JOURNAL OF
EMERGING INVESTIGATORS

Article

A comparative analysis of machine learning approaches for prediction of breast cancer

Shreya Nag and Jaydeep Nag
Plano West Senior High School, Plano, Texas

SUMMARY

One of the most dreadful diseases for women and their health is breast cancer. Breast cancer death rates are higher than those for any other cancer, aside from lung cancer. Machine learning and deep learning techniques can be used to predict the early onset of breast cancer. The main objective of this analysis was to determine whether machine learning algorithms can be used to predict the onset of breast cancer with more than 90% accuracy. Based on research with supervised machine learning algorithms, Gaussian Naïve Bayes, K Nearest Algorithm, Random Forest, and Logistic Regression were considered because they offer a wide variety of classification methods and also provide high accuracy and performance. We hypothesized that all these algorithms would provide accurate results, and Random Forest and Logistic Regression would provide better accuracy and performance than Naïve Bayes and K Nearest Neighbor. The Wisconsin Breast Cancer dataset from the UC Irvine repository was used to perform a comparison between the supervised machine learning algorithms of Gaussian Naïve Bayes, K Nearest Neighbor, Random Forest, and Logistic Regression. Based on the results, the Random Forest algorithm performed best among the four algorithms in malignant prediction (accuracy = 98%), and Logistic Regression algorithm performed best among the four algorithms in benign prediction (accuracy = 99%). All the algorithms performed well in the prediction of benign versus malignant cancer, with more than 90% accuracy based on their F1-score. The study results can be used for further research in prediction of cancer using machine learning algorithms.

INTRODUCTION

According to the National Breast Cancer Foundation, an estimated 276,480 new cases of invasive breast cancer will be diagnosed in women in the United States in 2020. Additionally, approximately 42,170 women in the U.S. are expected to pass away in 2020 due to breast cancer (1). Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women (1). Breast cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a tumor. A tumor does not mean cancer — tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as magnetic resonance imaging (MRI), mammogram, ultrasound, and biopsy are commonly used to diagnose breast cancer.

The early diagnosis and prognosis of a cancer type are a critical necessity in cancer research, as it can facilitate the subsequent clinical management and treatment of patients (2). One of the major challenges for medical science is the timely detection and diagnosis of breast cancer before it turns severe and requires emergency treatment. These treatments may be expensive and may not be always successful (2). The developments in computer sciences and technology, specifically in artificial intelligence (AI) and machine learning (ML), can help in addressing these challenges. Use of AI technology will result in reduction of costs in providing care, along with making the care faster and more efficient. The use of AI and ML technologies, will cause changes in the medical profession with greater focus on tasks related to creativity and critical thinking than time-consuming repetitive tasks (3). There have been many research studies which have suggested that AI techniques can perform as well as or better than humans at key healthcare tasks, such as diagnosing diseases. It has been observed that the ML algorithms are in various instances outperforming radiologists at spotting malignant tumors. The AI and ML technologies are also being used by researchers to create cohorts for clinical trials, which can be otherwise costly (3).

ML is the scientific study of algorithms and statistical models that computer systems utilize to improve their performance in completing a specific task. The science of ML is related to computational statistics, which specializes in making predictions using computers. Machine learning algorithms create a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task (4).

ML algorithms can broadly be categorized according to the purposes they are designed for. The primary categories of ML algorithms include: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Supervised learning is classified into two broad categories of algorithms. The first category is classification; a classification problem is defined when the output variable is a category, such as “malignant” or “benign”, “disease” or “no disease”. The second category is regression, which

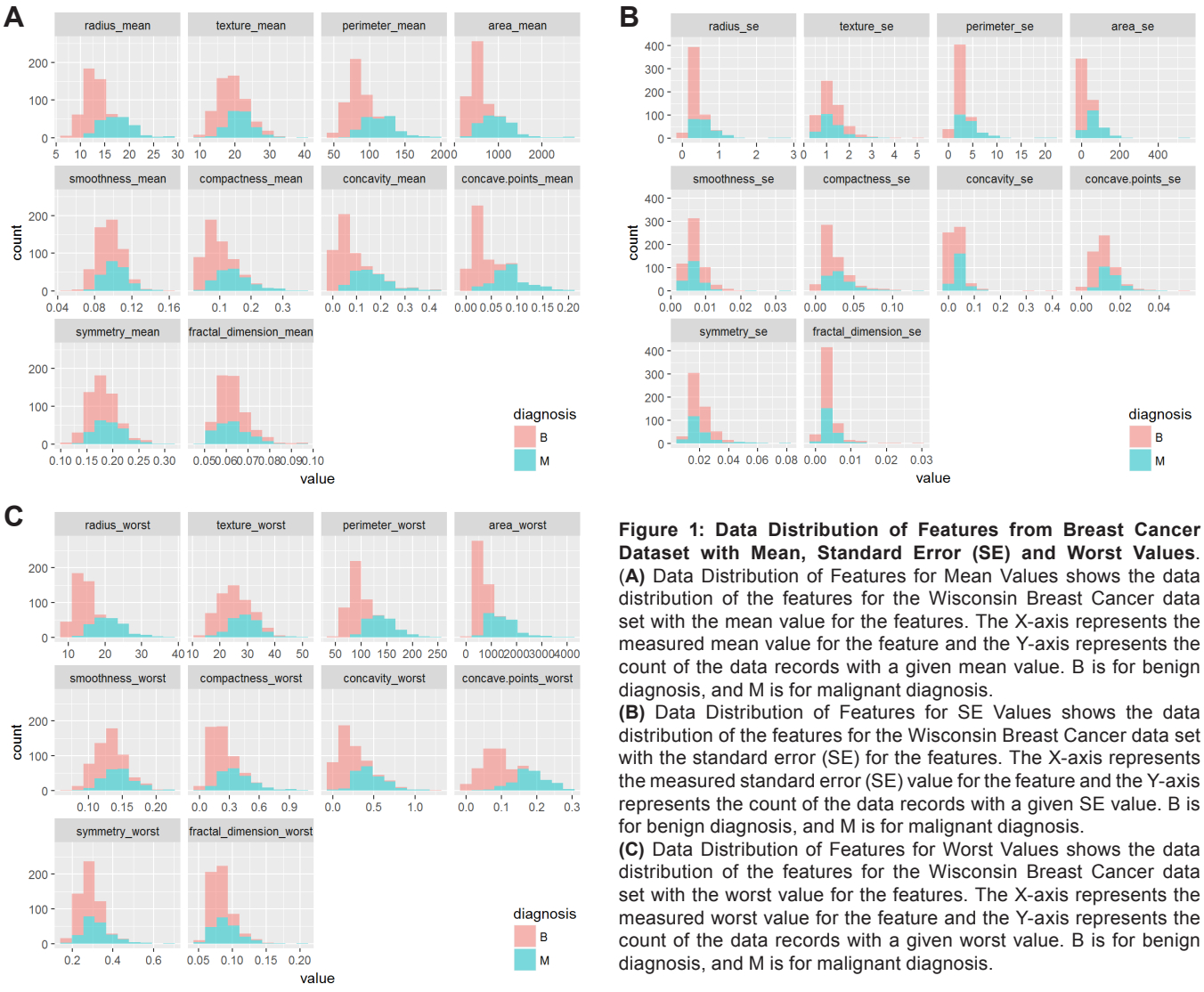

Figure 1: Data Distribution of Features from Breast Cancer Dataset with Mean, Standard Error (SE) and Worst Values. **(A)** Data Distribution of Features for Mean Values shows the data distribution of the features for the Wisconsin Breast Cancer data set with the mean value for the features. The X-axis represents the measured mean value for the feature and the Y-axis represents the count of the data records with a given mean value. B is for benign diagnosis, and M is for malignant diagnosis.
(B) Data Distribution of Features for SE Values shows the data distribution of the features for the Wisconsin Breast Cancer data set with the standard error (SE) for the features. The X-axis represents the measured standard error (SE) value for the feature and the Y-axis represents the count of the data records with a given SE value. B is for benign diagnosis, and M is for malignant diagnosis.
(C) Data Distribution of Features for Worst Values shows the data distribution of the features for the Wisconsin Breast Cancer data set with the worst value for the features. The X-axis represents the measured worst value for the feature and the Y-axis represents the count of the data records with a given worst value. B is for benign diagnosis, and M is for malignant diagnosis.

is defined when the output variable is a numerical value. Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features such that the output values for new data can be predicted based on those relationships which it learned from the previous data sets (5). In this study, we focused on the supervised learning that involves the training of the machine using data which has been accurately classified and labeled (training data). After we tagged the training data with the correct output result, the machine was provided with a new set of data (test data) which had not been marked with any output labels. The supervised learning algorithm's job was to analyze the unlabeled data and produce the correct output labels.

The main objective of this analysis was to determine whether machine learning algorithms can be used to predict the onset of breast cancer based on features from histopathological images with more than 90% accuracy. The four supervised learning algorithms that were selected for the analysis work are Naïve Bayes Classifier, K Nearest Neighbor, Random Forest, and Logistic Regression. We selected these algorithms to provide a broader perspective and variety for the comparison study due to their own unique characteristics and performance reasons. For any given machine learning analysis, there are no standard guidelines for model selection. The model selection is based on past work using these algorithms which provide some guidance with regard to performance and result accuracy. Our study confirmed our hypothesis that all four classification algorithms are able to predict the onset of breast cancer with more than 90% accuracy.

RESULTS

The Wisconsin Breast Cancer dataset from the UCI Machine Learning repository has 10 real-valued values. The mean, standard error (SE), and "worst" (mean of the three largest values) of the 10 real-valued features are computed for each image, resulting in 30 features. There are thus 32

attributes in the data set which includes the ID, the diagnosis, and the 30 real-value input features. All 30 features were used by the four algorithms for their classification (8). In this analysis we aimed to observe which features were most helpful in predicting malignant or benign cancer and to see general trends that would aid us in model selection and the parameter selection. The data distribution of features from the breast cancer dataset, with the mean, standard error (SE) and worst values were plotted using the Python Matplotlib library (Figure 1).

Comparison of feature distribution by malignancy showed that there was no perfect separation between any of the features. We noted the value (X-axis parameter) of the feature at which the count (Y-axis parameter) was highest for both benign diagnosis and malignant diagnosis. For any given feature, we also observed the separation of occurrence of the peak for benign curve and the peak for malignant curve and also the shape of the curve for benign and malignant. There are fairly good separations for benign curve versus malignant curve for worst values of concave points, worst values of radius, worst values of perimeter, mean values of area, and mean values of perimeter. Also, there are fairly close superpositions for some of the values, like standard error values of symmetry, standard values of smoothness, and standard error values of fractal dimension. Feature scaling was used to bring all features to the same level of magnitudes, and the data was transformed to fit within a specific scale. We used the Python package Matplotlib to conduct a data distribution analysis for multiple features. This helped determine which value was the most prevalent for each feature. The Gini impurity is used to assess the feature importance since it has lower computational cost than entropy, which requires calculating the logarithmic function. We used the Gini impurity calculation method to determine the feature importance, with the features having the most importance shown at the top (Figure 2). Based on research of the dataset using the Python NumPy, Pandas, and Matplotlib libraries, the most important features were identified as area, perimeter, concave points, and radius.

	Predicted: NO	**Predicted: YES**
Actual: NO	True Negatives (TN)	False Positives (FP)
Actual: YES	False Negatives (FN)	True Positives (TP)

Table 1: Confusion matrix.

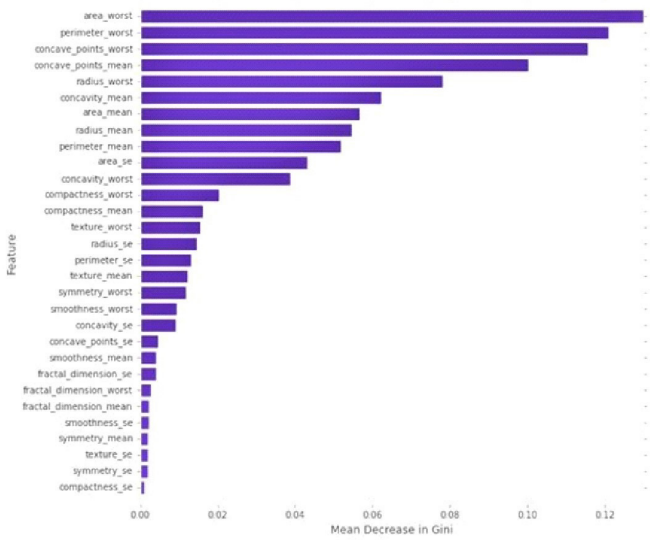

Figure 2: Relative Feature Importance Using Gini Impurity Calculation. The X-axis shows the mean decrease in Gini impurity value and the Y-axis shows the Features list. The figure shows that the top five features in order of importance are the area_worst, perimeter_worst, concave_points_worst, concave_points_mean, and radius_worst.

The breast cancer data set is split into training data and test data. The training data is used to train the algorithm. Once the algorithm gets trained, the test data is used to test the accuracy and performance of the algorithm in predicting an outcome, which for the breast cancer data would be to predict 'benign' versus 'malignant' cancer. We split the breast cancer dataset into training data and test data, using the train_test_split method of the Python Scikit Learn library. The training data was used to train the system and the test data was used to test the model's prediction. We then processed the data using four different classification algorithms of machine learning using the Scikit-Learn library: Gaussian Naïve Bayes, K Nearest Neighbor, Random Forest, and Logistic Regression.

To describe the performance of a classification model (or 'classifier') on a set of test data, a confusion matrix was used (Table 1). The actual value indicates the correct actual outcome. The predicted value indicates the prediction of the classification model. The goal is to reliably predict true positives and true negatives.

The Classification Report visualizer displays the precision, recall, F1, and support scores for the model. Precision is the ability of a classifier to avoid labeling a negative instance as positive. For each class, it is the ratio of true positives (TP) to the sum of true positives (TP) and false positives (FP). Recall is the ability of a classifier to find all positive instances. For each class, it is defined as the ratio of TP to the sum of TP and FN. F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are usually lower than accuracy measures as they embed precision and recall into their computation. Support is the number of actual occurrences of the class in the specified dataset. The F1 score calculation is used to determine the accuracy of the algorithm.

The precision, recall, F1-score and support values are calculated for the Naïve Bayes algorithm with the Scikit-learn library (Table 2). Test data size of 0.25 (25%) was used to measure the algorithm performance. We found that the weighted average precision was 0.95 and the weighted average F1-score was 0.95. With Naïve Bayes, the precision

Confusion Matrix				
	Predicted: NO	Predicted: YES		
Actual: NO	TN = 51	FP = 3		
Actual: YES	FN = 4	TP = 85		
Classification Report				
Test Size = 0.25	Precision	Recall	F1-score	Support
malignant	0.93	0.94	0.94	54
benign	0.97	0.96	0.96	89
accuracy			0.95	143
macro avg	0.95	0.95	0.95	143
weighted avg	0.95	0.95	0.95	143

Table 2: Confusion Matrix and Classification Report Calculation with Naïve Bayes algorithm.

Confusion Matrix				
	Predicted: NO	Predicted: YES		
Actual: NO	TN = 51	FP = 3		
Actual: YES	FN = 3	TP = 86		
Classification Report				
Test Size = 0.25	Precision	Recall	F1-score	Support
malignant	0.94	0.94	0.94	54
benign	0.97	0.97	0.97	89
accuracy			0.96	143
macro avg	0.96	0.96	0.96	143
weighted avg	0.96	0.96	0.96	143

Table 3: Confusion Matrix and Classification Report Calculation with K Nearest Neighbor algorithm.

Confusion Matrix				
	Predicted: NO	Predicted: YES		
Actual: NO	TN = 51	FP = 3		
Actual: YES	FN = 1	TP = 88		
Classification Report				
Test Size = 0.25	Precision	Recall	F1-score	Support
malignant	0.98	0.94	0.96	54
benign	0.97	0.99	0.98	89
accuracy			0.97	143
macro avg	0.97	0.97	0.97	143
weighted avg	0.97	0.97	0.97	143

Table 4: Confusion Matrix and Classification Report Calculation with Random Forest algorithm.

Confusion Matrix				
	Predicted: NO	Predicted: YES		
Actual: NO	TN = 51	FP = 3		
Actual: YES	FN = 4	TP = 85		
Classification Report				
Test Size = 0.25	Precision	Recall	F1-score	Support
malignant	0.96	0.98	0.97	54
benign	0.99	0.98	0.98	89
accuracy			0.98	143
macro avg	0.98	0.98	0.98	143
weighted avg	0.98	0.98	0.98	143

Table 5: Confusion Matrix and Classification Report Calculation with Logistic Regression algorithm.

was 0.93 for malignant prediction, and 0.97 for benign prediction. The accuracy calculation was 0.95. An accuracy calculation of 0.95 equates to 95% accuracy.

The precision, recall, F1-score and support values are calculated for the K Nearest Neighbor algorithm with the Scikit-learn library (Table 3). We found that the weighted average precision was 0.96 and the weighted average F1-score was 0.96. With K Nearest Neighbor, the precision was 0.94 for malignant prediction, and 0.97 for benign prediction. The accuracy calculation was 0.96. An accuracy calculation of 0.96 equates to 96% accuracy.

The precision, recall, F1-score and support values are calculated for the Random Forest algorithm with the Scikit-learn library (Table 4). We found that the weighted average precision was 0.97 and the weighted average F1-score was 0.97. With Random Forest, the precision was 0.98 for malignant prediction, and 0.97 for benign prediction. The accuracy calculation was 0.97. An accuracy calculation of 0.97 equates to 97% accuracy.

The precision, recall, F1-score and support values are calculated for the Logistic Regression algorithm with the Scikit-learn library (Table 5). We found that the weighted average precision was 0.98 and the weighted average F1-score was 0.98. With Logistic Regression, the precision was 0.96 for malignant prediction, and 0.99 for benign prediction. The accuracy calculation was 0.98. An accuracy calculation of 0.98 equates to 98% accuracy.

ROC (Receiver Operating Characteristic) curves are used to see how any predictive model can distinguish between the true positives and negatives. In order to do this, a model needs to not only correctly predict a positive as a positive, but also predict a negative as a negative. The ROC curve does this by plotting sensitivity, the probability of predicting a real positive will be a positive, against 1-specificity, the probability of predicting a real negative will be a positive. The best decision rule is high on sensitivity and low on 1-specificity.

AUC (Area under the ROC Curve) measures the two-dimensional area underneath the ROC curve. AUC is one of the most important evaluation metrics for checking any classification model's performance. AUC ranges in value

from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. Higher value of AUC indicates higher performance of the classification model.

We used the Scikit-learn and the Python Matplotlib libraries to plot the ROC curves for the Naïve Bayes, K Nearest Neighbor, Random Forest and the Logistic Regression classification algorithms (Figure 3). The ROC curve was used by each of the 4 classification algorithms to calculate the AUC value. The performance of the classification algorithms was analyzed using the Classification Report and the AUC value for the ROC curve. The calculated AUC value for the Naïve Bayes algorithm was 0.9760, 0.9701 for the K Nearest Neighbor algorithm, 0.9815 for the Random Forest algorithm, and 0.9854 for the Logistic Regression algorithm. Based on the calculated AUC values, the Logistic Regression and Random Forest algorithms performed better than Naïve Bayes and K Nearest Neighbor algorithms.

Multiple test runs were performed to test the performance and behavior of the classification algorithms, using different amounts of training and test data. The test runs used different percentage splits of the total data set into training data and the test data. We conducted the test runs with different splits of the test data (Table 6).

Based on experimentation, we observed that Gaussian Naïve Bayes performs significantly better with a test size = 0.2 compared to a test size = 0.3. K Nearest Neighbor performs marginally better with a test size = 0.3 compared to a test size = 0.2. Random Forest performs marginally better with a test size = 0.2 compared to a test size = 0.3. Logistic Regression performs marginally better with a test size = 0.3 compared to a test size = 0.2. The accuracy of the classification algorithm with the different test runs was determined using the F1-scores and was observed for comparison purposes.

The selected supervised classification algorithms — Gaussian Naïve Bayes (average accuracy = 96%), K Nearest Neighbor (average accuracy = 95%), Random Forest (average accuracy = 97%), and Logistic Regression (average accuracy = 98%) — all performed well in prediction of benign versus malignant cancer. Random Forest algorithm performed best among the four algorithms in malignant prediction (accuracy = 98%) and Logistic Regression algorithm performed best among the four algorithms in benign prediction (accuracy = 99%).

DISCUSSION

It is important to accurately classify cancer patients into high or low risk groups so that the medical professionals can determine the most effective methods of treatment. Machine learning techniques can be utilized to model the

		Test Run #1 Test size: 0.20	Test Run #2 Test size: 0.23	Test Run #3 Test size: 0.25	Test Run #4 Test size: 0.27	Test Run #5 Test size: 0.30
Gaussian Naïve Bayes	malignant	0.98	0.94	0.93	0.93	0.92
	benign	0.96	0.97	0.97	0.95	0.94
	accuracy	0.96	0.95	0.95	0.94	0.94
	macro avg	0.97	0.95	0.95	0.94	0.93
	weighted avg	0.97	0.95	0.95	0.94	0.94
K Nearest Neighbor	malignant	0.93	0.94	0.94	0.95	0.95
	benign	0.96	0.96	0.97	0.97	0.96
	accuracy	0.95	0.95	0.96	0.96	0.96
	macro avg	0.94	0.95	0.96	0.96	0.96
	weighted avg	0.95	0.95	0.96	0.96	0.96
Random Forest	malignant	0.98	0.98	0.98	0.98	0.97
	benign	0.96	0.96	0.97	0.96	0.95
	accuracy	0.96	0.97	0.97	0.97	0.96
	macro avg	0.97	0.97	0.97	0.97	0.96
	weighted avg	0.97	0.97	0.97	0.97	0.96
Logistic Regression	malignant	0.98	0.96	0.96	0.97	0.97
	benign	0.97	0.98	0.99	0.99	0.99
	accuracy	0.97	0.97	0.98	0.98	0.98
	macro avg	0.97	0.97	0.98	0.98	0.98
	weighted avg	0.97	0.97	0.98	0.98	0.98

Table 6: Performance metrics of the classification algorithms with different test data sizes.

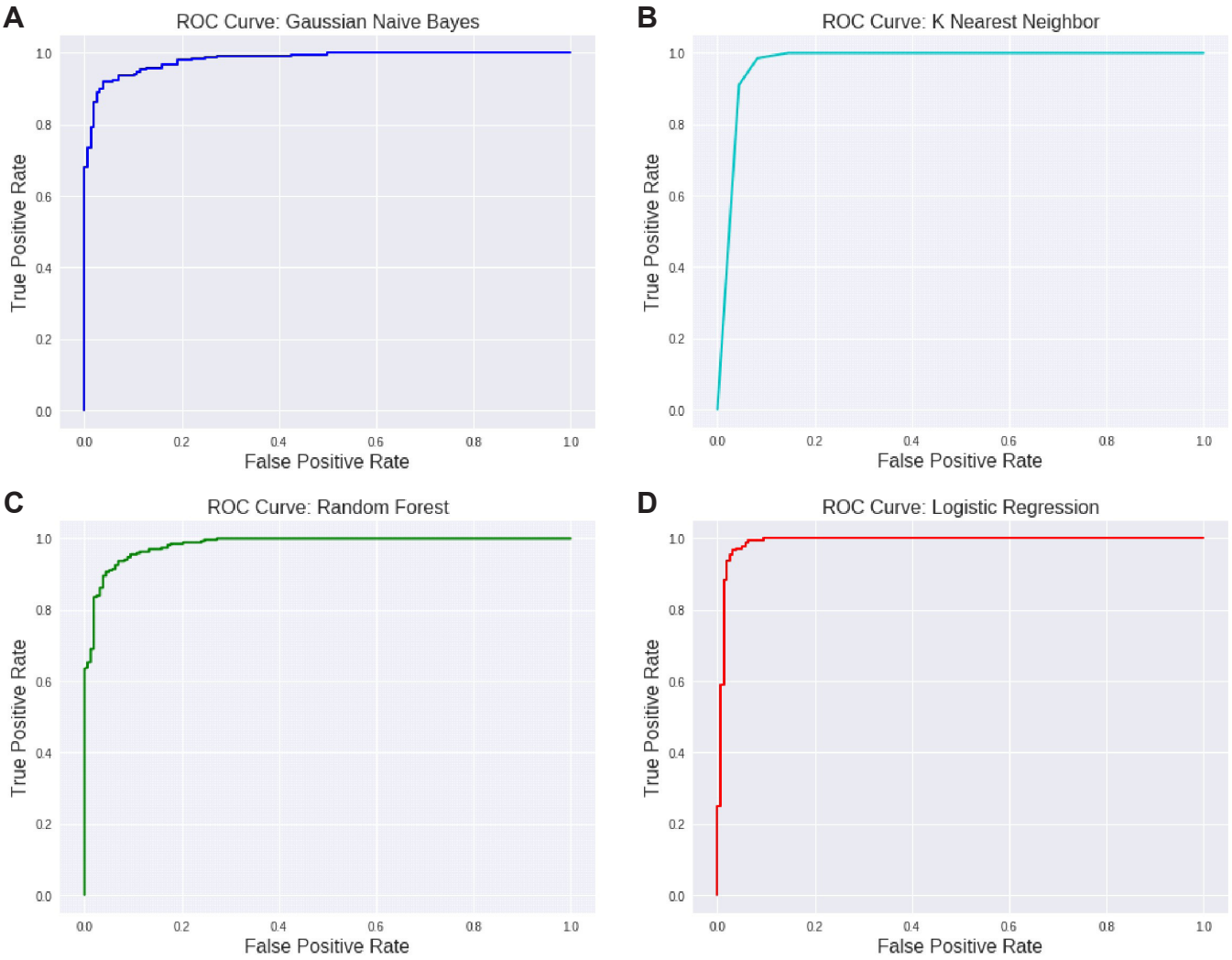

Figure 3: Plot of ROC curve for the four classification algorithms. (A) ROC curve with Naïve Bayes Algorithm shows False Positive Rate (FPR) along X-axis and True Positive Rate (TPR) along Y-axis. The plot is shown in blue color. **(B)** ROC curve with K Nearest Neighbor algorithm shows False Positive Rate (FPR) along X-axis and True Positive Rate (TPR) along Y-axis. The plot is shown in cyan color. **(C)** ROC curve with Random Forest algorithm shows False Positive Rate (FPR) along X-axis and True Positive Rate (TPR) along Y-axis. The plot is shown in green color. **(D)** ROC curve with Logistic Regression algorithm shows False Positive Rate (FPR) along X-axis and True Positive Rate (TPR) along Y-axis. The plot is shown in red color.

progression of cancerous conditions. In addition, the ability of the machine learning tools to detect key features from complex datasets provides significant benefits in analysis of clinical data. Our study confirmed our hypothesis that all four classification algorithms are able to predict the onset of breast cancer with more than 90% accuracy. Our results also showed that Random Forest and Logistic Regression provide better accuracy and performance than Naïve Bayes and K Nearest Neighbor. Comparing the performance of the two best classification algorithms, Random Forest had better accuracy in malignant prediction and Logistic Regression had better accuracy in benign prediction. The Logistic Regression algorithm provided the most accurate results overall; however, Random Forest algorithm was more accurate in malignant cancer prediction.

The current study showed that machine learning classification algorithms can be used in malignant versus benign prediction of cancer with high accuracy. However, appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice. The Wisconsin Breast Cancer data set has a total of 569 data instances, of which 357 are classified to have 'benign' outcome, and 212 are classified to have 'malignant' outcome (8). Based on our results, the most important features were identified as area, perimeter, concave points, and radius. With the limited set of data instances, it is possible to have data biases. In general, training data for machine learning projects has to be representative of the real world. This is important because using this data is how the machine learns to do its job. Data bias can occur in a range of areas, from human reporting and selection bias to algorithmic and interpretation bias. It is possible that there may be measurement bias if

the data collected for training differs from that collected in the real world, or if faulty measurements at data collection time result in data distortion. Based on the understanding of source of the data, it is highly unlikely that there would be measurement bias because the dataset was created from real patient clinical tests. It is however possible that some of the valuable data may be deleted in the pre-processing stage and that may result in exclusion bias. Based on the source of the dataset there could be inherent racial bias with the data skewed in favor of particular demographics, or association bias if the data for the machine learning model reinforces and/or multiplies a cultural bias. Based on the source of the dataset it is highly unlikely that there would be racial bias in the data gathering process.

Our study only compared the efficiency and performance of the four classification algorithms and did not involve comparison with clinical standards of care and biopsy results. This can be part of a future study to determine how the classification algorithms perform in comparison with the outcome from the biopsy and other clinical data used by pathologists. Also, the study used the available breast cancer data set which uses clinical data available from biopsy tests. The classification algorithms can be applied to any other data set which has additional input data features with values and the output prediction label. The data from mammography and ultrasound images can be included in the input data set to predict the outcome of malignant versus benign breast cancer. We also observed that Gaussian Naïve Bayes and Random Forest perform better with smaller test data size, and K Nearest Neighbor and Logistic Regression perform marginally better with larger test data size. There is no clear explanation for these behaviors and it requires future exploration.

Our study can be used as a guidance for machine learning algorithm selection and designing a system that can be used to predict the onset of breast cancer based on any new patient lab and clinical test results data, as specified in the Wisconsin Breast Cancer data set. This would aid diagnosis of patients in a timely manner, before the disease turns severe and needs emergency treatment. Future applications of this research would consist of applying the same techniques in predictions of other malignancies. AI and ML tools will help doctors and medical professionals make better diagnostic decisions, improve treatment outcomes, and reduce medical errors. A variety of these AI and ML tools and techniques can be utilized for the development of predictive models for different types of cancer and other diseases, resulting in more effective and accurate decision making.

METHODS

Scikit-learn is a free open-source software machine learning library for the Python programming language. Scikit-learn has the support for various types of ML algorithms that are related to supervised learning and un-supervised learning, and interoperates with several Python libraries. The Python libraries include NumPy library for working with matrices and math operations, SciPy library for scientific and technical computing, Matplotlib library for data visualization, and Pandas library for data handling, manipulation, and analysis (9).

For the prediction of the onset of a disease, the Wisconsin Breast Cancer (Diagnostic) dataset was used. This is a dataset that is available from the UCI Machine Learning Repository of the Center for Machine Learning and Intelligent Systems. The datasets are available on public domain and available for research purposes. In the breast cancer dataset, the features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass (8). They describe characteristics of the cell nuclei present in the image. The Wisconsin Breast Cancer dataset has a total 569 number of data instances, of which 357 are classified to have 'benign' outcome and 212 are classified to have 'malignant' outcome (8). In the dataset, 1 means the cancer is malignant and 0 means the cancer is benign. Ten real-valued features are computed for each cell nucleus. The features included in the computation are radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1).

The data exploration and visualization were performed to determine the distribution of the features, as encapsulated in the dataset. We used the Python Pandas and Matplotlib libraries to conduct a data distribution analysis for multiple features, which helped determine which value is the most prevalent for each feature.

Gini Impurity is a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance were randomly classified according to the distribution of class labels from the data set. Gini Impurity provides us with the probability of misclassifying an observation. When the Gini Impurity is lower, there is lower likelihood of misclassification. We carefully analyzed the dataset using Gini Impurity calculations to determine the features which are most helpful in predicting whether cancer is malignant or benign, to aid in the process of model selection and parameter selection.

Feature scaling is used to bring all the features to the same scale. Without feature scaling, the model tends to give higher weights to higher values and lower weights to lower values irrespective of the units of values. There are two types of feature scaling used in Machine learning such as min-max normalization, and standardization. Prior to training of the classification algorithms, we used feature scaling to bring all features to the same level of magnitudes, and the data for the features were transformed and normalized to fit within the scale of 0–1.

The breast cancer data set is a public dataset available

from the University of California, Irvine Machine Learning Repository (8). In the current study, the dataset was obtained from the repository, and then programmatically split into training data and test data. The training data was used for training of the classification algorithms. Test data was used to validate the performance of the classification algorithms after they have been trained. We used the train_test_split method of the Python Scikit Learn library to split the breast cancer dataset into training data and test data.

Using the same dataset for both training and testing leaves room for miscalculations, and thus increases the chances of inaccurate predictions. The split of the training data and the test data is dependent on consideration given to avoid overfitting or underfitting the model. A higher percentage of the training data could cause overfitting, and the model would show almost perfect accuracy when handling training data, but it can be inaccurate when handling new data. A lower percentage of training data could cause underfitting, and the model may not accurately fit the training data, resulting in inaccurate predictions. With less training data, the parameter estimates have greater variance. With less testing data, the performance statistic will have greater variance.

The ideal split is considered to be a training data: test data split ratio of 75:25 (10). The testing was performed with average split ratio of 75:25, and other test runs were performed with split ratios between 70:30 and 80:20 to validate that we are able to obtain comparable results. The data was processed using four different classification algorithms of machine learning using the Scikit-Learn library: Gaussian Naïve Bayes, K Nearest Neighbors, Random Forest, and Logistic Regression.

Bayes Theorem describes the probability of an event based on prior knowledge of conditions. Mathematically Bayes Theorem can be expressed as:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

where A and B are events, P(A) and P(B) are the probabilities of the event and are independent from each other. P(A|B) is the probability of A if B has occurred, and P(B|A) is the probability of B if A has occurred. Naïve Bayes is a machine learning algorithm which is based on the Bayes Theorem. In Naïve Bayesian classifiers the assumption is that there are no dependencies amongst attributes. Naive Bayes classifiers are computationally fast when making decisions and do not require large amount of data for the learning process (6).

The K Nearest Algorithm is a data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it. The logic of the K Nearest Neighbor is to compute a distance value between the item to be classified and every item in the training data set. The calculated distance is then used to pick the K closest data points (i.e. the items with K lowest distances). A majority vote is conducted among those data points and the dominating classification in that pool is decided as the final classification. The K Nearest Neighbor uses the principle of lazy learning, in which the algorithm performs local approximation, and all computation happens at the time of final classification (6).

The Random Forest algorithm consists of a large number of individual decision trees that work together in synchronization. The Random Forest classifier creates a set of decision trees, which are generated from randomly selected subset of training set. Each individual decision tree in the random forest returns a class prediction. The algorithm decides the final classification of the test data by computing the aggregated vote count from the different decision trees (7).

The Logistic Regression algorithm is used to predict a binary outcome based on a set of independent variables. Logistic Regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. Logistic Regression uses the logistic/sigmoid function to measure the relationship between the dependent variable and one or more independent variables. The output of the logistic regression will be a probability ($0 \le x \le 1$), and can be used to predict the binary 0 or 1 as the output (if $x<0.5$, output= 0, else output=1) (7).

The current study aimed to test the hypothesis that the Random Forest and Logistic Regression algorithms would perform better than Gaussian Naïve Bayes and K Nearest Neighbor algorithms. To test this hypothesis, experiments were performed with the four classification algorithms and the accuracy and performance were measured using the Classification Report and ROC curve with the input dataset.
The loading, handling, manipulating, and computations of the data are handled by Pandas, NumPy, and SciPy libraries and the visualizing of data are handled by the Matplotlib library.
The accuracy of the predictions is checked by calculating the number of the correct predictions relative to the total number of predictions made for each of the classification algorithms. Multiple trials with different splits of training data and test data are conducted to observe the accuracy of the predictions. Testing of the algorithms is also conducted with the new test data.

The Python Graph plotting library is used to display the results of each classification algorithm. The Confusion Matrix, Classification Reports are generated for each classification algorithm using the Scikit Learn (sklearn.metrics) library. The ROC curves are generated using the Scikit Learn (sklearn.metrics) library and the Python Matplotlib (matplotlib.pyplot) library. The results are compared to determine the best classification algorithm for the data set. The ROC curves are based on the following definitions for sensitivity and specificity.
Sensitivity is the proportion of actual positive cases which are correctly identified and is calculated as TP/(TP+FN). Specificity is the proportion of actual negative cases which are correctly identified and is calculated as TN/(TN+FP).

REFERENCES

1. Breast Cancer Facts." National Breast Cancer Foundation, 23 Oct. 2020, www.nationalbreastcancer.org/breast-cancer-facts.
2. Kourou, Konstantina, et al. "Machine Learning Applications in Cancer Prognosis and Prediction." Computational and Structural Biotechnology Journal, vol. 13, 2015, pp. 8–17., doi:10.1016/j.csbj.2014.11.005.
3. Meskó, Bertalan, et al. "Will Artificial Intelligence Solve the Human Resource Crisis in Healthcare?" BMC Health Services Research, vol. 18, no. 1, 2018, doi:10.1186/s12913-018-3359-4.
4. Chaurasia, Vikas, and Saurabh Pal. "Early Prediction of Heart Diseases Using Data Mining Techniques." SSRN Electronic Journal, vol. 1, 2013, pp. 208–217.
5. Domingos, Pedro. "A Few Useful Things to Know about Machine Learning." Communications of the ACM, vol. 55, no. 10, 2012, pp. 78–87., doi:10.1145/2347736.2347755.
6. Ashari, Ahmad, et al. "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool." International Journal of Advanced Computer Science and Applications, vol. 4, no. 11, 2013, doi:10.14569/ijacsa.2013.041105.
7. Kirasich, Kaitlin, et al. "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets." SMU Data Science Review, vol. 1, no. 9, ser. 3, 2018. 3.
8. Dua, D. "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]." Irvine, CA: University of California, School of Information and Computer Science. n.p.: n.p., 12 . 28 Dec. 2020.
9. Pedregosa, Fabian, et al. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, no. 85, 2011, pp. 2825–2830.
10. Guyon, I. "A Scaling Law for the Validation-Set Training-Set Size Ratio." Semantic Scholar, 1 Jan. 1997, www.semanticscholar.org/paper/A-Scaling-Law-for-the-Validation-Set-Training-Set-Guyon/452e6c05d46e061290fefff8b46d0ff161998677.

Article submitted: September 9, 2020
Article accepted: October 9, 2020
Article published: May 11, 2021