Breast Cancer Dataset Analytics

Kevin Daimi
Department of Electrical, Computer Engineering and Computer Science
University of Detroit Mercy
Detroit, USA
daimikj@udmercy.edu

Noha Hazzazi
Electrical Engineering and Computer Science
Howard University
Washington DC, USA
noha.hazzazi@howard.edu

***Abstract*—Breast cancer is a disease that causes the cells of the breast to uncontrollably grow. It is the most occurring cancer in females worldwide. The type of breast cancer is governed by to breast cells that turn into cancer. Breast cancer can begin in different parts of the breast including lobules, ducts, and connective tissue. The clinical prognostic (the likelihood or expected development of a disease) stage depends on a number of factors including tumor size, lymph node status, whether the cancer has spread to other parts of the body, the cancer grade, Estrogen status, and Progesterone status. In this paper, the clinical prognostic stage (referred to as 6^{th} Stage in this study) will be predicted using both a Python program and the Weka tool. The three algorithms, Neural Networks, Support Vector Machines, and Random Forest will be applied to the SEER Breast Cancer Dataset to classify the 6^{th} Stage, which includes five classes.**

Keywords–Classification; Breast Cancer; Neural Networks; Support Vector Machines; Random Forest; Python; Weka.

I. INTRODUCTION

Breast cancer kicks off when cells in the breast start to grow out of control. These out of control cells usually develop a tumor that can be observed with an x-ray or sensed as a lump. The tumor is characterized as malignant (cancer) if the cells grow into surrounding tissues or spread to distant areas of the body. Breast cancer spreads when the cancer cells move to the blood or lymph system and are transferred to other parts of the body. Breast cancer occurs mainly in women, but men can also get it. Most breast cancers are introduced in the ducts that carry milk to the nipple (ductal cancers). Others develop in the glands that produce breast milk (lobular cancers) [1]. Breast cancer can span different ages. It is rare in very young women, but it has distinctive aspects that are not spotted in older patients. Young age breast cancer has aggressive biological characteristics and is liable to be diagnosed at an advanced stage providing poorer outcomes as compared to breast cancer in older premenopausal and postmenopausal women [2]. Cancer treatment is costly, especially in developing and under developing countries. Cost-effectiveness analyses can offer valuable information for planning and developing a breast cancer control policy. Such analyses can drive budget development, substantiate allocation of limited resources to national breast cancer control programs, improve breast cancer education, and pinpoint the most effective approaches of carrying out diagnostic and treatment services [3]. There are a number of breast cancer detection techniques. Among these are the mammographic images analysis. Kashyap, Bajpai, and Khanna [4] proposed a method to segment and classify abnormalities found in mammograms. They concluded that the removal of improved preprocessed and improved inverted preprocessed image enhances the detection of suspicious region in mammograms. Furthermore, edges of the skeptical region are sharpened by including thorough coefficients of the wavelet decomposition with the filtered image. Nugroho, Faisal, Soesanti, and Choridah [5] attempted to implement and analyze the contrast enhancement and feature selection technique to build a CAD(Computer Aided Design) system to differentiate normal, benign, and malignant. Preprocessing was required to improve the poor quality of images and remove the pieces added by the preprocessing step. The region of interest (suspicious area) was segmented and then extracted by texture feature method. The high dimensionality of features was identified by a feature selection technique. The digital mammogram images were taken from the Private Database of Oncology Clinic Kotabaru Yogyakarta. The dataset involved 40 mammogram images with 14 benign cases, 6 malignant cases, and 20 normal cases. Further mammograms analysis approaches can be found in [6]–[9]. Wang [10] provided an overview of recent advances in microwave sensors for biomedical imaging applications focusing on breast cancer detection. The electric characteristics of biological tissues at microwave spectrum, microwave imaging approaches, microwave biosensors, and current challenges in the field were also covered.

Breast cancer detection, diagnosis, and analysis paved the path for many machine learning applications. Gupta and Gupta [11] applied the machine learning techniques; Linear regression, Random Forest, Multi-layer Perceptron, and Decision Trees (DT) to the Wisconsin Breast Cancer Dataset. This dataset is made up of 569 samples (rows). Only the performance measures including accuracy, recall, and precision were summarized. No details for further conclusions were provided. Agarap [12] compared the performance of Linear regression, Multi-layer Perceptron, K-Nearest Neighbor (KNN), Softmax Regression, and Support Vector Machines (SVM) when applied to the Wisconsin Breast Cancer Dataset. The tool used was not specified and no details were provided. It is not that clear how the paper concluded that SVM performed the best. The two studies above involved a binary classification (benign or malignant tumor) of breast cancer tumor. Binary classification of the breast cancer using Naive Bayes and K-Nearest Neighbor was carried out in [13]. Their results showed that KNN achieved the highest accuracy of 97.51% with the lowest error rate (96.19%) than Naive Bayes. No details were provided apart from the well-known details of the algorithms. Bazazeh and Shubair [14] studied breast cancer binary classification using Support Vector Machines, Random

 ISBN: 978-1-61208-816-7

Forest, and Bayesian Networks. They also applied them to the Wisconsin Breast Cancer Dataset. They concluded Bayesian Networks had the best performance. A hybridized classifier for breast cancer diagnosis was proposed in [15]. They combined Self-Organizing Maps (unsupervised artificial neural network) method with the supervised classifier Stochastic Gradient Descent (SGD) to perform binary (cancer/no cancer) classification on the Wisconsin Breast Cancer Dataset. They compared the results of this combination with the outcomes of Decision Trees (DTs), Random Forests (RF) and Support Vector Machine (SVM). They concluded their combination resulted in excellent accuracy. Another approach was based on first using image processing techniques to prepare the mammography images for the feature and pattern extraction phase, and then feeding the extracted features to Back Propagation Neural Networks (BPNN) and Logistic Regression (LR) models [16]. They concluded BPNN performed the best. Similar approaches with different algorithms were proposed in [17]–[19]. A number of studies concentrate on using Convolutional Neural Networks for the analysis of breast tumors. Pawar and Patil [20] used Backpropagation Neural Network and compared the results with Radial Basis Function Network. Using the Wisconsin Breast Cancer Dataset and relying on MATLAB, it was concluded that a neural network with nine neurons in the hidden layer provided an accuracy of 99%. Convolutional Neural Networks was applied to the detection of breast cancer using Mammograms-MIAS dataset with 322 mammograms in which 189 images were normal and 133 abnormal breasts tumors [21]. The authors stressed that the experimental results depicting the efficacy of deep learning for breast cancer detection in mammogram images was promising and suggested using deep learning for various medical imaging. Further work on applying Artificial Neural Networks, Convolutional Neural Networks, and both Convolutional Neural Networks and Support Vector Machines to classifying breast cancer as either cancerous (malignant), non-cancerous (benign) could be found in [22]–[24].

All of the above studies concentrated on binary classification. In other words, they aimed at determining whether a tumor is benign or malignant. In general, they relied on the Wisconsin Breast Cancer Dataset with 569 samples. In this paper, a multi-class classification using the SEER Breast Cancer Dataset [25] with 4024 samples will be implemented. Members of IEEE can download this dataset. The attribute that will be the goal of this classification is the 6th Stage (S_Stage). It has five classes, as explained in Section II. Three classification methods were used: Neural Networks, Support Vector Machines, and Random Forest. Those were first included in a Python program and then run through the Weka tool [26]. Analysis of the outcomes are then provided. The remainder of the paper is organized as follows: Section II describes the SEER Breast Cancer Dataset and its preparation. Section III deals with the classification of the 6th Stage using a Python program. The classification of the 6th Stage using Weka is presented in Section IV. Section V deliberates on possible predictions on the SEER Dataset. The discussion of the outcomes is depicted in Section VI, and the paper is concluded in Section VII.

II. SEER BREAST CANCER DATASET

The Breast Cancer Dataset contains fifteen attributes (columns) and 4024 rows. Some attributes have been renamed for programming purpose. Those are enclosed in parentheses below.

A. Dataset Description

The attributes used in dataset are described below. The type of data and the values they can take are also stated. A sample of this dataset is presented in Table **??**. The rows appear as columns.

1) *Age: This represents the age of the patient. It is a continuous numerical attribute.*
2) *Race: A nominal attribute that has three values: White, Black, and Other (American Indian / AK Native, Asian / Pacific Islander) (See table 9).*
3) *Marital Status (M_Status): A nominal attribute with five different values: Married (including common law), Divorced, Single (never married), Widowed, and Separated (See table* **??***).*
4) *T-Stage: The letter "T" followed by a number (1-4) refers to the size and location of the tumor including how much the tumor has grown into nearby tissues. This Nominal variable takes the values; T1, T2, T3, and T4 (See table* **??***) [27].*
5) *N-Stage: The letter "N" followed by a number (1-3) stands for lymph nodes. Most often, the more lymph nodes with cancer, the larger the number assigned. N1, N2, and N3 are the possible values for this nominal attribute (See table* **??***).*
6) *6 th Stage (S_Stage): This nominal attribute takes the values IIA, IIB, IIIA, IIIB, and IIIC in this database. There are other values that are not included. The values (stages) of this attribute are based on a number of factors including type (invasive/inflammatory) of breast cancer, tumor found, tumor size, breast cancer cells found in the lymph nodes, number of auxiliary lymph nodes that the cancer spread to, number of lymph nodes near the breastbone that the cancer spread to, cancer spreading to the chest wall and/or skin of the breast, and redness/swelling/ulcer/warmness in large portion of the breast skin. Details of this categorization can be found in [27]. This attribute represents the class for this study (See table* **??***)*
7) *Grade: The grade refers to the amount of cancer cells that look like healthy cells when observed under a microscope. There are four grades (I-IV) for this nominal attribute: Well differentiated, Moderately differentiated, Poorly differentiated, and Undifferentiated; anaplastic (See table* **??***).*
8) *A Stage (A_Stage): The A Stage has two values: 'Reginal' indicating a neoplasm that has spread directly into surrounding organs or tissues, and 'Distant' indicating a neoplasm has spread to parts far from the primary tumor (See table* **??***).*
9) *Tumor Size (Tumor_Size): represents the size of the tumor in centimeters. It is a continuous numerical attribute.*
10) *Estrogen Status (E_Status): This nominal attribute has two values, positive if the breast cancer has*

 ISBN: 978-1-61208-816-7

estrogen receptors, and negative otherwise. These receptors are proteins that allow normal and some cancerous breast cells to grow (See table **??***).*

11) *Progesterone Status (P_Status): This nominal attribute has two values; positive if the breast cancer has progesterone receptors, and negative otherwise. Progesterone receptors enable normal and some cancerous breast cells to grow.*
12) *Regional Node Examined (RN_Exam): Represents the total number of regional lymph nodes that were removed and examined by the pathologist.*
13) *Regional Node Positive (RN_Pos): A continuous value that reflects the exact number of regional lymph nodes examined by the pathologist and found to contain metastases (development of secondary malignant growths at a distance from a primary site of cancer).*
14) *Survival Months (S_Months): A continuous attribute indicating the number of months a patient will survive.*
15) *Status: The status of the patient. The nominal attribute has two values; Dead and Alive.*

TABLE 1. SAMPLE DATASET

Attribute	Row1	Row2	Row3
Age	43	67	58
Race	Other	White	Black
M_Status	Married	Divorced	Widowed
T_Stage	T2	T2	T1
N_Stage	N3	N1	N1
S_Stage	IIIC	IIB	IIA
Grade	II	III	II
A_Stage	Regional	Regional	Regional
Tumor_Size	40	25	11
E_Status	Positive	Positive	Positive
P_Status	Positive	Positive	Positive
RN_Exam	19	4	16
RN_Pos	11	1	1
S_Months	1	2	9
Status	Alive	Dead	Alive

B. Dataset Preparation

The preparation and cleaning of the dataset went through a number of steps. These steps are explained below. Some sample Python code will be shown.

1) *The Status column that has values "Dead" and "Alive" was completely removed. This has no impact on the classification.*

```
BC= pd.read\_csv(`BreastCancer2.csv',
    ↪ encoding=``latin-1")

BC1=BC.drop([`Status'], axis=1)
BC1.to\_csv(`BreastCancer3.csv')
```

2) *In this step, the nominal values were replaced by numbers as in the following Tables. The Python code for T_Type will be shown. The rest have similar code.*

Note that Estrogen Status and Progesterone Status values are similar. This should explain the absence of Progesterone Status values table.

```
initialization;
with (open('BreastCancer4.csv', 'w')) as predictfile:
  writer = csv.writer(predictfile, delimiter=',') x = 1
while x < 3409 do
    instructions;
    if lines[x][2] == 'White': then
        lines[x][2] = '1';
        if lines[x][2] == 'Black': then
            1
        else
            i
        end
        nes[x][2] = '2';
    else
        if lines[x][2] == 'Other (American Indian/AK
          Native, Asian/Pacific Islander)': then
            1
        else
            i
        end
        nes[x][2] = '3';
    end
    x = x +1
end
```

Algorithm 1: Python code for T_Type

TABLE 2. RACE VALUES

Race	Value
White	1
Black	2
Other	3

TABLE 3. MARTIAL STATUS VALUES

M_Status	Value
Married	1
Divorced	2
Single	3
Widowed	4
Separated	5

TABLE 4. T-STAGE VALUES

T-Stage	Value
T1	1
T2	2
T3	3
T4	4

TABLE 5. N-STAGE VALUES

N-Stage	Value
N1	1
N2	2
N3	3

TABLE 6. 6^{th}-STAGE VALUES

Grade	Value
IIA	1
IIB	2
IIIA	3
IIIB	4
IIIC	5

TABLE 7. GRADE VALUES

Grade	Value
Grade I: well differentiated	1
Grade II: moderately differentiated	2
Grade III: poorly differentiated	3
Grade IV: undifferentiated	4

C. Understanding the Datasets

To better understand the dataset, the method "describe" was used as below. Table 10 provides insight into the dataset.

TABLE 8. A-STAGE VALUES

A-Stage	Value
Regional	1
Distant	2

TABLE 9. ESTROGEN STATUS VALUES

E_Status	Value
Positive	1
Negative	2

Percentile 25, 50, and 75 have been omitted in this table. Note that the actual maximum for Age, Tumor Size. Regional Node Examined, Regional Node Positive, and Survival months are 69, 140, 61, 41, and 107, respectively.

TABLE 10. DATASET STATISTICS

Attribute	Mean	STD	Min	Max
Age	53.968361	8.968991	30.000000	69.00000
Race	1.231440	0.580434	1.000000	3.000000
M_Status	1.646986	1.010138	1.000000	5.000000
T_Stage	1.783259	0.764173	1.000000	4.000000
N_Stage	0.693169	0.693169	1.000000	3.000000
S_Stage	2.320877	1.266084	1.000000	5.000000
Grade	2.150722	0.638458	1.000000	4.000000
A_Stage	1.022920	0.149666	1.000000	2.000000
Tumor Size	30.40059	20.95136	1.000000	140.0000
E_Status	1.067015	0.250080	1.000000	2.000000
P_Status	1.173642	0.378849	1.000000	2.000000
RN_Exam	14.358744	8.095945	1.000000	61.00000
RN_Pos	4.157200	5.110535	1.000000	46.00000
S_Months	71.286746	22.926034	1.000000	107.0000

```
BCS= pd.read\_csv(`FullNormalizedBC.csv')\\
df=BCS.describe(include = `all')\\
tp = dict(df)\\
print( tp, ``\n")
```

III. CLASSIFYING 6TH STAGE USING PYTHON

The 6th Stage (S_Stage) will be classified using Neural Network (MLPClassifier), Support Vector Machine (SVC), and Random Forest techniques. The dataset contains 1303 rows for class 1, 1126 for class 2, 1049 for class 3, 66 for class 4, and 470 for class 5. 70% of the dataset is used for training, and 30% for testing for both Sections III and IV. The resulting classification model will be saved and used to classify unseen data.

A. S_Stage Classification Using Neural Networks

After executing training and testing on the dataset, the resulting model is saved and then used to classify the 6th Stage for a dataset of 10 rows that does not contain values for S_Stage (no column existed for this attribute). The simple code used will be depicted below. It applies to all the methods with the expectation of fitting the method to the training data. Therefore, only this part of code will be shown for the other methods in B and C below.

```
# Fitting Linear Regression to the Training set
   from sklearn.neural_network import
       ↪ MLPClassifier
lm = MLPClassifier(solver=`lbfgs', alpha=1e-5,
    ↪ hidden_layer_sizes=(150, 10), random_state
    ↪ =1)
lm.fit(X_train, y_train)
#Testing
classifications = lm.predict(X_test)
```

```
# Saving model to disk to predict unkown 6\
    ↪ textsuperscript{th} Stage
pickle.dump(lm, open(`NN_BC_model.pkl',`wb'))
# Loading the model
model = pickle.load(open(`NN_BC_model.pkl',`rb'))
```

Table 11 depicts classified S_Stage values for ten randomly selected rows of the test data together with the actual values of this attribute in the same rows of the test data. The abbreviations NN, SVM, and RF will be used to denote Neural Networks, Support Vector Machines, and Random Forest, respectively.

TABLE 11. ACTUAL AND PREDICTED VALUES OF S_STAGE

Row #	Actual	NN-Pred	SVM_Pred	RF_Pred	
1213	1	1	1	1	
3181	3	2	1	3	
3228	3	3	2	3	
560	1	1	1	1	
1707	2	2	2	2	
1624	2	1	2	2	
3223	1	1	1	1	
102	2	2	2	2	
2719	5	5	3	5	
906	5	5	3	5	height

The accuracy of the NN classifier on training set is 0.79, and the accuracy of the NN classifier on test set is 0.79. The Confusion Matrix is given below. It summarizes the results of classifications based on the test data. Each column represents the classifications for one of the classes. It is obvious there is a problem with class 4. The dataset has only 67 rows containing the value 4. Out of 135 class 5, only 126 were classified correctly (TP=126), and 9 incorrectly classified (FN=9). There were also 14 test data rows that were incorrectly classified as class 5 (FP=14). This will leave TN=1056.

$$\begin{array}{c} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \end{array} \\ \begin{pmatrix} 338 & 43 & 0 & 0 & 0 \\ 53 & 189 & 91 & 0 & 3 \\ 1 & 19 & 301 & 0 & 5 \\ 0 & 0 & 21 & 0 & 1 \\ 1 & 2 & 11 & 0 & 126 \end{pmatrix} \end{array}$$

The classification report is given in Table 12 below. Note that *Accuracy* is the number of correct predictions divided by the total number of predictions and multiplied by 100 to get a percentage. *Recall* is the ratio of the total number of correctly classified positive examples divide by the total number of positive examples. High recall is an indication that the class is correctly identified. Dividing the total number of correctly classified positive rows by the total number of predicted positive rows provides *Precision*. High precision indicates examples identified as positive are in fact positive (This indicates small FP). *F1 Score* (or *F Measure*) is a measurement that includes both recall and precision. In other words, it is the weighted average of both. The F-Measure will always be close to the minimum of Precision and Recall. Finally, Support depicts the number of examples of the true response that lie in that class. The Python code to print all these is as follows:

```
print(`Accuracy of NN classifier on training set:
    ↪ {:.1f}'
.format(lm.score(X_train, y_train)))
```

```
print(`Accuracy of NN classifier on test set:
    ↪ {:.1f}'
.format(lm.score(X_test, y_test)))
print(``\n")
print(``Confusion Matrix and classification
    ↪ Report for NN", ``\n")
print(confusion_matrix(y_test, classifications),
    ↪ ``\n")
print(classification_report(y_test,
    ↪ classifications), ``\n")
```

TABLE 12. NN CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
1	0.86	0.89	0.87	381
2	0.75	0.56	0.64	336
3	0.71	0.92	0.8	326
4	0	0	0	22
5	0.93	0.9	0.92	140

Finally, the models were applied to ten rows that have not been used before with either training or testing step. These ten rows are shown in Table 13. The actual classes that were not provided to the models are: 5, 2, 1, 2, 3, 1, 5, 2, 2, 4.

TABLE 13. UNSEEN DATA FOR CLASSIFYING S_STAGE

Dataset Rows												
[53	1	1	4	3	2	1	043	1	1	13	10	034]
[49	1	1	2	1	2	1	035	1	1	14	2	070]
[46	1	3	1	1	2	1	013	1	1	9	2	093]
[65	1	1	2	1	2	1	035	1	1	7	1	093]
[62	1	1	3	1	3	1	120	1	2	7	1	086]
[52	1	1	1	1	2	1	014	1	1	10	2	104]
[53	1	1	3	3	3	1	140	1	1	41	15	051]
[53	1	2	2	1	2	1	035	1	1	14	1	064]
[60	3	1	2	1	1	1	023	1	1	13	3	074]
[62	1	2	4	2	2	1	140	1	1	9	8	089]

The classifications for the three methods obtained by running the three models are provided in Table 14. Note that row 1 in Table 14 represents the classification for S_Stage using NN, SVM, and RF methods for row 1 of Table 13, and so on.

B. S_Stage Classification Using Support Vector Machines

For this method, the actual and predicted values using the test data are given in Table 11 (columns 2 and 4). The predictions of unseen data could be found in Table 14, column 3. Here, only the code for fitting the model will be shown. The rest is similar to code of Section A above.

```
from sklearn import svm
from sklearn.svm import SVC
lm= svm.SVC()
lm.fit(X_train, y_train)
SVC(C=1.0, cache_size=200, coef0=0.0, degree=3,
    ↪ decision_function_shape=`ovo', kernel=`rbf
    ↪ ', max_iter=-1, shrinking=True,
tol=0.001, verbose=False)
```

TABLE 14. S_STAGE CLASSIFICATIONS FOR UNSEEN DATA

For Row #	NN	SVM	RF
1	5	3	5
2	3	2	2
3	1	1	1
4	2	2	2
5	3	3	3
6	1	1	1
7	5	3	5
8	2	2	2
9	2	2	2
10	3	3	4

TABLE 15. SVM CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
1	0.93	0.81	0.87	381
2	0.84	0.73	0.78	336
3	0.49	0.86	0.62	326
4	0	0	0	22
5	1	0.01	0.01	140

The ac curacies of SVM classifier on training and testing sets are 1.0 and 0.69, respectively. The Confusion Matrix, and the classification report are listed below. Class 5 has only 1 correct classification (TP=1) and 139 incorrectly classified (FP=139).

$$\begin{array}{c} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \end{array} \\ \begin{pmatrix} 309 & 12 & 60 & 0 & 0 \\ 53 & 244 & 83 & 0 & 0 \\ 1 & 30 & 281 & 0 & 0 \\ 0 & 3 & 19 & 0 & 0 \\ 0 & 3 & 136 & 0 & 0 \end{pmatrix} \end{array}$$

C. S_Stage Classification Using Random Forest

Predictions for test data, and predictions for new data (Table 13) are depicted in Table 11 (column 5), and Table 14 (column 4) ,respectively. The Python code is given below. The accuracy of RF classifier on training and testing sets are 1.0 and 1.0 ,respectively.

```
# Fitting Random Forest to the Training set
from sklearn.ensemble import
    ↪ RandomForestClassifier
lm=RandomForestClassifier(n_estimators=1000,
    ↪ max_depth=10,
random_state=0)
lm.fit(X_train, y_train)
```

The Confusion Matrix is demonstrated below. Table 16 shows the Random forest Classification Report. Here, Class 5 has TP=140 with no FN and FP.

$$\begin{array}{c} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \end{array} \\ \begin{pmatrix} 381 & 0 & 0 & 0 & 0 \\ 0 & 336 & 0 & 0 & 0 \\ 0 & 0 & 326 & 0 & 0 \\ 0 & 0 & 0 & 22 & 0 \\ 0 & 0 & 0 & 0 & 140 \end{pmatrix} \end{array}$$

TABLE 16. RANDOM FOREST CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
1	1	1	1	381
2	1	1	1	336
3	1	1	1	326
4	1	1	1	22
5	1	1	1	140

IV. CLASSIFICATION WITH WEKA

The same Breast Cancer dataset is used for classification using Weka. Neural Networks (Multi-Layer Perceptron), Support Vector Machines (SMOreg), and Random Forest algorithms are adopted. The dataset was split into 70% for training and 30% for testing as was the case in Section III above. The values of S_Stage were converted from numeric to nominal using *NumbericToNominal* filter to get the Confusion Matrix and the classification report for both Neural Networks and

Random Forest. This was not allowed with SVM (SMOreg). The same data of Table 13 is used by the models to predict unseen examples. The Weka tables that are equivalent to Tables 11 and 14 are Tables 17 and 18 ,respectively and are given below. Some of the results of SMOreg are rounded to get whole numbers. The statistics for each model are presented in their respective subsections (A-C).

TABLE 17. ACTUAL AND PREDICTED VALUES OF S_STAGE

Row #	Actual	NN-Pred	SVM_Pred	RF_Pred
1	1	1	1	1
4	3	3	1	3
7	3	3	2	3
8	1	1	1	1
10	2	2	2	2
21	2	2	2	2
22	1	1	1	1
23	2	2	2	2
24	5	5	6	5
25	5	5	3	5

TABLE 18. S_STAGE CLASSIFICATIONS FOR UNSEEN DATA

For Row #	NN	SVM	RF
1	5	6	5
2	2	2	2
3	1	1	1
4	2	2	2
5	3	3	3
6	1	1	1
7	5	5	5
8	2	2	2
9	2	2	2
10	4	5	4

A. Neural Network Saistic Using Weka

The accuracy of NN on the testing set is 1.0. The classification report, Confusion Matrix and further statistics produced by Weka for NN are as below. Note that the *Mean Absolute Error* measures the average of the absolute errors in a set of predictions, *Root Mean Squared Error* is the square root of the average of squared differences between predictions and actual observations, *Relative Absolute Error* is the sum of the absolute differences between the predictions and actual observations divided by the sum of the absolute differences between the average of the observation and the observations, and *Root Relative Squared Error* is the square root of the sum of the squared differences between the predictions and actual observations divided by the sum of the squared differences between the average of the observations and the observations.

$$\begin{array}{c} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \end{array} \\ \begin{pmatrix} 321 & 0 & 0 & 0 & 0 \\ 0 & 335 & 0 & 0 & 0 \\ 0 & 0 & 326 & 0 & 0 \\ 0 & 0 & 0 & 18 & 0 \\ 0 & 0 & 0 & 0 & 143 \end{pmatrix} \end{array}$$

TABLE 19. NN CLASSIFICATION REPORT FROM WEKA

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	382
2	1.00	1.00	1.00	335
3	1.00	1.00	1.00	326
4	1.00	1.00	1.00	18
5	1.00	1.00	1.00	143

Mean Absolute Error 0.0018
Root Mean Squared Error 0.0038
Relative Absolute Error 0.6114%
Root Relative Squared Error 0.9792%

B. Support Vector Machine Statistics Using Weka

Weka provided the following statistics for SMOreg. It did not allow producing the Confusion Matrix and the classification report.

Mean Absolute Error 0.1938
Root Mean Squared Error 0.4630
Relative Absolute Error 18.744%
Root Relative Squared Error 36.5768%

C. Random Forest Statistics Using Weka

The statistics and classification report provided by Weka for Random Forest are illustrated below.

$$\begin{array}{c} \\ \mathbf{1} \\ \mathbf{2} \\ \mathbf{3} \\ \mathbf{4} \\ \mathbf{5} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} \end{array} \\ \begin{pmatrix} 382 & 0 & 0 & 0 & 0 \\ 0 & 335 & 0 & 0 & 0 \\ 0 & 0 & 326 & 0 & 0 \\ 0 & 0 & 0 & 18 & 0 \\ 0 & 0 & 0 & 0 & 143 \end{pmatrix} \end{array}$$

TABLE 20. RF CLASSIFICATION REPORT FROM WEKA

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	382
2	1.00	1.00	1.00	335
3	1.00	1.00	1.00	326
4	1.00	1.00	1.00	018
5	1.00	1.00	1.00	143

Mean Absolute Error 0.0057
Root Mean Squared Error 0.0289
Relative Absolute Error 1.9552%
Root Relative Squared Error 37.5397%

V. BREAST CANCER DATASET PREDICTIONS

An attempt has been made to perform some predictions on the SEER Breast Cancer Dataset. These included predicting Survival Months and Tumor Size. Linear Regression resulted in a very high Mean Squared Error (372.01). This forced the normalization of all attributes' values to make them between 0 and 1. However, the results are also discouraging as can be seen in the tables below. Although, the errors look very small, but because the values of the attributes are very small, these errors are still large. Moreover, the comparison of actual and predicted values for both S_Months and Tumor_Size revealed large difference.

TABLE 21. MEAN SQUARED ERROR

Method	Mean Squared Error
Linear Regression	0.032551797301851120
Bayesian Ridge	0.032272116790858174
Support Vector Machine	0.032785887677874580

TABLE 22. MEAN SQUARED ERROR FOR SURVIVAL MONTHS

Method	Mean Squared Error
Linear Regression	0.005817131
Bayesian Ridge	0.005812664
Support Vector Machine	0.006662543

VI. CLASSIFICATION OUTCOME DISCUSSION

A. Discussion Based on Python Results

- For the classification using the test set (Table 11), SVM has four incorrect classifications, and NN has just two with one nearer to the actual value than SVM. Here, RF was the best followed by NN. SVM did not perform well enough.
- As mentioned above, the actual classes that were hidden from the three models are: 5, 2, 1, 2, 3, 1, 5, 2, 2, 4. Table 14 reveals that RF was able to get them all, NN missed two, and SVM missed three actual classes.
- By comparing the three confusion matrices for NN, SVM, and RF, it is obvious that RF has the highest TPs for class 1 (381) followed by NN (336). For class 2, RF was superior (336) and SVM followed with TP=224. RF is leading with TP=326 and then NN with TP=301. Both NN and SVM failed to grant any class 4 as TP, but RF got a TP count of 22. Remember, there are only 66 rows for class 4 in the dataset. Once more, RF leads for class 5. SVM correctly classified only one class 5.
- By comparing Tables 12, 15, and 16, it is perceived that RF is superior with regards to Precision, Recall, and F1-Score. NN and SVM did not perform well with class 4, but NN scored better with class 5.

B. Discussion Based on Weka Results

- The predicted values for SVM in Table 17 missed classifying four classes compared to the actual values during testing. Even worse, SVM produced a class value equal to 6, which does not exist in the dataset. Both NN and RF matched all the actual values.
- For the unseen dataset (Table 18), both NN and RF achieved all the values of the classes. However, SVM missed two values and supplied the value 6 for class 5, and 5 for class 4.
- By observing the confusion metrices for NN and RF, it is clear they both performed very well. They have equal TPs for all the classes without any FP, TN, and FN. Note that confusion matrices and classification reports are only issued by Weka when classifying Nominal attributes. SVM did not allow classification on nominal S_Stage, but only numeric S_Stage attribute. This should explain why they were not included in this discussion.
- The same applies to the classification reports of NN and RF. Precision, Recall, and F1-Score are all perfect.
- The Mean absolute Error and the Root Mean Squared Error for NN (0.0018 and 0.0038, respectively) are the smallest, and SVM has the highest errors of 0.1938 and 0.4630, respectively.

C. Discussion Based on Python and Weka Results

- For SVM, the Python program produced both the Confusion Matrix and classification report. This was not allowed in Weka.
- From Tables 11 and 17, SVM missed four classifications and introduced the value 6 which is not a valid class in the database. RF performed equally well using both Python and Weka. NN performed better with Weka.
- For the unseen data (Tables 14 and 18), SVM missed the most values (correct classes) using both Weka and Python program. However, it missed more classes with Weka. Class 6, which never exist was introduced in Weka but not with the Python Program. RF got the classification correct for both approaches, while NN missed two classes using the program and none with Weka.
- Using the confusion matrices, both NN and RF were able to get all TP values with no FP, FN, and TN values. However, with the Python program, RF almost achieved the same counts for all classes as in Weka with a slight difference not exceeding 3 with no FP, FN, and TN values.
- In Weka, NN performed well with all the classes, but it had an issue with class 4 using Python. RF performed equally well in both.

VII. CONCLUSION

Based on the discussion above, it can be concluded that Random Forest technique is the best for classification of the underlying Breast Cancer Dataset. However, more data and attributes are needed to provide even better classification. The analysis also revealed that Weka outperformed the Python program with regards to Neural Networks. It is further concluded that SVM did not perform as good as the other two techniques using this dataset and the selected attribute. Furthermore, further classification work could be carried out using other attributes including T_Stage, N_Stage, Grade, and A_Stage.

The results of applying prediction to Tumor Size and Survival Months were misleading and characterized by high prediction errors. However, analytics using prediction could be pursued if further data and attributes are added with the possibility of removing some of the nominal attributes.

REFERENCES

[1] "What is breast cancer? breast cancer definition," accessed on August 2020. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

[2] H.-B. Lee and W. Han, "Unique features of young age breast cancer and its management," vol. 17, no. 4, pp. 301–307. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4278047/

[3] M. T. Groot, R. Baltussen, C. A. Uyl-de Groot, B. O. Anderson, and G. N. Hortobagyi, "Costs and health effects of breast cancer interventions in epidemiologically different regions of africa, north america, and asia," vol. 12, pp. S81–S90. [Online]. Available: http://doi.wiley.com/10.1111/j.1075-122X.2006.00206.x

[4] K. L. Kashyap, M. K. Bajpai, and P. Khanna, "Breast cancer detection in digital mammograms," in 2015 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7294523/

 ISBN: 978-1-61208-816-7

[5] Hanung Adi Nugroho, Faisal N, Indah Soesanti, and Lina Choridah, "Analysis of digital mammograms for detection of breast cancer," in in Proc. the 2014 International Conference on Computer, Control, Informatics and Its Applications (IC3INA), 2014, pp. 25–29.

[6] R. Sangeetha and K. S. Murthy, "A novel approach for detection of breast cancer at an early stage using digital image processing techniques," in 2017 International Conference on Inventive Systems and Control (ICISC). IEEE, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/document/8068625/

[7] N. El Atlas, M. El Aroussi, and M. Wahbi, "Computer-aided breast cancer detection using mammograms: A review," in 2014 Second World Conference on Complex Systems (WCCS). IEEE, pp. 626–631. [Online]. Available: http://ieeexplore.ieee.org/document/7060995/

[8] B. Hela, M. Hela, H. Kamel, B. Sana, and M. Najla, "Breast cancer detection: A review on mammograms analysis techniques," in 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/6563999/

[9] T. Cahoon, M. Sutton, and J. Bezdek, "Breast cancer detection using image processing techniques," in Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000 (Cat. No.00CH37063), vol. 2. IEEE, pp. 973–976. [Online]. Available: http://ieeexplore.ieee.org/document/839171/

[10] L. Wang, "Microwave sensors for breast cancer detection," vol. 18, no. 2, p. 655. [Online]. Available: http://www.mdpi.com/1424-8220/18/2/655

[11] M. Gupta and B. Gupta, "A comparative study of breast cancer diagnosis using supervised machine learning techniques," in 2018 Second International Conference on Computing Methodologies and Communication (ICCMC). IEEE, pp. 997–1002. [Online]. Available: https://ieeexplore.ieee.org/document/8487537/

[12] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in Proceedings of the 2nd International Conference on Machine Learning and Soft Computing - ICMLSC '18. ACM Press, pp. 5–9. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3184066.3184080

[13] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). IEEE, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/8391453/

[14] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA). IEEE, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/document/7818560/

[15] D. Mittal, D. Gaurav, and S. Sekhar Roy, "An effective hybridized classifier for breast cancer diagnosis," in 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, pp. 1026–1031. [Online]. Available: http://ieeexplore.ieee.org/document/7222674/

[16] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast cancer detection using k-nearest neighbor machine learning algorithm," in 2016 9th International Conference on Developments in eSystems Engineering (DeSE). IEEE, pp. 35–39. [Online]. Available: http://ieeexplore.ieee.org/document/7930620/

[17] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in 2010 5th International Symposium on Health Informatics and Bioinformatics. IEEE, pp. 114–120. [Online]. Available: http://ieeexplore.ieee.org/document/5478895/

[18] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," vol. 13, pp. 8–17. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2001037014000464

[19] I. I. Esener, S. Ergin, and T. Yuksel, "A new ensemble of features for breast cancer diagnosis," in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, pp. 1168–1173. [Online]. Available: http://ieeexplore.ieee.org/document/7160452/

[20] P. S. Pawar and D. R. Patil, "Breast cancer detection using neural network models," in 2013 International Conference on Communication Systems and Network Technologies. IEEE, pp. 568–572. [Online]. Available: http://ieeexplore.ieee.org/document/6524463/

[21] S. Charan, M. J. Khan, and K. Khurshid, "Breast cancer detection in mammograms using convolutional neural network," in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). IEEE, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8346384/

[22] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," vol. 6, pp. 24 680–24 693. [Online]. Available: https://ieeexplore.ieee.org/document/8353225/

[23] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," vol. 7, p. e6201.

[24] M. H.-M. Khan, "Automated breast cancer diagnosis using artificial neural network (ANN)," in 2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS). IEEE, pp. 54–58. [Online]. Available: http://ieeexplore.ieee.org/document/8311589/

[25] J. Teng, "SEER breast cancer data," accessed on August 2020. [Online]. Available: https://ieee-dataport.org/open-access/seer-breast-cancer-data

[26] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufmann. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

[27] Stages of cancer. Accessed on August 2020. [Online]. Available: https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/stages-cancer

 ISBN: 978-1-61208-816-7