records and the limitations of the machine where it was done. Finally, the calculator was developed with the trained models and integrated to a web framework. WEKA and Spring MVC's merge was easier since both utilize the programming language Java.

A predicted survival is obtain from the application by visiting the BOSOM Calculator's page. There are instructions to help the user fill the form with the appropriate content. Once a user's form is valid, they are redirected to the results page containing a consolidation of the entered form data and the predicted survival. A PDF feature is available for the user to seek further medical attention with the results of the application.

In terms of performance, the calculator faired well in comparison with the complete dataset. Each has average accuracies of 88.27784% and 90.7174% respectively for the five ensemble voting results. This is comparable to the LCOC's performance against the complete lung cancer dataset at 79.8% and 80.36%. This proves the usefulness of the BOSOM Calculator in the context of its variables and their predictions.

VIII. Recommendations

There were technical debts and better methodologies that could have been implemented during the development of the BOSOM application but because of knowledge, time and financial constraints, they have been removed from the process flow in favor of creating the application within the given timeframe.

First is a wider scope of understanding cancer coding for standards such as the American Joint Committee on Cancer (AJCC) and Extent of Disease (EOD). The related variables in Table 7 were supposed to be merged but an interview with an SEER representative revealed that further interactions between variables and values must be noted as each differs per cancer type [63]. Addition of these could lead to a betetr dataset and in turn, a more accurate predictive model.

Next is the exlporation of other WEKA classifiers for creating the calculator's models. This study strictly followed the required five classifiers to train the breast cancer model and since it is known that these classifiers behave differently from dataset to dataset, future researches could try evaluating all WEKA classifiers with their dataset to find the ones that exhibited the highest accuracy and lowest length of time to finish training [6].

The framework of Agrawal et. al.'s study made it possible for the development of a breast cancer calculator hence other cancer types and conditions could have their own survival prediction calculators. The most important component however is securing a comprehensive dataset similar to SEER's.

The online presence of the application makes it accessible to most individuals with a device and Internet connection. It is expected to generate results for around five hundred or less individuals during its course of deployment and this can be used to add a storage feature for all the values submitted by users together with the results. Once a certain quota is reached, a system administrator could retrain the models with new data to increase its performance.

The supplements page can be converted into a dynamic source of breast cancer related information wherein a search engine tracker could load the latest news and articles from local and international websites, and a constant update on the social media information of known organizations and support groups.

Finally, access to a better machine than the one used in this study, as discussed in Subsection D..2 of Section D. in Chapter IV., is encouraged. An ideal machine for data mining would have a random access memory of at least 8GB, a hard disk drive of at least 7200 rpms and 500GB of space, and a 64 bit multi-core processor to support the first two hardwares. Inclusion of more records from the SEER dataset is ideal since the study opted to 100,000 wherein the original planned number of records was 200,000.

IX. References

[1] J. Ferlay, H. Shin, F. Bray, D. Forman, C. Mathers, and D. Parkin, "GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC Cancerbase No. 10," 2010. Accessed on Sept. 7, 2013.

[2] H.-C. Lin, "Learning Accurate Regressors for Predicting Survival Times of Individual Cancer Patients," m. sci. thesis, University of Alberta, Edmonton, Alberta, Canada, 2011. Accessed on Sept. 4, 2013.

[3] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data," in *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (BIOKDD)*, (New York, NY, USA), pp. 1–9, ACM, 2011.

[4] "Understanding Cancer Prognosis," May 2011. Last reviewed on May 11, 2012; Accessed on Jul. 19, 2013.

[5] G. C. Wishart, E. M. Azzato, D. C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, and P. D. Pharoah, "PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer," *Breast Cancer Research*, vol. 12, pp. 1–10, Jan. 2010.

[6] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 3 ed., 2011.

[7] "Breast Cancer," 2013. Accessed on Jul. 19, 2013.

[8] M. T. Redaniel, A. Laudico, M. R. Mirasol-Lumague, A. Gondos, D. Pulte, C. Mapua, and H. Brenner, "Cancer survival discrepancies in developed and developing countries: comparisons between the Philippines and the United States," *Cancer Research UK*, vol. 100, pp. 858–862, Mar 2009.

[9] A. V. Laudico, V. Medina, M. R. Mirasol-Lumague, C. A. Mapua, M. T. M. Redanie, F. G. Valenzuela, and E. Pukkala, "2010 Philippine Cancer Facts and Estimates," 2010. Accessed on Jul. 19, 2013.

[10] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, pp. 113–127, June 2005.

[11] A. E. Millen, M. Pettinger, J. L. Freudenheim, R. D. Langer, C. A. Rosenberg, Y. Mossavar-Rahmani, C. M. Duffy, D. S. Lane, A. McTiernan, L. H. Kuller, A. M. Lopez, and J. Wactawski-Wende, "Incident Invasive Breast Cancer, Geographic Location of Residence, and Reported Average Time Spent Outside," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 18, pp. 495–507, February 2009.

[12] "SEER*Stat Database: Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2012 Sub (1973-2010 varying) - Linked To County Attributes - Total U.S.., 1969-2010 Counties." Released on Apr. 2013; Accessed on Aug. 4, 2013.

[13] "SEER Program Coding and Staging Manual 2013," January 2013. Accessed on Sep. 10, 2013.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.

[15] A. Deveria, "Can I use...," 2010. Accessed on Sept. 11, 2013.

[16] A. Bellaachia and E. Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques," in *9th Workshop on Mining Scientific and Engineering Datasets in conjunction with the 6th SIAM International Conference on Data Mining*, (Bethesda, Maryland, USA), pp. 1–4, April 2006.

[17] A. Endo, T. Shibata, and H. Tanaka, "Comparison of Seven Algorithm to Predict Breast Cancer Survival," *Biomedical Soft Computing and Human Sciences*, vol. 13, pp. 11–16, Feb. 2008.

[18] K. Rajesh and S. Anand, "Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 1, pp. 72–77, April 2012.

[19] P. M. Ravdin, L. A. Siminoff, G. J. Davis, M. B. Mercer, J. Hewlett, N. Gerson, and H. L. Parker, "Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women with Early Breast Cancer," *Journal of Clinical Oncology*, vol. 19, no. 4, pp. 980–991, 2001.

[20] "Adjuvant! Online," 2011. Accessed on Jul. 17, 2013.

[21] T. Gegg-Harrison, M. Zhang, N. Meng, Z. Sun, and P. Yang, "Porting a cancer treatment prediction to a mobile device," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 6218–6221, 2009.

[22] A. Agrawal, S. Misra, R. Narayanan, L. Polepedd, and A. Choudhary, "Lung Cancer Outcome Calculator," 2011. Accessed on Jul. 2, 2013.

[23] "Breast Cancer in Men," 2012. Last medical review on Sep. 21, 2012; last revision on Feb. 26, 2013; accessed on Sep. 7, 2013.

[24] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, pp. 37–54, 1996.

[25] D. Michie, D. J. Spiegelhalter, C. C. Taylor, and J. Campbell, eds., *Machine Learning: Neural and Statistical Classification*. Upper Saddle River, NJ, USA: Ellis Horwood, 1994.

[26] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[27] C. Rudin, "Decision Trees," 2012. Sept. 21, 2013.

[28] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, (San Francisco, CA, USA), pp. 124–133, Morgan Kaufmann Publishers Inc., 1999.

[29] H. K. Sok, M. Chowdhury, M.-L. Ooi, and S. Demidenko, "Using the ADTree for feature reduction through knowledge discovery," in *Instrumentation and Measurement Technology Conference (I2MTC), 2013 IEEE International*, pp. 1040–1044, 2013.

[30] W. Zhang, F. Zeng, X. Wu, X. Zhang, and R. Jiang, "A Comparative Study of Ensemble Learning Approaches in the Classification of Breast Cancer Metastasis," in *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS '09. International Joint Conference on*, pp. 242–245, 2009.

[31] D. Dittman, T. Khoshgoftaar, R. Wald, and A. Napolitano, "Random Forest: A Reliable Tool For Patient Response Prediction," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, pp. 289–296, 2011.

[32] N. de Freitas, "Machine learning - random forests," 2013. Accessed on Sept. 21, 2013.

[33] L. Breiman and R. E. Schapire, "Random Forests," in *Machine Learning*, pp. 5–32, 2001.

[34] N. Chu, L. Ma, X. Chen, Z. Che, and Y. Hu, "Ensemble learning for synthesis of the four diagnostics of tcm," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, pp. 843–847, 2011.

[35] I. Palit and C. K. Reddy, "Scalable and Parallel Boosting with MapReduce," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 10, pp. 1904–1916, 2012.

[36] Y. Freund and R. E. Schapire, "A Short Introduction to Boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, pp. 771–780, September 1999.

[37] W. Yang and G. Toderici, "Discriminative Tag Learning on YouTube Videos with Latent Sub-tags," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3217–3224, 2011.

[38] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 832–844, 1998.

[39] Y. Cai, Q. Zhu, and X. Cheng, "Semi-Supervised Short Text Categorization Based on Random Subspace," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 4, pp. 470–473, 2010.

[40] U. Turhal, S. Babur, C. Avci, and A. Akbas, "Performance improvement for diagnosis of colon cancer by using ensemble classification methods," in *Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 International Conference on*, pp. 271–275, 2013.

[41] G. Salama, M. Abdelhalim, and M. Zeid, "Experimental Comparison of Classifiers for Breast Cancer Diagnosis," in *Computer Engineering Systems (ICCES), 2012 Seventh International Conference on*, pp. 180–185, 2012.

[42] K. Selvakuberan, M. Indradevi, and R. Rajaram, "Combined Feature Selection and classification – A novel approach for the categorization of web pages," *Journal of Information and Computing Science*, vol. 3, no. 2, pp. 83–89, 2008.

[43] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, p. 427–437, May 2009.

[44] "Overiew of the SEER Program." Accessed on Jul. 19, 2013.

[45] R. Robu and C. Hora, "Medical data mining with extended WEKA," in *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on*, pp. 347–350, 2012.

[46] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *Biomedical Soft Computing and Human Sciences*, vol. 11, no. 1, 2009.

[47] "Making predictions." Accessed on January, 19, 2014.

[48] "Serialization." Accessed on January, 19, 2014.

[49] "How to access source code of a WEKA model file." Accessed on March 5, 2014.

[50] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1995.

[51] rj45, "Simple Example of MVC (Model View Controller) Design Pattern for Abstraction," April 2008. Accessed on Sept. 7, 2013.

[52] "Cancer Staging," May 2013. Last reviewed on May 3, 2013; Accessed on Jul. 19, 2013.

[53] K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, and A. Zeileis, "Rweka: R/Weka interface," October 2013.

[54] R. Kirkby and B. Pfahringer, "Class ADTree." Accessed on November, 30, 2013.

[55] E. Frank, "Class J48." Accessed on November, 30, 2013.

[56] R. Kirkby, "Class RandomForest." Accessed on November, 30, 2013.

[57] L. Trigg and E. Frank, "Class LogitBoost." Accessed on November, 30, 2013.

[58] B. Pfahringer and P. Reutemann, "Class RandomSubSpace." Accessed on November, 30, 2013.

[59] "Web application framework," 2013. Accessed on Feb. 27, 2014.

[60] J. O'Hanley, "Introduction to WEB4J: Web development for minimalists," October 2008. Accessed on Feb. 28, 2014.

[61] R. Johnson, J. Hoeller, K. Donald, C. Sampaleanu, R. Harrop, T. Risberg, A. Arendsen, D. Davison, D. Kopylenko, M. Pollack, T. Templier, E. Vervaet, P. Tung, B. Hale, A. Colyer, J. Lewis, C. Leau, M. Fisher, S. Brannen, R. Laddad, A. Poutsma, C. Beams, T. Abedrabbo, A. Clement, D. Syer, O. Gierke, R. Stoyanchev, P. Webb, R. Winch, and B. Clozel, "Web MVC framework," 2014. Accessed on Nov. 25, 2013.

[62] "Agila (Computer Science Development Server)," 2013. Accessed on Jul. 17, 2013.

[63] "Collaborative Stage for TNM7 - Revised 06/30/2011," 2011. Accessed on October 1, 2013.

X. Appendix

A. Forms

Last Name: VOID
SEER ID: VOID
Request Type: VOID

SURVEILLANCE, EPIDEMIOLOGY, AND END RESULTS PROGRAM
Data-Use Agreement for the 1973-2010 File

It is of utmost importance to protect the identities of cancer patients. Every effort has been made to exclude identifying information on individual patients from the computer files. Certain demographic information - such as sex, race, etc. - has been included for research purposes. All research results must be presented or published in a manner that ensures that no individual can be identified. In addition, there must be no attempt either to identify individuals from any computer file or to link with a computer file containing patient identifiers.

In order for the Surveillance, Epidemiology, and End Results Program to provide access to its Research Data File to you, it is necessary that you agree to the following provisions.

1. I will not use - or permit others to use - the data in any way other than for statistical reporting and analysis for research purposes. I must notify the SEER Program if I discover that there has been any other use of the data.
2. I will not present or publish data in which an individual patient can be identified. I will not publish any information on an individual patient, including any information generated on an individual case by the case listing session of SEER*Stat. In addition, I will avoid publication of statistics for very small groups.
3. I will not attempt either to link - or permit others to link - the data with individually identified records in another database.
4. I will not attempt to learn the identity of any patient whose cancer data is contained in the supplied file(s).
5. If I inadvertently discover the identity of any patient, then (a) I will make no use of this knowledge, (b) I will notify the SEER Program of the incident, and (c) I will inform no one else of the discovered identity.
6. I will not either release - or permit others to release - the data - in full or in part - to any person except with the written approval of the SEER Program. In particular, all members of a research team who have access to the data must sign this data-use agreement.
7. I will use appropriate safeguards to prevent use or disclosure of the information other than as provided for by this data-use agreement. If accessing the data from a centralized location on a time sharing computer system or LAN with SEER*Stat or another statistical package, I will not share my logon name or password with any other individuals. I will also not allow any other individuals to use my computer account after I have logged on with my logon name and password.
8. For all software provided by the SEER Program, I will not copy it, distribute it, reverse engineer it, profit from its sale or use, or incorporate it in any other software system.
9. I will cite the source of information in all publications. The appropriate citation is associated with the data file used. (Please see either Suggested Citations on the SEER*Stat Help menu or the Readme.txt associated with the ASCII text version of the SEER data.)

My signature indicates that I agree to comply with the above stated provisions.

_____
Signature

_____
DATE Date

Figure 52: SEER Research Data-Use Agreement form

B. Source Code

NOTE: Several components of the source codes were modified in order to be presented properly in this section. URLs, file paths and variable names are several examples of these components.

Source Code 8: R preprocessing script: data loading

```r
require(reshape)
require(plyr)
options(max.print=1000000000)

dat <- read.csv("path/to/dataset.csv",
  colClasses=c(
    "ageDiagNom"="factor",
    "ageNom"="factor",
    "basisDiag"="factor",
    "behav1"="factor",
    "diagConf"="factor",
    "er"="factor",
    "ext2"="factor",
    "female"="factor",
    "firstMalPrimInd"="factor",
    "grade1"="factor",
    "histBehav1"="factor",
    "histGroup"="factor",
    "histInd"="factor",
    "laterality"="factor",
    "ln2"="factor",
    "m3"="factor",
    "marital"="factor",
    "monDiag"="factor",
    "n3"="factor",
    "numBenTum"="factor",
    "numMalTum"="factor",
    "placeBirthGroup"="factor",
    "pr"="factor",
    "primSite"="factor",
    "raceGroup"="factor",
    "rad"="factor",
    "radSeqSurg"="factor",
    "reasonNoCancerSurg"="factor",
    "regNodeExamNom"="factor",
    "regNodePosNom"="factor",
    "stage3"="factor",
    "sumStage"="factor",
    "surgOthRegDis1"="factor",
    "surgPrimSite1"="factor",
    "t3"="factor",
    "tumSizeNom2"="factor",
    "vsr"="factor",
    "yrBirth"="factor",
    "yrDiag"="factor"))

# convert for WEKA ARFF
dat[,c("ageDiagNum","ageNum","numPrim","regNodeExamNum",
  "regNodePosNum","time","tumSizeNum2")] <- as.numeric(
    as.character(
      unlist(dat[,c("ageDiagNum","ageNum","numPrim",
        "regNodeExamNum","regNodePosNum","time","tumSizeNum2")])))

dat$timeNot <- ifelse(dat$time < 24, 1, 0)
dat$time2 <- ifelse(dat$time >= 24, 1, 0)
dat$time4 <- ifelse(dat$time >= 48, 1, 0)
dat$time6 <- ifelse(dat$time >= 72, 1, 0)
dat$time8 <- ifelse(dat$time >= 96, 1, 0)
dat$time10 <- ifelse(dat$time >= 120, 1, 0)
```

Source Code 9: R preprocessing script: data transformation

```r
require(reshape)
require(plyr)
require(RWeka)
options(max.print=1000000000)

# drop unneeded columns
# http://ewens.tepper.cmu.edu/2011/05/17/
# simple-r-functions-to-keep-or-remove-data-frame-columns/
filterDataframe <- function(df) {
  keeps <- c("ageDiagNum",
    "behav1", "diagConf", "er", "ext2", "female",
          "firstMalPrimInd", "grade1",
    "histGroup", "laterality", "m3",
    "n3", "numMalTum", "numPrim", "pr",
    "primSite", "raceGroup", "rad", "radSeqSurg",
          "reasonNoCancerSurg",
    "regNodeExamNom", "regNodeExamNum", "regNodePosNom",
          "regNodePosNum", "stage3",
    "sumStage", "surgPrimSite1", "t3", "tumSizeNom2", "tumSizeNum2",
    "timeNot", "time2", "time4", "time6", "time8", "time10")
  return (df[,names(df) %in% keeps])
}

# replace all NA's with a blank
# http://r.789695.n4.nabble.com/How-to-replace-all-NA-values-
# in-a-data-frame-with-another-not-0-value-tp2125458p2125516.html
dat2 <- as.matrix(filterDataframe(dat))
temp <- which(is.na(dat2)==TRUE | dat2=="Blank(s)")
dat2[temp] <- ""
dat3 <- as.data.frame(dat2)
rm(dat2,temp)

# get rows by column condition
# http://stackoverflow.com/a/5391697/1685185
timeNot <- dat3[dat3,"timeNot"] == 1,,drop=FALSE]
time2 <- dat3[dat3[,"time2"] == 1,,drop=FALSE]
time4 <- dat3[dat3[,"time4"] == 1,,drop=FALSE]
time6 <- dat3[dat3[,"time6"] == 1,,drop=FALSE]
time8 <- dat3[dat3[,"time8"] == 1,,drop=FALSE]
time10 <- dat3[dat3[,"time10"] == 1,,drop=FALSE]

dat3.100 <- rbind.fill(
  timeNot[1:10000,],
  time2[1:10000,],
  time4[1:10000,],
  time6[1:20000,],
  time8[1:20000,],
  time10[1:30000,])

write.arff(dat3.100, "path/to/dataset.arff", eol="\n")
rm(dat3.100,dat3.150)
```

Source Code 10: Training.java

```java
import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.FileReader;
import java.io.IOException;
import java.io.ObjectOutputStream;
import java.io.OutputStreamWriter;
import java.io.Writer;
import java.util.Random;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;

import weka.classifiers.Classifier;
import weka.classifiers.Evaluation;
import
       weka.classifiers.evaluation.output.prediction.AbstractOutput;
import weka.classifiers.evaluation.output.prediction.CSV;
import weka.classifiers.meta.LogitBoost;
import weka.classifiers.meta.RandomSubSpace;
import weka.classifiers.rules.ZeroR;
import weka.classifiers.trees.ADTree;
import weka.classifiers.trees.J48;
import weka.classifiers.trees.RandomForest;
import weka.core.Instances;
import weka.core.Range;
import weka.filters.Filter;
import weka.filters.unsupervised.attribute.Remove;

/**
 * Parts adapted from
 * http://weka.wikispaces.com/file/view/CrossValidationMultipleRuns
 * .java/82916745/CrossValidationMultipleRuns.java
 */
public class Training {

  protected final static Log logger =
        LogFactory.getLog(Training.class);
  private static String PATH_DATA = "path/to/data/folder";
  private static String PATH_RESULTS = "path/to/results/folder";
  private static String DATASET_NAME = "bosom.100k";
  private static String DATASET_TIME = "2";
```

```
42  private static final String HEADER_CSV = "Iteration #,Correctly
        Classified Instances,"
43      + "Incorrectly Classified Instances,Correctly Classified
            Instances (%),"
44      + "Incorrectly Classified Instances (%),Kappa statistic,Mean
            absolute error,"
45      + "Root mean squared error,Relative absolute error,Root
            relative squared error,"
46      + "Coverage of cases (0.95 level),Mean rel. region size (0.95
            level),"
47      + "TP Rate (0),FP Rate (0),Precision (0),Recall/Sensitivity
            (0),Specificity (0),"
48      + "F-Measure (0),MCC (0),ROC Area (0),PRC Area (0),TP Rate
            (1),FP Rate (1),"
49      + "Precision (1),Recall/Sensitivity (1),Specificity
            (1),F-Measure (1),"
50      + "MCC (1),ROC Area (1),PRC Area (1),"
51      + "TP (0),TN (0),FP (0),FN (0),TP (1),TN (1),FP (1),FN (1)\n";
52
53  public static void main(String[] args) throws Exception {
54
55    Instances train = removeAttributes(getData(DATASET_NAME));
56    train.setClassIndex(train.numAttributes() - 1);
57
58    ZeroR zr = new ZeroR();
59    zr.buildClassifier(train);
60    assemble(zr, train, DATASET_NAME + "." + DATASET_TIME + "." +
          "zr");
61
62    RandomForest rf = new RandomForest();
63    rf.buildClassifier(train);
64    assemble(rf, train, DATASET_NAME + "." + DATASET_TIME + "." +
          "rf");
65
66    LogitBoost lb = new LogitBoost();
67    lb.buildClassifier(train);
68    assemble(lb, train, DATASET_NAME + "." + DATASET_TIME + "." +
          "lb");
69
70    RandomSubSpace rs = new RandomSubSpace();
71    rs.buildClassifier(train);
72    assemble(rs, train, DATASET_NAME + "." + DATASET_TIME + "." +
          "rs");
73
74    J48 j48 = new J48();
75    j48.buildClassifier(train);
76    assemble(j48, train, DATASET_NAME + "." + DATASET_TIME + "." +
      "j48");
77
78    ADTree adt = new ADTree();
79    adt.buildClassifier(train);
80    assemble(adt, train, DATASET_NAME + "." + DATASET_TIME + "." +
81    "adt");
82  }
83
84  private static Instances getData(String fileName) throws
        Exception {
85    BufferedReader bufferedReader = null;
86    bufferedReader = new BufferedReader(new FileReader(PATH_DATA +
          fileName
87        + ".ARFF"));
88    Instances train = new Instances(bufferedReader);
89    bufferedReader.close();
90    return train;
91  }
92
93  private static Instances removeAttributes(Instances instance)
94      throws Exception {
95    // Indices of attributes to remove
96    // http://weka.wikispaces.com/Use+Weka+in+your+Java+code#Filter
97    String[] optionsRemove = new String[] {
98        "-R",
99        "3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22,
              23, 24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36"
100           };
101   Remove remove = new Remove();
102   remove.setOptions(optionsRemove);
103   remove.setInputFormat(instance);
104
105   Instances instanceFiltered = Filter.useFilter(instance, remove);
106   return instanceFiltered;
107 }
108
109 private static void assemble(Classifier classifier, Instances
        instance,
110     String fileName) throws Exception {
111   logger.info("Training data with " +
          classifier.getClass().toString()
112       + "\n");
113
114   CSV predictionOutput = (CSV) getAbstractObject();
115   predictionOutput.setNumDecimals(6);
116
117   Evaluation evaluation = getEvaluation(classifier, instance,
118       predictionOutput, fileName);
119
120   StringBuffer resultBuffer = new StringBuffer();
121   resultBuffer.append("=== Run information ===\n\n");
122   resultBuffer.append(getRunInformation(classifier, instance));
123   resultBuffer.append("=== Classifier model (full training set)
          ===\n\n");
124   resultBuffer.append(getClassifierModelResult(classifier));
125   resultBuffer
126       .append("=== Predictions on test data ===\n\nsee associated
              CSV file \n");
127   resultBuffer.append(getEvaluationResults(evaluation));
128
129   saveResultBuffer(resultBuffer.toString(), "/main-results/" +
          fileName
130       + ".result.buffer.TXT");
131   saveResultBuffer(predictionOutput.getBuffer().toString(),
132       "/main-preds/" + fileName + "-preds.CSV");
133   saveTrainedModel(classifier, fileName);
134 }
135
136 private static AbstractOutput getAbstractObject() {
137   StringBuffer predictionSB = new StringBuffer();
138   CSV output = new CSV();
139   output.setBuffer(predictionSB);
140   output.setOutputDistribution(true);
141   return output;
142 }
143
144 private static Evaluation getEvaluation(Classifier classifier,
145     Instances instance, AbstractOutput output, String fileName)
146     throws Exception {
147   int numCvIters = 10;
148   int numFolds = 10;
149   Evaluation evaluation = null;
150   Range attributesToShow = null;
151   Boolean outputDistributions = new Boolean(true);
152
153   String preds = "";
154   StringBuffer cvIterResultSB = new StringBuffer();
155   cvIterResultSB.append(HEADER_CSV);
156
157   // 10 x 10 CV
158   for (int i = 1; i <= numCvIters; i++) {
159     int seed = i + 15;
160     Random rand = new Random(seed);
161     Instances randomData = new Instances(instance);
162     randomData.randomize(rand);
163
164     // do CV
165     evaluation = new Evaluation(randomData);
166     evaluation.crossValidateModel(classifier, instance, numFolds,
            rand,
167         output, attributesToShow, outputDistributions);
168
169     cvIterResultSB.append(getAssembledCvCsvResults(i,
            evaluation));
170   }
171
172   saveResultBuffer(cvIterResultSB.toString(), "/cv-results/" +
          fileName
173       + ".folds.results.CSV");
174
175   return evaluation;
176 }
177
178
179 private static String getAssembledCvCsvResults(int foldNum,
180     Evaluation evaluation) throws Exception {
181   StringBuffer resultSB = new StringBuffer();
182
183   resultSB.append(foldNum + ",");
184   resultSB.append(evaluation.correct() + ",");
185   resultSB.append(evaluation.incorrect() + ",");
186   resultSB.append(evaluation.pctCorrect() + ",");
187   resultSB.append(evaluation.pctIncorrect() + ",");
188   resultSB.append(evaluation.kappa() + ",");
189   resultSB.append(evaluation.meanAbsoluteError() + ",");
190   resultSB.append(evaluation.rootMeanSquaredError() + ",");
191   resultSB.append(evaluation.relativeAbsoluteError() + ",");
192   resultSB.append(evaluation.rootRelativeSquaredError() + ",");
193   resultSB.append(evaluation.coverageOfTestCasesByPredictedRegions()
194       + ",");
195   resultSB.append(evaluation.sizeOfPredictedRegions() + ",");
196
197   resultSB.append(evaluation.truePositiveRate(0) + ",");
198   resultSB.append(evaluation.falsePositiveRate(0) + ",");
199   resultSB.append(evaluation.precision(0) + ",");
200   resultSB.append(evaluation.recall(0) + ",");
201   resultSB.append(evaluation.numTrueNegatives(0)
202       / (evaluation.numTrueNegatives(0) + evaluation
203           .numFalsePositives(0)) + ",");
204   resultSB.append(evaluation.fMeasure(0) + ",");
205   resultSB.append(evaluation.matthewsCorrelationCoefficient(0) +
          ",");
206   resultSB.append(evaluation.areaUnderROC(0) + ",");
207   resultSB.append(evaluation.areaUnderPRC(0) + ",");
208   resultSB.append(evaluation.truePositiveRate(1) + ",");
209   resultSB.append(evaluation.falsePositiveRate(1) + ",");
210   resultSB.append(evaluation.precision(1) + ",");
211   resultSB.append(evaluation.recall(1) + ",");
212   resultSB.append(evaluation.numTrueNegatives(1)
213       / (evaluation.numTrueNegatives(1) + evaluation
214           .numFalsePositives(1)) + ",");
```

```
    resultSB.append(evaluation.fMeasure(1) + ",");
    resultSB.append(evaluation.matthewsCorrelationCoefficient(1) +
          ",");
    resultSB.append(evaluation.areaUnderROC(1) + ",");
    resultSB.append(evaluation.areaUnderPRC(1) + ",");

    resultSB.append(evaluation.numTruePositives(0) + ",");
    resultSB.append(evaluation.numTrueNegatives(0) + ",");
    resultSB.append(evaluation.numFalsePositives(0) + ",");
    resultSB.append(evaluation.numFalseNegatives(0) + ",");

    resultSB.append(evaluation.numTruePositives(1) + ",");
    resultSB.append(evaluation.numTrueNegatives(1) + ",");
    resultSB.append(evaluation.numFalsePositives(1) + ",");
    resultSB.append(evaluation.numFalseNegatives(1) + "\n");

    return resultSB.toString();
  }

  private static String getRunInformation(Classifier classifier,
      Instances instance) {
    String resultString = "";
    StringBuilder resultSB = new StringBuilder();

    resultString += "Scheme:\t\t" + classifier.getClass() + "\n"
        + "Relation:\t" + instance.relationName() + "\n"
        + "Instances:\t" + instance.numInstances() + "\n"
        + "Attributes:\t" + instance.numAttributes() + "\n";

    for (int i = 0; i < instance.numAttributes(); i++) {
      resultSB.append("\t\t\t" + instance.attribute(i).name() +
          "\n");
    }

    resultString += resultSB.toString() + "\n"
        + "Test mode:\t10-fold cross-validation" + "\n\n";
    return resultString;
  }

  private static String getClassifierModelResult(Classifier
      classifier) {
    return classifier.toString() + "\n" + "Time taken to build
        model:\t"
        + "\n\n";
  }

  private static String getEvaluationResults(Evaluation evaluation)
      throws Exception {
    return evaluation.toSummaryString(true) + "\n"
        + evaluation.toClassDetailsString() + "\n"
        + evaluation.toMatrixString() + "\n";
  }

  private static void saveResultBuffer(String results, String
      fileName) {
    Writer writer = null;
    String fullFilePathAndName = PATH_RESULTS + fileName;
    try {
      writer = new BufferedWriter(new OutputStreamWriter(
          new FileOutputStream(fullFilePathAndName), "UTF-8"));
      writer.write(results);
      results = null;
      logger.info("Successfully created file [" +
          fullFilePathAndName
          + "]\n");
    } catch (IOException e) {

    } finally {
      try {
        writer.close();
      } catch (Exception e) {

      }
    }
  }

  private static void saveTrainedModel(Classifier classifier,
      String fileName)
      throws FileNotFoundException, IOException {
    String fullFilePathAndName = PATH_RESULTS + "/models/" +
        fileName
        + ".MODEL";
    ObjectOutputStream oos = new ObjectOutputStream(new
        FileOutputStream(
        fullFilePathAndName));
    logger.info("Successfully created model [" + fullFilePathAndName
        + "]\n");
    oos.writeObject(classifier);
    classifier = null;
    oos.flush();
    oos.close();
  }

}
```

Source Code 11: Spring `application-context.xml` file

```
<beans xmlns="http://www.springframework.org/schema/beans"
  xmlns:context="http://www.springframework.org/schema/context"
  xmlns:mvc="http://www.springframework.org/schema/mvc"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:tx="http://www.springframework.org/schema/tx"
  xsi:schemaLocation="
    http://www.springframework.org/schema/beans
    http://www.springframework.org/schema/beans/spring-beans-3.2.5xsd
    http://www.springframework.org/schema/context
    http://www.springframework.org/schema/context/spring-context-3.2.5.xsd
    http://www.springframework.org/schema/mvc
    http://www.springframework.org/schema/mvc/spring-mvc-3.2.5.xsd">

  <!-- Auto-detect components -->

  <context:component-scan
        base-package="ph.edu.upm.agila.gtmeren.bosom" />

  <bean id="calcService"
    class="ph.edu.upm.agila.gtmeren.bosom.service.impl.CalcServiceImpl"
        />
  <bean id="calcArffService"
    class="ph.edu.upm.agila.gtmeren.bosom.service.impl.CalcArffServiceImpl"
        />
  <bean id="calcModelService"
    class="ph.edu.upm.agila.gtmeren.bosom.service.impl.CalcModelServiceImpl"
        />

  <mvc:annotation-driven />

  <!-- Application Message Bundle -->
  <bean id="messageSource"
    class="org.springframework.context.support.ReloadableResourceBundleMessageSource">
    <property name="basename" value="/WEB-INF/messages" />
    <property name="cacheSeconds" value="3000" />
  </bean>

</beans>
```

Source Code 12: Spring `spring-servlet.xml` file

```
<?xml version="1.0" encoding="UTF-8"?>

<beans xmlns="http://www.springframework.org/schema/beans"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:p="http://www.springframework.org/schema/p"
  xmlns:context="http://www.springframework.org/schema/context"
  xmlns:mvc="http://www.springframework.org/schema/mvc"
        xmlns:util="http://www.springframework.org/schema/util"
  xsi:schemaLocation="http://www.springframework.org/schema/beans
    http://www.springframework.org/schema/beans/spring-beans-3.0.xsd
    http://www.springframework.org/schema/context
    http://www.springframework.org/schema/context/spring-context-3.0.xsd
        http://www.springframework.org/schema/mvc
        http://www.springframework.org/schema/mvc/spring-mvc-3.0.xsd
        http://www.springframework.org/schema/util
        http://www.springframework.org/schema/util/spring-util.xsd">

    <!--This tag allows for mapping the DispatcherServlet to "/"
          (all extensions etc) -->
    <mvc:default-servlet-handler />

    <context:component-scan
          base-package="ph.edu.upm.agila.gtmeren.bosom" />
  <context:component-scan
        base-package="ph.edu.upm.agila.gtmeren.bosom.controller" />
  <context:component-scan
        base-package="ph.edu.upm.agila.gtmeren.bosom.domain" />
  <context:component-scan
        base-package="ph.edu.upm.agila.gtmeren.bosom.pdf" />
  <context:component-scan
        base-package="ph.edu.upm.agila.gtmeren.bosom.service" />

  <mvc:annotation-driven />

    <mvc:resources mapping="/resources/**" location="/,
          /resources/**, classpath:/WEB-INF/resources" />
    <mvc:resources mapping="/css/**" location="/, /resources/css/*"
          />
    <mvc:resources mapping="/js/**" location="/, /resources/js/*" />
    <mvc:resources mapping="/images/**" location="/,
          /resources/images/*" />
    <mvc:resources mapping="/fonts/**" location="/,
          /resources/fonts/*" />

    <mvc:resources mapping="/classes/**"
          location="/WEB-INF/classes/*" />
    <mvc:resources mapping="/calc/reports/**" location="/,
          /WEB-INF/reports/*, /reports/*" />

  <bean id="viewResolver"
    class="org.springframework.web.servlet.view.UrlBasedViewResolver">
    <property name="viewClass"
      value="org.springframework.web.servlet.view.JstlView" />
    <property name="prefix" value="/WEB-INF/jsp/" />
    <property name="suffix" value=".jsp" />
```

```
  </bean>

  <context:property-placeholder
        location="classpath:file.locations.properties" />

  <bean id="messageSource"
    class="org.springframework.context.support.ReloadableResourceBundleMessageSource">
    <property name="basenames">
      <list>
        <value>classpath:messages.validation</value>
      </list>
    </property>
  </bean>

  <bean id="viewResolver1"
    class="org.springframework.web.servlet.view.ResourceBundleViewResolver">
    <property name="order" value="1" />
    <property name="basename" value="pdf" />
  </bean>

</beans>
```

Source Code 13: Spring web.xml file

```
<?xml version="1.0" encoding="UTF-8"?>
<web-app xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="http://java.sun.com/xml/ns/javaee"
        xmlns:web="http://java.sun.com/xml/ns/javaee/web-app_3_1.xsd"
  id="WebApp_ID" version="3.0">

  <display-name>bosom</display-name>

  <welcome-file-list>
    <welcome-file>index.html</welcome-file>
    <welcome-file>index.htm</welcome-file>
    <welcome-file>index.jsp</welcome-file>
    <welcome-file>default.html</welcome-file>
    <welcome-file>default.htm</welcome-file>
    <welcome-file>default.jsp</welcome-file>
  </welcome-file-list>

  <servlet>
    <servlet-name>spring</servlet-name>
    <servlet-class>org.springframework.web.servlet.DispatcherServlet
    </servlet-class>
    <load-on-startup>1</load-on-startup>
  </servlet>

  <servlet-mapping>
    <servlet-name>spring</servlet-name>
    <url-pattern>/</url-pattern>
  </servlet-mapping>

  <session-config>
    <!-- Disables URL-based sessions (no more 'jsessionid' in the
          URL using
      Tomcat) -->
    <tracking-mode>COOKIE</tracking-mode>
  </session-config>

  <!-- Error pages -->
  <location>/WEB-INF/jsp/error/400.jsp</location>
  <error-page>
    <error-code>400</error-code>
  </error-page>

  <error-page>
    <error-code>403</error-code>
    <location>/WEB-INF/jsp/error/403.jsp</location>
  </error-page>

  <error-page>
    <error-code>404</error-code>
    <location>/WEB-INF/jsp/error/404.jsp</location>
  </error-page>

  <error-page>
    <error-code>405</error-code>
    <location>/WEB-INF/jsp/error/405.jsp</location>
  </error-page>

  <error-page>
    <error-code>500</error-code>
    <location>/WEB-INF/jsp/error/500.jsp</location>
  </error-page>

  <error-page>
    <exception-type>java.lang.Exception</exception-type>
    <location>/WEB-INF/jsp/error/exception.jsp</location>
  </error-page>

</web-app>
```

Source Code 14:
ph/edu/upm/agila/gtmeren/bosom/domain/WekaData.java

```
package ph.edu.upm.agila.gtmeren.bosom.domain;

import java.io.Serializable;
import java.math.BigDecimal;

import javax.validation.constraints.NotNull;

import org.hibernate.validator.constraints.NotEmpty;
import org.hibernate.validator.constraints.Range;

@SuppressWarnings("serial")
public class WekaData implements Serializable {

  @NotNull
  @Range(min = 1, max = 150)
  private Integer ageDiagNum;

  @NotEmpty
  private String raceGroup;

  @NotEmpty
  private String stage3;

  @NotEmpty
  private String m3;

  @NotEmpty
  private String reasonNoCancerSurg;

  @NotEmpty
  private String ext2;

  private BigDecimal time2;
  private BigDecimal time4;
  private BigDecimal time6;
  private BigDecimal time8;
  private BigDecimal time10;

  public Integer getAgeDiagNum() {
    return ageDiagNum;
  }

  public void setAgeDiagNum(Integer ageDiagNum) {
    this.ageDiagNum = ageDiagNum;
  }

  public String getRaceGroup() {
    return raceGroup;
  }

  public void setRaceGroup(String raceGroup) {
    this.raceGroup = raceGroup;
  }

  public String getStage3() {
    return stage3;
  }

  public void setStage3(String stage3) {
    this.stage3 = stage3;
  }

  public String getM3() {
    return m3;
  }

  public void setM3(String m3) {
    this.m3 = m3;
  }

  public String getReasonNoCancerSurg() {
    return reasonNoCancerSurg;
  }

  public void setReasonNoCancerSurg(String reasonNoCancerSurg) {
    this.reasonNoCancerSurg = reasonNoCancerSurg;
  }

  public String getExt2() {
    return ext2;
  }

  public void setExt2(String ext2) {
    this.ext2 = ext2;
  }

  public BigDecimal getTime2() {
    return time2;
  }

  public void setTime2(BigDecimal time2) {
    this.time2 = time2;
  }

  public BigDecimal getTime4() {
    return time4;
  }

```

```
  public void setTime4(BigDecimal time4) {
    this.time4 = time4;
  }

  public BigDecimal getTime6() {
    return time6;
  }

  public void setTime6(BigDecimal time6) {
    this.time6 = time6;
  }

  public BigDecimal getTime8() {
    return time8;
  }

  public void setTime8(BigDecimal time8) {
    this.time8 = time8;
  }

  public BigDecimal getTime10() {
    return time10;
  }

  public void setTime10(BigDecimal time10) {
    this.time10 = time10;
  }

  @Override
  public int hashCode() {
    final int prime = 31;
    int result = 1;
    result = prime * result
        + ((ageDiagNum == null) ? 0 : ageDiagNum.hashCode());
    result = prime * result + ((ext2 == null) ? 0 :
           ext2.hashCode());
    result = prime * result + ((m3 == null) ? 0 : m3.hashCode());
    result = prime * result
        + ((raceGroup == null) ? 0 : raceGroup.hashCode());
    result = prime
        * result
        + ((reasonNoCancerSurg == null) ? 0 : reasonNoCancerSurg
            .hashCode());
    result = prime * result + ((stage3 == null) ? 0 :
           stage3.hashCode());
    return result;
  }

  @Override
  public boolean equals(Object obj) {
    if (this == obj)
      return true;
    if (obj == null)
      return false;
    if (getClass() != obj.getClass())
      return false;
    WekaData other = (WekaData) obj;
    if (ageDiagNum == null) {
      if (other.ageDiagNum != null)
        return false;
    } else if (!ageDiagNum.equals(other.ageDiagNum))
      return false;
    if (ext2 == null) {
      if (other.ext2 != null)
        return false;
    } else if (!ext2.equals(other.ext2))
      return false;
    if (m3 == null) {
      if (other.m3 != null)
        return false;
    } else if (!m3.equals(other.m3))
      return false;
    if (raceGroup == null) {
      if (other.raceGroup != null)
        return false;
    } else if (!raceGroup.equals(other.raceGroup))
      return false;
    if (reasonNoCancerSurg == null) {
      if (other.reasonNoCancerSurg != null)
        return false;
    } else if (!reasonNoCancerSurg.equals(other.reasonNoCancerSurg))
      return false;
    if (stage3 == null) {
      if (other.stage3 != null)
        return false;
    } else if (!stage3.equals(other.stage3))
      return false;
    return true;
  }

  @Override
  public String toString() {
    return "WekaData [ageDiagNum=" + ageDiagNum + ", raceGroup="
        + raceGroup + ", stage3=" + stage3 + ", m3=" + m3
        + ", reasonNoCancerSurg=" + reasonNoCancerSurg + ", ext2="
        + ext2 + ", time2=" + time2 + ", time4=" + time4 + ",
              time6="
        + time6 + ", time8=" + time8 + ", time10=" + time10 + "]";
  }

}
```

Source Code 15:
ph/edu/upm/agila/gtmeren/bosom/controller/AboutController.java

```
package ph.edu.upm.agila.gtmeren.bosom.controller;

import org.springframework.stereotype.Controller;
import org.springframework.ui.ModelMap;
import org.springframework.web.bind.annotation.RequestMapping;

@Controller
public class AboutController {

  @RequestMapping(value = { "/about", "/about/bosom" })
  public String showAboutBosomPage(ModelMap model) {
    model.addAttribute("title", "About - BOSOM Calculator");
    model.addAttribute("pageName", "about");
    model.addAttribute("pageTitleHeader", "Calculating Breast
          Cancer");
    model.addAttribute("pageTitleSubheader",
        "Learn more about the BOSOM Calculator and its components");
    return "about/bosom";
  }

  @RequestMapping(value = "/about/site")
  public String showAboutSitePage(ModelMap model) {
    model.addAttribute("title", "About - Site");
    model.addAttribute("pageName", "about");
    model.addAttribute("pageTitleHeader", "Site General
          Information");
    model.addAttribute("pageTitleSubheader",
        "Technical information and acknowledgements to the site's
             backbone technology");
    return "about/site";
  }

}
```

Source Code 16:
ph/edu/upm/agila/gtmeren/bosom/controller/CalcController.java

```
package ph.edu.upm.agila.gtmeren.bosom.controller;

import javax.servlet.http.HttpServletRequest;
import javax.servlet.http.HttpServletResponse;
import javax.validation.Valid;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.springframework.beans.factory.annotation.Autowired;
import org.springframework.stereotype.Controller;
import org.springframework.ui.Model;
import org.springframework.ui.ModelMap;
import org.springframework.validation.BindingResult;
import org.springframework.web.bind.annotation.ModelAttribute;
import org.springframework.web.bind.annotation.RequestMapping;
import org.springframework.web.bind.annotation.RequestMethod;

import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;
import ph.edu.upm.agila.gtmeren.bosom.service.CalcService;

@Controller
public class CalcController {

  protected final Log logger = LogFactory.getLog(getClass());

  @Autowired
  private CalcService calcService;

  @RequestMapping(value = "/calc", method = RequestMethod.GET)
  public String showGet(@ModelAttribute("wekaData") WekaData
        wekaData,
      Model model, HttpServletRequest request,
      HttpServletResponse response) {
    model.addAttribute("title", "Calculator");
    model.addAttribute("pageName", "calc");
    model.addAttribute("pageTitleHeader", "BOSOM Calculator");
    model.addAttribute("pageTitleSubheader",
        "Evaluate your survival prediction");

    return "calc/form";
  }

  @RequestMapping(value = "/calc", method = RequestMethod.POST)
  public String showPost(
      @ModelAttribute("wekaData") @Valid WekaData wekaData,
      final BindingResult bindingResult, ModelMap model,
      HttpServletRequest request, HttpServletResponse response)
      throws Exception {

    if (bindingResult.hasErrors()) {
      model.addAttribute("title", "Calculator");
      model.addAttribute("pageName", "calc");
      model.addAttribute("pageTitleHeader", "BOSOM Calculator");
```

```
    model.addAttribute("pageTitleSubheader",
        "Evaluate your survival prediction");
    model.addAttribute("pageTitleSubheader",
        "Evaluate your survival prediction");
    model.addAttribute("alertStrongContent",
        "You have errors in the form.");
    model.addAttribute("alertContent",
        "Please review the information you provided before
            submission.");
    model.addAttribute("alertType", "danger");

    return "calc/form";
  }

  model.addAttribute("title", "Results");
  model.addAttribute("pageName", "calc");
  model.addAttribute("pageTitleHeader", "Predictive results");
  model.addAttribute("pageTitleSubheader",
      "Our predictive model's interpretation of your survival.");
  model.addAttribute("isFlotUsed", true);

  logger.info("Form data: " + wekaData + "\n");
  model.addAttribute("wekaData", wekaData);
  model.addAttribute("predictionsMap",
      calcService.evaluate(wekaData, request));

  model.addAttribute("pdfLocation",
      PdfController.getPdfFilePath(wekaData, request, response));

  return "calc/results";
 }

}
```

Source Code 17:
ph/edu/upm/agila/gtmeren/bosom/controller/PdfController.java

```
package ph.edu.upm.agila.gtmeren.bosom.controller;

import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.IOException;
import java.sql.SQLException;

import javax.servlet.ServletContext;
import javax.servlet.http.HttpServletRequest;
import javax.servlet.http.HttpServletResponse;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.joda.time.DateTime;

import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;
import ph.edu.upm.agila.gtmeren.bosom.pdf.PdfBuilder;
import ph.edu.upm.agila.gtmeren.bosom.pdf.PdfConcatenator;

import com.itextpdf.text.Document;
import com.itextpdf.text.DocumentException;
import com.itextpdf.text.PageSize;
import com.itextpdf.text.pdf.PdfWriter;

public class PdfController {

  protected final static Log logger =
        LogFactory.getLog(PdfController.class);

  private static DateTime TIME_NOW_RAW = DateTime.now();
  private static String NAME_DOCUMENT = TIME_NOW_RAW
      .toString("ddMMyyyyHHmmss");
  private static String NAME_FILE_ORIG = "ORIG-" + NAME_DOCUMENT +
        ".pdf";
  private static String NAME_FILE_CONCAT = NAME_DOCUMENT + ".pdf";

  public static String getPdfFilePath(WekaData wekaData,
      HttpServletRequest request, HttpServletResponse response)
      throws DocumentException, IOException, SQLException {
    logger.info("Start of PDF creation");

    Document document = new Document(PageSize.A4);
    PdfWriter pdfWriter = null;
    try {
      pdfWriter = PdfWriter.getInstance(document, new
            FileOutputStream(
          getPdfFile(request, NAME_FILE_ORIG)));
    } catch (FileNotFoundException e1) {
      e1.printStackTrace();
    }

    document.open();
    PdfBuilder.assemble(document, pdfWriter, wekaData);
    document.close();
    pdfWriter.close();

    PdfConcatenator.concatenate(
        getPdfFile(request, NAME_FILE_CONCAT).getAbsolutePath(),
        getPdfFile(request, NAME_FILE_ORIG).getAbsolutePath(),
        getPdfFile(request, "/bosom-info.pdf").getAbsolutePath());

    logger.info("End of PDF creation");

    /*
     * file.getAbsolutePath().toString(); currently mapped to
           /calc/reports/
     * via spring-servlet.xml
     */
    return "calc/reports/" + NAME_FILE_CONCAT;
  }

  public static File getPdfFile(HttpServletRequest request, String
        fileName) {
    ServletContext servletContext = request.getSession()
        .getServletContext();
    String filePath =
          servletContext.getRealPath("/WEB-INF/reports/") + "/"
        + fileName;
    File file = new File(filePath);
    file.deleteOnExit();
    logger.info("PDF file path: " +
          file.getAbsolutePath().toString());
    return file;
  }

}
```

Source Code 18:
ph/edu/upm/agila/gtmeren/bosom/controller/SupplementsController.java

```
package ph.edu.upm.agila.gtmeren.bosom.controller;

import javax.servlet.http.HttpServletRequest;
import javax.servlet.http.HttpServletResponse;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.springframework.stereotype.Controller;
import org.springframework.ui.ModelMap;
import org.springframework.web.bind.annotation.RequestMapping;
import org.springframework.web.bind.annotation.RequestMethod;

@Controller
public class SupplementsController {

  protected final static Log logger = LogFactory
      .getLog(SupplementsController.class);

  @RequestMapping(value = "/supplements", method =
        RequestMethod.GET)
  public String showSupplementsPage(ModelMap model,
      HttpServletRequest request, HttpServletResponse response) {
    model.addAttribute("title", "Supplements");
    model.addAttribute("pageName", "supplements");
    model.addAttribute("pageTitleHeader", "Supplemental Links");
    model.addAttribute("pageTitleSubheader",
        "Local and international institutions and groups dedicated "
            + "to breast cancer research and prevention.");

    return "supplements";
  }

}
```

Source Code 19:
ph/edu/upm/agila/gtmeren/bosom/pdf/ChartBuilder.java

```
package ph.edu.upm.agila.gtmeren.bosom.pdf;

import java.awt.BasicStroke;
import java.awt.Color;
import java.awt.Paint;
import java.math.BigDecimal;

import org.jfree.chart.ChartFactory;
import org.jfree.chart.JFreeChart;
import org.jfree.chart.StandardChartTheme;
import org.jfree.chart.axis.NumberAxis;
import org.jfree.chart.axis.NumberTickUnit;
import org.jfree.chart.plot.CategoryPlot;
import org.jfree.chart.plot.PlotOrientation;
import org.jfree.chart.renderer.category.BarRenderer;
import org.jfree.chart.renderer.category.CategoryItemRenderer;
import org.jfree.chart.renderer.category.StandardBarPainter;
import org.jfree.data.category.DefaultCategoryDataset;

import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;
import com.itextpdf.text.Font;

public class ChartBuilder {

  protected static JFreeChart createBarChart(WekaData wekaData) {

    BigDecimal[] dataset = { wekaData.getTime2(),
        wekaData.getTime4(),
        wekaData.getTime6(), wekaData.getTime8(),
            wekaData.getTime10() };
```

```
29
30    // http://www.wirelust.com/2008/03/17/
31    // creating-an-itext-pdf-with-embedded-jfreechart/
32    DefaultCategoryDataset chartData = new DefaultCategoryDataset();
33    for (int i = 0; i < dataset.length; i++) {
34      chartData.setValue(dataset[i], "Population", (i + 1) * 2 +
              "");
35    }
36
37    JFreeChart chart = ChartFactory.createBarChart("",
38        "Time period (years)", "Predicted survival (%)", chartData,
39        PlotOrientation.VERTICAL, false, true, false);
40    chart.setBackgroundPaint(Color.WHITE);
41    ChartFactory.setChartTheme(StandardChartTheme.createLegacyTheme());
42
43    final CategoryPlot plot = chart.getCategoryPlot();
44    ((BarRenderer) plot.getRenderer())
45        .setBarPainter(new StandardBarPainter());
46
47    plot.setBackgroundPaint(Color.WHITE);
48
49    plot.setDomainGridlinesVisible(true);
50    plot.setRangeGridlinesVisible(true);
51
52    plot.setDomainGridlineStroke(new BasicStroke(0.25f));
53    plot.setRangeGridlineStroke(new BasicStroke(0.25f));
54
55    plot.setDomainGridlinePaint(new Color(204, 204, 204));
56    plot.setRangeGridlinePaint(new Color(204, 204, 204));
57
58    java.awt.Font fontGraphLabel = new java.awt.Font("Helvetica",
59        Font.NORMAL, 8);
60    java.awt.Font fontGraphTicks = new java.awt.Font("Helvetica",
61        Font.NORMAL, 6);
62    plot.getDomainAxis().setLabelFont(fontGraphLabel);
63    plot.getRangeAxis().setLabelFont(fontGraphLabel);
64    plot.getDomainAxis().setTickLabelFont(fontGraphTicks);
65    plot.getRangeAxis().setTickLabelFont(fontGraphTicks);
66
67    CategoryItemRenderer categoryItemRenderer = new
            CustomRenderer();
68    plot.setRenderer(categoryItemRenderer);
69
70    final NumberAxis yAxis = (NumberAxis) plot.getRangeAxis();
71    yAxis.setStandardTickUnits(NumberAxis.createIntegerTickUnits());
72
73    int yLimit = (int)
            getRoundedUpMultipleOfTen(getHighestArrayValue(dataset));
74    yAxis.setRange(0, yLimit);
75    if (yLimit > 40) {
76      yAxis.setTickUnit(new NumberTickUnit(10));
77    } else {
78      yAxis.setTickUnit(new NumberTickUnit(5));
79    }
80
81    final BarRenderer barRenderer = (BarRenderer)
            plot.getRenderer();
82    barRenderer.setDrawBarOutline(true);
83    barRenderer.setShadowVisible(false);
84
85    barRenderer.setSeriesOutlinePaint(0, new Color(0, 0, 0));
86    barRenderer.setSeriesOutlineStroke(0, new BasicStroke(2f));
87
88    return chart;
89  }
90
91  private static class CustomRenderer extends BarRenderer {
92
93    private static final long serialVersionUID =
            6826676370155152948L;
94    private Paint[] colors;
95    int transparency = 95;
96
97    // http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/
98    public CustomRenderer() {
99      this.colors = new Paint[] { new Color(1, 158, 115,
              transparency),
100         new Color(240, 228, 66, transparency),
101         new Color(0, 114, 178, transparency),
102         new Color(213, 94, 0, transparency),
103         new Color(204, 121, 167, transparency) };
104   }
105
106   public Paint getItemPaint(final int row, final int column) {
107     return this.colors[column % this.colors.length];
108   }
109 }
110
111 private static double getHighestArrayValue(BigDecimal[] dataset) {
112   double max = 0;
113   for (int i = 1; i < dataset.length; i++) {
114     if (dataset[i].doubleValue() > max) {
115       max = dataset[i].doubleValue();
116     }
117   }
118   return max;
119 }
120
121 private static double getRoundedUpMultipleOfTen(double number) {
122   return ((number + 9) / 10) * 10;
123 }
124 }
```

Source Code 20:
ph/edu/upm/agila/gtmeren/bosom/pdf/PdfBuilder.java

```
1  package ph.edu.upm.agila.gtmeren.bosom.pdf;
2
3  import java.awt.Graphics2D;
4
5  import org.apache.commons.logging.Log;
6  import org.apache.commons.logging.LogFactory;
7  import org.jfree.chart.JFreeChart;
8  import org.joda.time.DateTime;
9  import org.joda.time.format.DateTimeFormat;
10 import org.joda.time.format.DateTimeFormatter;
11 import org.springframework.stereotype.Controller;
12
13 import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;
14
15 import com.itextpdf.awt.DefaultFontMapper;
16 import com.itextpdf.text.BaseColor;
17 import com.itextpdf.text.Chunk;
18 import com.itextpdf.text.Document;
19 import com.itextpdf.text.DocumentException;
20 import com.itextpdf.text.Element;
21 import com.itextpdf.text.Font;
22 import com.itextpdf.text.Font.FontFamily;
23 import com.itextpdf.text.Paragraph;
24 import com.itextpdf.text.Phrase;
25 import com.itextpdf.text.pdf.PdfContentByte;
26 import com.itextpdf.text.pdf.PdfPCell;
27 import com.itextpdf.text.pdf.PdfPTable;
28 import com.itextpdf.text.pdf.PdfTemplate;
29 import com.itextpdf.text.pdf.PdfWriter;
30 import com.itextpdf.text.pdf.draw.LineSeparator;
31
32 @Controller
33 public class PdfBuilder {
34
35   protected final static Log logger =
           LogFactory.getLog(PdfBuilder.class);
36
37   private static DateTimeFormatter DATE_TIME_FORMATTER =
           DateTimeFormat
38       .forPattern("hh:mm:ss a, dd MMMM yyyy");
39   private static DateTime TIME_NOW_RAW = DateTime.now();
40   private static String TIME_NOW_STR = TIME_NOW_RAW
41       .toString(DATE_TIME_FORMATTER);
42
43   private static BaseColor COLOR_BOOT_BLACK_3 = new BaseColor(85,
           85, 85);
44   private static BaseColor COLOR_BOOT_GREY = new BaseColor(221,
           221, 221);
45   private static BaseColor COLOR_BOOT_BLUE_LIGHT = new
           BaseColor(146, 188,
46       224);
47   private static BaseColor COLOR_BOOT_RED_LIGHT = new
           BaseColor(235, 165, 163);
48
49   private static BaseColor COLOR_GREY = new BaseColor(160, 160,
           160);
50
51   private static Font DOC_TITLE_HEAD = new
           Font(FontFamily.HELVETICA, 20,
52       Font.BOLD, COLOR_BOOT_GREY);
53   private static Font DOC_TITLE_SUB_HEAD = new
           Font(FontFamily.HELVETICA, 14,
54       Font.NORMAL, COLOR_BOOT_BLUE_LIGHT);
55   private static Font DOC_TITLE_SPECIAL = new
           Font(FontFamily.HELVETICA, 20,
56       Font.BOLD, COLOR_BOOT_RED_LIGHT);
57   private static Font DOC_LIST_HEAD = new
           Font(FontFamily.HELVETICA, 16,
58       Font.BOLD, COLOR_BOOT_BLACK_3);
59   private static Font DOC_TEXT_REG = new Font(FontFamily.HELVETICA,
           10,
60       Font.NORMAL, COLOR_BOOT_BLACK_3);
61   private static Font DOC_TABLE_REG = new
           Font(FontFamily.HELVETICA, 8,
62       Font.NORMAL, COLOR_BOOT_BLACK_3);
63   private static Font DOC_TABLE_BOLD = new
           Font(FontFamily.HELVETICA, 8,
64       Font.BOLD, COLOR_BOOT_BLACK_3);
65   private static Font DOC_TEXT_SUBDUED = new
           Font(FontFamily.HELVETICA, 8,
66       Font.NORMAL, COLOR_GREY);
67
68   private static LineSeparator LINE_SEPARATOR = new
           LineSeparator(0.5f, 100f,
69       COLOR_BOOT_GREY, 3, 0.5f);
70
71   private static float LINE_HEIGHT = 12.0f;
72
73   private static int WIDTH_GRAPH = 350;
74   private static int HEIGHT_GRAPH = 230;
75
```

```
public static void assemble(Document document, PdfWriter pdfWriter,
    WekaData wekaData) throws DocumentException {
  addMetaData(document);
  addTitlePage(document);
  addResultsSection(document, pdfWriter, wekaData);
}

private static void addMetaData(Document document) {
  document.addTitle("BOSOM Calculator Report");
  document.addSubject("Breast cancer survival prediction");
  document.addKeywords("Breast cancer, data mining, survival, prediction, machine learning");
  document.addAuthor("Guest " + TIME_NOW_RAW.toString("ddMMyyyyHHmmss"));
  document.addCreator("Troy Meren");
}

private static void addTitlePage(Document document)
    throws DocumentException {
  Paragraph paragraph = new Paragraph("BOSOM", DOC_TITLE_HEAD);
  paragraph.setSpacingAfter(1f);
  document.add(paragraph);

  paragraph = new Paragraph(
      "Breast Cancer Outcome - Survival Online Measurement Calculator",
      DOC_TITLE_SUB_HEAD);
  paragraph.setSpacingAfter(3f);
  document.add(paragraph);

  paragraph = new Paragraph("Generated on: " + TIME_NOW_STR,
      DOC_TEXT_SUBDUED);
  document.add(paragraph);

  document.add(Chunk.NEWLINE);
  document.add(LINE_SEPARATOR);

  paragraph = new Paragraph("CALCULATOR RESULTS REPORT",
      DOC_TITLE_SPECIAL);
  paragraph.setAlignment(Element.ALIGN_CENTER);
  paragraph.setSpacingAfter(15f);

  document.add(paragraph);
}

private static void addResultsSection(Document document,
    PdfWriter pdfWriter, WekaData wekaData) throws DocumentException {
  addEnteredDataSection(document, wekaData);
  addPredictedSurvivalSection(document, wekaData);
  addGraphPredictedSurvival(document, pdfWriter, wekaData);
}

private static void addEnteredDataSection(Document document,
    WekaData wekaData) throws DocumentException {

  Paragraph title = new Paragraph();
  title.add(new Paragraph("Entered data", DOC_LIST_HEAD));
  title.setSpacingAfter(10f);
  document.add(title);

  Paragraph text = new Paragraph();
  text.add(new Paragraph(
      "Here are the breast cancer values you provided in the calculator.",
      DOC_TEXT_REG));
  text.setLeading(LINE_HEIGHT);
  text.setSpacingAfter(5f);
  document.add(text);

  PdfPTable table = new PdfPTable(3);
  float[] columnWidths = { 5, 35, 60 };
  table.setWidthPercentage(100f);
  table.setWidths(columnWidths);

  PdfPCell cell = new PdfPCell(new Phrase("#", DOC_TABLE_BOLD));
  cell.setHorizontalAlignment(Element.ALIGN_CENTER);
  cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
  cell.setBorderColor(COLOR_BOOT_GREY);
  table.addCell(cell);

  cell = new PdfPCell(new Phrase("Variable", DOC_TABLE_BOLD));
  cell.setHorizontalAlignment(Element.ALIGN_CENTER);
  cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
  cell.setBorderColor(COLOR_BOOT_GREY);
  table.addCell(cell);

  cell = new PdfPCell(new Phrase("Value provided", DOC_TABLE_BOLD));
  cell.setHorizontalAlignment(Element.ALIGN_CENTER);
  cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
  cell.setBorderColor(COLOR_BOOT_GREY);

  table.addCell(cell);
  table.setHeaderRows(1);

  String[] enteredDataVars = { "Age at time of diagnosis",
      "Race of patient", "Stage of cancer", "Spread of etastasis",
      "Details of cancer-directed surgery", "Extension of primary tumor" };

  String[] enteredDataVals = {
      wekaData.getAgeDiagNum().toString(),
      wekaData.getRaceGroup(), wekaData.getStage3(),
      wekaData.getM3(), wekaData.getReasonNoCancerSurg(),
      wekaData.getM3() };

  for (int i = 0; i < enteredDataVars.length; i++) {
    cell = new PdfPCell(new Phrase(i + 1 + "", DOC_TABLE_REG));
    cell.setBorderColor(COLOR_BOOT_GREY);
    cell.setHorizontalAlignment(Element.ALIGN_CENTER);
    cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
    table.addCell(cell);

    cell = new PdfPCell(new Phrase(enteredDataVars[i], DOC_TABLE_REG));
    cell.setBorderColor(COLOR_BOOT_GREY);
    cell.setHorizontalAlignment(Element.ALIGN_LEFT);
    cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
    table.addCell(cell);

    cell = new PdfPCell(new Phrase(enteredDataVals[i], DOC_TABLE_REG));
    cell.setBorderColor(COLOR_BOOT_GREY);
    cell.setHorizontalAlignment(Element.ALIGN_LEFT);
    cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
    table.addCell(cell);
  }

  document.add(table);

  text = new Paragraph();
  text.setSpacingAfter(5f);
  document.add(text);

  text = new Paragraph();
  text.add(new Paragraph(
      "They are used to calculate the predicted survival rate given in the other sections.",
      DOC_TEXT_REG));
  text.setLeading(LINE_HEIGHT);
  text.setSpacingAfter(25f);
  document.add(text);
}

private static void addPredictedSurvivalSection(Document document,
    WekaData wekaData) throws DocumentException {

  Paragraph title = new Paragraph();
  title.add(new Paragraph("Predicted survival", DOC_LIST_HEAD));
  title.setSpacingAfter(10f);
  document.add(title);

  Paragraph text = new Paragraph();
  text.add(new Paragraph(
      "Here are the predicted survivals as determined by our models based from "
          + "past breast cancer patient records. "
          + "These are from two to ten years, with two years of interval for uniformity.",
      DOC_TEXT_REG));
  text.setSpacingAfter(5f);
  text.setLeading(LINE_HEIGHT);
  document.add(text);

  PdfPTable table = new PdfPTable(2);
  float[] columnWidths = { 50, 50 };
  table.setWidthPercentage(100f);
  table.setWidths(columnWidths);

  PdfPCell cell = new PdfPCell(new Phrase("Time period", DOC_TABLE_BOLD));
  cell.setHorizontalAlignment(Element.ALIGN_CENTER);
  cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
  cell.setBorderColor(COLOR_BOOT_GREY);
  table.addCell(cell);

  cell = new PdfPCell(new Phrase("Survival", DOC_TABLE_BOLD));
  cell.setHorizontalAlignment(Element.ALIGN_CENTER);
  cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
  cell.setBorderColor(COLOR_BOOT_GREY);
  table.addCell(cell);

  table.setHeaderRows(1);

  String[] predictions = { wekaData.getTime2().toString(),
      wekaData.getTime4().toString(),
      wekaData.getTime6().toString(),
      wekaData.getTime8().toString(),
      wekaData.getTime10().toString() };

  for (int i = 1; i <= predictions.length; i++) {
    cell = new PdfPCell(new Phrase(i * 2 + " years", DOC_TABLE_REG));
    cell.setBorderColor(COLOR_BOOT_GREY);
    cell.setHorizontalAlignment(Element.ALIGN_CENTER);
    cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
```

```
      table.addCell(cell);

      cell = new PdfPCell(new Phrase(predictions[i - 1] + "%",
          DOC_TABLE_REG));
      cell.setBorderColor(COLOR_BOOT_GREY);
      cell.setHorizontalAlignment(Element.ALIGN_CENTER);
      cell.setVerticalAlignment(Element.ALIGN_MIDDLE);
      table.addCell(cell);
    }

    document.add(table);

    text = new Paragraph();
    text.setSpacingAfter(5f);
    document.add(text);

    text = new Paragraph();
    text.add(new Paragraph(
        "Some of the values for each time period might not conform
             to the "
            + "inverse relationship of survival prediction and time
                 due to the data used.",
        DOC_TEXT_REG));
    text.setSpacingAfter(25f);
    text.setLeading(LINE_HEIGHT);
    document.add(text);
  }

  private static void addGraphPredictedSurvival(Document document,
      PdfWriter pdfWriter, WekaData wekaData) throws
           DocumentException {
    Paragraph title = new Paragraph();
    title.add(new Paragraph("Graph of predicted survival",
        DOC_LIST_HEAD));
    title.setSpacingAfter(10f);
    document.add(title);

    Paragraph text = new Paragraph();
    text.add(new Paragraph(
        "Here is a chart representation of the predicted survival
             computed by our models.",
        DOC_TEXT_REG));
    text.setLeading(LINE_HEIGHT);
    text.setSpacingAfter(10f);
    document.add(text);

    addBarChartToPdf(wekaData, pdfWriter);
  }

  // http://www.wirelust.com/ 2008/03/17/ creating-an-itext-pdf-
       with-embedded-jfreechart/
  private static void addBarChartToPdf(WekaData wekaData, PdfWriter
       pdfWriter) {
    logger.info("Start of charts creation");

    JFreeChart chart = ChartBuilder.createBarChart(wekaData);

    PdfContentByte dc = pdfWriter.getDirectContent();

    PdfTemplate tp = dc.createTemplate(800, HEIGHT_GRAPH);
    @SuppressWarnings("deprecation")
    Graphics2D g2 = tp.createGraphics(800, HEIGHT_GRAPH,
        new DefaultFontMapper());

    java.awt.geom.Rectangle2D r2D = new
        java.awt.geom.Rectangle2D.Double(
        75, 0, WIDTH_GRAPH, HEIGHT_GRAPH);
    chart.draw(g2, r2D);
    g2.dispose();

    dc.addTemplate(tp, 38, pdfWriter.getVerticalPosition(true)
        - HEIGHT_GRAPH);

    logger.info("End of charts creation");
  }

}
```

Source Code 21: ph/edu/upm/agila/gtmeren/bosom/pdf/PdfConcatenator.java

```
package ph.edu.upm.agila.gtmeren.bosom.pdf;

/*
 * This class is part of the book "iText in Action - 2nd Edition"
 * written by Bruno Lowagie (ISBN: 9781935182610)
 * For more info, go to: http://itextpdf.com/examples/
 * This example only works with the AGPL version of iText.
 */

import java.io.FileOutputStream;
import java.io.IOException;
import java.sql.SQLException;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;

import com.itextpdf.text.Document;
import com.itextpdf.text.DocumentException;
import com.itextpdf.text.pdf.PdfCopy;
import com.itextpdf.text.pdf.PdfReader;

public class PdfConcatenator {

  protected final static Log logger =
        LogFactory.getLog(PdfConcatenator.class);

  /**
   * Main method.
   *
   * @param args
   * no arguments needed
   * @throws DocumentException
   * @throws IOException
   * @throws SQLException
   */
  public static String concatenate(String concatPath, String...
        files) throws IOException,
      DocumentException, SQLException {

    logger.info("Preparation for PDF concatenation \n");

    for(int i = 0; i < files.length; i++) {
      logger.info("File # " + i + ": " + files[i]);
    }

    Document document = new Document();
    PdfCopy copy = new PdfCopy(document, new
          FileOutputStream(concatPath));
    document.open();
    PdfReader reader;
    int n;

    for (int i = 0; i < files.length; i++) {
      reader = new PdfReader(files[i]);
      n = reader.getNumberOfPages();
      for (int page = 0; page < n;) {
        copy.addPage(copy.getImportedPage(reader, ++page));
      }
      copy.freeReader(reader);
      reader.close();
    }
    document.close();

    return concatPath;
  }
}
```

Source Code 22: ph/edu/upm/agila/gtmeren/bosom/service/CalcArffService.java

```
package ph.edu.upm.agila.gtmeren.bosom.service;

import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;
import weka.core.Instances;

public interface CalcArffService {

  public Instances getInstances(WekaData wekaData);

}
```

Source Code 23: ph/edu/upm/agila/gtmeren/bosom/service/CalcModelService.java

```
package ph.edu.upm.agila.gtmeren.bosom.service;

import java.util.Map;

import javax.servlet.http.HttpServletRequest;

import weka.classifiers.Classifier;
import weka.core.Instances;

public interface CalcModelService {

  public Classifier getClassifier(String filePath,
        HttpServletRequest request);

  public Map<String, Map<String, Map<String, Object>>>
        getPredictions(
      Instances instances, HttpServletRequest request);

}
```

Source Code 24: ph/edu/upm/agila/gtmeren/bosom/service/CalcService.java

```
package ph.edu.upm.agila.gtmeren.bosom.service;

import java.util.Map;

import javax.servlet.http.HttpServletRequest;
```

```java

import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;

public interface CalcService {

  public Map<String, Map<String, Map<String, Object>>> evaluate(
      WekaData wekaData, HttpServletRequest request);

}
```

Source Code 25: ph/edu/upm/agila/gtmeren/bosom/service/impl/CalcArffServiceImpl.java

```java
package ph.edu.upm.agila.gtmeren.bosom.service.impl;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.springframework.stereotype.Service;

import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;
import ph.edu.upm.agila.gtmeren.bosom.service.CalcArffService;
import weka.core.Attribute;
import weka.core.DenseInstance;
import weka.core.FastVector;
import weka.core.Instances;
import weka.core.Utils;

@SuppressWarnings("deprecation")
@Service("calcArffService")
public class CalcArffServiceImpl implements CalcArffService {

  protected final Log logger = LogFactory.getLog(getClass());

  /**
   * @Source http://weka.wikispaces.com/Creating+an+ARFF+file
   */
  public Instances getInstances(WekaData wekaData) {

    FastVector<Attribute> atts;
    Instances data;
    double[] vals;

    // 1. set up attributes
    atts = new FastVector<Attribute>();

    atts.addElement(new Attribute("ageDiagNum"));

    FastVector<String> raceGroupAttVals = new FastVector<String>();
    raceGroupAttVals.addElement("Black");
    raceGroupAttVals.addElement("Other");
    raceGroupAttVals.addElement("Unknown");
    raceGroupAttVals.addElement("White");
    atts.addElement(new Attribute("raceGroup", raceGroupAttVals));

    FastVector<String> stage3AttVals = new FastVector<String>();
    stage3AttVals.addElement("0");
    stage3AttVals.addElement("I");
    stage3AttVals.addElement("IIA");
    stage3AttVals.addElement("IIB");
    stage3AttVals.addElement("IIIA");
    stage3AttVals.addElement("IIIB");
    stage3AttVals.addElement("IIIC");
    stage3AttVals.addElement("IIINOS");
    stage3AttVals.addElement("IV");
    stage3AttVals.addElement("UNK Stage");
    atts.addElement(new Attribute("stage3", stage3AttVals));

    FastVector<String> m3AttVals = new FastVector<String>();
    m3AttVals.addElement("M0");
    m3AttVals.addElement("M1");
    m3AttVals.addElement("MX");
    atts.addElement(new Attribute("m3", m3AttVals));

    FastVector<String> reasonNoCancerSurgAttVals = new
          FastVector<String>();
    reasonNoCancerSurgAttVals
        .addElement("Not performed, patient died prior to
            recommended surgery");
    reasonNoCancerSurgAttVals.addElement("Not recommended");
    reasonNoCancerSurgAttVals
        .addElement("Not recommended, contraindicated due to other
            conditions");
    reasonNoCancerSurgAttVals
        .addElement("Recommended but not performed, patient
            refused");
    reasonNoCancerSurgAttVals
        .addElement("Recommended but not performed, unknown
            reason");
    reasonNoCancerSurgAttVals
        .addElement("Recommended, unknown if performed");
    reasonNoCancerSurgAttVals.addElement("Surgery performed");
    reasonNoCancerSurgAttVals
        .addElement("Unknown; death certificate or autopsy only
            case");
    atts.addElement(new Attribute("reasonNoCancerSurg",
        reasonNoCancerSurgAttVals));

    FastVector<String> ext2AttVals = new FastVector<String>();
    ext2AttVals.addElement("00");
    ext2AttVals.addElement("05");
    ext2AttVals.addElement("10");
    ext2AttVals.addElement("11");
    ext2AttVals.addElement("13");
    ext2AttVals.addElement("14");
    ext2AttVals.addElement("15");
    ext2AttVals.addElement("16");
    ext2AttVals.addElement("17");
    ext2AttVals.addElement("18");
    ext2AttVals.addElement("20");
    ext2AttVals.addElement("21");
    ext2AttVals.addElement("23");
    ext2AttVals.addElement("24");
    ext2AttVals.addElement("25");
    ext2AttVals.addElement("26");
    ext2AttVals.addElement("27");
    ext2AttVals.addElement("28");
    ext2AttVals.addElement("30");
    ext2AttVals.addElement("31");
    ext2AttVals.addElement("33");
    ext2AttVals.addElement("34");
    ext2AttVals.addElement("35");
    ext2AttVals.addElement("36");
    ext2AttVals.addElement("37");
    ext2AttVals.addElement("38");
    ext2AttVals.addElement("40");
    ext2AttVals.addElement("50");
    ext2AttVals.addElement("60");
    ext2AttVals.addElement("70");
    ext2AttVals.addElement("80");
    ext2AttVals.addElement("85");
    ext2AttVals.addElement("99");
    atts.addElement(new Attribute("ext2", ext2AttVals));

    FastVector<String> time2AttVals = new FastVector<String>();
    time2AttVals.addElement("0");
    time2AttVals.addElement("1");
    atts.addElement(new Attribute("time2", time2AttVals));

    FastVector<String> time4AttVals = new FastVector<String>();
    time4AttVals.addElement("0");
    time4AttVals.addElement("1");
    atts.addElement(new Attribute("time4", time4AttVals));

    FastVector<String> time6AttVals = new FastVector<String>();
    time6AttVals.addElement("0");
    time6AttVals.addElement("1");
    atts.addElement(new Attribute("time6", time6AttVals));

    FastVector<String> time8AttVals = new FastVector<String>();
    time8AttVals.addElement("0");
    time8AttVals.addElement("1");
    atts.addElement(new Attribute("time8", time8AttVals));

    FastVector<String> time10AttVals = new FastVector<String>();
    time10AttVals.addElement("0");
    time10AttVals.addElement("1");
    atts.addElement(new Attribute("time10", time10AttVals));

    // 2. create Instances object
    data = new Instances("SeerBreastCancer", atts, 0);

    // 3. fill with data
    vals = new double[data.numAttributes()];

    vals[0] = wekaData.getAgeDiagNum();
    vals[1] = raceGroupAttVals.indexOf(wekaData.getRaceGroup());
    vals[2] = stage3AttVals.indexOf(wekaData.getStage3());
    vals[3] = m3AttVals.indexOf(wekaData.getM3());
    vals[4] = reasonNoCancerSurgAttVals.indexOf(wekaData
        .getReasonNoCancerSurg());
    vals[5] = ext2AttVals.indexOf(wekaData.getExt2());
    vals[6] = Utils.missingValue();
    vals[7] = Utils.missingValue();
    vals[8] = Utils.missingValue();
    vals[9] = Utils.missingValue();
    vals[10] = Utils.missingValue();

    // add
    data.add(new DenseInstance(1.0, vals));

    logger.info("\nCalcArffServiceImpl: creating Instances data\n"
        + data.toString() + "\n");

    return data;
  }

}
```

Source Code 26: ph/edu/upm/agila/gtmeren/bosom/service/impl/CalcModelServiceImpl.java

```java
package ph.edu.upm.agila.gtmeren.bosom.service.impl;

import java.io.IOException;
```

```
import java.io.InputStream;
import java.util.LinkedHashMap;
import java.util.Map;

import javax.annotation.Resource;
import javax.servlet.ServletContext;
import javax.servlet.http.HttpServletRequest;
import javax.xml.ws.WebServiceContext;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.springframework.stereotype.Service;

import ph.edu.upm.agila.gtmeren.bosom.service.CalcModelService;
import weka.classifiers.Classifier;
import weka.core.Instances;

@Service("calcModelService")
public class CalcModelServiceImpl implements CalcModelService {

  protected final static Log logger = LogFactory
      .getLog(CalcModelServiceImpl.class);
  private static final String[] NAME_CLASSIFIERS = { "adt", "lb",
        "j48",
      "rf", "rs" };
  private static final String[] NAME_CLASSIFIERS_LOCATION = {
        "time2",
      "time4", "time6", "time8", "time10" };

  @Resource
  private WebServiceContext wsContext;

  public Classifier getClassifier(String filePath,
        HttpServletRequest request) {
    String path = "/WEB-INF/models/" + filePath + ".MODEL";

    Classifier classifier = null;

    ServletContext servletContext = request.getSession()
        .getServletContext();
    InputStream inputStream = null;
    try {
      inputStream = servletContext.getResourceAsStream(path);

      logger.info("\nCalcModelServiceImpl: reading model files \n"
          + "Path: " + path + "\n");

      classifier = (Classifier) weka.core.SerializationHelper
          .read(inputStream);
    } catch (IOException e) {
      e.printStackTrace();
    } catch (Exception e) {
      e.printStackTrace();
    } finally {
      if (inputStream != null) {
        try {
          inputStream.close();
        } catch (IOException ioe) {
          ioe.printStackTrace();
        }
      }
    }

    return classifier;
  }

  private static Map<String, Object> predict(Instances instances,
      Classifier classifier, int attributeIndex) {
    Map<String, Object> map = new LinkedHashMap<String, Object>();
    int instanceIndex = 0; // do not change, equal to row 1 of ARFF
    double[] percentage = { 0 };
    double outcomeValue = 0;

    instances.setClassIndex(attributeIndex);

    try {
      outcomeValue =
            classifier.classifyInstance(instances.instance(0));
      percentage = classifier.distributionForInstance(instances
          .instance(instanceIndex));
    } catch (Exception e) {
      e.printStackTrace();
    }

    map.put("Class", outcomeValue);

    map.put("Percentage", percentage[1]);
    logger.info("CalcModelServiceImpl: predicting class and its
          percentage distribution\n"
        + "Classifier: "
        + classifier.getClass().toString()
        + "\n"
        + "Class [0=Dead,1=Alive]: "
        + outcomeValue
        + "\n"
        + "Percentage [0]: "
        + percentage[0]
        + "\n"
        + "Percentage [1]: " + percentage[1] + "\n");

    return map;
  }

  public Map<String, Map<String, Map<String, Object>>>
        getPredictions(
      Instances instances, HttpServletRequest request) {
    Map<String, Map<String, Map<String, Object>>> container = new
          LinkedHashMap<String, Map<String, Map<String,
          Object>>>();
    Map<String, Map<String, Object>> content;

    for (int i = 0; i < NAME_CLASSIFIERS_LOCATION.length; i++) {
      content = new LinkedHashMap<String, Map<String, Object>>();
      for (int j = 0; j < NAME_CLASSIFIERS.length; j++) {

        String path = NAME_CLASSIFIERS_LOCATION[i] + "/"
            + NAME_CLASSIFIERS[j];

        Map<String, Object> predictions = predict(instances,
            getClassifier(path, request), i + 6);

        content.put(NAME_CLASSIFIERS[j], predictions);
      }
      container.put(NAME_CLASSIFIERS_LOCATION[i], content);
    }

    return container;
  }
}
```

Source Code 27: ph/edu/upm/agila/gtmeren/bosom/service/impl/CalcServiceImpl.java

```
package ph.edu.upm.agila.gtmeren.bosom.service.impl;

import java.math.BigDecimal;
import java.util.HashMap;
import java.util.LinkedHashMap;
import java.util.Map;

import javax.servlet.http.HttpServletRequest;

import org.apache.commons.logging.Log;
import org.apache.commons.logging.LogFactory;
import org.springframework.beans.factory.annotation.Autowired;
import org.springframework.stereotype.Service;

import ph.edu.upm.agila.gtmeren.bosom.domain.WekaData;
import ph.edu.upm.agila.gtmeren.bosom.service.CalcArffService;
import ph.edu.upm.agila.gtmeren.bosom.service.CalcModelService;
import ph.edu.upm.agila.gtmeren.bosom.service.CalcService;
import weka.core.Instances;

@Service("calcService")
public class CalcServiceImpl implements CalcService {

  protected final Log logger = LogFactory.getLog(getClass());

  private CalcArffService calcArffService;
  private CalcModelService calcModelService;

  @Autowired
  public void setCalcArffService(CalcArffService calcArffService) {
    this.calcArffService = calcArffService;
  }

  @Autowired
  public void setCalcModelService(CalcModelService
        calcModelService) {
    this.calcModelService = calcModelService;
  }

  @Override
  public Map<String, Map<String, Map<String, Object>>>
        evaluate(WekaData wekaData, HttpServletRequest request) {

    Instances instances = calcArffService.getInstances(wekaData);

    Map<String, Map<String, Map<String, Object>>> predictions =
          calcModelService
        .getPredictions(instances, request);
    Map<String, Double> meanMap = getMeanPredictions(predictions);

    wekaData.setTime2(new
          BigDecimal(meanMap.get("time2")).setScale(2,
        BigDecimal.ROUND_HALF_UP));
    wekaData.setTime4(new
          BigDecimal(meanMap.get("time4")).setScale(2,
        BigDecimal.ROUND_HALF_UP));
    wekaData.setTime6(new
          BigDecimal(meanMap.get("time6")).setScale(2,
        BigDecimal.ROUND_HALF_UP));
    wekaData.setTime8(new
          BigDecimal(meanMap.get("time8")).setScale(2,
        BigDecimal.ROUND_HALF_UP));
    wekaData.setTime10(new
          BigDecimal(meanMap.get("time10")).setScale(2,
```

```
        BigDecimal.ROUND_HALF_UP));

    return predictions;
  }

  private Map<String, Double> getMeanPredictions(
      Map<String, Map<String, Map<String, Object>>> map) {
    Map<String, Double> meanMap = new LinkedHashMap<String,
        Double>();
    BigDecimal meanPrediction = new BigDecimal(0);

    for (Map.Entry<String, Map<String, Map<String, Object>>> entry1
        : map
        .entrySet()) {
      Map<String, Map<String, Object>> entry1Map =
          entry1.getValue();

      logger.info("\nCalcServiceImpl: extracting data per time
          period\n"
          + "Time Period: " + entry1.getKey() + "\n" + "Data: "
          + entry1Map + "\n");

      double sumPercentages = 0;
      for (Map.Entry<String, Map<String, Object>> entry2 : entry1Map
          .entrySet()) {
        HashMap<String, Object> entry2Map = (HashMap<String,
            Object>) entry2
            .getValue();
        sumPercentages += (Double) entry2Map.get("Percentage");
      }

      meanPrediction = new BigDecimal(
          String.valueOf((sumPercentages / 5) * 100)).setScale(2,
          BigDecimal.ROUND_HALF_UP);

      logger.info("\nCalcServiceImp: computing prediction means"
          + "Time Period: " + entry1.getKey() + "\n" + "Sum: "
          + (sumPercentages * 100) + "\n" + "Mean: "
          + meanPrediction.doubleValue() + "\n");

      meanMap.put(entry1.getKey(), meanPrediction.doubleValue());
    }
    return meanMap;
  }

}
```

Source Code 28: bosom/WEB-INF/classes/file.locations.properties

```
# location of desired files
# dependent to spring-servlet.xml
# <context:property-placeholder
      location="classpath:/file.locations.properties"/>
dir.server=/home/gtmeren/tomcat7/resources/
dir.local=D:/resources/
```

Source Code 29: bosom/WEB-INF/classes/messages.validation.properties

```
NotEmpty.wekaData = This field must not be empty.

NotNull.wekaData.ageDiagNum = "Age of patient in years at time of
      diagnosis (1 - 150 only)" must not be empty.
Range.wekaData.ageDiagNum = "Age of patient in years at time of
      diagnosis (1 - 150 only)" must be between 1 and 150 only.
typeMismatch.java.lang.Integer = "Age of patient in years at time
      of diagnosis (1 - 150 only)" must be a number (integer).

NotEmpty.wekaData.raceGroup = "Race of patient" must not be empty.
NotEmpty.wekaData.stage3 = "Stage of Cancer (AJCC 6th Edition)"
      field must not be empty.
NotEmpty.wekaData.m3 = "Spread of metastasis" must not be empty.
NotEmpty.wekaData.reasonNoCancerSurg = "Details of cancer-directed
      surgery" must not be empty.
NotEmpty.wekaData.ext2 = "Extension of primary tumor code" must
      not be empty.
```

Source Code 30: bosom/index.jsp

```
<%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/header.jsp"%>

<div class="jumbotron" id="page-index">
  <div class="overlay">
    <h1>Welcome</h1>
    <p class="lead">This is the Breast Cancer Outcome - Survival
        Online
      Measurement Calculator's website.</p>
    <p>
      <a class="btn btn-lg btn-info" href="<c:url
          value="/about"/>">Learn
      more &rarr;</a>
    </p>
  </div>
</div>

<div class="row marketing hide">
<%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

Source Code 31: bosom/WEB-INF/jsp/includes/header.jsp

```
<!DOCTYPE html>
<!--[if lt IE 7 ]><html class="ie ie6" lang="en"> <![endif]-->
<!--[if IE 7 ]><html class="ie ie7" lang="en"> <![endif]-->
<!--[if IE 8 ]><html class="ie ie8" lang="en"> <![endif]-->
<!--[if (gte IE 9)|!(IE)]><!-->
<html lang="en">
<!--<![endif]-->
<head>

<!-- Basic Page Needs -->
<meta charset="utf-8" />

<title>BOSOM
    <c:set var="title" value="${title}" />
    <c:choose>
        <c:when test="${not empty title}">
            | <c:out value="${title}" />
        </c:when>
        <c:when test="${empty title}">
            | Welcome
        </c:when>
    </c:choose>
</title>

<meta name="description" content="Breast cancer prediction
      calculator using WEKA models and SEER data" />
<meta name="author" content="Troy Meren" />

<!-- Mobile Specific Metas -->
<meta name="viewport" content="width=device-width,
      initial-scale=1, maximum-scale=1">

<!-- Grand JavaScript -->
<script type="text/javascript" src="<c:url
      value="/resources/js/jquery.js"/>"></script>
<script type="text/javascript" src="<c:url
      value="/resources/js/bootstrap.min.js"/>"></script>

<!-- HTML5 Shim and Respond.js IE8 support of HTML5 elements and
      media queries -->
<!-- WARNING: Respond.js doesn't work if you view the page via
      file:// -->
<!--[if lt IE 9]>
  <script type="text/javascript" src="<c:url
      value="/resources/js/html5shiv.js"/>"></script>
  <script type="text/javascript" src="<c:url
      value="/resources/js/respond.min.js"/>"></script>
  <script
      src="https://oss.maxcdn.com/libs/html5shiv/3.7.0/html5shiv.js"></script>
  <script
      src="https://oss.maxcdn.com/libs/respond.js/1.4.2/respond.min.js"></script>
<![endif]-->

<c:set var="isFlotUsed" value="${isFlotUsed}" />
<c:choose>
    <c:when test="${isFlotUsed == true}">
    <script type="text/javascript" src="<c:url
        value="/resources/js/flot.js/jquery.flot.min.js"/>"></script>
    <script type="text/javascript" src="<c:url
        value="/resources/js/flot.js/jquery.flot.categories.min.js"/>"></script>
    <script type="text/javascript" src="<c:url
        value="/resources/js/flot.js/jquery.flot.axislabels.js"/>"></script>
    </c:when>
</c:choose>

<!--[if lte IE 8]>
    <script language="javascript" type="text/javascript"
        src="/resources/js/flot.js/excanvas.min.js"></script>
<![endif]-->

<!-- Stylesheets -->
<link rel="stylesheet" type="text/css" href="<c:url
      value="/resources/css/bootstrap.min.css" />" />
<link rel="stylesheet" type="text/css" href="<c:url
      value="/resources/css/bootstrap-theme.min.css" />" />
<link rel="stylesheet" type="text/css" href="<c:url
      value="/resources/css/jumbotron-narrow.css" />" />
<link rel="stylesheet" type="text/css" href="<c:url
      value="/resources/css/custom.css" />" />

<!-- Favicons -->
<link rel="shortcut icon" href="<c:url
      value="/resources/images/favicon.ico"/>" />
<link rel="apple-touch-icon" href="<c:url
      value="/resources/images/apple-touch-icon.png"/>" />
<link rel="apple-touch-icon" sizes="72x72" href="<c:url
      value="/resources/images/apple-touch-icon-72x72.png"/>" />
```

```
<link rel="apple-touch-icon" sizes="114x114" href="<c:url
      value="/resources/images/apple-touch-icon-114x114.png"/>" />

</head>

<body data-spy="scroll" data-target="#bosom-scrollspy">

 <div class="container container-fluid">

   <div class="header">

          <h3 id="site-title">
             <img src="<c:url
                    value="/resources/images/logos/up.jpeg" />"
               class="img-responsive img-inline-left" alt="WEKA
                     logo" style="width:100px"
               data-image-source="http://up.edu.ph/
                     wp-content/uploads/ 2013/04/
                     seal-flattened.jpg" />
             BOSOM
             <small>
                Breast Cancer Outcome - Survival Online
                      Measurement Calculator
             </small>
          </h3>

     <ul class="nav nav-pills nav-justified" id="header-nav">

              <li
        <c:if test="${empty pageName}">
          class="active"
          </c:if>
             >
        <a href="<c:url value="/"/>"
           aria-label="Home"
           data-tooltip="Home"
           title="Home">
              Home
        </a>
       </li>

       <li class="dropdown
         <c:if test="${(pageName == 'about')}">
                  active
                  </c:if>
              ">
                     <a id="dropdown-about" role="button"
                           data-toggle="dropdown"
                        href="<c:url value="/about/bosom"/>">
             About <b class="caret"></b>
           </a>

                  <ul role="menu" class="dropdown-menu"
                        aria-labelledby="dropdown-about">
                     <li role="presentation">
                        <a role="menuitem" href="<c:url
                              value="/about/bosom"/>">
                           BOSOM Calculator
                        </a>
                     </li>
                     <li role="presentation" class="divider"></li>
                     <li role="presentation">
                        <a role="menuitem" href="<c:url
                              value="/about/site"/>">
                           BOSOM Site
                        </a>
                     </li>
                  </ul>
       </li>

       <li
         <c:if test="${(pageName == 'calc')}">
                     class="active"
                  </c:if>
              >
        <a href="<c:url value="/calc"/>" aria-label="BOSOM
              Calculator"
         data-tooltip="BOSOM Calculator" title="BOSOM Calculator">
              BOSOM
           Calculator </a>
       </li>

       <li
                  <c:if test="${(pageName == 'supplements')}">
                     class="active"
                  </c:if>
              >
                  <a href="<c:url value="/supplements"/>"
                     aria-label="Supplements"
                     data-tooltip="Supplements"
                     title="Supplements">
                        Supplements
                  </a>
              </li>

     </ul>

   </div>
```

Source Code 32:
bosom/WEB-INF/jsp/includes/footer.jsp

```
      </div> <!-- end div.row -->

      <div class="footer container-fluid">

         <div class="row">

            <div class="col-xs-12 col-sm-6 col-md-8">
               <div class="row">
                  <h4>Information</h4>
                  <p>
                     Developed by
                     <a
                           href="https://twitter.com/troymeren">Troy
                           Meren</a>
                     &copy; 2013 - 2014
                  </p>
               </div>

               <hr/>

               <div class="row">
                  <dl>
                     <dt>
                        <a href="http://www.upm.edu.ph/">
                           University of the Philippines Manila
                        </a>
                     </dt>
                     <dd>College of Arts and Sciences</dd>
                     <dd>Department of Physical Sciences and
                           Mathematics</dd>
                     <dd>Mathematics and Computing Sciences
                           Unit</dd>
                     <dd>Padre Faura St., Ermita, Manila</dd>
                  </dl>
               </div>
            </div>


            <div class="col-xs-6 col-md-4" id="site-map">

               <h4>Site Map</h4>

               <ul class="list-unstyled">
                  <li>
                     <a href="<c:url value="/"/>"
                        aria-label="Home"
                        data-tooltip="Home"
                        title="Home">
                           Home
                     </a>
                  </li>

                  <li>
                     <a href="<c:url value="/about/bosom"/>"
                        aria-label="About"
                        data-tooltip="About"
                        title="About">
                        About
                     </a>
                     <ul>
                        <li><a href="<c:url
                              value="/about/bosom"/>">BOSOM
                              Calculator</a></li>
                        <li><a href="<c:url
                              value="/about/site"/>">BOSOM
                              Site</a></li>
                     </ul>
                  </li>

                  <li>
                     <a href="<c:url value="/calc"/>"
                           aria-label="BOSOM Calculator"
                        data-tooltip="BOSOM Calculator"
                              title="BOSOM Calculator"> BOSOM
                           Calculator </a>
                  </li>

                  <li>
                     <a href="<c:url value="/supplements"/>"
                        aria-label="Supplements"
                        data-tooltip="Supplements"
                        title="Supplements">
                           Supplements
                     </a>
                  </li>
               </ul>

            </div>

         </div>
      </div>

   </div> <!-- div.container -->

   <script type="text/javascript">
   $(document).ready(function() {
       // nasty scrollspy
```

```
        $(window).resize(function() {
            if ($(this).width() > 720) {
                $(".scrollspy-nav").affix({ offset: { top: 15 } });
            }
        });
    });
    </script>

</body>
</html>
```

Source Code 33:
bosom/WEB-INF/jsp/includes/page-header.jsp

```
<div class="page-header">
  <h1>
    <c:out value="${pageTitleHeader}"/>
    <small><c:out value="${pageTitleSubheader}"/></small>
  </h1>
</div>

<div class="row">
```

Source Code 34:
bosom/WEB-INF/jsp/includes/taglibs.jsp

```
<%@ page isELIgnored="false"%>
<%@ page language="java" contentType="text/html;
      charset=ISO-8859-1" pageEncoding="ISO-8859-1"%>

<%@ taglib prefix="form"
      uri="http://www.springframework.org/tags/form"%>
<%@ taglib prefix="spring"
      uri="http://www.springframework.org/tags"%>
<%@ taglib prefix="c" uri="http://java.sun.com/jsp/jstl/core"%>
<%@ taglib prefix="fn"
      uri="http://java.sun.com/jsp/jstl/functions" %>
<%@ taglib prefix="fmt" uri="http://java.sun.com/jsp/jstl/fmt" %>
```

Source Code 35:
bosom/WEB-INF/jsp/about-bosom.jsp

```
<%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/header.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/page-header.jsp"%>

<div class="row">

  <div class="col-xs-12 col-sm-2 col-md-3 bosom-scrollspy-nav">

    <ul class="nav nav-tabs nav-stacked scrollspy-nav"
          id="about-nav">

          <li>
              <a href="#breast-cancer" class="list-group-item
                    stack-first"
                  aria-label="About: Breast cancer"
                  data-tooltip="About: Breast cancer"
                  title="About: Breast cancer">
                  Breast cancer
              </a>
          </li>


          <li>
              <a href="#data-mining" class="list-group-item"
                aria-label="About: Data mining"
                data-tooltip="About: Data mining"
                title="About: Data mining">
                Data mining
              </a>
          </li>

          <li>
              <a href="#seer-data" class="list-group-item
                    stack-last"
                aria-label="About: SEER data"
                data-tooltip="About: SEER data"
                title="About: SEER data">
                SEER data
              </a>
          </li>

          <li>
              <a href="#predictive-survival"
                    class="list-group-item"
                aria-label="About: Predicting survival"
                data-tooltip="About: Predicting survival"
                title="About: Predicting survival">
                Predicting survival
              </a>
          </li>

    </ul>


</div>

<div class="col-xs-12 col-sm-10 col-md-9 bosom-scrollspy-content">

      <div class="bosom-section" id="breast-cancer">

          <h2>Breast Cancer</h2>
          <p>
              It starts from healthy breast cells that undergo
                    mutation defects.
              Normally, unhealthy and dead cells are either
                    repaired or replaced completely
              in order to preserve the rest of the group but these
                    "defected" cells
              continue to develop and eventually affecting the
                    healthy cells.
              This causes <em>tumors</em>, or the mass group of
                    defected cells
              that if left untreated, could spread to the other
                    parts of the body
              [<a href="#bosom-ref-acs">1</a>].
          </p>
          <p>
              Breast cancer has been found to be the leading type
                    of cancer in women worldwide.
              2012 Global Cancer (GLOBOCAN) statistics show that
                    breast cancer scored the
              highest in incidence and second highest in mortality
                    in both sexes
              [<a href="#bosom-ref-globocan">2</a>].
          </p>
          <p>
              Efforts have been made worldwide to increase
                    awareness of the public
              to the causes and preventions of this cancer. The
                    challenge to eradicate the
              negative reputation of this disease has driven
                    organizations and governments
              to encourage everyone to be proactive in dealing
                    with breast cancer.
              In relation, early detection has been found to be
                    effective in treating early stages.
              Men and women who suspect to have abnormal lumps or
                    feels pain in their
              breast area are advised to go see a specialist for
                    proper diagnosis that could
              save their lives.
          </p>
      </div>

      <div class="bosom-section" id="data-mining">

          <h2>Data mining</h2>
          <p>
              Data mining is the discipline of finding patterns
                    and relationships within data or
              records that could lead to a sensible purpose to
                    help understanding the entire
              body of data.
          </p>
          <p>
              Mathematical and computing algorithms are applied to
                    data in order to obtain
              these patterns and relationships. The results could
                    be in the form of a rule-bases system,
              mirroring a human's reasoning method, or with
                    weights or scores, assigned
              to the records and parameters with high significance
                    in the dataset.
          </p>
          <p>
              Today, major websites like Facebook, Twitter and
                    Google, employ large-scale servers
              to store data from users worldwide in real time.
              Search strings, status posts, and tweets among others
              are constantly analyzed to discover what the users
                    are currently enjoying the most,
              for example. Results from this could be applied to
                    add new features geared to
              improve their website's appeal to the public.
          </p>
      </div>

      <div class="bosom-section" id="seer-data">

          <h2>SEER data</h2>
          <p>
              Our breast cancer data comes from the Surveillance,
                    Epidemiology, and End Results
              Program (SEER) of the National Cancer Institute of
                    the US Department of Health and
              Human Services. SEER is the government agency
                    responsible for collecting cancer
              data from key locations in the USA. These records
                    are free for public access but
              an interested party must submit an accomplished
                    research data-use agreement
              first to inform the agency of their intent.
```

```
            </p>
        </div>


        <div class="bosom-section" id="predictive-survival">

            <h2>Predicting survival</h2>
            <p>
                In order to predict a patient's survival, the SEER breast cancer data where
                analyzed by data mining algorithms. Survival within two, four, six, eight, and ten
                years where calculated from around 100,000 records between 1998 and 2003.
            </p>

        </div>

        <div class="panel panel-info">
            <div class="panel-heading">References</div>

            <div class="panel-body">
                <ol>
                    <li id="bosom-ref-acs">
                        <p>
                            "<em>Breast Cancer</em>". American Cancer Society. 2013.
                            Available from:
                            <tt>
                                <a href="http://www.cancer.org/acs/groups/cid/documents/webcontent/003090-pdf.pdf">
                                http://www.cancer.org/acs/groups/cid/documents/webcontent/003090-pdf.pdf </a>
                            </tt>.
                            Accessed on 19 July 2013.
                        </p>
                    </li>
                    <li id="bosom-ref-globocan">
                        <p>
                            Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C,
                            Rebelo M, Parkin DM, Forman D, Bray, F.
                            "<em>Population Fact Sheets</em>".
                            GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide:
                            IARC CancerBase No. 11 [Internet].
                            Lyon, France: International Agency for Research on Cancer; 2013.
                            Available from:
                            <tt>
                                <a href="http://globocan.iarc.fr">
                                    http://globocan.iarc.fr
                                </a>
                            </tt>.
                            Accessed on 28 February 2014.
                        </p>
                    </li>
                </ol>
            </div>

        </div>


        <div class="clearfix"></div>

    </div>

    <div class="clearfix"></div>

    <%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

Source Code 36: bosom/WEB-INF/jsp/about-site.jsp

```
<%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/header.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/page-header.jsp"%>

<div class="row">

    <div class="col-xs-12 col-sm-2 col-md-3 bosom-scrollspy-nav">

        <ul class="nav nav-tabs nav-stacked scrollspy-nav" id="about-nav">
            <li>
                <a href="#about-site-calculator" class="list-group-item stack-first"
                    aria-label="About: Breast cancer calculator"
                    data-tooltip="About: Breast cancer calculator"
                    title="About: Breast cancer calculator">
                    Calculator
                </a>
            </li>

            <li>
                <a href="#about-site-backend" class="list-group-item"
                    aria-label="About: Website backend"
                    data-tooltip="About: Website backend"
                    title="About: Website backend">
                    Website backend
                </a>
            </li>

            <li>
                <a href="#about-site-frontend" class="list-group-item"
                    aria-label="About: Website frontend"
                    data-tooltip="About: Website frontend"
                    title="About: Website frontend">
                    Website frontend
                </a>
            </li>

            <li>
                <a href="#about-site-developer" class="list-group-item"
                    aria-label="About: Developer"
                    data-tooltip="About: Developer"
                    title="About: Developer">
                    Developer
                </a>
            </li>

        </ul>
    </div>


    <div class="col-xs-12 col-sm-10 col-md-9 bosom-scrollspy-content">

        <div class="bosom-section" id="about-site-calculator">

            <h2>Calculator</h2>
            <p>
                The breast cancer data from the <a href="http://seer.cancer.gov/">Surveillance, Epidemiology, and End Results </a> program were used to create the models to calculate a prediction of a patient's survival.
            </p>

            <p>
                <img src="<c:url value="/resources/images/logos/weka.png"/>" class="img-responsive img-inline-left" alt="WEKA logo"
                    data-image-source="http://www.cs.waikato.ac.nz/ml/weka/citing.html" />

                The <a href="http://www.cs.waikato.ac.nz/ml/">University of Waikato Machine Learning Group </a>'s open-source machine learning software <a href="http://www.cs.waikato.ac.nz/ml/weka/">Waikato Environment for Knowledge Analysis </a> provided the tool to create models.
            </p>

            <p>
                The Java API has built-in algorithms and modules for preprocessing, modeling and forecasting that are helpful in general data mining and artificial intelligence projects. Its components are free to modify for more specific tasks that are not currently implemented in the official releases.
            </p>

        </div>

        <div class="bosom-section" id="about-site-backend">

            <h2>Website backend</h2>

            <h3>Framework</h3>
            <p>
                <img src="<c:url value="/resources/images/logos/spring-source.gif"/>" class="img-responsive img-inline-left" alt="Spring Source logo"
                    data-image-source="http://blog.architexa.com/v2012/10/the-decline-of-spring/" />

                The open source Java web framework <a href="http://spring.io/">Spring MVC </a> is used to serve the user interface and data from the WEKA models.
            </p>
            <p>
                This Java-based model-view-controller framework is known for its reliability and maintainability in development of websites proven by its
```

```
                    &amp;quot;separation of concern&amp;quot; paradigm as
                    seen in its components.
            </p>
            <p>
                Kindly refer to the <a href="http://docs.spring.io/
                    spring/docs/current/
                    spring-framework-reference/html/mvc.html">
                    introductory documentation </a> on its
                    theoretical flow and principles and how to
                    get started programming with Spring and the
                    web.
            </p>

            <h3>Server</h3>
            <p>
                This website is hosted by University of the
                    Philippines Manila's
                <a href="http://agila.upm.edu.ph/">
                    Agila Computer Science Development Server
                 </a>. More information can be found in its
                <a href="http://agila.upm.edu.ph/ docs/doku.php">
                    wiki page
                 </a>.
            </p>

        </div>

        <div class="bosom-section" id="about-site-frontend">

            <h2>Website frontend</h2>

            <h3>User interface</h3>
            <p>
                We employed <a
                    href="http://getbootstrap.com/">Twitter's
                    Bootstrap </a> user interface framework for
                    faster and better deployment of the
                    application given its variety of helper
                    design components and interactive modules.
                    User interaction and website visibility to
                    most devices and browsers are greatly
                    improved by this framework.
            </p>

            <h3>Graphs</h3>
            <p>
                The Calculator's graph in the results page is made
                    using <a
                    href="http://www.flotcharts.org/">Flot </a>,
                    a cross-browser interactive plotting
                    JavaScript library. It is capable of
                    generating various chart types - line, bar
                    and pie and can be further expanded for more
                    specific usage.
            </p>

            <h3>Images</h3>
            <p>
                Images seen in this site are not my property unless
                    stated.
            </p>

            <ul>
                <li>
                    <strong><a href="http://nos.twnsnd.co/"> New Old
                        Stock </a></strong> provides free vintage
                        photos from public archives. Its usage
                        policies are stated <a
                        href="http://www.flickr.com/
                        commons/usage/"> here </a>.
                </li>
                <li>
                    <strong><a href="http://unsplash.com/"> Unsplash
                        </a></strong> provides high-resolution
                        photos under the <a
                        href="http://creativecommons.org/
                        publicdomain/zero/1.0/"> &amp;quot;Public
                        Domain Dedication&amp;quot; </a> license.
                </li>
            </ul>

        </div>

        <div class="bosom-section" id="about-site-developer">

            <h2>Developer</h2>
            <p>
                I am GilTroy Meren, an undergraduate computer
                    science student
                from the University of the Philippines Manila.
            </p>

            <p>
                You can contact me through the following:
            </p>

            <ul>
                <li>
                    <strong>E-mail: </strong>
                    <tt>gpmeren@up.edu.ph</tt>
                </li>
                <li>
                    <strong>Twitter: </strong>
                    <a href="https://twitter.com/troymeren">
                        <tt>@troymeren</tt>
                     </a>
                </li>
            </ul>

        </div>



        <div class="panel panel-info">
            <div class="panel-heading">References</div>

            <div class="panel-body">
              <p>
                  Additional credit to the images and other
                      resources used in this specific page.
              </p>

              <ul>
                  <li>
                      <p>
                          WEKA logo. "Citing Weka". <em>Waikato
                              Enviroment for Knowledge Analysis
                              website</em>.
                          Waikato Enviroment for Knowledge Analysis.
                          Machine Learning Group, Department of
                              Computer Science,
                          The University of Waikato, New Zealand.
                          <tt>
                              <a href="http://www.cs.waikato.ac.nz/
                                  ml/weka/citing.html">
                                  href="http://www.cs.waikato.ac.nz/
                                      ml/weka/citing.html
                              </a>
                          </tt>.
                      </p>

                    <p>
                      Licensed under Creative Commons
                        <a href="http://creativecommons.org/
                              licenses/by-sa/2.5/">
                            Attribution-ShareAlike 2.5 Generic
                        </a>.
                      </p>
                  </li>

                  <li>
                      <p>
                        Spring Source logo. "The Decline Of Spring?".
                              <em>Working with Large Codebases</em>.
                        Architexa, Inc.
                        <tt>
                             <a href="http://blog.architexa.com/
                                  2012/10/the-decline-of-spring/">
                                http://blog.architexa.com/
                                    2012/10/the-decline-of-spring/
                            </a>
                          </tt>.
                      </p>
                  </li>
              </ul>
            </div>

        </div>

        <div class="clearfix"></div>

    </div>

<div class="clearfix"></div>

<%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

Source Code 37: bosom/WEB-INF/jsp/calc/form.jsp

```
<%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/header.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/page-header.jsp"%>

<c:if test="${not empty alertContent}">
  <div class="col-lg-12 col-sm-12 col-xs-12">
    <div
      class="alert alert-<c:out value="${alertType}"/>
            alert-dismissable">
      <button type="button" class="close" data-dismiss="alert"
        aria-hidden="true">&amp;times;</button>
      <strong><c:out value="${alertStrongContent}" /></strong>
      <c:out value="${alertContent}" />
    </div>
  </div>
</c:if>

```

```
<div class="col-lg-5 col-sm-4 col-xs-12">
  <h4>Reminders</h4>
  <p>
    In order to provide you with the predicted breast cancer survival,
    the form provided in this page must be accomplished completely and correctly.
    If any alerts or error messages show after submitting, kindly follow their
    instruction to successfully answer the form.
  </p>

  <p>
    There are guides provided
    (seen as
    <span class="label label-info">
        <span class="glyphicon glyphicon-info-sign"></span>
        More info
    </span>)
    beside each item to help you understand. Note that most of the terms provided
    are in medical jargon - please ask a doctor for these values' definition.
  </p>

  <p>
    The predicted survival provided by the BOSOM Calculator does not directly
    correspond to a legitimate diagnosis. It is strongly advised to consult a
    doctor or cancer specialist to interpret and guide the patient regarding the
    relationships of the input fields and their values to the predictions.
  </p>

</div>

<div class="col-lg-7 col-sm-8 col-xs-12 details">

  <h4>Please provide answers to the following items:</h4>

  <p>
    Click on each item to either type in your answer or choose from the values provided.
  </p>

    <div class="panel panel-default">
        <form:form id="form-calc" modelAttribute="wekaData" method="POST" role="form" action="calc">
            <div class="panel-body">


                <c:set var="ageDiagNumErrors"><form:errors path="ageDiagNum"/></c:set>
                <c:set var="raceGroupErrors"><form:errors path="raceGroup"/></c:set>
                <c:set var="stage3Errors"><form:errors path="stage3"/></c:set>
                <c:set var="m3Errors"><form:errors path="m3"/></c:set>
                <c:set var="reasonNoCancerSurgErrors"><form:errors path="reasonNoCancerSurg"/></c:set>
                <c:set var="ext2Errors"><form:errors path="ext2"/></c:set>

                <div class="form-group
                    <c:if test="${not empty ageDiagNumErrors}">has-error</c:if>
                ">
                    <label for="form-ageDiagNum" class="block">
                        Age of patient in years <br/>
                        at time of diagnosis (1 - 150 only)
                    </label>
                    <form:input class="form-control" path="ageDiagNum"
                        id="form-ageDiagNum" required="required" autocomplete="off"
                        min="1" max="150" pattern="\d+"/>
                    <form:errors path="ageDiagNum" element="p" cssClass="help-block bg-danger" text-danger"/>
                </div>

                <div class="form-group
                    <c:if test="${not empty raceGroupErrors}">has-error</c:if>
                ">
                    <label for="form-raceGroup" class="block">Race of patient</label>
                    <form:select class="form-control" path="raceGroup"
                        id="form-raceGroup" required="required">
                        <form:option value=""></form:option>
                        <form:option value="Black">Black</form:option>
                        <form:option value="White">White</form:option>
                        <form:option value="Other">Otherwise</form:option>
                        <form:option value="Unknown">Unknown race</form:option>
                    </form:select>
                    <form:errors path="raceGroup" element="p" cssClass="help-block bg-danger" text-danger"/>
                </div>

                <div class="form-group
                    <c:if test="${not empty stage3Errors}">has-error</c:if>
                ">
                    <label for="form-stage3" class="block">Stage of cancer (AJCC 6th Edition)</label>
                    <form:select class="form-control" path="stage3" id="form-stage3"
                        required="required">
                        <form:option value=""></form:option>
                        <form:option value="0">0</form:option>
                        <form:option value="I">I</form:option>
                        <form:option value="IIA">IIA</form:option>
                        <form:option value="IIB">IIB</form:option>
                        <form:option value="IIIA">IIIA</form:option>
                        <form:option value="IIIB">IIIB</form:option>
                        <form:option value="IIIC">IIIC</form:option>
                        <form:option value="IIINOS">IIINOS</form:option>
                        <form:option value="IV">IV</form:option>
                        <form:option value="UNK Stage">Unknown stage</form:option>
                    </form:select>
                    <form:errors path="stage3" element="p" cssClass="help-block bg-danger" text-danger"/>
                </div>

                <div class="form-group
                    <c:if test="${not empty m3Errors}">has-error</c:if>
                ">
                    <label for="form-m3" class="block">Spread of metastasis</label>
                    <button type="button" class="btn btn-info btn-xs pull-right"
                        data-toggle="modal" data-target="#modal-m3">
                        <span class="glyphicon glyphicon-info-sign"></span> More info
                    </button>
                    <form:select class="form-control" path="m3" id="form-m3"
                        required="required">
                        <form:option value=""></form:option>
                        <form:option value="M0">M0 (No distant metastasis)</form:option>
                        <form:option value="M1">M1 (Distant metastasis)</form:option>
                        <form:option value="MX">MX (Distant metastasis cannot be assessed)</form:option>
                    </form:select>
                    <form:errors path="m3" element="p" cssClass="help-block bg-danger" text-danger"/>
                </div>

                <div class="form-group
                    <c:if test="${not empty reasonNoCancerSurgErrors}">has-error</c:if>
                ">
                    <label for="form-reasonNoCancerSurg" class="block">
                        Details of cancer-directed surgery
                    </label>
                    <form:select class="form-control" path="reasonNoCancerSurg"
                        id="form-reasonNoCancerSurg" required="required">
                        <form:option value=""></form:option>
                        <form:option value="Not performed, patient died prior to recommended surgery">
                            Not performed and patient died prior to recommended surgery
                        </form:option>
                        <form:option value="Not recommended">
                            Not recommended only
                        </form:option>
                        <form:option value="Not recommended, contraindicated due to other conditions">
                            Not recommended and contraindicated due to other conditions
                        </form:option>
                        <form:option value="Recommended but not performed, patient refused">
                            Recommended but not performed, patient refused
                        </form:option>
```

```
                    <form:option value="Recommended but not performed, unknown reason">Recommended but not performed for unknown reasons</form:option>
                    <form:option value="Recommended, unknown if performed">Recommended but unknown if performed</form:option>
                    <form:option value="Surgery performed">Surgery performed</form:option>
                    <form:option value="Unknown; death certificate or autopsy only case">Unknown OR death certificate or autopsy-only case</form:option>
                </form:select>
                <form:errors path="reasonNoCancerSurg" element="p" cssClass="help-block bg-danger text-danger"/>
            </div>

            <div class="form-group <c:if test="${not empty ext2Errors}">has-error</c:if>">
                <label for="form-ext2" class="block">Extension of primary tumor code</label>
                <button type="button" class="btn btn-info btn-xs pull-right" data-toggle="modal" data-target="#modal-ext2"><span class="glyphicon glyphicon-info-sign"></span> More info</button>
                <form:select class="form-control" path="ext2" id="form-ext2" required="required">
                    <form:option value=""></form:option>
                    <form:option value="00">0</form:option>
                    <form:option value="05">5</form:option>

                    <form:option value="10">10</form:option>
                    <form:option value="11">11</form:option>
                    <form:option value="13">13</form:option>
                    <form:option value="14">14</form:option>
                    <form:option value="15">15</form:option>
                    <form:option value="16">16</form:option>
                    <form:option value="17">17</form:option>
                    <form:option value="18">18</form:option>

                    <form:option value="20">20</form:option>
                    <form:option value="21">21</form:option>
                    <form:option value="23">23</form:option>
                    <form:option value="24">24</form:option>
                    <form:option value="25">25</form:option>
                    <form:option value="26">26</form:option>
                    <form:option value="27">27</form:option>
                    <form:option value="28">28</form:option>

                    <form:option value="30">30</form:option>
                    <form:option value="31">31</form:option>
                    <form:option value="33">33</form:option>
                    <form:option value="34">34</form:option>
                    <form:option value="35">35</form:option>
                    <form:option value="36">36</form:option>
                    <form:option value="37">37</form:option>
                    <form:option value="38">38</form:option>

                    <form:option value="40">40</form:option>
                    <form:option value="50">50</form:option>
                    <form:option value="60">60</form:option>
                    <form:option value="70">70</form:option>
                    <form:option value="80">80</form:option>
                    <form:option value="85">85</form:option>
                    <form:option value="99">99</form:option>
                </form:select>
                <form:errors path="ext2" element="p" cssClass="help-block bg-danger text-danger"/>
            </div>

        </div>

        <div class="panel-footer">
            <button type="submit" class="btn btn-primary" data-toggle="modal" data-target="#modal-submit" id="calc-btn-submit" data-loading-text="<span class='glyphicon glyphicon-refresh'></span> Submitting your form ...">Submit</button>
            <button type="reset" class="btn btn-danger">Clear</button>
        </div>

        <!--
        <div class="modal fade" id="modal-submit" tabindex="-1" role="dialog" aria-labelledby="submit-label" aria-hidden="true">
            <div class="modal-dialog">
                <div class="modal-content">
                    <div class="modal-header">
                        <button type="button" class="close" data-dismiss="modal" aria-hidden="true">&times;</button>
                        <h3 class="modal-title" id="submit-label">Confirm form submission</h3>
                    </div>
                    <div class="modal-body">
                        <p>Are you sure with the answers you provided in the form?
                        There could be items you missed or answered incorrectly.</p>

                        <p>
                            Click
                            <kbd>Submit</kbd>
                            to continue the form submission or
                            <kbd>Cancel</kbd>
                            to go back to the BOSOM page and check the answers.
                        </p>
                        <div class="progress progress-striped active">
                            <div class="progress-bar" role="progressbar" aria-valuenow="0" aria-valuemin="0" aria-valuemax="100" style="width: 0">
                            <span class="sr-only">45% Complete</span>
                            </div>
                        </div>
                    </div>
                    <div class="modal-footer">
                        <button id="submit-modal-form" type="button" class="btn btn-default" data-dismiss="modal">Cancel</button>
                        <button id="modal-submit-btn" type="submit" class="btn btn-primary" data-loading-text="<span class='glyphicon glyphicon-refresh'></span> Submitting your form...">Submit</button>
                    </div>
                </div>
            </div>
        </div>
        -->

    </form:form>
</div>

</div>

<%@ include file="/WEB-INF/jsp/calc/modals.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

Source Code 38:
bosom/WEB-INF/jsp/calc/modals.jsp

```
<div class="modal fade" id="modal-m3" tabindex="-1" role="dialog" aria-labelledby="modal-m3-label" aria-hidden="true">
    <div class="modal-dialog">
        <div class="modal-content">
            <div class="modal-header">
                <button type="button" class="close" data-dismiss="modal" aria-hidden="true">&times;</button>
                <h4 class="modal-title" id="modal-m3-label">
                    Metastasis <br/>
                    <small>
                        <tt>Adjusted AJCC M 6th edition</tt>, SEER 18
                    </small>
                </h4>
            </div>

            <div class="modal-body">
                The information below were taken from
                <a target="_blank" href="http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage/6th/breast.html#m">
                "Adjusted AJCC 6 M (1988+)", <em>Breast Schema for 1988+ based on AJCC 6th edition</em>
                </a>

                <hr/>

                <div class="table-responsive">
                    <table class="table table-hover table-striped table-bordered">
```

```
                    <thead>
                        <tr>
                            <th>Code</th>
                            <th>Description</th>
                        </tr>
                    </thead>
                    <tbody>
                        <tr>
                            <td><tt>M0</tt></td>
                            <td>No distant metastasis</td>
                        </tr>
                        <tr>
                            <td><tt>M1</tt></td>
                            <td>Distant metastasis</td>
                        </tr>
                        <tr>
                            <td><tt>MX</tt></td>
                            <td>Distant metastasis cannot be
                                assessed</td>
                        </tr>
                    </tbody>
                </table>
            </div>

        </div>

        <div class="modal-footer">
            <button type="button" class="btn btn-default"
                data-dismiss="modal">Close</button>
        </div>
    </div>
  </div>
</div>

<div class="modal fade" id="modal-ext2" tabindex="-1" role="dialog"
    aria-labelledby="modal-ext2-label" aria-hidden="true">
    <div class="modal-dialog">
        <div class="modal-content">
            <div class="modal-header">
                <button type="button" class="close"
                    data-dismiss="modal"
                    aria-hidden="true">&times;</button>
                <h4 class="modal-title" id="modal-ext2-label">
                    Extension of tumor <br/>
                    <small>
                        <tt>EOD 10 - extend (1988-2003)</tt>, SEER 18
                    </small>
                </h4>
            </div>

            <div class="modal-body">
                The information below were taken from
                <a target="_blank" href="http://seer.cancer.gov/
                    archive/manuals/EOD10Dig.pub.pdf">
                    "Breast Extension", <em>SEER Extent of Disease
                        -- 1988: Codes and Coding Instructions,
                        1998</em>
                </a>
                (page 110).

                <hr/>

                <div class="table-responsive">
                    <table class="table table-hover table-striped
                        table-bordered">
                        <thead>
                            <tr>
                                <th>Code</th>
                                <th>Description</th>
                            </tr>
                        </thead>
                        <tbody>
                            <tr>
                                <td><tt>00</tt></td>
                                <td>
                                    IN SITU: Noninfiltrating;
                                        intraductal <br/>
                                    WITHOUT infiltration; lobular
                                        neoplasia
                                </td>
                            </tr>
                            <tr>
                                <td><tt>05</tt></td>
                                <td>
                                    Paget's disease (WITHOUT
                                        underlying tumor)
                                </td>
                            </tr>
                            <tr>
                                <td><tt>10</tt></td>
                                <td>
                                    Confined to breast tissue and fat
                                        including nipple and/or
                                        areola
                                </td>
                            </tr>
                            <tr>
                                <td><tt>11</tt></td>
                                <td>
                                    Entire tumor reported as invasive
                                        (no in situ component
                                        reported)
                                </td>
                            </tr>
                            <tr>
                                <td><tt>13</tt></td>
                                <td>
                                    Invasive and in situ components
                                        present, size of invasive
                                        component stated and coded
                                        in tumor Size
                                </td>
                            </tr>
                            <tr>
                                <td><tt>14</tt></td>
                                <td>
                                     Invasive and in situ components
                                        present, size of entire
                                        tumor coded in Tumor Size
                                        (size of invasive
                                        component not stated) AND
                                        in situ described as
                                        minimal (less than 25%)
                                </td>
                            </tr>
                            <tr>
                                <td><tt>15</tt></td>
                                <td>
                                    Invasive and in situ components
                                        present, size of entire
                                        tumor coded in Tumor Size
                                        (size of invasive component
                                        not stated) AND in situ
                                        described as extensive (25%
                                        or more)
                                </td>
                            </tr>
                            <tr>
                                <td><tt>16</tt></td>
                                <td>
                                    Invasive and in situ components
                                        present, size of entire
                                        tumor coded in Tumor Size
                                        (size of invasive component
                                        not stated) AND proportions
                                        of in situ and invasive not
                                        known
                                </td>
                            </tr>
                            <tr>
                                <td><tt>17</tt></td>
                                <td>
                                    Invasive and in situ components
                                        present, unknown size of
                                        tumor (Tumor Size coded
                                        <tt>999</tt>)
                                </td>
                            </tr>
                            <tr>
                                <td><tt>18</tt></td>
                                <td>
                                    Unknown if invasive and in situ
                                        components present, unknown
                                        if tumor size represents
                                        mixed tumor or a "pure"
                                        tumor
                                </td>
                            </tr>
                            <tr>
                                <td><tt>20</tt></td>
                                <td>
                                    Invasion of subcutaneous tissue
                                        <br/>
                                    Skin infiltration of primary
                                        breast including skin of
                                        nipple and/or areola <br/>
                                    Local infiltration of dermal
                                        lymphatics adjacent to
                                        primary tumor involving
                                        skin by direct extension
                                </td>
                            </tr>
                            <tr>
                                <td><tt>21</tt></td>
                                <td>
                                    Entire tumor reported as invasive
                                        (no in situ component
                                        reported)
                                </td>
                            </tr>
                            <tr>
                                <td><tt>23</tt></td>
                                <td>
                                    Invasive and in situ components
                                        present, size of invasive
                                        component stated and coded
                                        in tumor Size
                                </td>
```

```
</tr>
<tr>
    <td><tt>24</tt></td>
    <td>
        Invasive and in situ components present, size of entire tumor coded in Tumor Size (size of invasive component not stated) AND in situ described as minimal (less than 25%)
    </td>
</tr>
<tr>
    <td><tt>25</tt></td>
    <td>
        Invasive and in situ components present, size of entire tumor coded in Tumor Size (size of invasive component not stated) AND in situ described as extensive (25% or more)
    </td>
</tr>
<tr>
    <td><tt>26</tt></td>
    <td>
        Invasive and in situ components present, size of entire tumor coded in Tumor Size (size of invasive component not stated) AND proportions of in situ and invasive not known
    </td>
</tr>
<tr>
    <td><tt>27</tt></td>
    <td>
        Invasive and in situ components present, unknown size of tumor (Tumor Size coded <tt>999</tt>)
    </td>
</tr>
<tr>
    <td><tt>28</tt></td>
    <td>
        Unknown if invasive and in situ components present, unknown if tumor size represents mixed tumor or a "pure" tumor
    </td>
</tr>
<tr>
    <td><tt>30</tt></td>
    <td>
        Invasion of (or fixation to) pectoral fascia or muscle; deep fixation; attachment or fixation to pectoral muscle or underlying tissue
    </td>
</tr>
<tr>
    <td><tt>31</tt></td>
    <td>
        Entire tumor reported as invasive (no in situ component reported)
    </td>
</tr>
<tr>
    <td><tt>33</tt></td>
    <td>
        Invasive and in situ components present, size of invasive component stated and coded in tumor Size
    </td>
</tr>
<tr>
    <td><tt>34</tt></td>
    <td>
        Invasive and in situ components present, size of entire tumor coded in Tumor Size (size of invasive component not stated) AND in situ described as minimal (less than 25%)
    </td>
</tr>
<tr>
    <td><tt>35</tt></td>
    <td>
        Invasive and in situ components present, size of entire tumor coded in Tumor Size (size of invasive component not stated) AND in situ described as extensive (25% or more)
    </td>
</tr>
<tr>
    <td><tt>36</tt></td>
    <td>
        Invasive and in situ components present, size of entire tumor coded in Tumor Size (size of invasive component not stated) AND proportions of in situ and invasive not known
    </td>
</tr>
<tr>
    <td><tt>37</tt></td>
    <td>
        Invasive and in situ components present, unknown size of tumor (Tumor Size coded <tt>999</tt>)
    </td>
</tr>
<tr>
    <td><tt>38</tt></td>
    <td>
        Unknown if invasive and in situ components present, unknown if tumor size represents mixed tumor or a "pure" tumor
    </td>
</tr>
<tr>
    <td><tt>40</tt></td>
    <td>
        Invasion of (or fixation to) chest wall, ribs, intercostal or serratus anterior muscles
    </td>
</tr>
<tr>
    <td><tt>50</tt></td>
    <td>
        Extensive skin involvement: Skin edema, peau d'orange, "pigskin", en cuirasse, lenticular nodule(s), inflammation of skin, erythema, ulceration of skin of breast, satellite nodule(s) in skin of primary breast
    </td>
</tr>
<tr>
    <td><tt>60</tt></td>
    <td>
        (<tt>50</tt>) + (<tt>40</tt>)
        <ul>
            <li>
                Extensive skin involvement: Skin edema, peau d'orange, "pigskin", en cuirasse, lenticular nodule(s), inflammation of skin, erythema, ulceration of skin of breast, satellite nodule(s) in skin of primary breast
            </li>
            <li>
                Invasion of (or fixation to) chest wall, ribs, intercostal or serratus anterior muscles
            </li>
        </ul>
    </td>
</tr>
<tr>
    <td><tt>70</tt></td>
    <td>
        Inflammatory carcinoma, incl. diffuse (beyond that directly overlying the tumor) dermal lymphatic permeation or infiltration
```

```
                            </td>
                        </tr>
                        <tr>
                            <td><tt>80</tt></td>
                            <td>
                                FURTHER contiguous extension: Skin over sternum, upper abdomen, axilla or opposite breast
                            </td>
                        </tr>
                        <tr>
                            <td><tt>85</tt></td>
                            <td>
                                Metastasis:
                                <ul>
                                    <li>Bone, other than adjacent rib</li>
                                    <li>Lung</li>
                                    <li>Breast, contralateral - if stated as metastatic</li>
                                    <li>Adrenal gland</li>
                                    <li>Ovary</li>
                                    <li>Satellite nodule(s) in skin other than primary breast</li>
                                </ul>
                            </td>
                        </tr>
                        <tr>
                            <td><tt>99</tt></td>
                            <td>
                                UNKNOWN if extension or metastasis
                            </td>
                        </tr>
                    </tbody>
                </table>
            </div>

        </div>

        <div class="modal-footer">
            <button type="button" class="btn btn-default" data-dismiss="modal">Close</button>
        </div>
    </div>
  </div>
</div>
```

Source Code 39:
bosom/WEB-INF/jsp/calc/results.jsp

```
<%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/header.jsp"%>
<%@ include file="/WEB-INF/jsp/includes/page-header.jsp"%>

<div class="row">

    <div class="col-xs-12 col-sm-2 col-md-3 bosom-scrollspy-content">

        <ul class="nav nav-tabs nav-stacked scrollspy-nav" id="about-nav">
            <li>
                <a href="#entered-data" class="list-group-item stack-first"
                    aria-label="Results: Entered data"
                    data-tooltip="Results: Entered data"
                    title="Results: Entered data">
                    Entered data
                </a>
            </li>

            <li>
                <a href="#predicted-survival" class="list-group-item"
                    aria-label="Results: Table for predicted survival"
                    data-tooltip="Results: Table for predicted survival"
                    title="Results: Table for predicted survival">
                    Table for predicted survival
                </a>
            </li>

            <li>
                <a href="#graph-for-predicted-survival" class="list-group-item"
                    aria-label="Results: Graph for predicted survival"
                    data-tooltip="Results: Graph for predicted survival"
                    title="Results: Graph for predicted survival">
                    Graph for predicted survival
                </a>
            </li>

            <li>
                <a href="#predictive-modeling" class="list-group-item"
                  aria-label="About: Predictive modeling"
                  data-tooltip="About: Predictive modeling"
                  title="About: Predictive modeling">
                  Predictive modeling
                </a>
            </li>

            <li>
                <a href="#export-as-pdf" class="list-group-item stack-last"
                    aria-label="Results: Export as PDF"
                    data-tooltip="Results: Export as PDF"
                    title="Results: Export as PDF">
                    Export results as PDF
                </a>
            </li>
        </ul>
    </div>

    <div class="col-xs-12 col-sm-10 col-md-9 bosom-scrollspy-content">

        <div class="bosom-section" id="entered-data">
            <h2>Entered data</h2>
            <p>Here are the breast cancer values you provided in the calculator.</p>

            <div class="table-responsive">
                <table class="table table-hover table-striped table-bordered cell-vertical-middle">
                    <thead>
                        <tr>
                            <th>&#35;</th>
                            <th>Variable</th>
                            <th>Value provided</th>
                        </tr>
                    </thead>

                    <tbody>
                        <tr>
                            <td>1</td>
                            <td>Age of patient at diagnosis</td>
                            <td>
                                <c:out value=" ${wekaData.ageDiagNum} " />
                            </td>
                        </tr>

                        <tr>
                            <td>2</td>
                            <td>Race of patient</td>
                            <td>
                                <c:out value=" ${wekaData.raceGroup} " />
                            </td>
                        </tr>

                        <tr>
                            <td>3</td>
                            <td>Cancer stage (AJCC 6th Edition)</td>
                            <td>
                                <c:out value=" ${wekaData.stage3} " />
                            </td>
                        </tr>

                        <tr>
                            <td>4</td>
                            <td>Presence of distant metastasis (M of TNM staging 6th edition)</td>
                            <td>
                                <c:out value=" ${wekaData.m3} " />
                            </td>
                        </tr>

                        <tr>
                            <td>5</td>
                            <td>Reason for no cancer surgery</td>
                            <td>
                                <c:out value=" ${wekaData.reasonNoCancerSurg} " />
                            </td>
                        </tr>

                        <tr>
                            <td>6</td>
                            <td>Extension</td>
                            <td>
                                <c:out value="${wekaData.ext2} " />
                            </td>
                        </tr>
                    </tbody>
                </table>
            </div>
        </div>

        <div class="bosom-section" id="predicted-survival">
```

```
                <h2>Table for predicted survival</h2>
                <p>
                    Here are the predicted survivals as determined by our models
                    based from past breast cancer patient records.
                    These are from two to ten years, with two years of interval for uniformity.
                </p>

                <div class="table-responsive">
                    <table class="table table-hover table-striped table-bordered cell-vertical-middle">
                        <thead>
                            <tr>
                                <th>Time period</th>
                                <th>Prediction<br>model</th>
                                <th>Predicted<br>survival (%)</th>
                                <th>Mean of predicted <br>survivals (%)</th>
                            </tr>
                        </thead>

                        <tbody>
                          <c:forEach items= "${predictionsMap}" var= "predictionsMap" varStatus= "predictionsLoop">

                              <c:forEach items= "${predictionsMap.value}" var="model" varStatus="modelLoop">
                              <tr>
                                  <c:choose>
                                      <c:when test="${modelLoop.count < 2}">
                                          <td rowspan="5">
                                              ${predictionsLoop.count * 2} years</td>
                                      </c:when>
                                  </c:choose>

                                      <td
                                          style="text-transform:uppercase">
                                          ${model.key} </td>
                                      <td>
                                         <fmt:formatNumber type="number"
                                            maxFractionDigits="2"
                                                minFractionDigits="2"
                                            value="${model.value.Percentage * 100}" />
                                      </td>

                                  <c:choose>
                                      <c:when test="${modelLoop.count < 2}">
                                          <c:choose>
                                              <c:when test= "${predictionsLoop.count == 1}">
                                                  <td rowspan="5"
                                                      id="results-time-2">
                                                      ${wekaData.time2}
                                                  </td>
                                              </c:when>
                                              <c:when test= "${predictionsLoop.count == 2}">
                                                  <td rowspan="5"
                                                      id="results-time-4">
                                                      ${wekaData.time4}
                                                  </td>
                                              </c:when>
                                              <c:when test= "${predictionsLoop.count == 3}">
                                                  <td rowspan="5"
                                                      id="results-time-6">
                                                      ${wekaData.time6}
                                                  </td>
                                              </c:when>
                                              <c:when test= "${predictionsLoop.count == 4}">
                                                  <td rowspan="5"
                                                      id="results-time-8">
                                                      ${wekaData.time8}
                                                  </td>
                                              </c:when>
                                              <c:when test= "${predictionsLoop
                                                  .count == 5}">
                                                  <td rowspan="5"
                                                      id="results-time-10">
                                                      ${wekaData.time10}
                                                  </td>
                                              </c:when>
                                          </c:choose>
                                      </c:when>
                                  </c:choose>
                              </tr>
                          </c:forEach>

                      </c:forEach>

                    </tbody>
                </table>
            </div>
        </div>

        <div class="bosom-section" id="graph-for-predicted-survival">
            <h2>Graph for predicted survival</h2>

            <div id="results-graph-graph" style="min-height: 300px"></div>
            <div id="results-graph-legend"></div>
        </div>

        <div class="bosom-section" id="export-as-pdf">
            <h2>Export report as PDF file </h2>

            <p>
                Clicking the button below might do any of the following, based on your
                browser, its version and your device:
            </p>

            <ul>
                <li>
                    open a new browser tab that will show the PDF file that you can choose
                    to save or print right away;
                </li>
                <li>
                    it will be automatically saved; or
                </li>
                <li>
                    a <tt>Save As</tt> prompt will ask you if you want to save the file in
                    a location in your machine.
                </li>
            </ul>

            <p class="callout callout-warning bg-warning text-warning">
                The generated PDF file is only available for each BOSOM form submission.
                Please download and save it in your device or take note of the results.
                It will not be available after you leave the Results page.
                You can always try again by answering the
                <a href="<c:url value="/calc"/>">Calculator</a> again.
            </p>

            <div class="form-group">
                <a href="<c:out value="${pdfLocation}"/>" class="btn btn-info btn-lg btn-block">
                    <span class="glyphicon glyphicon-save"></span>
                    View report in PDF
                </a>
            </div>

            <p>
                You can keep the file as a reference for further analysis and interpretation
                by a licensed oncologist or breast cancer specialist to help you understand
                and assess the results better.
            </p>

        </div>

    <div class="clearfix"></div>

    </div> <!-- contents -->

<div class="clearfix"></div>

<c:set var="isFlotUsed" value="${isFlotUsed}" />
<c:choose>
    <c:when test="${isFlotUsed == true}">

    <script type="text/javascript">
    $(document).ready(function() {
      // flot.js
       var data = [
           { data: [["2", $('#results-time-2').html()]], color: "#009E73" },
           { data: [["4", $('#results-time-4').html()]], color: "#F0E442" },
           { data: [["6", $('#results-time-6').html()]], color: "#0072B2" },
           { data: [["8", $('#results-time-8').html()]], color: "#D55E00" },
           { data: [["10", $('#results-time-10').html()]], color: "#CC79A7" }
       ];
```

```
279
280         var flotContainer = $('#results-graph-graph');
281         $.plot(flotContainer, data, {
282             xaxes: [
283               { position: 'bottom', axisLabel: 'Time period (years)'
                      }
284             ],
285             yaxes: [
286               { position: 'left', axisLabel: 'Predicted survival
                      (%)',
287                 axisLabelPadding: 10 }
288             ],
289             grid: { hoverable: true, show: true },
290             series: {
291                 bars: { show: true, barWidth: 0.75, align: "center" }
292             },
293             xaxis: { mode: "categories" }
294         });
295     });
296     </script>
297
298     </c:when>
299 </c:choose>
300
301 <%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

Source Code 40:
bosom/WEB-INF/jsp/supplements.jsp

```
1  <%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>
2  <%@ include file="/WEB-INF/jsp/includes/header.jsp"%>
3  <%@ include file="/WEB-INF/jsp/includes/page-header.jsp"%>
4
5  <h2>Local hospitals and NGO's</h2>
6
7  <h3>Hospitals</h3>
8
9  <ul>
10     <li>
11         Cancer Institute, <a href="http:// www.
                pgh.gov.ph/en/">University of the Philippines -
                Philippine General Hospital</a>
12         <ul>
13             <li>
14                 <a href="https:// www. facebook .com/pages/
                        Philippine-General-Hospital/
                        104067952962987">UP-PGH Facebook page</a>
15             </li>
16         </ul>
17     </li>
18
19     <li>
20         <a href="http:// www. stlukesmedicalcenter
                .com.ph/aboutus/institutes/cancer-institute">Cancer
                Institute</a>, St. Luke's Medical Center
21         <ul>
22             <li>
23                 <a href="https:// www. facebook
                        .com/StLukesMedicalCenterOfficial">Facebook
                        page</a>
24             </li>
25         </ul>
26     </li>
27
28     <li>
29         <a href="http:// www. themedicalcity .com/services/
                centers_of_excellence/ cancer-center/
                contact-details">Cancer Center</a>, The Medical City
30         <ul>
31             <li>
32                 <a href="https:// www. facebook
                        .com/MetroMedicalCenter">Facebook page</a>,
                        as &amp;quot;Metropolitan Medical Center&amp;quot;
33             </li>
34         </ul>
35     </li>
36
37     <li>
38         <a href="http:// www. usthospital
                .com.ph/bci/bci.php">Benavides Cancer Institute</a>,
                University of Sto. Tomas Hopsital
39     </li>
40 </ul>
41
42 <h3>Non-governmental organizations</h3>
43 <ul>
44     <li>
45         Philippine Cancer Society Inc.
46         <ul>
47             <li>
48                 <a href="http:// philcancer.org.ph/">official
                        website</a>
49             </li>
50             <li>
51                 <a href="https:// www. facebook .com/pages/
                        Philippine-Cancer-Society-INC/
                        216156488399561">Facebook page</a>
52             </li>
53         </ul>
54     </li>
55
56     <li>
57         Philippine Breast Cancer Network
58         <ul>
59             <li>
60                 <a href="http:// www. pbcn.org/">official website</a>
61             </li>
62             <li>
63                 <a href="http:// pbcn.blogspot .com/">official
                        blog</a>
64             </li>
65             <li>
66                 <a href="https:// www. facebook
                        .com/groups/philippinebreastcancernetwork/">Facebook
                        (support) group</a>
67             </li>
68         </ul>
69
70     </li>
71
72     <li>
73         Philippine Foundation for Breast Cancer Inc.
74         <ul>
75             <li>
76                 <a href="http:// kasuso.org/">official website</a>
77             </li>
78             <li>
79                 <a href="https:// www. facebook
                        .com/kasusongpinay">Facebook page</a>
80             </li>
81         </ul>
82     </li>
83
84     <li>
85         Cancer Treatment and Support Foundation Inc.
86         <ul>
87             <li>
88                 <a href="http://
                        the-cancer-foundation.org/">official
                        website</a>
89             </li>
90             <li>
91                 <a href="https:// www. facebook .com/
                        CancerTreatmentAndSupportFoundationInc">Facebook
                        page</a>
92             </li>
93         </ul>
94
95     </li>
96
97     <li>
98         ICanServe Foundation Inc.
99         <ul>
100            <li>
101                <a href="http:// www.
                       icanservefoundation.org/">official website</a>
102            </li>
103            <li>
104                <a href="http:// www. icanserve-foundation.blogspot
                       .com/">blog</a>
105            </li>
106            <li>
107                <a href="https:// www. facebook .com/pages/
                       ICanServe-Foundation-Inc/
                       183543015002337">Facebook page</a>
108            </li>
109            <li>
110                <a href="https:// twitter .com/icanserve">Twitter
                       account</a>
111
112            </li>
113        </ul>
114    </li>
115 </ul>
116
117 <h3>Other helpful resources</h3>
118 <ul>
119    <li>
120        <a href="http:// beatingcancers.rxpinoy
               .com/index.php">Beating Cancers</a> by RxPinoy
121        <ul>
122            <li>
123                &amp;quot;<a href="http:// beatingcancers.rxpinoy
                       .com/groups_local.php">Local Cancer Support
                       Groups</a>&amp;quot;
124            </li>
125        </ul>
126    </li>
127 </ul>
128
129 <h2>International programs</h2>
130 <p>
131     The following websites/organizations are listed in the spirit
            of providing additional information and resources for
            anyone interested in learning and understanding cancer
            and breast cancer.
132 </p>
```

```
<p>
    These mostly provide general information pages containing
           symptoms, prevention and statistics while some have
           options to for direct contact through e-mail, calls and
           other means available.
</p>

<ul>
    <li>
        <a href="http:// www. cancer.gov/">National Cancer
               Institute</a>, National Institutes of Health,
               Department of Health and Human Services, USA
        <ul>
            <li>
                <a href="http:// www.
                       cancer.gov/cancertopics/types/breast"> Breast
                       cancer general organization </a>
            </li>
        </ul>
    </li>

    <li>
        <a href="http:// seer.cancer.gov/"> Surveillance,
               Epidemiology and End Results program </a>, National
               Cancer Institute, National Institutes of Health,
               Department of Health and Human Services, USA
        <ul>
            <li>
                <a href="http://
                       seer.cancer.gov/statfacts/html/breast.html">
                       Breast cancer general information </a>
            </li>
        </ul>
    </li>

    <li>
        <a href="http:// www.
               breastcancer.org/">breastcancer.org</a>,
               Pennsylvania, USA
    </li>
    <li>
        <a href="http:// www. nationalbreastcancer.org/">National
               Breast Cancer Foundation, Inc.</a>, Frisco, Texas,
               USA
    </li>
    <li>
        <a href="http:// thebreastcancersite.greatergood
               .com/clickToGive/bcs/home"> The Breast Cancer Site
               </a>, USA
    </li>
    <li>
        <a href="http:// www. breastcancercare.org.uk/">Breast
               Cancer Care</a>, United Kingdom
    </li>
</ul>

<%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

Source Code 41: bosom/WEB-INF/jsp/error/404.jsp

```
<%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>

<c:set var="title" value="This page does not exist" />
<c:set var="pageName" value="error" />

<%@ include file="/WEB-INF/jsp/includes/header.jsp"%>

<div class="jumbotron jumbotron-error" id="page-404">
    <div class="overlay">
        <h1>This page does not exist.</h1>
        <p class="lead">
            The page you are trying to visit is not part of this
                   website.
            We have the <a href="#header-nav">navigation menu</a> at
                   the top
            and a <a href="#site-map">site map</a> in the bottom
            to help you go around this website.
        </p>
        <p>
            Please click <a href="/bosom">this link</a> to go back
                   to the home page.
        </p>
    </div>
</div>

<div class="row hide">

<%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

Source Code 42:
bosom/WEB-INF/jsp/error/exception.jsp

```
<%@ include file="/WEB-INF/jsp/includes/taglibs.jsp"%>

<c:set var="title" value="Something wrong happened" />

<%@ include file="/WEB-INF/jsp/includes/header.jsp"%>

<div class="jumbotron jumbotron-error" id="page-exception">
    <div class="overlay">
        <h1>Something wrong happened.</h1>
        <p class="lead">
            We are genuinely sorry for this.
            Don't worry, it's on us. We'll fix it as soon as we can.
        </p>
        <p>
            Please e-mail the developer at
                   <tt>gpmeren+bosom@up.edu.ph</tt> to report this
                   occurrence.
            Kindly state what you were doing i.e., answering the
                   form so we can find ans solve
            the problem in less time.
        </p>
        <p>
            Please click <a href="/">this link</a> to go back to the
                   home page.
        </p>
    </div>
</div>

<div class="row hide">

<%@ include file="/WEB-INF/jsp/includes/footer.jsp"%>
```

C. Tables

Table 12: BOSOM application's JAR file dependencies

Name	Version
Apache Commons Logging	1.1.3
Hibernate Validator	4.2.0
iTextPDF	5.5.0
iTextPDF (Javadocs)	5.5.0
iTextPDF (sources)	5.5.0
JCommon	1.0.17
JFreeChart	1.0.17
Joda-Time	2.2
JSTL (JavaServer Pages Tag Library)	1.2
JSTL (JavaServer Pages Tag Library) (sources)	1.2
Simple Logging Façade for Java (SLF4J)	1.7.6
Spring Beans	3.2.5
Spring Beans (Javadocs)	3.2.5
Spring Beans (sources)	3.2.5
Spring Context	3.2.5
Spring Context (Javadocs)	3.2.5
Spring Context (sources)	3.2.5
Spring Core	3.2.5
Spring Core (Javadocs)	3.2.5
Spring Core (sources)	3.2.5
Spring Expression	3.2.5
Spring Expression (Javadocs)	3.2.5
Spring Expression (sources)	3.2.5
Spring Web	3.2.5
Spring Web (Javadocs)	3.2.5
Spring Web (sources)	3.2.5
Spring Web	3.2.5
Spring Web (Javadocs)	3.2.5
Spring Web (sources)	3.2.5
Bean Validation API	1.0.0

Table 13: Filter commands for extracting breast cancer data from SEER*Stat

Category	Variable	Condition	Filtering values
Age at Diagnosis	Age recode with <1 year olds	=	'15-19 years', '20-24 years', '25-29 years', '30-34 years', '35-39 years', '40-44 years', '45-49 years', '50-54 years', '55-59 years', '60-64 years', '65-69 years', '70-74 years', '75-79 years', '80-84 years', '85+ years', 'Unknown'
Race, Sex, Year Dx, Registry, County	Year of diagnosis	=	'1998', '1999', '2000', '2001', '2002', '2003'
Site and Morphology	Site recode ICD-O-3/WHO 2008	=	' Breast'
Site and Morphology	Behavior recode for analysis	=	'Benign', 'Borderline malignancy', 'In situ', 'Malignant'
Site and Morphology	Primary Site - labeled	=	'C50.0-Nipple', 'C50.1-Central portion of breast', 'C50.2-Upper-inner quadrant of breast', 'C50.3-Lower-inner quadrant of breast', 'C50.4-Upper-outer quadrant of breast', 'C50.5-Lower-outer quadrant of breast', 'C50.6-Axillary tail of breast', 'C50.8-Overlapping lesion of breast', 'C50.9-Breast, NOS'
Site and Morphology	Grade	=	'Well differentiated; Grade I', 'Moderately differentiated; Grade II', 'Poorly differentiated; Grade III', 'Undifferentiated; anaplastic; Grade IV'

Continued on next page

Table 13 – continued from previous page

Category	Variable	Condition	Filtering values
Site and Morphology	Diagnostic Confirmation	=	' Positive histology', ' Positive exfoliative cytology, no positive histology', ' Pos hist AND immunophenotyping AND/OR pos genetic studies', ' Positive microscopic confirm, method not specified', 'Positive laboratory test/marker study', 'Direct visualization without microscopic confirmation', 'Radiography without microscopic confirm', 'Clinical diagnosis only', 'Unknown'
Stage - LRD (Summary and Historic)	Summary stage 2000 (1998+)	=	'In situ', 'Localized', 'Regional', 'Distant', 'Unknown/unstaged'
Therapy	Radiation	=	'None', 'Beam radiation', 'Radioactive implants', 'Radioisotopes', 'Combination of beam with implants or isotopes', 'Radiation, NOS method or source not specified', 'Refused', 'Recommended, unknown if administered', 'Unknown'
Extent of Disease - CS	ER Status Recode Breast Cancer (1990+)	=	'Positive', 'Negative', 'Borderline', 'Unknown'
Extent of Disease - CS	PR Status Recode Breast Cancer (1990+)	=	'Positive', 'Negative', 'Borderline', 'Unknown'
Cause of Death (COD) and Follow-up	COD to site recode	=	'Alive', ' Breast'
Cause of Death (COD) and Follow-up	Survival months	=	0-455
Dates	Year of birth	!=	'Blank(s)'

Continued on next page

Table 13 – continued from previous page

Category	Variable	Condition	Filtering values
Stage - TNM	Adjusted AJCC 6th T (1988+)	!=	'NA', 'Blank(s)'
Stage - TNM	Adjusted AJCC 6th N (1988+)	!=	'NA', 'Blank(s)'
Stage - TNM	Adjusted AJCC 6th M (1988+)	!=	'NA', 'Blank(s)'
Extent of Disease - Historic	EOD 10 - extent (1988-2003)	!=	'Blank(s)'
Extent of Disease - Historic	EOD 10 - nodes (1988-2003)	!=	'Blank(s)'
Extent of Disease - Historic	EOD 10 - size (1988-2003)	!=	'Blank(s)'
Site and Morphology	Histologic Type ICD-O-3	=	8000-8005, 8010-8015, 8020-8022, 8030-8035, 8041, 8043, 8050-8052, 8070-8076, 8078, 8140-8141, 8143, 8147, 8190, 8200-8201, 8211, 8230-8231, 8251, 8255, 8260-8261, 8310, 8314-8315, 8320, 8323, 8401, 8440, 8480-8481, 8490, 8500-8504, 8507-8508, 8510, 8512-8514, 8520-8525, 8530, 8540-8541, 8543, 8550-8551, 8560, 8562, 8570-8575, 8800-8806, 8810-8811, 8813-8815, 8850-8855, 8857-8858, 8890-8891, 8894-8896, 8935, 8980-8982, 8990-8991, 9020, 9120, 9130, 9133, 9580-9581, 9590-9591, 9596, 9650-9655, 9659, 9661-9665, 9667, 9670-9671, 9673, 9675,

Continued on next page

Table 13 – continued from previous page

Category	Variable	Condition	Filtering values
			9680, 9684, 9687-9688, 9690-9691, 9695, 9698-9699, 9701-9702, 9705, 9712, 9714, 9719, 9724, 9727-9729, 9731, 9734, 9740-9741, 9750-9751, 9754-9759, 9811-9818, 9823, 9831, 9837, 9965, 9967, 9971, 9975
Site and Morphology	ICD-O-3 Hist/behav	=	See Appendix D. in page 170 for the entire filter command.

Table 14: Parameters of the five WEKA classifiers used to train the SEER breast cancer datasets

Classifier	Parameters			
	Code	**Name**	**Description**	**Default**
Alternating decision tree	-B	Number of boosting iterations	Sets the number of boosting iterations to perform. Each iteration will add 3 nodes (1 split + 2 prediction) to the tree unless merging occurs.	10
	-E	Search path	Sets the type of search to perform when building the tree. The default option is -3 (expand all paths); -2 (expand the heaviest path); -1 (expand the best z-pure path); and >=0 (expand a random path).	-3
	-D	Sets whether the tree is to save instance data - the model will take up more memory if it does.		False
J48 decision tree	-U	Use unpruned tree	Whether pruning is performed.	False
	-O	Do not collapse tree	Whether parts are removed that do not reduce training error.	True
	-C	The confidence factor used for pruning (smaller values incur more pruning).		0.25
	-M	The minimum number of instances per leaf.		2
	-R	Use reduced error pruning	Whether reduced-error pruning is used instead of C.4.5 pruning.	False
	-N	Number of folds	Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.	10
	-B	Use binary splits only	Whether to use binary splits on nominal attributes when building the trees.	False

Continued on next page

Table 14 – continued from previous page

Classifier	Parameters			
	Code	**Name**	**Description**	**Default**
	`-S`	Do not perform subtree raising	Whether to consider the subtree raising operation when pruning.	`True`
	`-L`	Do not clean up after the tree has been built	No information available	`False`
	`-A`	Laplace smoothing for predicted probabilities	Whether counts at leaves are smoothed based on Laplace.	`False`
	`-J`	Do not use MDL correction for info gain on numeric attributes	Whether MDL correction is used when finding splits on numeric attributes.	`True`
	`-Q`	The seed used for randomizing the data when reduced-error pruning is used.		`1`
Random Forest	`-I`	The number of trees to be generated.		`10`
	`-K`	The number of attributes to be used in random selection (see RandomTree).		`0`
	`-S`	The random number seed to be used.		`1`
	`-depth`	The maximum depth of the trees, 0 for unlimited.		`0`
	`-num-slots`	The number of execution slots (threads) to use for constructing the ensemble.		`1`
	`-D`	Debug mode	If set to true, classifier may output additional info to the console.	`False`

Continued on next page

Table 14 – continued from previous page

Classifier	Parameters			
	Code	Name	Description	Default
LogitBoost	-Q	Resampling over reweighting for boosting	Whether resampling is used instead of reweighting.	False
	-P	Percent of weight mass	Weight threshold for weight pruning (reduce to 90 for speeding up learning process).	100
	-F	Number of folds for internal cross-validation (default 0 means no cross-validation is performed).		0
	-R	Number of runs for internal cross-validation		1
	-L	Threshold on the improvement of the likelihood		-Double.MAX_VALUE
	-H	Shrinkage parameter (use small value like 0.1 to reduce overfitting).		1
	-S	The random number seed to be used.		1
	-I	The number of iterations to be performed.		10
	-D	Debug mode	If set to true, classifier may output additional info to the console.	False
	-W	The base classifier to be used.		weka.classifiers.trees.DecisionStump
Random SubSpace	-P	Size of each subSpace: if less than 1 as a percentage of the number of attributes, otherwise the absolute number of attributes.		0.5
	-S	The random number seed to be used.		1
	-I	The number of iterations to be performed.		10
	-D	If set to true, classifier may output additional info to the console.		False
	-W	The base classifier to be used.		weka.classifiers.trees.REPTree
	Exclusive to underlying classifier weka.classifiers.trees.REPTree			

Continued on next page

Table 14 – continued from previous page

Classifier	Parameters			
	Code	**Name**	**Description**	**Default**
	-M		The minimum total weight of the instances in a leaf.	2
	-V		The minimum proportion of the variance on all the data that needs to be present at a node in order for splitting to be performed in regression trees.	0.001
	-N	Number of folds for reduced error pruning (default 3).	Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules.	3
	-S		The seed used for randomizing the data.	1
	-P	Pruning	Whether pruning is performed.	False
	-L		Maximum tree depth (default -1, no maximum)	-1

Table 15: Modification of SEER variable "Sequence number"

Number of malignant tumors		Basis of diagnosis	
Code	**Recode**	**Code**	**Recode**
One primary only	0	Diagnosis year	1
1st of 2 or more primaries	1	Diagnosis year	1
2nd of 2 or more primaries	2	Diagnosis year	1
3rd of 3 or more primaries	3	Diagnosis year	1
4th of 4 or more primaries	4	Diagnosis year	1
5th of 5 or more primaries	5	Diagnosis year	1
6th of 6 or more primaries	6	Diagnosis year	1
7th of 7 or more primaries	7	Diagnosis year	1
8th of 8 or more primaries	8	Diagnosis year	1
9th of 9 or more primaries	9	Diagnosis year	1
10th of 10 or more primaries	10	Diagnosis year	1
11th of 11 or more primaries	11	Diagnosis year	1
12th of 12 or more primaries	12	Diagnosis year	1
13th of 13 or more primaries	13	Diagnosis year	1
14th of 14 or more primaries	14	Diagnosis year	1
15th of 15 or more primaries	15	Diagnosis year	1
16th of 16 or more primaries	16	Diagnosis year	1
17th of 17 or more primaries	17	Diagnosis year	1
18th of 18 or more primaries	18	Diagnosis year	1
19th of 19 or more primaries	19	Diagnosis year	1
20th of 20 or more primaries	20	Diagnosis year	1
21st of 21 or more primaries	21	Diagnosis year	1
22nd of 22 or more primaries	22	Diagnosis year	1
23rd of 23 or more primaries	23	Diagnosis year	1
24th of 24 or more primaries	24	Diagnosis year	1
25th of 25 or more primaries	25	Diagnosis year	1
26th of 26 or more primaries	26	Diagnosis year	1
27th of 27 or more primaries	27	Diagnosis year	1
28th of 28 or more primaries	28	Diagnosis year	1
29th of 29 or more primaries	29	Diagnosis year	1
30th of 30 or more primaries	30	Diagnosis year	1
31st of 31 or more primaries	31	Diagnosis year	1
32nd of 32 or more primaries	32	Diagnosis year	1
33rd of 33 or more primaries	33	Diagnosis year	1
34th of 34 or more primaries	34	Diagnosis year	1
35th of 35 or more primaries	35	Diagnosis year	1
36th of 36 or more primaries	36	Diagnosis year	1
37th of 37 or more primaries	37	Diagnosis year	1
38th of 38 or more primaries	38	Diagnosis year	1

Continued on next page

Table 15 – continued from previous page

Number of malignant tumors		Basis of diagnosis	
Code	**Recode**	**Code**	**Recode**
39th of 39 or more primaries	39	Diagnosis year	1
40th of 40 or more primaries	40	Diagnosis year	1
41st of 41 or more primaries	41	Diagnosis year	1
42nd of 42 or more primaries	42	Diagnosis year	1
43rd of 43 or more primaries	43	Diagnosis year	1
Unknown seq num - federally required in situ or malig tumors	99	Diagnosis year	1

Number of benign tumors		Basis of diagnosis	
Code	**Recode**	**Code**	**Recode**
Only one state registry-defined neoplasm	0	State/province defined	2
1st of 2 or more state registry-defined neoplasms	1	State/province defined	2
2nd of 2 or more state registry-defined neoplasms	2	State/province defined	2
3rd of 3 or more state registry-defined neoplasms	3	State/province defined	2
4th of 4 or more state registry-defined neoplasms	4	State/province defined	2
5th of 5 or more state registry-defined neoplasms	5	State/province defined	2
6th of 6 or more state registry-defined neoplasms	6	State/province defined	2
8th of 8 or more state registry-defined neoplasms	8	State/province defined	2
15th of 15 or more state registry-defined neoplasms	15	State/province defined	2
Unknown seq num - state registry-defined neoplasms	99	State/province defined	2

Table 16: Breast cancer-related variables selected from the SEER*Stat database

Category	SEER Name	Recode Name	Type
Age at Diagnosis	Age recode with < 1 year olds	ageDiagNom	nominal
Race and Age (case data only)	Age at diagnosis	ageDiagNum	numeric
Race and Age (case data only)	Age recode with single ages and 85+	ageNom	nominal
Race and Age (case data only)	Age recode with single ages and 85+	ageNum	numeric
Multiple Primary Fields	Sequence number	basisDiag	nominal
Site and Morphology	Behavior recode for analysis	behav1	nominal
Site and Morphology	Behavior code ICD-O-3 (1973+)	behav2	nominal
Site and Morphology	Diagnostic Confirmation	diagConf	nominal
Extent of Disease - CS	ER Status Recode Breast Cancer (1990+)	er	nominal
Extent of Disease - CS	CS extension (2004+)	ext1	nominal
Extend of Disease - Historic	EOD 10 - extent (1988-2003)	ext2	nominal
Race, Sex, Year Dx, Registry, County	Sex	female	nominal
Multiple Primary Fields	First malignant primary indicator	firstMalPrimInd	nominal
Site and Morphology	Grade	grade1	nominal
Site and Morphology	ICD-O-3 Hist/behav	histBehav1	nominal
Site and Morphology	ICD-O-3 Hist/behav, malignant	histBehav2	nominal
Site and Morphology	Histology recode - broad groupings	histGroup	nominal
Site and Morphology	Histologic Type ICD-O-3	histInd	nominal
Extent of Disease - CS	Laterality (1973+)	laterality	nominal
Extent of Disease - CS	CS lymph nodes (2004+)	ln1	nominal
Extend of Disease - Historic	EOD 10 - nodes (1988-2003)	ln2	nominal
Stage TNM	Derived AJCC M, 7th ed (2010+)	m1	nominal
Stage TNM	Derived AJCC M, 6th ed (2004+)	m2	nominal

Continued on next page

Table 16 – continued from previous page

Category	SEER Name	Recode Name	Type
Stage TNM	Adjusted AJCC 6th M (1988+)	m3	nominal
Other	Marital status at diagnosis	marital	nominal
Dates	Month of diagnosis recode	monDiag	nominal
Stage TNM	Derived AJCC N, 7th ed (2010+)	n1	nominal
Stage TNM	Derived AJCC N, 6th ed (2004+)	n2	nominal
Stage TNM	Adjusted AJCC 6th N (1988+)	n3	nominal
Multiple Primary Fields	Sequence number	numBenTum	nominal
Multiple Primary Fields	Sequence number	numMalTum	nominal
Multiple Primary Fields	Number of primaries	numPrim	numeric
Geographic Locations	Place of birth	placeBirthGroup	nominal
Geographic Locations	Place of birth	placeBirthInd	nominal
Extent of Disease - CS	PR Status Recode Breast Cancer (1990+)	pr	nominal
Site and Morphology	Primary Site	primSite	nominal
Race, Sex, Year Dx, Registry, County	Race recode (White, Black, Other)	raceGroup	nominal
Race and Age (case data only)	Race/ethnicity	raceInd	nominal
Therapy	Radiation	rad	nominal
Therapy	Radiation sequence with surgery	radSeqSurg	nominal
Therapy	Reason no cancer-directed surgery	reasonNoCancerSurg	nominal
Extent of Disease - CS	Regional nodes examined (1988+)	regNodeExamNom	nominal
Extent of Disease - CS	Regional nodes examined (1988+)	regNodeExamNum	numeric
Extent of Disease - CS	Regional nodes positive (1988+)	regNodePosNom	nominal
Extent of Disease - CS	Regional nodes positive (1988+)	regNodePosNum	numeric
Therapy	RX Summ--Scope Reg LN Sur (2003+)	scopeRegLnSurg1	nominal
Therapy	Scope of reg lymph nd surg (1998-2002)	scopeRegLnSurg2	nominal

Continued on next page

Table 16 – continued from previous page

Category	SEER Name	Recode Name	Type
Stage AJCC	Derived AJCC Stage Group, 7th ed (2010+)	stage1	nominal
Stage AJCC	Derived AJCC Stage Group, 6th ed (2004+)	stage2	nominal
Stage AJCC	Adjusted AJCC 6th (1988+)	stage3	nominal
Stage AJCC	SEER modified AJCC stage 3rd (1988-2003)	stage4	nominal
Stage - LRD (Summary and Historic)	Summary Stage 2000 (1998+)	sumStage	nominal
Therapy	RX Summ--Surg Oth Reg/Dis (2003+)	surgOthRegDis1	nominal
Therapy	Surgery of oth reg/dis sites (1998-2002)	surgOthRegDis2	nominal
Therapy	RX Summ--Surg Prim Site (1998+)	surgPrimSite1	nominal
Therapy	Surgery of primary site (1998-2002)	surgPrimSite2	nominal
Stage TNM	Derived AJCC T, 7th ed (2010+)	t1	nominal
Stage TNM	Derived AJCC T, 6th ed (2004+)	t2	nominal
Stage TNM	Adjusted AJCC 6th T (1988+)	t3	nominal
Extent of Disease - CS	CS tumor size (2004+)	tumSizeNom1	nominal
Extent of Disease - CS	CS tumor size (2004+)	tumSizeNum1	numeric
Cause of Death (COD) and Follow-up	Survival months	time	numeric
Cause of Death (COD) and Follow-up	Survival months	timeNot	nominal
Cause of Death (COD) and Follow-up	Survival months	time2	nominal
Cause of Death (COD) and Follow-up	Survival months	time4	nominal
Cause of Death (COD) and Follow-up	Survival months	time6	nominal
Cause of Death (COD) and Follow-up	Survival months	time8	nominal
Cause of Death (COD) and Follow-up	Survival months	time10	nominal
Extend of Disease - Historic	EOD 10 - size (1988-2003)	tumSizeNom2	nominal
Extend of Disease - Historic	EOD 10 - size (1988-2003)	tumSizeNum2	numeric
Cause of Death (COD) and Follow-up	Vital status recode (study cutoff used)	vsr	nominal

Continued on next page

Table 16 – continued from previous page

Category	SEER Name	Recode Name	Type
Dates	`Year of birth`	`yrBirth`	nominal
Race, Sex, Year Dx, Registry, County	`Year of diagnosis`	`yrDiag`	nominal

Table 17: Selected WEKA result buffer classifier details of models trained with the complete breast cancer dataset

Classifier	Classifier metrics		Outcome variable					Mean
			2	**4**	**6**	**8**	**10**	
ADT	**Tree size (number of nodes)**		31	31	31	31	31	31
	Leaves (number of predictor nodes)		21	21	21	21	21	21
J48	**Number of leaves**		3361	3965	3950	6538	22142	7991.2
	Size of tree		3884	4559	4572	7652	26473	9428
RF	**Out of bag error**		6.61%	6.71%	6.77%	9.25%	10.89%	8.05%
RS	**Size of tree (per CV fold)**	**1**	2300	2319	2839	5035	13204	5139.4
		2	2544	2921	3616	5269	15265	5923
		3	4173	3709	5147	6940	13825	6758.8
		4	2480	2966	3683	4925	10699	4950.6
		5	2660	2626	2805	3898	7837	3965.2
		6	2878	3194	3790	5825	15836	6304.6
		7	4178	4865	5649	7342	13014	7009.6
		8	1593	1616	2021	3006	6574	2962
		9	2845	2947	3272	5469	12337	5374
		10	1881	2475	2838	4689	12977	4972

Table 18: Selected WEKA result buffer classifier details of models trained with the subset breast cancer dataset

Classifier	Classifier metrics		Outcome variable					Mean
			2	**4**	**6**	**8**	**10**	
ADT	**Tree size (number of nodes)**		31	31	31	31	31	31
	Leaves (number of predictor nodes)		21	21	21	21	21	21
J48	**Number of leaves**		630	542	505	733	851	652.2
	Size of tree		795	704	669	984	1249	880.2
RF	**Out of bag error**		7.51%	7.66%	8.11%	11.82%	23.95%	11.81%
RS	**Size of tree (per CV fold)**	**1**	51	51	51	46	59	51.6
		2	110	143	179	212	147	158.2
		3	255	225	249	293	305	265.4
		4	36	40	36	28	36	35.2
		5	346	411	380	337	316	358
		6	124	142	160	162	92	136
		7	291	309	347	455	663	413
		8	166	152	174	166	168	165.2
		9	108	134	152	140	88	124.4
		10	211	267	309	379	317	296.6

Table 19: Time of execution of each classifier and outcome variable (time period) pair for predictive model creation and training on the complete dataset

Outcome variable	Classifier	Start	Documentation			Total
			Folds	Results Buffer	Model	
2	ZR	5:30:00	5:39:36	5:39:40	5:39:40	-
	RF	-	6:02:59	6:02:59	6:02:59	-
	LB	-	6:42:22	6:42:22	6:42:22	-
	RS	-	7:36:02	7:36:04	7:36:04	-
	J48	-	8:19:04	8:19:05	8:19:05	-
	ADT	-	9:37:53	9:37:53	9:37:53	4:07:53
4	ZR	11:06:03	11:14:19	11:14:19	11:14:19	-
	RF	-	11:40:52	11:40:52	11:40:53	-
	LB	-	12:25:17	12:25:18	12:25:18	-
	RS	-	1:29:25	1:29:27	1:29:28	-
	J48	-	2:08:58	2:08:58	2:08:58	-
	ADT	-	3:23:08	3:23:08	3:23:08	4:17:05
6	ZR	3:41:28	3:50:10	3:50:11	3:50:11	-
	RF	-	4:15:43	4:15:43	4:15:43	-
	LB	-	4:55:13	4:55:13	4:55:13	-
	RS	-	5:50:35	5:50:37	5:50:38	-
	J48	-	6:33:06	6:33:07	6:33:08	-
	ADT	-	7:47:45	7:47:46	7:47:46	4:06:18
8	ZR	7:49:35	7:57:47	7:57:48	7:57:48	-
	RF	-	8:23:57	8:23:57	8:23:57	-
	LB	-	9:05:19	9:05:20	9:05:20	-
	RS	-	10:03:55	10:03:57	10:03:57	-
	J48	-	10:54:59	10:54:59	10:54:59	-
	ADT	-	12:11:54	12:11:54	12:11:54	4:22:19
10	ZR	3:07:21	3:15:37	3:15:38	3:15:38	-
	RF	-	3:45:09	3:45:09	3:45:09	-
	LB	-	4:25:34	4:25:34	4:25:34	-
	RS	-	5:31:20	5:31:22	5:31:22	-
	J48	-	6:38:42	6:38:43	6:38:43	-
	ADT	-	7:45:14	7:45:14	7:45:14	4:37:53

Table 20: Time of execution of each classifier and outcome variable (time period) pair for predictive model creation and training on the subset dataset

Outcome variable	Classifier	Start	Documentation			Total
			Folds	Results Buffer	Model	
2	ZR	11:39:00	11:47:25	11:47:32	11:47:35	-
	RF	-	12:10:40	12:10:43	12:10:51	-
	LB	-	12:26:13	12:26:13	12:26:18	-
	RS	-	12:48:09	12:48:11	12:48:14	-
	J48	-	1:06:44	1:06:45	1:06:47	-
	ADT	-	1:30:52	1:30:52	13:30:53	1:51:53
4	ZR	1:33:52	1:42:14	1:42:17	1:42:18	-
	RF	-	2:03:19	2:03:19	2:03:21	-
	LB	-	2:17:22	2:17:23	2:17:23	-
	RS	-	2:36:27	2:36:27	2:36:29	-
	J48	-	2:51:02	2:51:03	2:51:03	-
	ADT	-	3:10:10	3:10:11	3:10:22	1:36:30
6	ZR	3:14:10	3:22:40	3:22:41	3:22:43	-
	RF	-	3:40:16	3:40:17	3:40:19	-
	LB	-	3:54:17	3:54:17	3:54:20	-
	RS	-	4:13:29	4:13:30	4:13:32	-
	J48	-	4:28:25	4:28:25	4:28:26	-
	ADT	-	4:46:45	4:46:46	4:46:48	1:32:38
8	ZR	4:49:34	4:58:08	4:58:10	4:58:14	-
	RF	-	5:18:53	5:18:53	5:18:56	-
	LB	-	5:33:41	5:33:43	5:33:44	-
	RS	-	5:55:24	5:55:26	5:55:28	-
	J48	-	6:13:34	6:13:34	6:13:35	-
	ADT	-	6:35:44	6:35:44	6:35:45	1:46:11
10	ZR	8:24:55	8:34:50	8:34:56	8:35:07	-
	RF	-	8:53:15	8:53:16	8:53:18	-
	LB	-	9:07:59	9:08:01	9:08:03	-
	RS	-	9:30:11	9:30:26	9:30:34	-
	J48	-	9:50:14	9:50:15	9:50:18	-
	ADT	-	10:10:36	10:10:37	10:10:45	1:45:50

Table 21: Attribute selection variables and their respective values as seen in the BOSOM Calculator

SEER Name	Form Name	Form Values
Age at diagnosis	Age of patient in years at time of diagnosis (1 - 150 only)	numeric; minimum=1, maximum=150
Race recode (White, Black, Other)	Race of patient	Black, White, Otherwise, Unknown
Adjusted AJCC 6th (1988+)	Stage of cancer (AJCC 6th Edition)	0, I, IIA, IIB, IIIA, IIIB, IIIC, IIINOS, IV, Unknown stage
Adjusted AJCC 6th M (1988+)	Spread of metastasis	M0 (No distant metastasis)
		M1 (Distant metastasis)
		MX (Distant metastasis cannot be assessed)
Reason no cancer-directed surgery	Details of cancer-directed surgery	Not performed and patient died prior to recommended surgery
		Not recommended only
		Not recommended and contraindicated due to other conditions
		Recommended but not performed, patient refused
		Recommended but not performed for unknown reasons
		Recommended but unknown if performed
		Surgery performed
		Unknown OR death certificate or autopsy-only case
EOD 10 - extent (1988 - 2003)	Extension of primary tumor code	0, 5, 10, 11, 13, 14, 15, 16, 17, 18, 20, 21, 23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 38, 40, 50, 60, 70, 80, 85, 99

Table 22: Performance metrics of the predictive models applied to the complete set of variables of the breast cancer dataset

Outcome variable	Classifier	ACC (%)	Dead			
			PRE (%)	REC (%)	SPE (%)	ROC (%)
2	ZR	90.0000	00.0000	00.0000	100.000	50.0000
	ADT	93.2775	74.7817	49.4570	98.1464	90.0625
	J48	94.4121	82.8895	55.5990	98.7247	85.4791
	RF	94.9214	86.5549	58.2650	98.9943	93.4213
	LB	93.3248	79.1369	45.1540	98.6771	90.0961
	RS	94.3695	86.8675	51.4780	99.1352	92.0089
	Mean	94.0611	82.0461	51.9906	98.7356	90.2136
4	ZR	89.5630	00.0000	00.0000	100.000	49.9888
	ADT	93.0468	75.8509	48.9738	98.1827	89.7862
	J48	94.2329	83.8294	55.4383	98.7537	85.4071
	RF	94.7784	87.6774	58.1422	99.0477	93.3559
	LB	93.0783	79.5791	45.3080	98.6451	89.8758
	RS	94.1983	87.5884	51.7448	99.1455	91.7218
	Mean	93.8669	82.9050	51.9214	98.7550	90.0294
6	ZR	88.8270	00.0000	00.0000	100.000	49.9894
	ADT	92.5689	77.1258	47.6121	98.2237	88.6730
	J48	93.8260	84.4045	54.8832	98.7244	85.1644
	RF	94.7661	88.3796	61.2029	98.9878	93.8393
	LB	92.5613	78.9243	45.6001	98.4683	88.6348
	RS	93.7962	88.3227	51.2512	99.1477	91.2060
	Mean	93.5037	83.4314	52.1099	98.7104	89.5035
8	ZR	84.3650	00.0000	00.0000	100.000	49.9905
	ADT	88.1321	77.0401	34.3256	98.1038	81.6831
	J48	90.4490	83.8936	48.1599	98.2863	80.1535
	RF	92.4864	88.7962	59.4442	98.6100	93.4189
	LB	88.2421	77.3110	35.0982	98.0910	81.4086
	RS	90.2852	89.3379	42.9971	99.0489	88.5966
	Mean	89.9190	83.2758	44.0050	98.4280	85.0521
10	ZR	64.0500	00.0000	00.0000	100.000	50.0000
	ADT	75.0322	71.9213	50.1138	89.0184	77.9428
	J48	86.1877	83.9272	76.1658	91.8128	87.8325
	RF	91.1437	89.6266	85.2295	94.4632	96.7200
	LB	74.0607	70.4801	47.9149	88.7358	77.3982
	RS	84.7572	89.8295	64.9555	95.8715	92.9891
	Mean	82.2363	81.1569	64.8759	91.9804	86.5765

Table 23: Results of attribute selection applied to the complete set of variables of the breast cancer dataset

Outcome variable	2		4		6		8		10	
Attributes	**NF**[1]	(%)[2]	**NF**	(%)	**NF**	(%)	**NF**	(%)	**NF**	(%)
ageDiagNum	10	100	10	100	10	100	10	100	10	100
behav1	0	0	0	0	0	0	0	0	0	0
diagConf	0	0	0	0	0	0	0	0	0	0
er	0	0	0	0	0	0	0	0	0	0
ext2	0	0	0	0	0	0	0	0	10	100
female	0	0	0	0	0	0	0	0	0	0
firstMalPrimInd	0	0	0	0	0	0	0	0	0	0
grade1	0	0	0	0	0	0	0	0	0	0
histGroup	0	0	0	0	0	0	0	0	0	0
laterality	0	0	0	0	0	0	0	0	0	0
m3	10	100	10	100	10	100	10	100	10	100
n3	0	0	0	0	0	0	0	0	0	0
numMalTum	0	0	0	0	0	0	0	0	0	0
numPrim	0	0	0	0	0	0	0	0	0	0
pr	0	0	0	0	0	0	0	0	0	0
primSite	0	0	0	0	0	0	0	0	0	0
raceGroup	10	100	10	100	10	100	0	0	0	0
rad	0	0	0	0	0	0	0	0	0	0
radSeqSurg	0	0	0	0	0	0	0	0	0	0
reasonNoCancerSurg	10	100	10	100	10	100	10	100	10	100
regNodeExamNom	0	0	0	0	0	0	0	0	0	0
regNodeExamNum	0	0	0	0	0	0	0	0	0	0
regNodePosNom	0	0	0	0	0	0	0	0	0	0
regNodePosNum	0	0	0	0	0	0	0	0	0	0
stage3	10	100	10	100	10	100	10	100	0	0
sumStage	0	0	0	0	0	0	0	0	0	0
surgPrimSite1	0	0	0	0	0	0	0	0	0	0
t3	0	0	0	0	0	0	0	0	0	0
tumSizeNom2	0	0	0	0	0	0	0	0	0	0
tumSizeNum2	0	0	0	0	0	0	0	0	0	0

[1]Number of cross-validation folds
[2]Score

Table 24: Performance metrics of the predictive models applied to the subset of variables of the breast cancer dataset

Outcome variable	Classifier	ACC (%)	Dead				Alive			
			PRE (%)	REC (%)	SPE (%)	ROC (%)	PRE (%)	REC (%)	SPE (%)	ROC (%)
2	ZR	90.0000	0.00000	0.00000	100.000	50.0000	90.0000	100.000	0.00000	50.0000
	ADT	92.7605	75.8918	40.4730	98.5702	87.2985	93.7120	98.5702	40.4730	87.2985
	J48	93.4582	80.4412	45.6940	98.7653	86.5520	94.2423	98.7653	45.6940	86.5520
	RF	93.4284	75.7522	50.4260	98.2064	83.2764	94.6891	98.2064	50.4260	83.2764
	LB	92.7298	74.0310	42.0530	98.3606	87.8374	93.8563	98.3606	42.0530	87.8374
	RS	93.0309	81.6650	39.1010	99.0231	88.3847	93.6039	99.0231	39.1010	88.3847
	Mean	93.0816	77.5562	43.5494	98.5851	86.6698	94.0207	98.5851	43.5494	86.6698
4	ZR	89.5630	0.00000	0.00000	100.000	49.9888	89.5630	100.000	0.00000	49.9888
	ADT	92.6136	76.0068	42.7134	98.4286	87.9761	93.6485	98.4286	42.7134	87.9761
	J48	93.2275	81.0147	45.8580	98.7476	86.3648	93.9944	98.7476	45.8580	86.3648
	RF	93.2532	76.8996	50.5394	98.2307	83.2279	94.4576	98.2307	50.5394	83.2279
	LB	92.4875	79.8442	37.4830	98.8973	87.1528	93.1389	98.8973	37.4830	87.1528
	RS	92.7944	82.6815	39.1731	99.0430	88.1094	93.3213	99.0430	39.1731	88.1094
	Mean	92.8752	79.2893	43.1534	98.6695	86.5662	93.7121	98.6695	43.1534	86.5662
6	ZR	88.8270	0.00000	0.00000	100.000	49.9894	88.8270	100.000	0.00000	49.9894
	ADT	92.1294	78.0708	41.1062	98.5473	86.6931	93.0085	98.5473	41.1062	86.6931
	J48	92.7841	82.3160	45.1087	98.7809	85.4995	93.4670	98.7809	45.1087	85.4995
	RF	92.7995	77.7195	49.8443	98.2026	83.4425	93.9635	98.2026	49.8443	83.4425
	LB	91.9433	75.4224	41.3801	98.3033	86.2630	93.0227	98.3033	41.3801	86.2630
	RS	92.2322	83.8149	37.7902	99.0801	87.1158	92.6806	99.0801	37.7902	87.1158

Continued on next page

Table 24 – continued from previous page

Outcome variable	Classifier	ACC (%)	Dead				Alive			
			PRE (%)	REC (%)	SPE (%)	ROC (%)	PRE (%)	REC (%)	SPE (%)	ROC (%)
	Mean	92.3777	79.4687	43.0459	98.5828	85.8028	93.2285	98.5828	43.0459	85.8028
8	ZR	84.3650	0.00000	0.00000	100.000	49.9905	84.3650	100.000	0.00000	49.9905
	ADT	87.7548	75.5737	32.0377	98.0806	81.0167	88.6198	98.0806	32.0377	81.0167
	J48	88.7188	81.0770	36.3256	98.4286	76.6770	89.2946	98.4286	36.3256	76.6770
	RF	89.0603	77.4855	42.3313	97.7204	80.2145	90.1414	97.7204	42.3313	80.2145
	LB	87.7135	73.6396	33.3681	97.7851	80.2597	88.7878	97.7851	33.3681	80.2597
	RS	88.1661	83.8551	30.1337	98.9210	81.1179	88.4260	98.9210	30.1337	81.1179
	Mean	88.2827	78.3262	34.8393	98.1871	79.8571	89.0539	98.1871	34.8393	79.8572
10	ZR	64.0500	0.00000	0.00000	100.000	50.0000	64.0500	100.000	0.00000	50.0000
	ADT	73.9971	73.1218	43.7530	90.9725	76.4059	74.2374	90.9725	43.7530	76.4059
	J48	75.7707	76.5428	47.0095	91.9138	77.7871	75.5520	91.9138	47.0095	77.7871
	RF	76.4751	75.0371	51.7925	90.3290	79.8997	76.9498	90.3290	51.7925	79.8997
	LB	73.9625	70.9862	46.6381	89.2991	76.5162	74.8842	89.2991	46.6381	76.5162
	RS	73.6544	80.8707	35.0006	95.3500	77.6229	72.3278	95.3500	35.0006	77.6229
	Mean	74.7720	75.3117	44.8387	91.5729	77.6464	74.7903	91.5729	44.8387	77.6464

D. Complete filter value for SEER variable **ICD-O-3 Hist/behav**

```
'8000/3: Neoplasm, malignant',
'8001/3: Tumor cells, malignant',
'8002/3: Malignant tumor, small cell type',
'8003/3: Malignant tumor, giant cell type',
'8004/3: Malignant tumor, spindle cell type',
'8005/3: Malignant tumor, clear cell type',
'8010/2: Carcinoma in situ, NOS',
'8010/3: Carcinoma, NOS',
'8011/3: Epithelioma, malignant',
'8012/3: Large cell carcinoma, NOS',
'8013/3: Large cell neuroendocrine carcinoma',
'8014/3: Large cell carcinoma with rhabdoid phenotype',
'8015/3: Glassy cell carcinoma',
'8020/3: Carcinoma, undifferentiated type, NOS',
'8021/3: Carcinoma, anaplastic type, NOS',
'8022/3: Pleomorphic carcinoma',
'8030/3: Giant cell and spindle cell carcinoma',
'8031/3: Giant cell carcinoma',
'8032/3: Spindle cell carcinoma',
'8033/3: Pseudosarcomatous carcinoma',
'8034/3: Polygonal cell carcinoma',
'8035/3: Carcinoma with osteoclast - like giant cells',
'8041/3: Small cell carcinoma, NOS',
'8043/3: Small cell carcinoma, fusiform cell',
'8050/2: Papillary carcinoma in situ',
'8050/3: Papillary carcinoma, NOS',
'8051/3: Verrucous carcinoma, NOS',
'8052/2: Papillary squamous cell carcinoma, non - invasive',
'8052/3: Papillary squamous cell carcinoma',
'8070/2: Squamous cell carcinoma in situ, NOS',
'8070/3: Squamous cell carcinoma, NOS',
'8071/3: Squamous cell carcinoma, keratinizing, NOS',
'8072/3: Squamous cell ca., large cell, nonkeratinizing',
'8073/3: Squamous cell ca., small cell, nonkeratinizing',
'8074/3: Squamous cell carcinoma, spindle cell',
'8075/3: Squamous cell carcinoma, adenoid',
'8076/2: Squamous cell CIS with questionable stromal invasion',
'8076/3: Squamous cell carcinoma, micro - invasive',
'8078/3: Squamous cell carcinoma with horn formation',
'8140/2: Adenocarcinoma in situ',
'8140/3: Adenocarcinoma, NOS',
'8141/3: Scirrhous adenocarcinoma',
'8143/3: Superficial spreading adenocarcinoma',
'8147/3: Basal cell adenocarcinoma',
'8190/3: Trabecular adenocarcinoma',
'8200/3: Adenoid cystic carcinoma',
'8201/2: Cribriform carcinoma in situ',
'8201/3: Cribriform carcinoma',
'8211/3: Tubular adenocarcinoma',
'8230/2: Duct carcinoma in situ, solid type',
'8230/3: Solid carcinoma, NOS',
'8231/3: Carcinoma simplex',
'8251/3: Alveolar adenocarcinoma',
'8255/3: Adenocarcinoma with mixed subtypes',
'8260/3: Papillary adenocarcinoma, NOS',
'8261/2: Adenocarcinoma in situ in villous adenoma',
'8261/3: Adenocarcinoma in villous adenoma',
'8310/3: Clear cell adenocarcinoma, NOS',
'8314/3: Lipid - rich carcinoma',
'8315/3: Glycogen - rich carcinoma',
'8320/3: Granular cell carcinoma',
'8323/3: Mixed cell adenocarcinoma',
'8401/3: Apocrine adenocarcinoma',
'8440/3: Cystadenocarcinoma, NOS',
'8480/3: Mucinous adenocarcinoma',
'8481/3: Mucin - producing adenocarcinoma',
'8490/3: Signet ring cell carcinoma',
'8500/2: Intraductal carcinoma, non - infiltrating, NOS',
'8500/3: Infiltrating duct carcinoma, NOS',
'8501/2: Comedocarcinoma, non - infiltrating',
'8501/3: Comedocarcinoma, NOS',
'8502/3: Secretory carcinoma of breast',
'8503/2: Noninfiltrating intraductal papillary adenocarcinoma',
'8503/3: Intraductal papillary adenocarcinoma with invasion',
'8504/2: Non - infiltrating intracystic carcinoma',
'8504/3: Intracystic carcinoma, NOS',
'8507/2: Intraductal micropapillary carcinoma',
'8508/3: Cystic hypersecretory carcinoma',
'8510/3: Medullary carcinoma, NOS',
'8512/3: Medullary carcinoma with lymphoid stroma',
'8513/3: Atypical medullary carcinoma',
'8514/3: Duct carcinoma, desmoplastic type',
'8520/2: Lobular carcinoma in situ',
'8520/3: Lobular carcinoma, NOS',
'8521/3: Infiltrating ductular carcinoma',
'8522/2: Intraductal and lobular in situ carcinoma',
'8522/3: Infiltrating duct and lobular carcinoma',
'8523/2: Intraductal with other types of carcinoma in situ',
'8523/3: Infiltrating duct mixed with other types of ca.',
'8524/3: Infiltrating lobular mixed with other types of ca.',
'8525/3: Polymorphous low grade adenocarcinoma',
'8530/3: Inflammatory carcinoma',
'8540/3: Paget disease, mammary',
'8541/3: Paget disease and infiltrating duct carcinoma',
'8543/3: Paget disease and intraductal carcinoma',
'8550/3: Acinar cell carcinoma',
'8551/3: Acinar cell cystadenocarcinoma',
'8560/3: Adenosquamous carcinoma',
'8562/3: Epithelial - myoepithelial carcinoma',
'8570/3: Adenocarcinoma with squamous metaplasia',
'8571/3: Adenocarcinoma w. cartilaginous \& osseous metaplasia',
'8572/3: Adenocarcinoma with spindle cell metaplasia',
'8573/3: Adenocarcinoma with apocrine metaplasia',
'8574/3: Adenocarcinoma with neuroendocrine differentiation',
'8575/3: Metaplastic carcinoma, NOS',
'8800/3: Sarcoma, NOS',
'8801/3: Spindle cell sarcoma',
'8802/3: Giant cell sarcoma',
'8803/3: Small cell sarcoma',
'8804/3: Epithelioid sarcoma',
'8805/3: Undifferentiated sarcoma',
'8806/3: Desmoplastic small round cell tumor',
'8810/3: Fibrosarcoma, NOS',
'8811/3: Fibromyxosarcoma',
'8813/3: Fascial fibrosarcoma',
'8814/3: Infantile fibrosarcoma',
'8815/3: Solitary fibrous tumor, malignant',
'8850/3: Liposarcoma, NOS',
'8851/3: Liposarcoma, well differentiated',
'8852/3: Myxoid liposarcoma',
'8853/3: Round cell liposarcoma',
'8854/3: Pleomorphic liposarcoma',
'8855/3: Mixed type liposarcoma',
'8857/3: Fibroblastic liposarcoma',
'8858/3: Dedifferentiated liposarcoma',
'8890/3: Leiomyosarcoma, NOS',
'8891/3: Epithelioid leiomyosarcoma',
'8894/3: Angiomyosarcoma',
'8895/3: Myosarcoma',
'8896/3: Myxoid leiomyosarcoma',
'8935/3: Stromal sarcoma, NOS',
'8980/3: Carcinosarcoma, NOS',
'8981/3: Carcinosarcoma, embryonal type',
'8982/3: Malignant myoepithelioma',
'8990/3: Mesenchymoma, malignant',
'8991/3: Embryonal sarcoma',
'9020/3: Phyllodes tumor, malignant',
'9120/3: Hemangiosarcoma',
'9130/3: Hemangioendothelioma, malignant',
'9133/3: Epithelioid hemangioendothelioma, malignant',
'9580/3: Granular cell tumor, malignant',
'9581/3: Alveolar soft part sarcoma',
'9590/3: Malignant lymphoma, NOS',
'9591/3: Malignant lymphoma, non - Hodgkin, NOS',
'9596/3: B - cell lymphoma, between diffuse large B and HL
      (composite HL and NHL)',
'9650/3: Classical Hodgkin lymphoma',
'9651/3: Lymphocyte - rich classical Hodgkin lymphoma',
'9652/3: Mixed cellularity classical Hodgkin lymphoma',
'9653/3: Lymphocyte - depleted classical Hodgkin lymphoma',
'9654/3: Hodgkin lymphoma, LD, diffuse fibrosis [OBS 2010+] see
      9651/3',
'9655/3: Hodgkin lymphoma, lymphocytic depleted, reticular',
'9659/3: Nodular lymphocyte predominant Hodgkin lymphoma',
'9661/3: Hodgkin granuloma [OBS 2010+] see 9651/3',
'9662/3: Hodgkin sarcoma [OBS 2010+] see 9651/3',
'9663/3: Nodular sclerosis classical Hodgkin lymphoma',
'9664/3: Hodgkin lymphoma, nodular sclerosis, cellular phase [OBS]
      see 9663/3',
'9665/3: Hodgkin lymphoma, nodular sclerosis, grade 1 [OBS 2010+]
      see 9663/3',
'9667/3: Hodgkin lymphoma, nodular sclerosis, grade 2 [OBS 2010+]
      see 9663/3',
'9670/3: Malignant lymphoma, small B lymphocytes, NOS [OBS 2012+]
      see 9823/3',
'9671/3: Lymphoplasmacytic lymphoma (NHL)',
'9673/3: Mantle cell lymphoma',
'9675/3: Malig lymphoma, mixed small \& large cell, diffuse [OBS
      2010+] see 9690/3',
'9680/3: Diffuse large B - cell (NHL) lymphoma (DLBCL)',
'9684/3: Malig. lymphoma, large B, diffuse, immunoblastic [OBS
      2012+] See 9680/3',
'9687/3: Burkitt lymphoma',
'9688/3: T - cell histiocyte - rich large B - cell lymphoma',
'9690/3: Follicular lymphoma, NOS',
'9691/3: Follicular lymphoma, grade 2',
'9695/3: Follicular lymphoma, grade 1',
'9698/3: Follicular lymphoma, grade 3',
'9699/3: Extranodal marginal zone lymphoma of mucosal - assoc.
      lymphoid tissue - MALT',
'9701/3: Sezary syndrome',
'9702/3: Peripheral (mature) T - cell lymphoma, NOS',
'9705/3: Angioimmunoblastic T - cell lymphoma',
'9712/3: Intravascular large B - cell lymphoma',
'9714/3: Anaplastic large cell (T - cell and Null cell) lymphoma,
      ALK - positive',
'9719/3: Extranodal NK - - /T - cell lymphoma, (nasal and ) nasal
```

```
       type',
178 '9724/3: Systemic EBV pos T - cell lymphoprolif dis of child',
179 '9727/3: Blastic plasmacytoid dendritic cell neoplasm (prec b -
       cell leuk/lymph)',
180 '9728/3: Precursor B - lymphoblastic lymphoma [OBS 2010+] See
       9811/3',
181 '9729/3: Precursor T - - cell lymphoblastic lymphoma, NOS [OBS
       2010+] See 9837/3',
182 '9731/3: Solitary plasmacytoma of bone/plasmacytoma, NOS',
183 '9734/3: Extraosseous (extramedullary) plasmacytoma',
184 '9740/3: Mast cell sarcoma',
185 '9741/3: Malignant mastocytosis',
186 '9750/3: Malignant histiocytosis [OBS 2010+] see 9751/3',
187 '9751/3: Langerhans cell histiocytosis (malignant) (NOS)',
188 '9754/3: Langerhans cell histiocytosis, disseminated [OBS 2010+]
       see 9751/3',
189 '9755/3: Histiocytic sarcoma',
190 '9756/3: Langerhans cell sarcoma',
191 '9757/3: Interdigitating dendritic cell sarcoma',
192 '9758/3: Follicular dendritic cell sarcoma',
193 '9811/3: B lymphoblastic leuk/lymph, NOS',
194 '9812/3: B lymphobl leuk/lymph w/t(9;22)(q34;q11.2); BCR - ABL1',
195 '9813/3: B lymphobl leuk/lymph w/t(v;11q23); MLL rearranged',
196 '9814/3: B lymphobl leuk/lymph w/t(12;21)(p13;q22);TEL - AML1',
197 '9815/3: B lymphoblastic leuk/lymph w/hyperdiploidy',
198 '9816/3: B lymphobl leuk/lymph w/hypodiploidy (hypodip ALL)',
199 '9817/3: B lymphobl leuk/lymph w/t(5;14)(q31;q32); IL3 - IGH',
200 '9818/3: B lymphobl leuk/lymph w/t(1;19)(q23;p13.3); E2A PBX1',
201 '9823/3: Chronic lymphocytic leukemia/small lymphocytic lymphoma
       (B - cell)',
202 '9831/3: T - cell large granular lymphocytic leukemia',
203 '9837/3: Adult T - cell leukemia/lymphoma',
204 '9965/3: Myeloid and lymphoid neoplasms w/PDGFRA rearrange',
205 '9967/3: Myeloid and lymphoid neoplasm w/FGFR1 abnormalities',
206 '9971/3: Polymorphic PTLD',
207 '9975/3: Myelodysp./myeloproliferative neoplasm unclass (malig
       only before 2010)'%
```

E. Complete console log of the prediction process in BOSOM Calculator once a user submits a validated calculator form

NOTE: Several components of this log file were modified in order to be presented properly in this section. URLs, file paths and variable names are several examples of these components.

```
Mar 12, 2014 7:50:12 PM
        ph.edu.upm.agila.gtmeren.bosom.controller.CalcController
        showPost
INFO: Form data: WekaData [ageDiagNum=50, raceGroup=Black,
        stage3=IIA, m3=M0, reasonNoCancerSurg=Surgery performed,
        ext2=11, time2=null, time4=null, time6=null, time8=null,
        time10=null]

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcArffServiceImpl getInstances
INFO:
CalcArffServiceImpl: creating Instances data
@relation SeerBreastCancer

@attribute ageDiagNum numeric
@attribute raceGroup {Black,Other,Unknown,White}
@attribute stage3 {0,I,IIA,IIB,IIIA,IIIB,IIIC,IIINOS,IV,'UNK
        Stage'}
@attribute m3 {M0,M1,MX}
@attribute reasonNoCancerSurg {'Not performed, patient died prior
        to recommended surgery','Not recommended','Not recommended,
        contraindicated due to other conditions','Recommended but
        not performed, patient refused','Recommended but not
        performed, unknown reason','Recommended, unknown if
        performed','Surgery performed','Unknown; death certificate
        or autopsy only case'}
@attribute ext2 {00, 05, 10, 11, 13, 14, 15, 16, 17, 18, 20, 21,
        23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 38, 40,
        50, 60, 70, 80, 85, 99}
@attribute time2 {0,1}
@attribute time4 {0,1}
@attribute time6 {0,1}
@attribute time8 {0,1}
@attribute time10 {0,1}

@data
50,Black,IIA,M0,'Surgery performed',11,?,?,?,?,?

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time2/adt.MODEL

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.ADTree
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.28012585481457003
Percentage [1]: 0.71987414518543

Mar 12, 2014 7:50:12 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time2/lb.MODEL

Mar 12, 2014 7:50:13 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.LogitBoost
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.07496357829362793
Percentage [1]: 0.9250364217063721

Mar 12, 2014 7:50:13 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time2/j48.MODEL

Mar 12, 2014 7:50:13 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.J48
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.05391520352283633
Percentage [1]: 0.9460847964771637

Mar 12, 2014 7:50:13 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time2/rf.MODEL

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.RandomForest
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.1370508658008658
Percentage [1]: 0.8629491341991342

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time2/rs.MODEL

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.RandomSubSpace
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.06297097851445468
Percentage [1]: 0.9370290214855453

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time4/adt.MODEL

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.ADTree
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.27356475076848163
Percentage [1]: 0.7264352492315184

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time4/lb.MODEL

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.LogitBoost
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.06509926614144576
Percentage [1]: 0.9349007338585542

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time4/j48.MODEL

Mar 12, 2014 7:50:14 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.J48
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.05747191328402868
Percentage [1]: 0.9425280867159713

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
```

```
Path: /WEB-INF/models/time4/rf.MODEL

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.RandomForest
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.1370508658008658
Percentage [1]: 0.8629491341991342

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time4/rs.MODEL

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.RandomSubSpace
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.06617540648857886
Percentage [1]: 0.9338245935114211

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time6/adt.MODEL

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.ADTree
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.20347839043151927
Percentage [1]: 0.7965216095684807

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time6/lb.MODEL

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.LogitBoost
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.13606290039657967
Percentage [1]: 0.8639370996034205

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time6/j48.MODEL

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.J48
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.06447242138542314
Percentage [1]: 0.9355275786145769

Mar 12, 2014 7:50:15 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time6/rf.MODEL

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.RandomForest
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.3668560606060606
Percentage [1]: 0.6331439393939393

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time6/rs.MODEL

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.RandomSubSpace
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.07324705810648151
Percentage [1]: 0.9267529418935185

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time8/adt.MODEL

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.ADTree
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.2781600367791818
Percentage [1]: 0.7218399632208182

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time8/lb.MODEL

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.LogitBoost
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.12541784931176994
Percentage [1]: 0.87458215068823

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time8/j48.MODEL

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.J48
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.10912888838706035
Percentage [1]: 0.8908711116129396

Mar 12, 2014 7:50:16 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time8/rf.MODEL

Mar 12, 2014 7:50:17 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.RandomForest
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.3668560606060606
Percentage [1]: 0.6331439393939393

Mar 12, 2014 7:50:17 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time8/rs.MODEL

Mar 12, 2014 7:50:17 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.RandomSubSpace
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.13041089920651855
Percentage [1]: 0.8695891007934815

Mar 12, 2014 7:50:17 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time10/adt.MODEL

Mar 12, 2014 7:50:17 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.ADTree
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.4755008971480796
Percentage [1]: 0.5244991028519205

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time10/lb.MODEL

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
```

```
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.LogitBoost
Class [0=Dead,1=Alive]: 0.0
Percentage [0]: 0.5160820111859176
Percentage [1]: 0.4839179888140825

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time10/j48.MODEL

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.J48
Class [0=Dead,1=Alive]: 0.0
Percentage [0]: 0.7432432432432432
Percentage [1]: 0.25675675675675674

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time10/rf.MODEL

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.trees.RandomForest
Class [0=Dead,1=Alive]: 0.0
Percentage [0]: 1.0
Percentage [1]: 0.0

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl getClassifier
INFO:
CalcModelServiceImpl: reading model files
Path: /WEB-INF/models/time10/rs.MODEL

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcModelServiceImpl predict
INFO: CalcModelServiceImpl: predicting class and its percentage
        distribution
Classifier: class weka.classifiers.meta.RandomSubSpace
Class [0=Dead,1=Alive]: 1.0
Percentage [0]: 0.3953288516455826
Percentage [1]: 0.6046711483544174

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImpl: extracting data per time period
Time Period: time2
Data: {adt={Class=1.0, Percentage=0.71987414518543},
        lb={Class=1.0, Percentage=0.9250364217063721},
        j48={Class=1.0, Percentage=0.9460847964771637},
        rf={Class=1.0, Percentage=0.8629491341991342},
        rs={Class=1.0, Percentage=0.9370290214855453}}

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImp: computing prediction meansTime Period: time2
Sum: 439.09735190536463
Mean: 87.82

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImpl: extracting data per time period
Time Period: time4
Data: {adt={Class=1.0, Percentage=0.7264352492315184},
        lb={Class=1.0, Percentage=0.9349007338585542},
        j48={Class=1.0, Percentage=0.9425280867159713},
        rf={Class=1.0, Percentage=0.8629491341991342},
        rs={Class=1.0, Percentage=0.9338245935114211}}

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImp: computing prediction meansTime Period: time4
Sum: 440.0637797516599
Mean: 88.01

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImpl: extracting data per time period
Time Period: time6
Data: {adt={Class=1.0, Percentage=0.7965216095684807},
        lb={Class=1.0, Percentage=0.8639370996034205},
        j48={Class=1.0, Percentage=0.9355275786145769},
        rf={Class=1.0, Percentage=0.6331439393939393},
        rs={Class=1.0, Percentage=0.9267529418935185}}

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImp: computing prediction meansTime Period: time6
Sum: 415.58831690739356
Mean: 83.12

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImpl: extracting data per time period
Time Period: time8
Data: {adt={Class=1.0, Percentage=0.7218399632208182},
        lb={Class=1.0, Percentage=0.87458215068823}, j48={Class=1.0,
        Percentage=0.8908711116129396}, rf={Class=1.0,
        Percentage=0.6331439393939393}, rs={Class=1.0,
        Percentage=0.8695891007934815}}

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImp: computing prediction meansTime Period: time8
Sum: 399.00262657094083
Mean: 79.8

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImpl: extracting data per time period
Time Period: time10
Data: {adt={Class=1.0, Percentage=0.5244991028519205},
        lb={Class=0.0, Percentage=0.4839179888140825},
        j48={Class=0.0, Percentage=0.25675675675675674},
        rf={Class=0.0, Percentage=0.0}, rs={Class=1.0,
        Percentage=0.6046711483544174}}

Mar 12, 2014 7:50:18 PM ph.edu.upm.agila.gtmeren.bosom.service.
        impl.CalcServiceImpl getMeanPredictions
INFO:
CalcServiceImp: computing prediction meansTime Period: time10
Sum: 186.9844996777177
Mean: 37.4

Mar 12, 2014 7:50:18 PM
        ph.edu.upm.agila.gtmeren.bosom.controller.PdfController
        getPdfFilePath
INFO: Start of PDF creation
Mar 12, 2014 7:50:19 PM
        ph.edu.upm.agila.gtmeren.bosom.controller.PdfController
        getPdfFile
INFO: PDF file path:
        C:\apache-tomcat\7.0.47\webapps\bosom\WEB-INF\reports\ORIG-12032014194823.pdf
Mar 12, 2014 7:50:19 PM
        ph.edu.upm.agila.gtmeren.bosom.pdf.PdfBuilder
        addBarChartToPdf
INFO: Start of charts creation
Mar 12, 2014 7:50:39 PM
        ph.edu.upm.agila.gtmeren.bosom.pdf.PdfBuilder
        addBarChartToPdf
INFO: End of charts creation
Mar 12, 2014 7:50:39 PM
        ph.edu.upm.agila.gtmeren.bosom.controller.PdfController
        getPdfFile
INFO: PDF file path:
        C:\apache-tomcat\7.0.47\webapps\bosom\WEB-INF\reports\12032014194823.pdf
Mar 12, 2014 7:50:39 PM
        ph.edu.upm.agila.gtmeren.bosom.controller.PdfController
        getPdfFile
INFO: PDF file path:
        C:\apache-tomcat\7.0.47\webapps\bosom\WEB-INF\reports\ORIG-12032014194823.pdf
Mar 12, 2014 7:50:39 PM
        ph.edu.upm.agila.gtmeren.bosom.controller.PdfController
        getPdfFile
INFO: PDF file path:
        C:\apache-tomcat\7.0.47\webapps\bosom\WEB-INF\reports\bosom-info.pdf
Mar 12, 2014 7:50:39 PM
        ph.edu.upm.agila.gtmeren.bosom.pdf.PdfConcatenator
        concatenate
INFO: Preparation for PDF concatenation

Mar 12, 2014 7:50:39 PM
        ph.edu.upm.agila.gtmeren.bosom.pdf.PdfConcatenator
        concatenate
INFO: File # 0:
        C:\apache-tomcat\7.0.47\webapps\bosom\WEB-INF\reports\ORIG-12032014194823.pdf
Mar 12, 2014 7:50:39 PM
        ph.edu.upm.agila.gtmeren.bosom.pdf.PdfConcatenator
        concatenate
INFO: File # 1:
        C:\apache-tomcat\7.0.47\webapps\bosom\WEB-INF\reports\bosom-info.pdf
Mar 12, 2014 7:50:40 PM
        ph.edu.upm.agila.gtmeren.bosom.controller.PdfController
        getPdfFilePath
INFO: End of PDF creation
```

F. Result buffers of selected trained predictive models

NOTE: The following section's format was modified in order to be presented properly in this section.

Source Code 43: Result buffer of the subset dataset's alternating decision tree model for predicting two-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.ADTree
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time2

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Alternating decision tree:

: 1.099
| (1)m3 = M0: 0.174
| | (2)stage3 = I: 0.588
| | (2)stage3 != I: -0.224
| | | (3)stage3 = 0: 1.117
| | | (3)stage3 != 0: -0.191
| | | | (4)stage3 = IIA: 0.471
| | | | (4)stage3 != IIA: -0.263
| | | | | (9)stage3 = IIB: 0.366
| | | | | (9)stage3 != IIB: -0.181
| | (6)ext2 = 10: 0.258
| | (6)ext2 != 10: -0.133
| | (7)reasonNoCancerSurg = Surgery performed: 0.033
| | (7)reasonNoCancerSurg != Surgery performed: -0.929
| (1)m3 != M0: -1.736
| (5)ageDiagNum < 77.5: 0.093
| | (10)raceGroup = Black: -0.341
| | (10)raceGroup != Black: 0.048
| (5)ageDiagNum >= 77.5: -0.64
| (8)stage3 = UNK Stage: 0.508
| (8)stage3 != UNK Stage: -0.039
Legend: -ve = 0, +ve = 1
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 92879 92.879 %
Incorrectly Classified Instances 7121 7.121 %
Kappa statistic 0.4817
K&B Relative Info Score -7821637.8404 %
K&B Information Score -36685.3406 bits -0.3669 bits/instance
Class complexity | order 0 46899.5594 bits 0.469 bits/instance
Class complexity | scheme 39291.3251 bits 0.3929 bits/instance
Complexity improvement (Sf) 7608.2343 bits 0.0761 bits/instance
Mean absolute error 0.2105
Root mean squared error 0.2649
Relative absolute error 116.9425 %
Root relative squared error 88.2985 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC ROC
                     Area PRC Area Class
               0.378 0.010 0.808 0.378 0.515 0.523 0.883 0.613 0
               0.990 0.622 0.935 0.990 0.962 0.523 0.883 0.980 1
Weighted Avg. 0.929 0.561 0.922 0.929 0.917 0.523 0.883 0.943

=== Confusion Matrix ===

    a b <-- classified as
 3776 6224 | a = 0
  897 89103 | b = 1
```

Source Code 44: Result buffer of the subset dataset's alternating decision tree model for predicting four-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.ADTree
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time4

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Alternating decision tree:

: 1.075
| (1)m3 = M0: 0.169
| | (2)stage3 = I: 0.576
| | (2)stage3 != I: -0.222
| | | (3)stage3 = 0: 1.114
| | | (3)stage3 != 0: -0.192
| | | | (4)stage3 = IIA: 0.47
| | | | (4)stage3 != IIA: -0.265
| | | | | (7)ext2 = 70: -0.958
| | | | | (7)ext2 != 70: 0.081
| | | | | | (9)stage3 = IIIC: -0.493
| | | | | | (9)stage3 != IIIC: 0.099
| | (6)reasonNoCancerSurg = Surgery performed: 0.035
| | | (8)ext2 = 10: 0.174
| | | (8)ext2 != 10: -0.108
| | (6)reasonNoCancerSurg != Surgery performed: -0.992
| (1)m3 != M0: -1.753
| (5)ageDiagNum < 77.5: 0.092
| | (10)raceGroup = Black: -0.342
| | (10)raceGroup != Black: 0.038
| (5)ageDiagNum >= 77.5: -0.641
Legend: -ve = 0, +ve = 1
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 92616 92.616 %
Incorrectly Classified Instances 7384 7.384 %
Kappa statistic 0.5087
K&B Relative Info Score -7107103.0236 %
K&B Information Score -34306.9792 bits -0.3431 bits/instance
Class complexity | order 0 48269.7048 bits 0.4827 bits/instance
Class complexity | scheme 40412.818 bits 0.4041 bits/instance
Complexity improvement (Sf) 7856.8868 bits 0.0786 bits/instance
Mean absolute error 0.2158
Root mean squared error 0.2696
Relative absolute error 115.4231 %
Root relative squared error 88.1885 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC ROC
                     Area PRC Area Class
               0.424 0.015 0.763 0.424 0.545 0.535 0.880 0.615 0
               0.985 0.576 0.936 0.985 0.960 0.535 0.880 0.979 1
Weighted Avg. 0.926 0.517 0.918 0.926 0.917 0.535 0.880 0.941

=== Confusion Matrix ===

    a b <-- classified as
 4429 6008 | a = 0
 1376 88187 | b = 1
```

Source Code 45: Result buffer of the subset dataset's alternating decision tree model for predicting six-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.ADTree
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time6

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Alternating decision tree:

: 1.037
| (1)m3 = M0: 0.16
| | (2)stage3 = I: 0.543
| | (2)stage3 != I: -0.218
| | | (3)stage3 = 0: 1.043
| | | (3)stage3 != 0: -0.19
| | | | (4)stage3 = IIA: 0.459
| | | | (4)stage3 != IIA: -0.264
| | | | | (7)ext2 = 70: -0.939
| | | | | (7)ext2 != 70: 0.078
| | | | | | (8)stage3 = IIIC: -0.51
| | | | | | (8)stage3 != IIIC: 0.097
| | | | | | (9)ext2 = 50: -0.768
| | | | | | (9)ext2 != 50: 0.048
| | (6)reasonNoCancerSurg = Surgery performed: 0.031
| | (6)reasonNoCancerSurg != Surgery performed: -0.965
| (1)m3 != M0: -1.762
| | (10)stage3 = IV: -0.265
| | (10)stage3 != IV: 0.695
| (5)ageDiagNum < 77.5: 0.087
| (5)ageDiagNum >= 77.5: -0.623
Legend: -ve = 0, +ve = 1
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 92124 92.124 %
Incorrectly Classified Instances 7876 7.876 %
Kappa statistic 0.5005
K&B Relative Info Score -6852624.1367 %
K&B Information Score -34613.6154 bits -0.3461 bits/instance
Class complexity | order 0 50511.2213 bits 0.5051 bits/instance
Class complexity | scheme 42804.3603 bits 0.428 bits/instance
Complexity improvement (Sf) 7706.861 bits 0.0771 bits/instance
Mean absolute error 0.2279
Root mean squared error 0.2794
Relative absolute error 114.8003 %
Root relative squared error 88.6984 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC ROC
                     Area PRC Area Class
               0.412 0.015 0.779 0.412 0.539 0.531 0.868 0.615 0
               0.985 0.588 0.930 0.985 0.957 0.531 0.868 0.974 1
Weighted Avg. 0.921 0.524 0.913 0.921 0.910 0.531 0.868 0.934

=== Confusion Matrix ===

    a b <-- classified as
 4607 6566 | a = 0
 1310 87517 | b = 1
```

Source Code 46: Result buffer of the subset dataset's alternating decision tree model for predicting eight-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.ADTree
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time8

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Alternating decision tree:

: 0.843
| (1)m3 = M0: 0.111
| | (2)stage3 = I: 0.312
| | (2)stage3 != I: -0.163
| | | (3)stage3 = 0: 0.562
| | | (3)stage3 != 0: -0.154
| | | | (6)stage3 = IIA: 0.299
| | | | (6)stage3 != IIA: -0.247
| | | (7)stage3 = IIIC: -0.461
| | | (7)stage3 != IIIC: 0.07
| | (4)ext2 = 10: 0.356
| | (4)ext2 != 10: -0.165
| | | (10)ext2 = 13: -0.272
| | | (10)ext2 != 13: 0.058
| (1)m3 != M0: -1.643
| (5)ageDiagNum < 77.5: 0.067
| (5)ageDiagNum >= 77.5: -0.553
| (8)reasonNoCancerSurg = Surgery performed: 0.017
| (8)reasonNoCancerSurg != Surgery performed: -0.538
| (9)stage3 = UNK Stage: 0.484
| (9)stage3 != UNK Stage: -0.029
Legend: -ve = 0, +ve = 1
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 87739 87.739 %
Incorrectly Classified Instances 12261 12.261 %
Kappa statistic 0.39
K&B Relative Info Score -4297288.6199 %
K&B Information Score -26880.0004 bits -0.2688 bits/instance
Class complexity | order 0 62550.565 bits 0.6255 bits/instance
Class complexity | scheme 56404.7916 bits 0.564 bits/instance
Complexity improvement (Sf) 6145.7734 bits 0.0615 bits/instance
Mean absolute error 0.2944
Root mean squared error 0.3369
Relative absolute error 111.6017 %
Root relative squared error 92.7491 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC ROC
                     Area PRC Area Class
               0.316 0.019 0.760 0.316 0.446 0.438 0.811 0.568 0
               0.981 0.684 0.886 0.981 0.931 0.438 0.811 0.947 1
Weighted Avg. 0.877 0.580 0.866 0.877 0.855 0.438 0.811 0.888

=== Confusion Matrix ===

    a b <-- classified as
 4936 10699 | a = 0
 1562 82803 | b = 1
```

Source Code 47: Result buffer of the subset dataset's alternating decision tree model for predicting ten-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.ADTree
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
```

```
      ext2
      time10

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Alternating decision tree:

: 0.289
| (1)ext2 = 10: 0.521
| (1)ext2 != 10: -0.238
| | (2)ext2 = 13: -0.724
| | (2)ext2 != 13: 0.095
| | | (3)ext2 = 85: -1.449
| | | (3)ext2 != 85: 0.062
| | | (7)ext2 = 11: -0.233
| | | (7)ext2 != 11: 0.089
| | | | (8)stage3 = IIIB: -0.754
| | | | (8)stage3 != IIIB: 0.029
| (4)ageDiagNum < 77.5: 0.04
| (4)ageDiagNum >= 77.5: -0.429
| (5)stage3 = IIIC: -0.715
| (5)stage3 != IIIC: 0.022
| | (9)stage3 = IIIA: -0.408
| | (9)stage3 != IIIA: 0.021
| | | (10)stage3 = IIB: -0.253
| | | (10)stage3 != IIB: 0.022
| (6)reasonNoCancerSurg = Surgery performed: 0.019
| (6)reasonNoCancerSurg != Surgery performed: -0.712
Legend: -ve = 0, +ve = 1
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 73992 73.992 %
Incorrectly Classified Instances 26008 26.008 %
Kappa statistic 0.3812
K&B Relative Info Score 1163557.7837 %
K&B Information Score 10963.8556 bits 0.1096 bits/instance
Class complexity | order 0 94226.7369 bits 0.9423 bits/instance
Class complexity | scheme 81839.2982 bits 0.8184 bits/instance
Complexity improvement (Sf) 12387.4387 bits 0.1239 bits/instance
Mean absolute error 0.4155
Root mean squared error 0.4361
Relative absolute error 90.2182 %
Root relative squared error 90.8862 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC ROC
                     Area PRC Area Class
               0.439 0.091 0.730 0.439 0.548 0.405 0.764 0.681 0
               0.909 0.561 0.743 0.909 0.817 0.405 0.764 0.821 1
Weighted Avg. 0.740 0.392 0.738 0.740 0.721 0.405 0.764 0.771

=== Confusion Matrix ===

    a b <-- classified as
 15795 20155 | a = 0
  5853 58197 | b = 1r
```

Source Code 48: Result buffer of the subset dataset's random forest model for predicting two-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.RandomForest
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time2

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 3
       random features.
Out of bag error: 0.0751


Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 93465 93.465 %
Incorrectly Classified Instances 6535 6.535 %
Kappa statistic 0.5721
K&B Relative Info Score 3244359.8228 %
K&B Information Score 15216.8187 bits 0.1522 bits/instance
Class complexity | order 0 46899.5594 bits 0.469 bits/instance
Class complexity | scheme 1373483.0448 bits 13.7348 bits/instance
Complexity improvement (Sf) -1326583.4853 bits -13.2658
       bits/instance
Mean absolute error 0.0954
Root mean squared error 0.2343
Relative absolute error 53.0078 %
Root relative squared error 78.1089 %
Coverage of cases (0.95 level) 97.577 %
Mean rel. region size (0.95 level) 62.4475 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC ROC
                     Area PRC Area Class
               0.503 0.017 0.763 0.503 0.606 0.587 0.832 0.628 0
               0.983 0.497 0.947 0.983 0.964 0.587 0.832 0.962 1
Weighted Avg. 0.935 0.449 0.928 0.935 0.929 0.587 0.832 0.928

=== Confusion Matrix ===

    a b <-- classified as
 5027 4973 | a = 0
 1562 88438 | b = 1
```

Source Code 49: Result buffer of the subset dataset's random forest model for predicting four-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.RandomForest
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time4

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 3
       random features.
Out of bag error: 0.0766


Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 93239 93.239 %
Incorrectly Classified Instances 6761 6.761 %
Kappa statistic 0.5746
K&B Relative Info Score 3329168.8347 %
K&B Information Score 16070.3631 bits 0.1607 bits/instance
Class complexity | order 0 48269.7048 bits 0.4827 bits/instance
Class complexity | scheme 1403413.9968 bits 14.0341 bits/instance
Complexity improvement (Sf) -1355144.292 bits -13.5514
       bits/instance
Mean absolute error 0.0991
Root mean squared error 0.2387
Relative absolute error 52.9991 %
Root relative squared error 78.0892 %
Coverage of cases (0.95 level) 97.514 %
Mean rel. region size (0.95 level) 63.2735 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

```

```
                TP Rate FP Rate Precision Recall F-Measure MCC ROC
                      Area PRC Area Class
                0.507 0.018 0.766 0.507 0.610 0.589 0.830 0.632 0
                0.982 0.493 0.945 0.982 0.963 0.589 0.830 0.959 1
Weighted Avg. 0.932 0.444 0.926 0.932 0.926 0.589 0.830 0.925

=== Confusion Matrix ===

    a b <-- classified as
 5287 5150 | a = 0
 1611 87952 | b = 1
```

Source Code 50: Result buffer of the subset dataset's random forest model for predicting six-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.RandomForest
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time6

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 3
       random features.
Out of bag error: 0.0811


Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 92803 92.803 %
Incorrectly Classified Instances 7197 7.197 %
Kappa statistic 0.5696
K&B Relative Info Score 3408869.4133 %
K&B Information Score 17218.7023 bits 0.1722 bits/instance
Class complexity | order 0 50511.2213 bits 0.5051 bits/instance
Class complexity | scheme 1318550.2137 bits 13.1855 bits/instance
Complexity improvement (Sf) -1268038.9925 bits -12.6804
       bits/instance
Mean absolute error 0.1065
Root mean squared error 0.2461
Relative absolute error 53.6323 %
Root relative squared error 78.1264 %
Coverage of cases (0.95 level) 97.482 %
Mean rel. region size (0.95 level) 64.5355 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

                TP Rate FP Rate Precision Recall F-Measure MCC ROC
                      Area PRC Area Class
                0.498 0.018 0.778 0.498 0.607 0.587 0.834 0.640 0
                0.982 0.502 0.940 0.982 0.960 0.587 0.834 0.958 1
Weighted Avg. 0.928 0.448 0.922 0.928 0.921 0.587 0.834 0.923

=== Confusion Matrix ===

    a b <-- classified as
 5561 5612 | a = 0
 1585 87242 | b = 1
```

Source Code 51: Result buffer of the subset dataset's random forest model for predicting eight-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.RandomForest
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time8

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 3
       random features.
Out of bag error: 0.1182


Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 89098 89.098 %
Incorrectly Classified Instances 10902 10.902 %
Kappa statistic 0.4937
K&B Relative Info Score 3009094.1601 %
K&B Information Score 18822.2062 bits 0.1882 bits/instance
Class complexity | order 0 62550.565 bits 0.6255 bits/instance
Class complexity | scheme 1231549.593 bits 12.3155 bits/instance
Complexity improvement (Sf) -1168999.028 bits -11.69 bits/instance
Mean absolute error 0.164
Root mean squared error 0.3003
Relative absolute error 62.1553 %
Root relative squared error 82.6884 %
Coverage of cases (0.95 level) 97.609 %
Mean rel. region size (0.95 level) 77.154 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

                TP Rate FP Rate Precision Recall F-Measure MCC ROC
                      Area PRC Area Class
                0.426 0.023 0.776 0.426 0.550 0.523 0.804 0.612 0
                0.977 0.574 0.902 0.977 0.938 0.523 0.804 0.934 1
Weighted Avg. 0.891 0.488 0.882 0.891 0.877 0.523 0.804 0.884

=== Confusion Matrix ===

    a b <-- classified as
 6656 8979 | a = 0
 1923 82442 | b = 1
```

Source Code 52: Result buffer of the subset dataset's random forest model for predicting ten-year breast cancer survival

```
=== Run information ===

Scheme:  class weka.classifiers.trees.RandomForest
Relation:
       BOSOM.100K-weka.filters.unsupervised.attribute.Remove-R3, 4,
       5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, 23, 24,
       25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35
Instances: 100000
Attributes: 7
      ageDiagNum
      raceGroup
      stage3
      m3
      reasonNoCancerSurg
      ext2
      time10

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 3
       random features.
Out of bag error: 0.2395


Time taken to build model:

=== Predictions on test data ===

see associated CSV file
=== Summary ===

Correctly Classified Instances 76545 76.545 %
Incorrectly Classified Instances 23455 23.455 %
Kappa statistic 0.453
K&B Relative Info Score 3343599.3106 %
K&B Information Score 31505.7322 bits 0.3151 bits/instance
Class complexity | order 0 94226.7369 bits 0.9423 bits/instance
Class complexity | scheme 1013471.2645 bits 10.1347 bits/instance
Complexity improvement (Sf) -919244.5276 bits -9.1924 bits/instance
```

```
Mean absolute error 0.3129
Root mean squared error 0.4067
Relative absolute error 67.9351 %
Root relative squared error 84.7544 %
Coverage of cases (0.95 level) 98.91 %
Mean rel. region size (0.95 level) 93.5155 %
Total Number of Instances 100000

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC ROC
                     Area PRC Area Class
               0.517 0.095 0.753 0.517 0.613 0.470 0.799 0.731 0
               0.905 0.483 0.769 0.905 0.832 0.470 0.799 0.848 1
Weighted Avg. 0.765 0.344 0.764 0.765 0.753 0.470 0.799 0.806

=== Confusion Matrix ===

    a b <-- classified as
 18583 17367 | a = 0
  6088 57962 | b = 1
```

XI. Acknowledgement

I would like to take this opportunity to give my deepest appreciation and gratitude to all the people who have helped in the progress of my research and completion of the special program.

First I want to give myself a pat in the back and big long hug for another milestone achieved. This experience made me push myself again to the limits of my abilities as a student, a programmer, a son and as a friend. I am proud of this accomplishment and I am looking forward to the next challenges that will come.

To *Dr. Vincent Peter C. Magboo*, I am blessed to have you as a mentor and adviser for my research. Thank you for the guidance and motivation on the topics I had difficulty of comprehending. I am grateful for this priceless experience I shared with you.

To *Mama* and *Papa*, thank you for the financial support for the supplies I needed during my research. Sorry *sa mga araw na masungit at hindi palakibo ako dahil gumagawa ng* thesis. To *Gellie*, thank you for waking me up whenever I ask to, even though you failed at most of them. Thank you for everything.

To *Prof. Ma. Sofia Criselda Poblador*, I am grateful for the time, effort and knowledge in statistics you have shared with me. Even though I was not able to use them in the end, I am still looking forward to learning more about them in the future.

To *Prof. Geoffrey Solano*, thank you for the motivation to finish this study. You have been an integral part of my undergraduate life and I am lucky to have you as one of my mentors in UP.

To *Prof. Richard Bryann Chua*, I want to thank you for accommodating my questions during the development of my research. I appreciate all the additional suggestions you provided and these have contributed to my research's improvement.

To *Mrs. Emma Alota* and the PGH Cancer Instute Pharmacy, thank you all the help during my data gathering period. I would never have the chance to have a consultation with *Dr. Lou Joel Tia* about the SEER and breast cancer dataset. Thank you again, *Tita*.

To *Ms. Eden Huelgas*, you'll never know how much of my research advanced because of your help. Thank you, ma'am. I will never forget all the kindness you showed me.

To *AJ Mendoza*, I owe you for motivating me in continuing my breast cancer prediction system research. Thank you very much for being there during the topic proposal season. You don't know how much the talk we had in the LRT helped me clear my thoughts to trust and proceed with this research. You're still as awesome as I remember way back when I was a freshman.

To *John Simon Tayson*, *Michelle dela Cruz*, *Hainah Kariza Leus*, *EJ Giray*, *Kim Isle Cortez*, *Azeil Louisse Codizar*, *Rheabedette Corpuz*, *Alyra Escutido*, *Sherwin Keith Saringan*, *Matthew Kendrick Co* and *Naris Morales*, I want to thank all of you for the contributions you have given to my research. Each one of you have paved the way to the gradual completion of my research. You guys are amazing.

To *Mr. Ankit Agrawal*, I am grateful for answering my questions and explaining the concepts I have not understood clearly during our correspondences. I still have a lot to learn but your guidance cleared the way. Your research and system served as an inspiration to finish this special problem. Thank you very much and I hope our paths would intersect again someday.

To *Mx. EdM*, *Mx. Walter* and *Mx. Rob* from `StackOverflow`, I may not know who you all are in real life but I want to personally thank you both for all the help in neural networks, predictive modeling, R statistics and WEKA programming. I would never be able to create the modules in my system without the answers, resources and explanations you provided before. I wish you all the best.

To *Mr. Bernie Terrado*, *Mr. Aldrich Co* and *Mr. Marvin Ignacio*, my thesis proposal and defense panel, thank you for all the knowledge you have shared with me during those events. You all have contributed into the improvement of my system and I appreciate them from the bottom of my heart.

To *Block 12 2010*, I love you all. Four years with you all was a priceless journey and I am hoping to spend the next century with you all again. You guys don't know how much you all mean to me.

To *Rizza Gonzaga*, thank you for everything. Thank you for tolerating my eclectic, unpredictable personality. Thank you for being my grandmother / big sister / best friend / partner-in-crime. The summer of 2012 would never be the same without you.

To *Ms. Estrellita Salorio*, I owe you a part of my life. You were the only one who pushed me out of my comfort zone and to take risks back then. You are one of the people who gave interest in who I am and what I have to offer. You'll never know the extent your motivation caused me after high school. You are always in my heart and I can't wait to show and discuss this with you in the future. Rest in peace, *Tchr. Lily*.