Supervised Machine Learning Algorithms To Predict Breast Cancer Diagnoses

Lisa Meng[1], MS; Shagun Gupta[1], MS

[1]University of Southern California
{lisameng, shagungu}@usc.edu

ABSTRACT

Background: Breast cancer is the second leading cause of death from cancer in women in the United States. Early detection is strongly correlated with better prognosis. Mammography is the first line of detection, but it is not perfect. Mammography has a false negative rate of 8-10% and it is difficult to determine the severity or type of cancer, as some abnormalities do not affect the woman's health. Therefore a biopsy is performed to confirm diagnosis. Fine-needle aspiration is the simplest and most common type of biopsy.

Objective: This study tests various supervised Machine Learning algorithms to determine which is most accurate in predicting breast cancer diagnoses with quantitative, continuous data.

Methods: The Wisconsin Breast Cancer dataset was used to train and test three Machine Learning models: Decision Trees, Neural Networks, and Support Vector Machines (with and without Principal Component Analysis). Each algorithm was implemented from scratch then validated with the Python scikit-learn library.

Results: All the presented Machine Learning algorithms performed proficiently (all approximately or excceding 90% test accuracy) on predicting diagnoses. From scratch, Support Vector Machines with PCA performed the best with 98.2% accuracy, Decision Trees performed as well with an accuracy of 95.3%, and Neural Networks attained an accuracy of 89.4%.

Conclusion: Three supervised Machine Learning algorithms were implemented and compared in order to discover which performed with the highest accuracy in predicting benign versus malignant diagnoses from fine-needle aspiration instances. Support Vector Machines with PCA performed the best with an accuracy of 98.2%.

KEY WORDS: Decision-Tree, Neural Networks, Principal Component Analysis, Support Vector Machines

INTRODUCTION

Currently, more than 3.5 million women in the United States have been diagnosed with breast cancer. In the U.S., breast cancer is the second leading cause of death from cancer in women, after lung cancer. More women are diagnosed with breast cancer than any other type, besides some skin cancers.[1] In 2020, it is estimated that approximately 30% of newly diagnosed cancers in women will be breast cancers, in which 276,280 cases will be invasive breast cancer and 48,530 cases will be non-invasive. This same year, 42,170 women in the U.S. are expected to die from breast cancer.[2]

Studies have shown significant correlation with diagnosing breast cancer at an early stage and increased recovery, suggesting that early detection greatly impacts prognosis.[3,4] According to the American Cancer Society, when breast cancer is detected early, and is in the localized stage, the 5-year relative survival rate is 99%. Therefore mammography, a noninvasive, low-dose x-ray procedure used to screen for abnormalities, is an important tool in the fight against breast cancer. However, mammography is not perfect.[5] Due to the difficulty in differentiating abnormal tissue in mammograms in women ages 40-50 years old, false positives (scans that appear abnormal in the absence of cancer) are most likely to occur on a woman's first mammogram because their breast tissue appear thicker than older women. Studies have used quantitative measurement to classify mammographic patterns and have consistently found that women with dense tissue in more than 60-75% of their breast are four to six times more at risk of breast cancer than those with no densities.[6] Therefore radiologists may misconstrue the appearance of thicker breast tissue on a mammogram for cancerous tissue. About 12% of women who are screened, radiologists determined their mammograms as abnormal, but only 5% of those women actually have cancer.[7] Therefore if an abnormality is detected on a mammogram, additional images, such as ultrasound or MRI, are then required. If the abnormality can not be ruled out from imaging studies, the next course of action is for the woman to undergo a biopsy procedure in order to remove a sample of breast tissue for further testing, as a biopsy is considered the "gold standard" because it is the only diagnostic procedure that can definitively diagnose if an abnormal area is cancerous.[8]

There are three types of biopsies to test for breast cancer: fine-needle aspiration, core-needle biopsy, surgical biopsy.[8] For this study, we will be focusing on fine-needle aspiration (FNA). FNA is elected if the lump is easily accessible, palpable, or the doctor suspects it to be a fluid-filled cyst.[8] FNA is performed by a physician with a small needle to obtain tissue samples and fluid from the solid or cystic breast lesions.[9] The cells from the samples are then examined by a pathologist under a microscope for cytological analysis in order to make a diagnosis.[10]

Due to the recent popularity of Artificial Intelligence (AI) and Machine Learning (ML), numerous studies have developed and tested several data mining and ML techniques for early breast cancer detection and classification.[11] Related studies have used the same dataset and implemented their own code of largely popular ML algorithms such as Random Forest,

k-Nearest-Neighbor, and Naïve Bayes.[12] Whereas other studies have used the same dataset and implemented an ensemble of algorithms using the Python scikit-learn library and Google TensorFlow in order to compare performances.[13,14,15] Studies have also been done with medical images, namely using deep learning algorithms to image process mammograms and slices of tissue from biopsies for classification.[16]

In order to determine which algorithm was optimal for breast cancer detection, our study executed three supervised learning algorithms: Decision Trees, Neural Networks, and Support Vector Machines (with and without Principal Component Analysis (PCA)). Each ML algorithm was implemented from scratch and validated with the Python scikit-learn library.

METHODS

Data

This study used the Wisconsin Breast Cancer dataset that was obtained from the UCI Machine Learning Repository.[17] The data set was obtained clean and did not require any pre-processing. For the implementation of the ML algorithms, the dataset was partitioned: 70% for the training phase and 30% for the testing phase. This partitioning was consistent for all algorithms for better comparison between the achieved results. The entire dataset consists of 30 variables total: 10 diagnostic features, each providing the mean, standard error, and "worst" (mean of the three largest values) calculations. The dataset also provided patients' ID number and the diagnosis of each case. The features of each case were computed from a digitized image of a FNA of a breast mass. In detail, the features were computed for each nucleus per image, which were manually segmented, and the mean, standard error, and "worst" were calculated over the range of isolated cells. These features describe characteristics of the cell nuclei present in the image.[17,18]

Specifically for our study, we only focused on the mean variables of the features, which are defined below[19]:

Id: Patient identification number (anonymized)
Diagnosis: Label. The diagnosis of breast tissues (M = malignant, B = benign)

1. **Radius_mean**: mean of the lengths of the radial line segments defined by the centroid of the tumor cell and the individual perimeter points
2. **Texture_mean**: standard deviation of gray-scale values in the component pixels
3. **Perimeter_mean**: mean circumference of the tumor cells
4. **Area_mean:** mean area of the tumor cells, which was measured by counting the number of pixels on the interior of the segmentation outline and adding one-half for each pixel that lays on the perimeter.

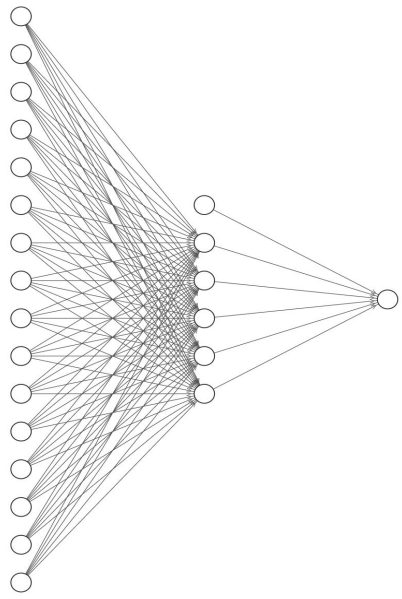
5. **Smoothness_mean**: mean of local variation in radius lengths, quantified by measuring the difference between the length of the radial line and the mean length of the lines surrounding it
6. **Compactness_mean**: mean of (perimeter^2 / area - 1.0) of each tumor cell. The value of this feature is minimized by a shape of a circular disk but increases with the irregularity of the cell boundary, in order to capture the measurement of "irregular".
7. **Concavity_mean**: mean of severity of concave portions/indentations of the contour of a cell nucleus. Concavity is calculated by counting the number of indentations, and drawing a line between non-adjacent perimeter points and measuring the extent to which the actual boundary of the nucleus lies on the inside of each line.
8. **Concave points_mean**: mean of the number of concave portions of the contour. Unlike Concavity, above, this feature only measures quantity, rather than magnitude, of contour cavities.
9. **Symmetry_mean**: Symmetry is measured by the length difference between lines perpendicular to the major axis (largest distance through the center of the tumor cell) to the cell boundary in both directions. The mean is then calculated by averaging the symmetry of individual cells per image.
10. **Fractal_dimension_mean**: mean for ("coastline approximation" - 1). Fractal dimension is an objective measurement of complex structures and allows for quantification of space-filling properties associated with the structure of interest. A higher value corresponds to a more irregular (complex) contour and thus to a higher probability of malignancy.[20] Fractal dimension is calculated by plotting the measurements of increasingly larger "rulers" on a log scale (as the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases) and evaluating the downward slope gives an approximation to the fractal dimension.

Algorithms

The following Machine Learning algorithms were applied to the above dataset to predict whether a tumor was Malignant or Benign. The hyper-parameters used for all the classifiers were manually assigned.

1. Neural Networks: Neural networks are a standard way of working with image classification data. However it is also used in combination with feature extraction where data is obtained or features are created by extracting data from the images like line segmentation, pixel distribution, radial features etc. Since the data is a direct measurement of the tumor features that would rather be recorded from a mammogram or cytological analysis, the problem statement was considered the same as an image classification problem with the feature extraction step implemented using PCA.

The neural network structure was finalised through experimentation. Due to availability of only 10 main features, a simple structure was implemented.



Input Layer Hidden Layer Output Layer

The most commonly used algorithm to implement neural networks is feedforward backpropagation algorithm. The algorithm in a stepwise manner iterates through the data row by row and for each iteration, considers each feature as a neuron in the input layer. The feedforward neural network is an acyclic network that only moves in the forward direction that is from the input to the hiffen to the output layers without any looping. Each layer receives information only from the preceding layer and forwards information to the succeeding layer. Every neuron converts the received information to information to be forwarded using an activation function. This normalises the received value in a range. The activation function used is the sigmoid function or the logistic function.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Layer l+1 receives information from Layer 1 using the following equation:

$a^{(l)} = \sigma(Wa^{l-1} + b)$ where $\sigma(x)$ is the activation function

The next step is the backpropagation step where the deltas or the error measures are calculated for Layer l using information received from Layer l+1

Summary: the equations of backpropagation

$$\delta^L = \nabla_a C \odot \sigma'(z^L) \tag{BP1}$$

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \tag{BP2}$$

$$\frac{\partial C}{\partial b^l_j} = \delta^l_j \tag{BP3}$$

$$\frac{\partial C}{\partial w^l_{j\,k}} = a^{l-1}_k \delta^l_j \tag{BP4}$$

After using the above equations to train the model by tuning parameters such as learning rate, number of epochs and the architecture of the neural network, the model is used to predict new sets of values that help in determining the performance of the model built. The higher the accuracy, the better the model.

2. SVM: Support Vector Machines are used to classify data in multi-dimensional space. It is based on the idea of finding a hyperplane that best separates the features into different domains. The role of an SVM model is to come up with an optimal solution that guarantees separation of data.

   The equation of the hyperplane is as follows:

$$\begin{aligned} y &= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \ldots . \\ &= w_0 + \sum_{i=1}^{m} w_i x_i \\ &= w_0 + w^T X \\ &= b + w^T X \end{aligned}$$

where W_i = vectors(W_0,W_1,W_2,W_3……W_m)
b = biased term (W0)
X = variable

Due to the availability of kernels that are not limited to linear separation, they fit the problem statement well. With many kernels available, we chose to apply the radial based kernel to our dataset which uses the following formula:

$$K(X_1, X_2) = exponent(-\gamma \|X_1 - X_2\|^2)$$

||X1 — X2 || = Euclidean distance between X1 & X2

SVM models can be tuned by tuning parameters such as C: Inverse of the strength of regularization and γ : Gamma (used only for RBF kernel). A model is built by finding the optimal hyperplane equation and using it to categorise new data points in the testing data.

3. Decision Trees: Decision Trees is a supervised learning method for classification tasks. Hence, it is used to predict and assign a class/category for new instances. The algorithm is given a set of classes and instances of each class. It then generates a model that when given a new instance will hypothesize its class. Decision Trees are most commonly used to classify discrete-valued data but our dataset deals with continuous values, therefore we have modified the general algorithm (that was taught in class) to fit our problem statement. When building Decision Trees, labelled data is required, thus it being a supervised learning technique. Decision Trees contain nodes (attribute-based decisions), branches (alternative values of the attribute), and leaves (each leaf is a class).

   A model is built by starting at the root node with the training dataset. An attribute that splits the dataset the best is selected and child nodes are created (i.e. splits more evening into subsets). When a node has all instances in the same class, it becomes a leaf node. This process is iterated until all nodes are leaves.

   In terms of mathematical logic, the criterion parameter for the information theory we assigned was Entropy. Entropy is the amount of uncertainty. It affects how the Decision Tree sets its boundaries. Entropy is used to calculate Information Gain, which is the difference in entropy between a parent node and a child node, thus measuring how much "information" a feature provides about a class. The attribute with the highest information gain will become the root nose and partition the dataset first. Both equations are shown below:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

   After using training data to build a model, when given a new test instance, a path is taken through the tree based on the value of its attributes. When a leaf is reached, that is the class assigned to the instance. The accuracy is calculated by total number of correct classifications divided by total number of instances.

4. PCA: A non-supervised feature extraction or dimensionality reduction technique used to build a new set of features that map not just the original data but also reflect the correlation between the various features. The algorithm can be broken down in the following steps:

   i. Take the whole dataset consisting of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.
   ii. Compute the mean for every dimension of the whole dataset.
   iii. Compute the covariance matrix of the whole dataset.
   iv. Compute eigenvectors and the corresponding eigenvalues.
   v. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W.
   vi. Use this d × k eigenvector matrix to transform the samples onto the new subspace.

   While not a necessary step, it is an accessory that has proven to improve the accuracy and performance of many machine learning algorithms. The algorithm is implemented only in combination with SVMs.

RESULTS

Neural Networks

Multiple models were created to experiment and understand the behaviour of the classification model. The labels were assigned as follows:

Malignant: 1
Benign: 0

The number of epochs were 2000 and weights were randomly initialised between -0.01 and 0.01. The data was normalised using the StandardScaler library in sklearn on the basis of mean values in each column. The parameters defined in the table below are mapped as follows:

Feature	**Abbreviation**
Radius_mean	RM
Perimeter_mean	PM
Area_mean	AM
Texture_mean	TM

Smoothness_mean	SM
Compactness_mean	CM
Concavity_mean	ConM
Concave points_mean	CPM
Symmetry_mean	SymM
Fractal_dimension_mean	FDM

Parameters used	**Learning Rate**	**Neurons in hidden layer**	**Accuracy (scratch)**	**Accuracy (library)**
RM, PM, AM, TM, SM, CM, ConM, CPM, SymM, FDM	1	5	82%	89%
RM, PM, AM, TM, SM, CM, ConM, CPM, SymM, FDM	1	3	88.23%	91%
RM, PM, AM, TM, SM, CM, ConM, CPM, SymM, FDM	0.1	3	77%	**92%**
RM, PM,AM, TM, SM, CM, ConM, CPM, SymM, FDM	0.1	5	85.88%	**92%**
RM, PM,AM, TM, SM, CM, FDM	1	3	**89.4%**	85%

Results for Row 4, when learning rate = 0.1 and number of neurons in hidden layer = 5 and all features are used:

```
testing accuracy  0.8588523529411764
[[104   3]
 [  4  59]]
              precision    recall  f1-score   support
```

```
         0.0       0.96      0.97      0.97       107
         1.0       0.95      0.94      0.94        63

    accuracy                           0.96       170
   macro avg       0.96      0.95      0.96       170
weighted avg       0.96      0.96      0.96       170
Accuracy: 0.92 (+/- 0.14)
```

The code implemented from scratch gave the best accuracy of 89.4% while the library gave the best accuracy of 92% when tested with 10-folds cross validation.

Decision Trees

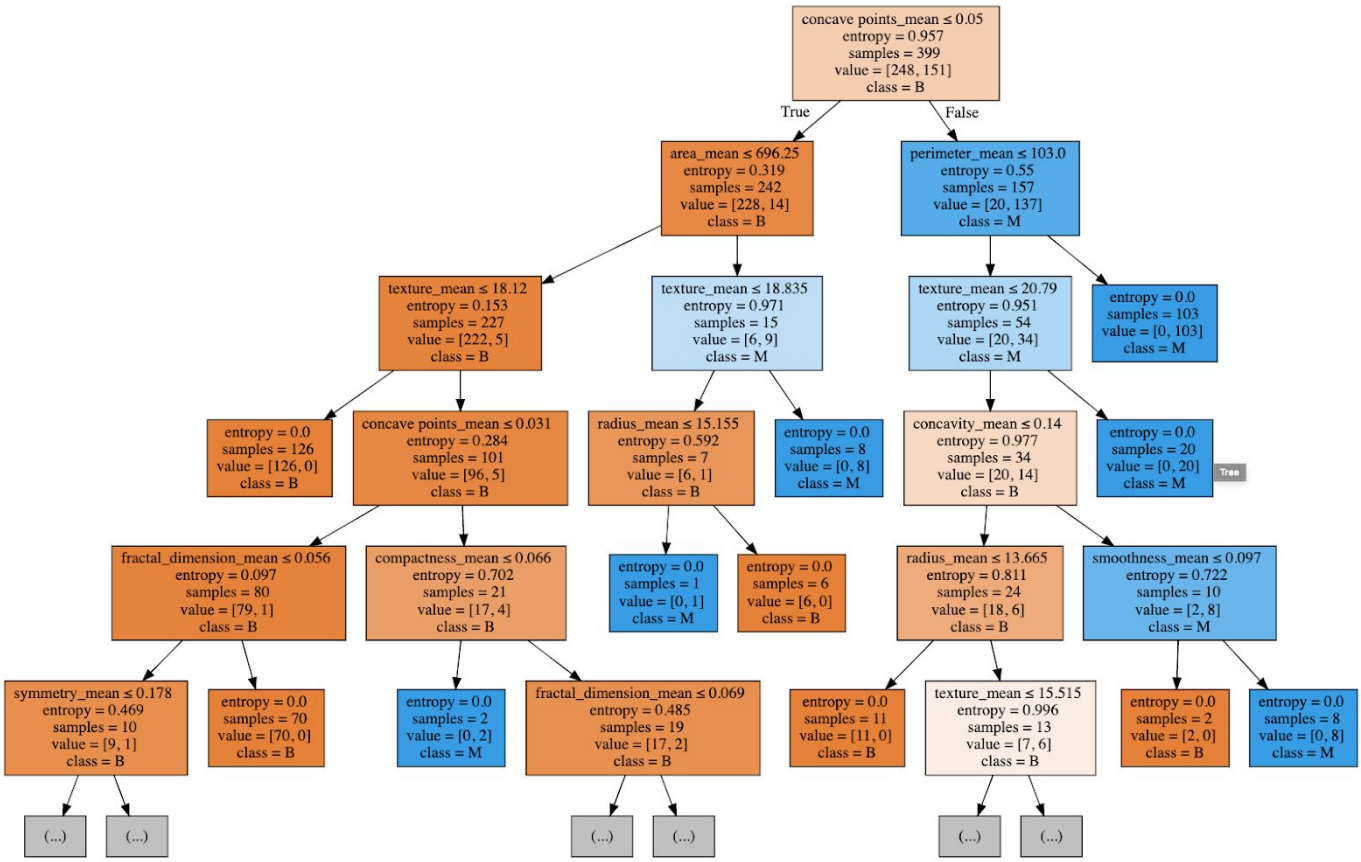
To be consistent, the labels were kept at 0 for Malignant and 1 for Benign.

Implementing from scratch, below is the construction of our Decision Tree, which yielded a 95.3% test accuracy.

```
Is concave points_mean <= 0.051019999999999996?
--> True:
  Is radius_mean <= 15.0?
  --> True:
    Is texture_mean <= 18.1?
    --> True:
      Predict {0.0: 127}
    --> False:
      Is area_mean <= 428.0?
      --> True:
        Predict {0.0: 45}
      --> False:
        Is smoothness_mean <= 0.08992?
        --> True:
          Is symmetry_mean <= 0.1773?
          --> True:
            Predict {0.0: 22}
          --> False:
            Is symmetry_mean <= 0.1784?
            --> True:
              Predict {1.0: 1}
            --> False:
              Predict {0.0: 5}
        --> False:
          Is compactness_mean <= 0.062879999999999999?
          --> True:
            Predict {1.0: 3}
          --> False:
            Is texture_mean <= 22.54?
            --> True:
              Predict {0.0: 13}
            --> False:
              Is radius_mean <= 13.38?
              --> True:
                Predict {0.0: 1}
```

```
              --> False:
                Predict {1.0: 2}
  --> False:
    Is texture_mean <= 19.46?
    --> True:
      Is radius_mean <= 15.61?
      --> True:
        Is texture_mean <= 13.21?
        --> True:
          Predict {0.0: 1}
        --> False:
          Predict {1.0: 3}
      --> False:
        Predict {0.0: 3}
    --> False:
      Predict {1.0: 5}
--> False:
  Is perimeter_mean <= 102.8?
  --> True:
    Is texture_mean <= 20.76?
    --> True:
      Is concave points_mean <= 0.08534?
      --> True:
        Is texture_mean <= 16.39?
        --> True:
          Predict {0.0: 11}
        --> False:
          Is smoothness_mean <= 0.09072999999999999?
          --> True:
            Predict {0.0: 4}
          --> False:
            Is concavity_mean <= 0.1065?
            --> True:
              Is radius_mean <= 13.05?
              --> True:
                Predict {0.0: 3}
              --> False:
                Is texture_mean <= 18.89?
                --> True:
                  Is concavity_mean <= 0.0858?
                  --> True:
                    Predict {0.0: 2}
                  --> False:
                    Predict {1.0: 2}
                --> False:
                  Predict {1.0: 4}
            --> False:
              Predict {1.0: 7}
      --> False:
        Predict {1.0: 7}
    --> False:
      Predict {1.0: 24}
  --> False:
    Predict {1.0: 104}
Accuracy: 95.29411764705881 %
```

To validate our algorithm, we also implemented a Decision Tree from the Python scikit-learn library. The library yielded an accuracy of 92% when testing with k-Means cross-validation (k=10) and the visualization is below:




```
Scores:
0.8823529411764706
0.8823529411764706
0.8823529411764706
0.9411764705882353
0.9411764705882353
0.8823529411764706
1.0
0.9411764705882353
0.9411764705882353
0.8823529411764706
Accuracy: 0.92 (+/- 0.08)
```

SVM

Multiple models were created to experiment and understand the behaviour of the classification model. The labels were assigned as follows:

Malignant: 1

Benign: -1

The data was normalised using the StandardScaler library in sklearn on the basis of mean values in each column. Sklearn's PCA class was applied to the normalised data before training on the SVM model. The PCA model was run for components ranging from 1 to 7 and the best performing model was selected.

Gamma	**C**	**Accuracy (scratch) (without PCA)**	**Accuracy (scratch) (with PCA) : no of components**	**Accuracy (library) (with PCA) : no of components**
0.063	1	94.35%	94.7% - 7	94.7% - 7
0.1	1	94.1%	93.6% - 5	93.6% - 5
0.1	2	92.9%	93.35% - 7	93.35% - 7
0.065	**2**	**97%**	**98.2% - 7**	**98.2% - 7**

```
 Without PCA
      pcost       dcost       gap    pres   dres
 0: -9.7325e+01 -1.7940e+03  6e+03  1e+00  4e-15
 1: -6.8605e+01 -7.2010e+02  8e+02  8e-02  4e-15
 2: -8.6169e+01 -2.1519e+02  1e+02  1e-02  4e-15
 3: -1.0168e+02 -1.5952e+02  6e+01  5e-03  4e-15
 4: -1.0917e+02 -1.3882e+02  3e+01  2e-03  4e-15
 5: -1.1464e+02 -1.2574e+02  1e+01  6e-04  4e-15
 6: -1.1707e+02 -1.2089e+02  4e+00  2e-04  4e-15
 7: -1.1788e+02 -1.1946e+02  2e+00  4e-05  4e-15
 8: -1.1839e+02 -1.1869e+02  3e-01  7e-06  4e-15
 9: -1.1850e+02 -1.1851e+02  8e-03  1e-07  4e-15
10: -1.1851e+02 -1.1851e+02  1e-04  1e-09  4e-15
11: -1.1851e+02 -1.1851e+02  3e-06  2e-11  4e-15
Optimal solution found.
Accuracy std (scratch): 0.9697058823529412
Accuracy : 0.9697058823529412
Components  1
     pcost       dcost       gap    pres   dres
 0: -1.6073e+02 -1.9082e+03  6e+03  1e+00  6e-15
 1: -1.2351e+02 -7.7018e+02  7e+02  4e-02  4e-15
 2: -1.4667e+02 -2.6385e+02  1e+02  6e-03  4e-15
 3: -1.6683e+02 -2.1276e+02  5e+01  2e-03  5e-15
 4: -1.7001e+02 -2.0480e+02  4e+01  1e-03  5e-15
 5: -1.7489e+02 -1.9203e+02  2e+01  5e-04  5e-15
 6: -1.7832e+02 -1.8501e+02  7e+00  1e-04  6e-15
 7: -1.7972e+02 -1.8256e+02  3e+00  3e-05  6e-15
 8: -1.8038e+02 -1.8149e+02  1e+00  1e-06  7e-15
```

```
 9: -1.8082e+02 -1.8097e+02  1e-01  1e-07  6e-15
10: -1.8088e+02 -1.8090e+02  2e-02  1e-08  6e-15
11: -1.8089e+02 -1.8089e+02  2e-04  1e-10  6e-15
Optimal solution found.
Accuracy (scratch): 0.9294117647058824
Accuracy : 0.9294117647058824
Components  2
     pcost       dcost       gap    pres   dres
 0: -1.2359e+02 -1.9715e+03  7e+03  1e+00  5e-15
 1: -8.9964e+01 -8.2158e+02  9e+02  7e-02  5e-15
 2: -1.0895e+02 -2.4113e+02  1e+02  1e-02  5e-15
 3: -1.2564e+02 -1.8619e+02  6e+01  4e-03  5e-15
 4: -1.3259e+02 -1.6841e+02  4e+01  2e-03  4e-15
 5: -1.3771e+02 -1.5600e+02  2e+01  8e-04  5e-15
 6: -1.4154e+02 -1.4825e+02  7e+00  2e-04  5e-15
 7: -1.4326e+02 -1.4544e+02  2e+00  2e-05  6e-15
 8: -1.4367e+02 -1.4473e+02  1e+00  5e-06  6e-15
 9: -1.4404e+02 -1.4421e+02  2e-01  9e-15  6e-15
10: -1.4412e+02 -1.4413e+02  1e-02  1e-14  6e-15
11: -1.4412e+02 -1.4412e+02  1e-04  3e-15  7e-15
Optimal solution found.
Accuracy (scratch): 0.9294117647058824
Accuracy : 0.9294117647058824
Components  3
     pcost       dcost       gap    pres   dres
 0: -1.1565e+02 -1.9626e+03  7e+03  1e+00  5e-15
 1: -8.2449e+01 -8.2032e+02  9e+02  8e-02  5e-15
 2: -1.0158e+02 -2.4616e+02  2e+02  1e-02  4e-15
 3: -1.2057e+02 -1.7577e+02  6e+01  4e-03  4e-15
 4: -1.2792e+02 -1.5502e+02  3e+01  1e-03  4e-15
 5: -1.3087e+02 -1.4915e+02  2e+01  8e-04  4e-15
 6: -1.3414e+02 -1.4166e+02  8e+00  2e-04  4e-15
 7: -1.3584e+02 -1.3861e+02  3e+00  1e-05  4e-15
 8: -1.3675e+02 -1.3733e+02  6e-01  4e-07  5e-15
 9: -1.3699e+02 -1.3703e+02  4e-02  3e-08  5e-15
10: -1.3701e+02 -1.3701e+02  1e-03  8e-10  5e-15
11: -1.3701e+02 -1.3701e+02  2e-05  1e-11  5e-15
Optimal solution found.
Accuracy (scratch): 0.9529411764705882
Accuracy : 0.9529411764705882
Components  4
     pcost       dcost       gap    pres   dres
 0: -1.1379e+02 -1.9617e+03  7e+03  1e+00  5e-15
 1: -8.0695e+01 -8.2553e+02  9e+02  8e-02  4e-15
 2: -9.9764e+01 -2.4673e+02  2e+02  1e-02  4e-15
 3: -1.1942e+02 -1.7489e+02  6e+01  4e-03  4e-15
 4: -1.2673e+02 -1.5583e+02  3e+01  2e-03  3e-15
 5: -1.3040e+02 -1.4637e+02  2e+01  5e-04  4e-15
```

```
 6: -1.3297e+02 -1.4110e+02  8e+00  2e-04  4e-15
 7: -1.3517e+02 -1.3689e+02  2e+00  6e-15  5e-15
 8: -1.3574e+02 -1.3609e+02  4e-01  5e-15  4e-15
 9: -1.3588e+02 -1.3590e+02  2e-02  3e-15  5e-15
10: -1.3589e+02 -1.3589e+02  3e-04  2e-15  5e-15
11: -1.3589e+02 -1.3589e+02  3e-06  4e-15  5e-15
Optimal solution found.
Accuracy (scratch): 0.9470588235294117
Accuracy : 0.9470588235294117
Components  5
     pcost       dcost       gap    pres   dres
 0: -1.0433e+02 -1.8868e+03  6e+03  1e+00  4e-15
 1: -7.3265e+01 -7.8098e+02  9e+02  9e-02  4e-15
 2: -9.0357e+01 -2.3680e+02  2e+02  1e-02  4e-15
 3: -1.0894e+02 -1.6564e+02  6e+01  4e-03  4e-15
 4: -1.1633e+02 -1.4672e+02  3e+01  2e-03  4e-15
 5: -1.2143e+02 -1.3518e+02  1e+01  6e-04  4e-15
 6: -1.2435e+02 -1.2905e+02  5e+00  1e-04  4e-15
 7: -1.2571e+02 -1.2665e+02  9e-01  5e-06  4e-15
 8: -1.2608e+02 -1.2615e+02  7e-02  4e-09  4e-15
 9: -1.2611e+02 -1.2611e+02  2e-03  1e-10  4e-15
10: -1.2611e+02 -1.2611e+02  3e-05  2e-12  4e-15
Optimal solution found.
Accuracy (scratch): 0.9705882352941176
Accuracy : 0.9705882352941176
Components  6
     pcost       dcost       gap    pres   dres
 0: -1.0134e+02 -1.8800e+03  6e+03  1e+00  4e-15
 1: -7.0566e+01 -7.7732e+02  9e+02  9e-02  4e-15
 2: -8.7258e+01 -2.2915e+02  2e+02  1e-02  4e-15
 3: -1.0601e+02 -1.5929e+02  6e+01  4e-03  4e-15
 4: -1.1495e+02 -1.3727e+02  2e+01  1e-03  4e-15
 5: -1.1794e+02 -1.3164e+02  1e+01  5e-04  4e-15
 6: -1.2118e+02 -1.2562e+02  4e+00  8e-05  5e-15
 7: -1.2248e+02 -1.2342e+02  9e-01  1e-14  5e-15
 8: -1.2279e+02 -1.2302e+02  2e-01  4e-16  4e-15
 9: -1.2289e+02 -1.2290e+02  3e-03  5e-15  4e-15
10: -1.2289e+02 -1.2289e+02  3e-05  1e-14  4e-15
Optimal solution found.
Accuracy (scratch): 0.9705882352941176
Accuracy : 0.9705882352941176
Components  7
     pcost       dcost       gap    pres   dres
 0: -9.7667e+01 -1.8192e+03  6e+03  1e+00  4e-15
 1: -6.8477e+01 -7.3570e+02  8e+02  8e-02  4e-15
 2: -8.5751e+01 -2.1730e+02  1e+02  1e-02  4e-15
 3: -1.0160e+02 -1.5994e+02  6e+01  5e-03  4e-15
 4: -1.0895e+02 -1.4007e+02  3e+01  2e-03  3e-15
```

```
 5: -1.1445e+02 -1.2675e+02  1e+01  8e-04  4e-15
 6: -1.1710e+02 -1.2129e+02  4e+00  1e-04  4e-15
 7: -1.1800e+02 -1.1976e+02  2e+00  3e-05  4e-15
 8: -1.1859e+02 -1.1886e+02  3e-01  4e-06  4e-15
 9: -1.1870e+02 -1.1871e+02  1e-02  2e-07  4e-15
10: -1.1870e+02 -1.1870e+02  2e-04  3e-09  4e-15
11: -1.1870e+02 -1.1870e+02  5e-06  4e-11  5e-15
Optimal solution found.
Accuracy (scratch): 0.9847058823529412
Accuracy : 0.9847058823529412
```

Results show that all the presented ML algorithms performed well (all approximately or excceding 90% test accuracy) on the classification task. The top performing algorithm implemented from scratch was SVM with PCA, reaching an accuracy of 98.2%. Decision Trees performed second best with an accuracy of 95.3%. Lastly, NN achieved an accuracy of 89.4%.

DISCUSSION

Our study presented three supervised Machine Learning algorithms: Decision Trees, SVM (with and without PCA), and Neural Networks. Comparing performances, SVM with PCA predicted diagnoses with the greatest accuracy. This study suggests that SVM with PCA is the optimal Machine Learning algorithm for predicting breast cancer diagnoses with fine-needle aspiration instances.

Ultimately, the results of this study highlight that Machine Learning is a useful tool that can be utilized to automate and standardize the process of diagnosing tissue samples from fine-needle aspiration biopsies. With an accuracy of 98.2%, our SVM with PCA algorithm can be implemented as a Computer-Aided Detection (CAD) system, also known as a decision support system, that will assist pathologists in deciding the correct diagnosis by recommending the predicted class (malignant vs. benign), in order to increase the pathologist's accuracy.

ACKNOWLEDGMENTS

We would like to express our deep gratitude to Professor Satish Thittamaranahalli Ka for his thorough lectures and his adaptability during this dire Covid-19 crisis. We would also like to thank TA Chaoyang He and Grader Shreyasi Chaudhary for their hard work and timeliness in grading the assignments. The dataset for this study was contributed by UCI Machine Learning Repository.[17]

CONTRIBUTIONS

Lisa and Shagun collectively discussed possible research topics for this paper and unanimously decided on predicting breast cancer diagnoses as our challenge problem. We also thoroughly examined all Machine Learning algorithms that were taught in this course and both agreed to use 3 supervised learning techniques to accomplish our classification task: Decision Trees, Neural Networks, and Support Vector Machines (with and without PCA), in order to compare and determine which algorithm is most proficient in predicting breast cancer diagnoses. Lisa researched the biology of breast cancer, breast cancer diagnostic tools/tests, and the medical workflow of diagnosing breast cancer in order to fully understand the dataset and the features, therefore she wrote the Background section. Shagun implemented the Neural Networks and SVM (with and without PCA) algorithms. Lisa implemented the Decision Trees algorithm. Please note that each algorithm was implemented from scratch and with a library for validation. This research paper was compiled by both Lisa and Shagun.

REFERENCES

1. https://www.cancer.net/cancer-types/breast-cancer/statistics. American Society of Clinical Oncology (ASCO). Breast Cancer: Statistics. 2020. Accessed 5/7/2020.
2. https://www.breastcancer.org/symptoms/understand_bc/statistics. Breastcancer.org. U.S. Breast Cancer Statistics. 2020. Accessed 5/7/2020.
3. Tabar L, Vitak B, Chen HH, Yen MF, Duffy SW, SmithRA. 2001. Beyond randomized controlled trials:Organized mammographic screening substantially reduces breast carcinoma mortality. Cancer 91:1724–1731.
4. Goldhirsch A, Colleoni M, Domenighetti G, GelberRD. 2003. Systemic treatments for women with breast cancer: Outcome with relation to screening for the disease. Ann Oncol 14: 1212–1214.
5. https://www.mayoclinic.org/tests-procedures/mammogram/expert-answers/mammogram-guidelines/faq-20057759. Mayo Foundation for Medical Education and Research (MFMER). Mammogram guidelines: What are they?. 2020. Accessed 5/7/2020.
6. Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ.1998. Mammographic densities and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 1998 Dec;7(12):1133-44.
7. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf. American Cancer Society. Cancer Facts & Figures 2020. 2020. Accessed 5/7/2020.
8. https://www.nationalbreastcancer.org/breast-cancer-biopsy. National Breast Cancer Foundation, Inc. Biopsy. 2020. Accessed 5/7/2020.
9. Casaubon JT, Regan JP. Fine Needle Aspiration Of Breast Masses. [Updated 2020 Feb 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470268/

10. Bukhari MH, Arshad M, Jamal S, Niazi S, Bashir S, Bakhshi IM, Shaharyar. Use of fine-needle aspiration in the evaluation of breast lumps. Patholog Res Int. 2011;2011:689521. doi: 10.4061/2011/689521. Epub 2011 Jun 21. PMID: 21789264; PMCID: PMC3135154.
11. Gayathri BM, Sumathi CP, Santhanam T, Vaishnav D. 2013. BREAST CANCER DIAGNOSIS USING MACHINE LEARNING ALGORITHMS -A SURVEY.
12. Sharma S, Aggarwal A, Choudhury T. Breast Cancer Detection Using Machine Learning Algorithms. 2018. International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.
13. Dhahri H, Maghayreh EA, Mahmood A, Elkilani W, Nagi MF. 2019. Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. Journal of Healthcare Engineering. doi: 10.1155/2019/4253641.
14. Agarap AFM. 2019. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. ICMLSC 2018, February 2–4, 2018, Phu Quoc Island, Viet Nam
15. Tahmooresi M, Afshar A, Rad B, Babak, Nowshath K, Bamiah M. 2018. Early Detection of Breast Cancer Using Machine Learning Techniques. Journal of Telecommunication, Electronic and Computer Engineering. 10. 21-27.
16. Shen L, Margolies LR, Rothstein JH et al. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep 9, 12495 (2019). https://doi.org/10.1038/s41598-019-48995-4
17. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29]. Irvine, CA: University of California, School of Information and Computer Science. Accessed 5/7/2020.
18. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Kaggle. Breast Cancer Wisconsin (Diagnostic) Data Set. 2016. Accessed 5/7/2020.
19. Street WN, Wolberg WH, Mangasarian OL. Nuclear Feature Extraction For Breast Tumor Diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, California, 1993.
20. Fuseler JW, Robichaux JP, Atiyah AI, Ramsdell AF. Morphometric and Fractal Dimension Analysis Identifies Early Neoplastic Changes in Mammary Epithelium of MMTV-cNeu Mice. Anticancer Research March 2014 vol. 34 no. 3 1171-1177.