Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Comparative Assessment of Machine Learning Algorithms for Detecting and Diagnosing Breast Cancer

Md Zahidul Islam[1] ✉ Md Nasiruddin[2], Shuvo Dutta[3], Rajesh Sikder[4], Chowdhury Badrul Huda[5] and Md Rasibul Islam[6]

[1]MBA Business Analytics, Gannon University, USA
[26]Department of management science and quantitative methods, Gannon University, USA
[3]Master of Arts in Physics, Western Michigan University, USA
[4]PhD Student in Information Technology, University of the Cumberland, KY, USA
[5]Department of Management, Master of Science in Project Management, ST. Francis College, USA
Corresponding Author: Md Zahidul Islam, **E-mail**: zahid.uiu007@gmail.com

| ABSTRACT

The principal goal of this study was to explore machine-learning techniques deployed for the early detection of breast cancer in the United States. Specifically, three algorithms were trained on a breast cancer dataset: Decision Tree, Random Forest, and Linear Regression. Each model was further evaluated for its performance, to ascertain the best model. Upon review, the Random Forest provided higher classification performance. It was postulated that the Random Forest offered higher accuracy models on the test data because Decision Trees and Linear Regression require more extensive data for them to be more precise in making high-precision predictions. Out of all the models, the Random Forest provided suitable accuracy on test data. Therefore, in this research scope, Random Forest was the most successful and proved effective in accurately identifying breast cancer malignancies. In that light, the proposed random forest can benefit healthcare organizations by facilitating in detection of breast cancer disease by identifying patients in high-risk groups at an early and more treatable stage of disease for improved outcomes and lower healthcare costs. Besides, Random Forest models can assist in identifying high-risk patients in advance for prompt treatment. In that regard, such detection saves lives and decreases long-term healthcare costs for the US government.

| KEYWORDS

Breast Cancer; Random Forest; Decision Tree; Linear Regression; Early Detection; Treatment plans1.

| ARTICLE INFORMATION

ACCEPTED: 01 June 2024 **PUBLISHED:** 13 June 2024 **DOI:** 10.32996/jcsts.2024.6.2.14

1. Introduction

Retrospectively, Meenalochini & Ramkumar, (2021), indicate that the exponential escalation of cancer breast cancer in the recent past has consequently made this ailment to be deemed as the leading cause of death globally for women. It is one of the most commonly diagnosed cancers among women and accounts for about 15% of all new cases of cancer in America, with over 1 million diagnoses every year. Mesleh (2021), states that breast cancer refers to the development of a type of cancer that begins in the breasts, forming cells that grow abnormally, especially the cells of the milk-producing ducts. In sporadic cases, it also affects men. Symptoms such as new lumps or thickening, changes in appearance, shape or size of the breast, inverted nipples, etc.

According to the *world cancer statistics,* breast cancer is the most prevalent cancer type, contributing to 11.7% of total cancer cases. Similarly, breast cancer was reported as the most common cancer type constituting 21.3%, among all cancer cases in the 2023 U.S. annual cancer report (Mohammed, et al., 2023). Breast cancer is the most commonly diagnosed cancer in women, and it

is also among the deadliest cancers globally. Breast cancer is divided into two based on the appearance of cells: invasive ductal carcinoma and ductal carcinoma in situ. Invasive ductal carcinoma is more lethal because it can metastasize to other tissues of the breasts, which results in death among most of the patients of breast cancer. In America, it ranks among the most common malignancies causing deaths for American women, after lung and colorectal cancer.

As per Sindhwani et al. (2023), breast cancer can be cured relatively better with less risk if detected early, minimizing the mortality rate by 25%. Major Causes of breast cancer include hormones, obesity, past radiation therapy, lifestyle, reproductive factors, or a family history of the disease. Currently, machine learning algorithms have been employed in diagnosing breast cancer in women since they are the most accurate and can predict the chances of cancer. The deployment of machine learning has increased the accuracy of cancer prediction by 15-20%. Essentially, machine learning comprises four phases for cancer classification: gathering data, selecting the suitable algorithm, training the algorithm, and testing. Precancers, when recognized early with the help of machine learning and the use of machine learning for early detection of breast cancer, have increased prognosis to be very successful since the use of fewer risky therapies can be used once the condition is found.

1.1 Problem Statement

In retrospect, Abunasser et al. (2023), contend that breast cancer is one of the threatening health problems in contemporary America, rated as one of the leading cancers diagnosed in women globally, with a high proportion of cases and deaths in the world every year. While the conventional has been deployed in terms of early detection and diagnostics, the disease remains difficult to detect in time. Many other factors and specific causes are not entirely understood. Besides, when detected in the late stages, it can be challenging to deal with, making it a high rate of mortality in cases of cancer. Sindhwani et al. (2023), hold that the optimization of modalities for early screening and risk prediction may lead to an essential decrease in breast cancer mortality, capturing the disease before it has extensively spread. Advanced machine learning and artificial intelligence techniques now hold promise for more precise identification of high-risk patients and early detection through imaging or blood tests, but through noninvasive methods; again, however, these techniques will require considerable refinement and testing to realize their potential in reducing the health and economic burden of breast cancer.

2. Literature Review

As per Bazazeh & Shubair (2023), in their study where they employed ensemble learning techniques such as AdaBoost, Random Forest, and XG-Boost to predict breast cancer. Their results indicated that random forest achieved 97% accuracy. According to Hamed et al., a new majority voting-based hybrid classifier was used to improve the power of breast cancer prediction, but the improvement was only 79%. Basker et al. (2022), indicated that support vector machines (SVM) attained 96.25% accuracy in predicting breast cancer using the Wisconsin breast cancer dataset. Agarap (2022), explored various machine-learning methods in breast cancer detection and found AdaBoost to be the potential algorithm with 98.77% accuracy. Ahmad et al. applied diverse machine learning methods to the Wisconsin breast cancer data and established that the random forest and SVM reached 96.5% and SVM 96.5%. In the same vein, Allugunti (2022), hybridized AdaBoost Random Forest in a hybrid algorithm for breast cancer classification, where the comparison of the method with the classifiers like SVM showed that the accuracy of the ensemble classifiers was boosted up to 9.8%, with at least a 4.3% growth.

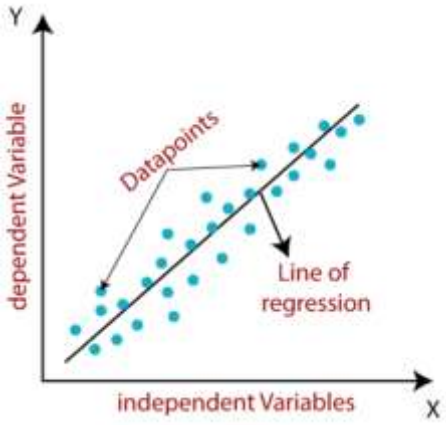
In the study of Bhise et al. (2021), they ascertained that the mortally related aspects of breast cancer in India were probed concerning several risk factors like demographic characteristics, lifestyle, water intake, etc. An ensemble algorithm named Bragboos, with an accuracy of 98.21%, was employed for the prediction of breast cancer in women from Malwa, India. By contrast, an ensemble algorithm by IJSRSWIT (2022) combined multiple classification methods and achieved the classification of benign and malignant tumors at the UC-Irvine machine learning repository. The combined stacked ensemble classifier showed an accuracy of 97.20%. Among other studies, Master (2019) compared machine learning-based methods of predicting breast cancer, including deep learning. He obtained the highest accuracy of 96.99%. Finally, Meenalochini & Ramkumar, (2021) found that a convolutional neural network obtained an accuracy of 97.66% to classify breast cancer.

2.1 Machine Learning Algorithms

As per Mohammed et al. (2023), Machine learning is subdivided into four key categories, most notably, supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. As regards supervised learning, the training data presented to the model is labeled so it knows the anticipated output, enabling it to learn correlations between inputs and outputs. Unsupervised learning searches for the hidden structure in unlabeled data without labeled responses. A sort of mix when there is a bit of labeled data coupled with a significant amount of unlabeled data forms semi-supervised learning. Reinforcement learning is an area that an agent learns when it interacts with a dynamic environment by trial and error. From the feedback in the form of rewards or punishments, which arise from the actions taken and reaching the goals over a significant volume of episodes, agents learn optimal decision policies. From this ongoing interaction, all potential system states are gradually learned.

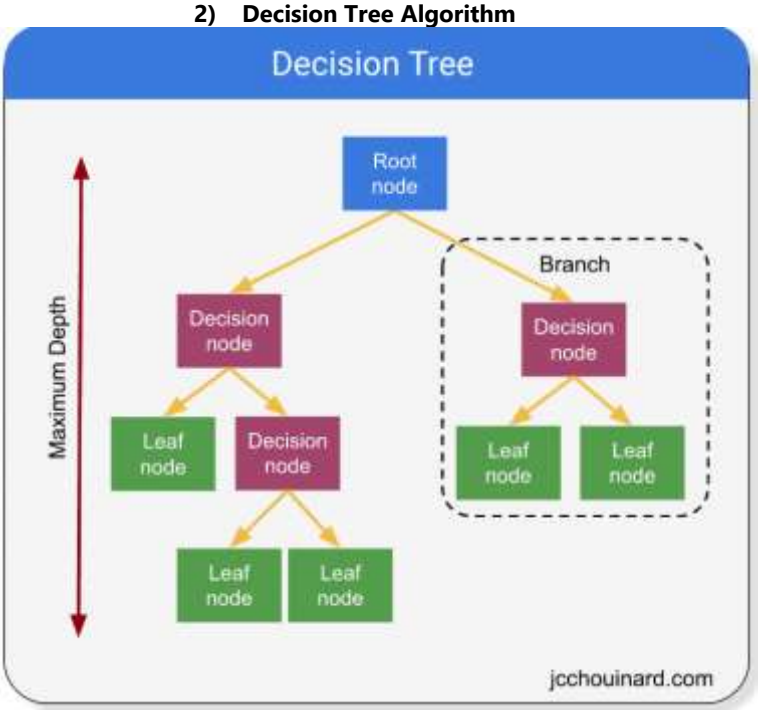
Mesleh (2021), argues that when applied to the classification of cancer, ML algorithms can clearly distinguish benign and malignant tumors, which can be very helpful in diagnosing cancer for the physician. Essential steps to follow in this process are identification of key features with them. Some popular techniques that are used for the classification of breast cancer, monitoring progression, treatment, and prediction include Support Vector Machines, Decision Trees, Random Forest, K Nearest Neighbor, AdaBoost, Naive Bayes, traditional Neural Networks, Probabilistic Neural Networks, Recurrent Neural Networks, and Conventional Neural Networks. Such machine learning algorithms demonstrate capability in the classification of cancer types and the assessment of cancer risk, providing decision support to oncologists in all kinds of clinical encounters. In this respect, feature selection will have paramount importance in machine learning models; techniques vary in their approach and performance for specific cancer modeling applications. In the present paper, the researcher will deploy Supervised machine learning models, most notably, Linear Regression, Decision Tree, and Random Forest.

1) Linear Regression Algorithm



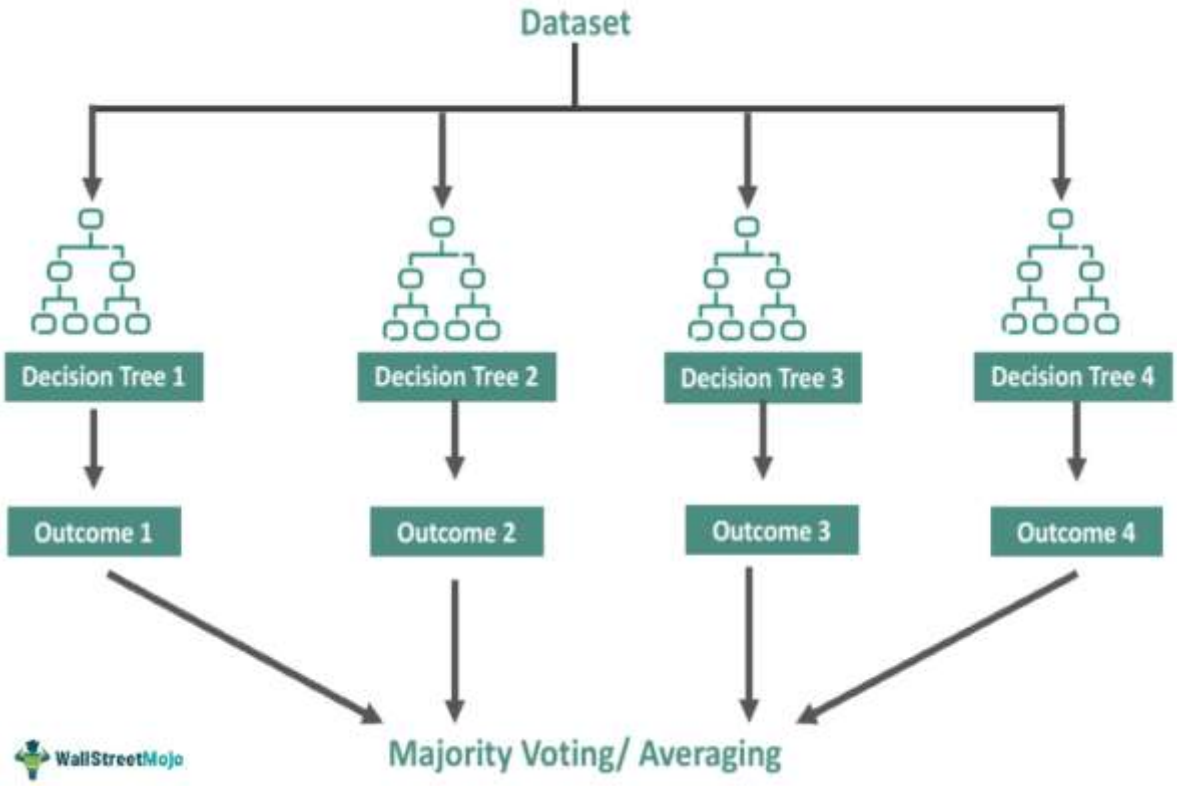
Sindhwani et al. (2023), indicate that linear regression is a supervised machine learning model, and regression is a predictive modeling algorithm. Models can be used to predict target values based on input variables. Linear regression is a predictive technique used to indicate the relation and forecast of the value between the variables. Several categories of regression models assume the relationship between dependent and independent variables and the inclusion of features in the input. This method predicts values of a dependent variable, y, based on independent variables, x. It identifies a linear or straight-line relationship between y and x. Therefore, this regression technique is linear because it fits a linear equation to the dataset of x and y values to model their relationship.

2) Decision Tree Algorithm



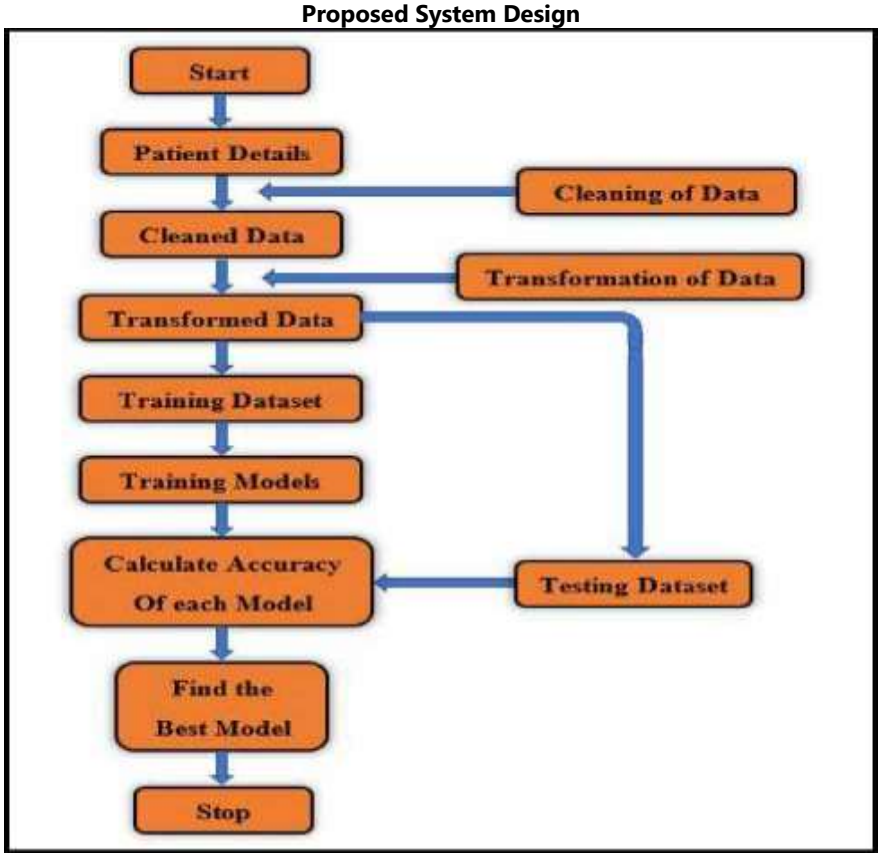
Meenalochini & Ramkumar, (2021), states that a decision tree is a kind of supervised machine-learning algorithm that can be used to solve both classification and regression problems. It takes the form of a tree structure to solve a problem. It contains the leaf nodes constituting a set of class labels together with the manipulated data and internal nodes constituting a feature. The objective is to create a model—a decision tree to predict the target variable from the features of the data by learning simple rules. Model structures of decision trees can adopt Boolean functions on discrete attributes. Choosing the correct attribute for each node at each level of the tree is most likely the main difficulty in forming a decision tree. Decision tree learning is one method of approximating a discrete-valued target function. It makes use of decision trees to make the learned patterns representable. Decision tree learning is a much-used and practical method in inductive reasoning, mainly because of the simplicity and interpretability it gives rise to.

3) Random Forest Algorithm



Master (2019), contends that Random Forest is a machine learning algorithm under a supervised category used in classification and regression problems. As the name implies, this model creates an ensemble of decision trees: multiple trees to build up a "forest." More trees build up a more substantial, robust forest model. The random forest builds multiple decision trees upon random data samples, takes a prediction from each one, and then picks the best outcome. This classifier is good with missing values and remains accurate even when a large percentage of the data is uncertain. The fact that it has many trees means there is no chance of overfitting since the individual tree can't become the majority of the model. A significant advantage of the random forest methodology is that it utilizes the strengths of many decision trees to generate a much more powerful predictive model.

Proposed System Design



3. Methodology

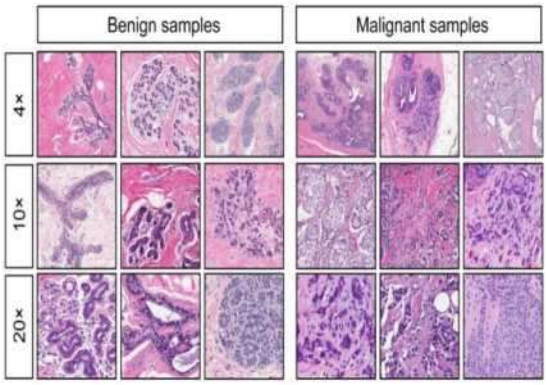
To conduct the experiments in this paper, the following system configuration was used: an Intel Pentium H.P laptop with a 2.60GHz i7 processor and 16GB of RAM, executing a 64-bit version of the Ubuntu 20.04.2 operating system. The machine learning classification algorithms were implemented using the sci-kit-learn library (Pro-AI- Robikul, 2024). Consequently, confusion matrix were applied to the three models to assess their respective performance.

3.1 Dataset Description

This research paper used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, containing 569 instances broken down into 357 for benign cases and 212 for malignant cases. It had two categories: malignant at 37.19% and benign at 62.63%. Attributes describing an instance are 32 integer-valued. Figure 2 shows details of the dataset and lists 3 classifiers, which are trained to predict breast cancer classifications using the Wisconsin BC dataset. As class appraisal, "clump thickness" is upheld (Pro-AI- Robikul, 2024).

3.2 Data preprocessing

The data from the Wisconsin Breast Cancer dataset was raw patient data in CSV format, with some attribute values tagged as NA, Null, NAN, etc. To clean the data, the researcher dropped the attributes that contained the unwanted missing or unknown values that had been found. They then derived counts of the Malignant (M) and Benign (B) classes, applying the .value counts() method to the 'diagnosis' column. This showed that, of the 569 records, 212 were malignant, and 357 were benign(Pro-AI- Robikul, 2024). By eliminating incomplete or suspect records and tallying the diagnosis breakdown, the researcher prepared the record set for further measures and modeling. Ten attributes are taken from the entire dataset to be applied in the experiments. These values of the attributes are observed to vary from 1 to 10. The class variable takes numeric values of 2 and 4 to indicate benign and malignant tumors, respectively.



Data Attributes

Uniformity of Cell Size
Clump thickness
Marginal adhesion
Uniformity of Cell Shape
Bare nuclei
Single epithelial cell size
Bare chromatin
Mitoses
Normal nucleoli
Class

3.3 Dataset Splitting

After the preprocessing of data, a clean dataset was partitioned into a training dataset and a testing dataset. The data was split into 75% for the training dataset and 25% for the testing dataset, respectively. Three machine learning algorithms—Linear Regression, Decision Tree, and Random Forest—were trained with only the training data. The testing set was used to check the performance of the model. Performance comparison based on predictive accuracies on testing data could show a researcher the model that would work best for breast cancer classification purposes without leaking information from the test data into the model selection process. This standard train-test-split approach allowed one to make an unbiased estimate of the performance of each machine learning algorithm regarding how well it will generalize to unseen examples(Pro-AI- Robikul, 2024).

4. Experimentation Results

Importing Libraries

```
import numpy as np
import seaborn as sns
import pandas as pd
from matplotlib import pyplot as plt
# Suppress warnings
import warnings
warnings.filterwarnings('ignore')
```

```
df = pd.read_csv("data.csv")
df
```

For this analysis, the breast cancer dataset was loaded into Python using the machine learning library sci-kit-learn. Subsequent data exploration, preprocessing, model training, and evaluation were all done using programs written in Python. The first step was to open and load the dataset to understand its structure and the attributes of the data. Below is an overview of the loaded data, which gives an idea of the number of samples, features, and some first attributes of potential interest before further processing or

the development of a model. This way, getting to know the raw dataset sets up the analyst well with an approach before moving into deeper data wrangling, analysis, and model training.

Output:

Out[4]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0
...	...	...	...	...	...	...	...	...	...
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0

569 rows × 33 columns

With the dataset successfully loaded into Python, data preprocessing tasks were done at this juncture. This was under cleaning, eliminating the records where the predictor variables were null or missing values. The transformation was also done by changing or formatting the variables accordingly for the predictive model being developed.

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

In [6]:
```
df.describe()
```

Output:

Out[6]:

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800

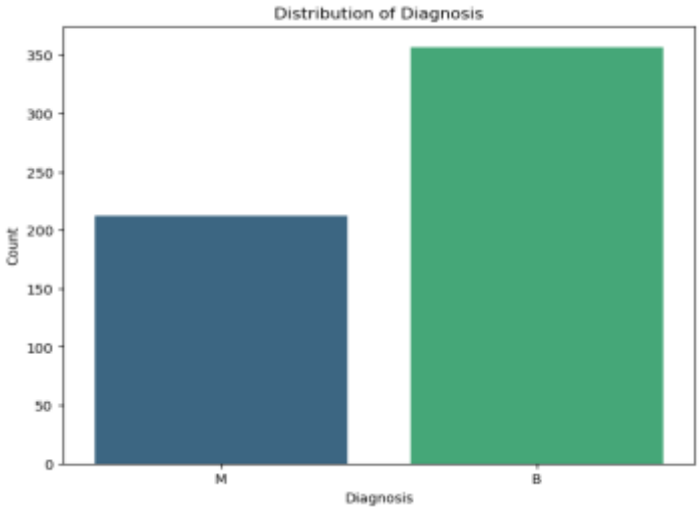
8 rows × 32 columns

In [7]:
```
df = df.drop(columns=['id', 'Unnamed: 32'])
df
```

To understand the general distribution of diagnoses with the dataset, the analyst performed snippets to generate an appropriate count plot for diagnosis. In this way, upon running the code, a histogram output can be observed. This helped in visualizing that the count-based diagnoses variable distribution could be simply represented, with the two classes, benign and malignant, both easily visualized as showcased below:

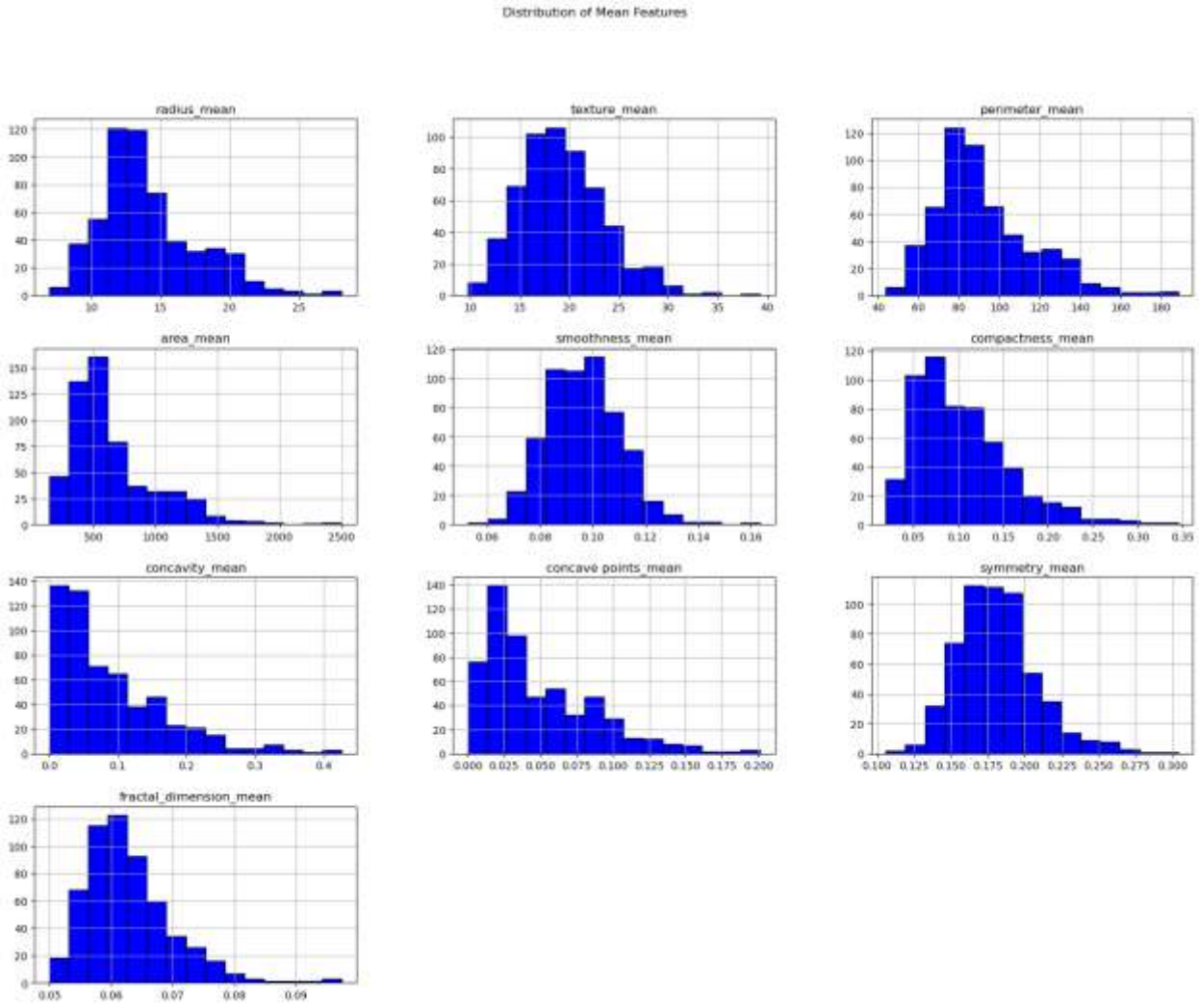
Output:

In [8]:
```
# Count plot for diagnosis
plt.figure(figsize=(8, 6))
sns.countplot(x='diagnosis', data=df, palette='viridis')
plt.title('Distribution of Diagnosis')
plt.xlabel('Diagnosis')
plt.ylabel('Count')
plt.show()
```



Subsequently, the analyst explored the dataset distribution mean features, particularly exploring the correlations between different variables as showcased below:

Output:

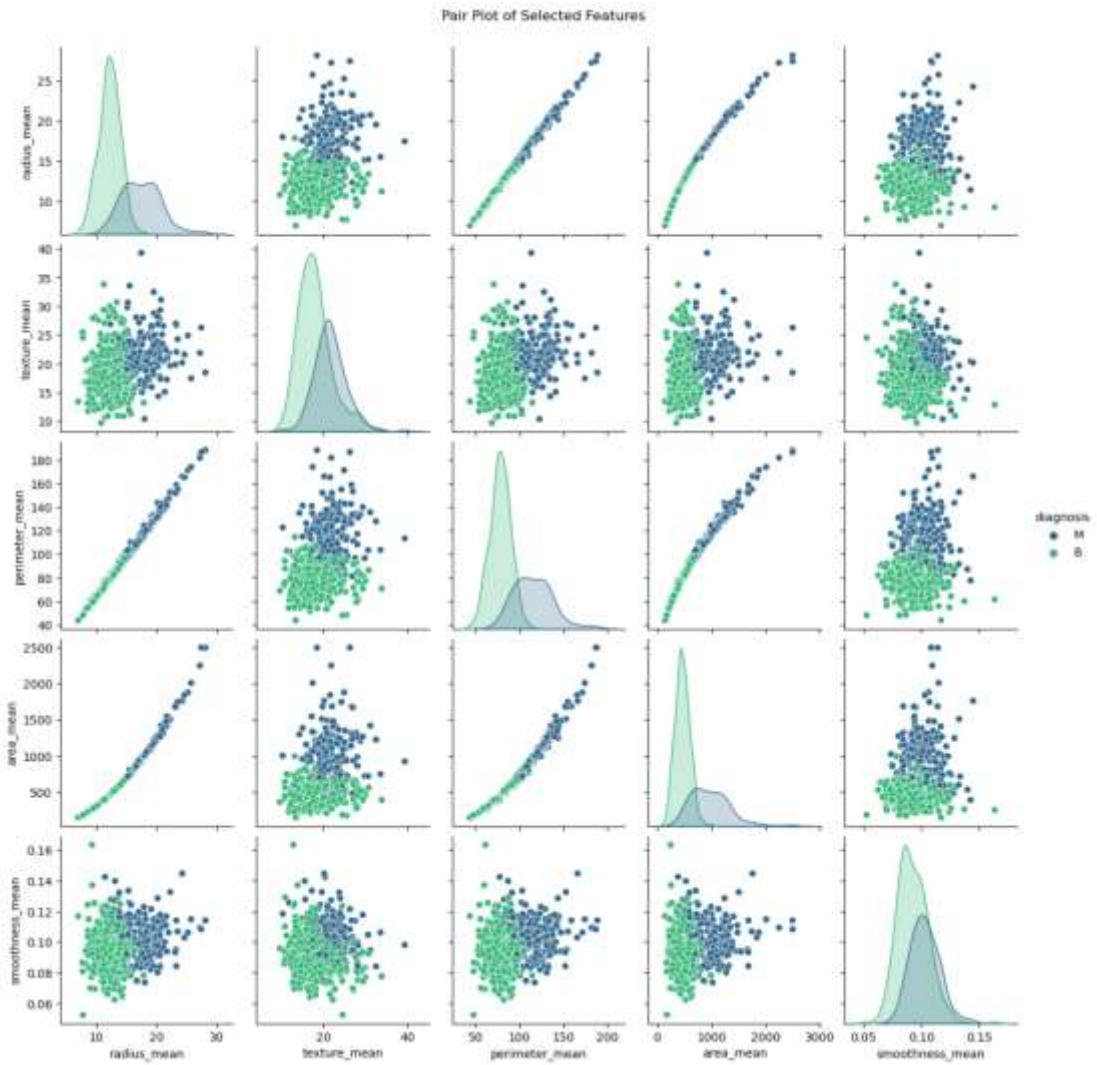


Furthermore, the analyst went ahead to determine the correlation between the distribution of sub-set features in the breast cancer dataset as showcased below:

```
subset_features = ['radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean']

sns.pairplot(df[subset_features + ['diagnosis']], hue='diagnosis', palette='viridis')
plt.suptitle('Pair Plot of Selected Features', y=1.02)
plt.show()
```

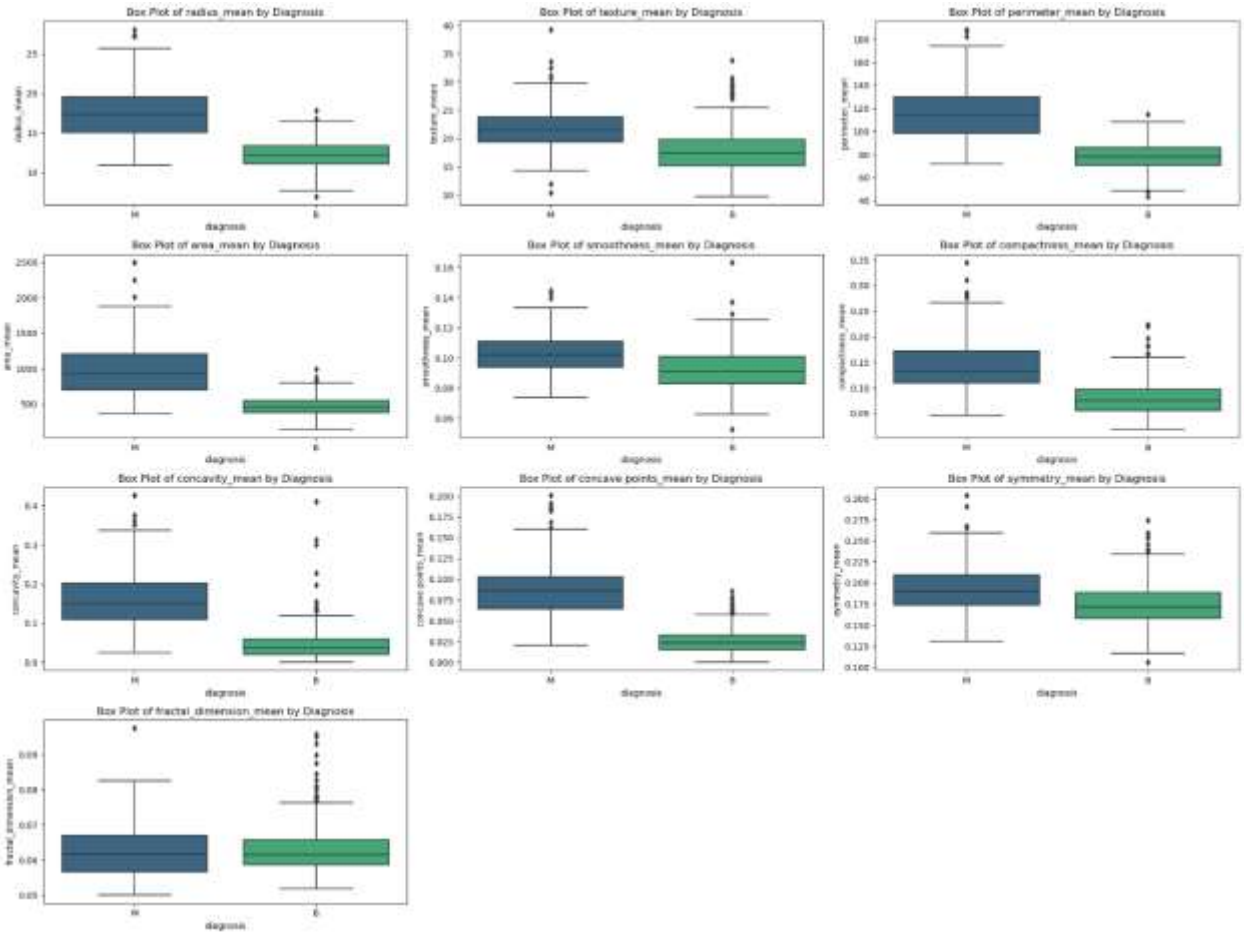
Output:



Apart from that, the analyst imposed a code snippet to generate box plots of the features by diagnosis as displayed below:

```
plt.figure(figsize=(20, 15))
for i, feature in enumerate(mean_features):
        plt.subplot(4, 3, i + 1)
        sns.boxplot(x='diagnosis', y=feature, data=df, palette='viridis')
        plt.title(f'Box Plot of {feature} by Diagnosis')
plt.tight_layout()
plt.show()
```

Output:

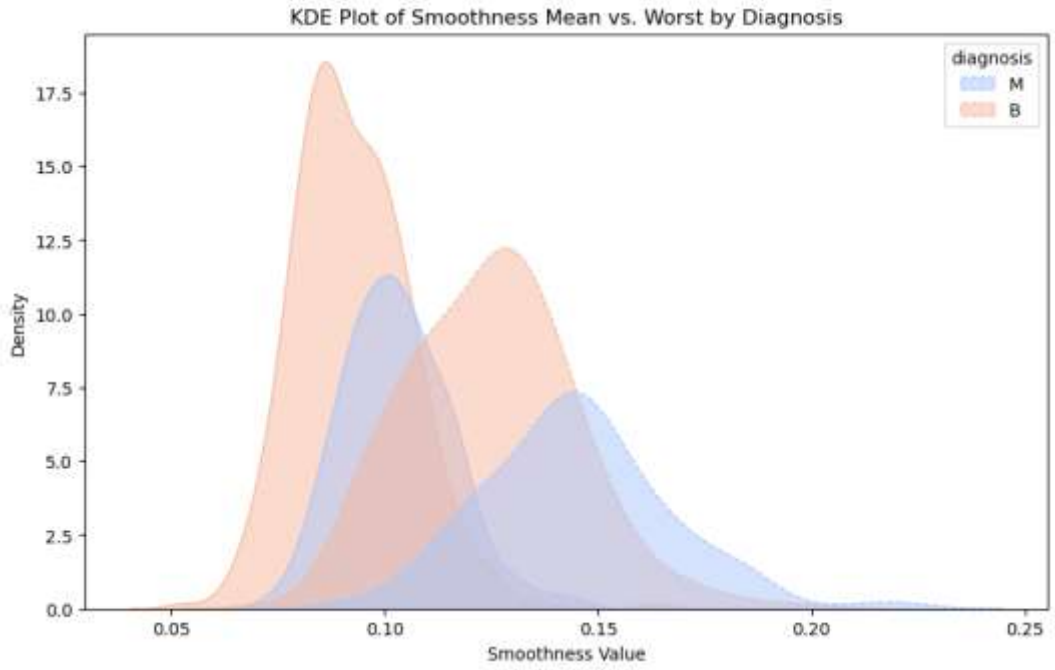


Last but not least, an appropriate snippet was performed to generate a chart graph displaying the KDE plot of smoothness vs. worst by diagnosis as exhibited below:

```
for feature in ['radius', 'texture', 'perimeter', 'area', 'smoothness']:
    plt.figure(figsize=(10, 6))
sns.kdeplot(data=df, x=f'{feature}_mean', hue='diagnosis', fill=True, palette='coolwarm', alpha=0.5)
sns.kdeplot(data=df, x=f'{feature}_worst', hue='diagnosis', fill=True, palette='coolwarm', alpha=0.5, linestyle='--')
plt.title(f'KDE Plot of {feature.capitalize()} Mean vs. Worst by Diagnosis')
plt.xlabel(f'{feature.capitalize()} Value')
plt.ylabel('Density')
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
<Figure size 1000x600 with 0 Axes>
<Figure size 1000x600 with 0 Axes>
<Figure size 1000x600 with 0 Axes>
```

Output:



4.1 Performance Measures

A confusion matrix was used in this study to accurately assess model performance on a classification problem. A confusion matrix summarizes prediction results by class regarding actual and correct and incorrect predictions. Mainly, a confusion matrix is a core summary of results containing these notations: TP, TN, FP, FN.

- TP stands for true positive, meaning an observation is positive and correctly predicted as positive.
- TN represents true negative; a negative case is correctly predicted as negative.
- FP is false positive — an observation is mistakenly predicted as positive when in fact it is negative.
- FN stands for False Negative - this is a genuinely positive observation that is mistakenly predicted as being negative.

These figures in a confusion matrix can theoretically give a perfect determination of how accurate the model is.

4.2 Classification Rate/Accuracy

Classification accuracy was computed using the expression shown below (i):

Accuracy = **(TP + TN)/(TP + TN + FP + FN)**
Where, **TP, TN, FP**, and **FN** refer to the values in the confusion matrix.

The data was split into training and test data sets to evaluate the different models. The 426 patient records were used to train the Linear Regression, Decision Tree, and Random Forest models developed through this technique, and the performance was matched against the rest of the 143 patient records acting as an independent test set. The model that was created was tested over the rest of the data points, acting as a test set. This helped estimate the classifier's accuracy in predicting new examples that were unseen during the training.

4.3 Confusion Matrix of the Linear Regression

		Predicted Cancer Status		
Actual Cancer Status	Total = 143	Benign	Malignant	
	Benign	TN = 86	FP = 4	90
	Malignant	FN = 4	TP = 49	53
		90	53	**Total = 143**

From Table 2 results, the Linear Regression model accurately classified 135 patient records as either benign or malignant. However, the model inaccurately predicted the test diagnoses of 8 patient records in contrast. Thus, the general performance at this stage is regarded as the correctly classified ones as True Positives and True Negatives. In contrast, the incorrect ones predicted are known to be False Positives and False Negatives. The test-set-based accuracy of the developed Linear Regression model, computed from equation (i), according to these counts, was found to be 94.41%.

Accuracy= (TN+TP)/ (TN+TP+FN+FP)
= (86+49)/ (86+49+4+4)
= (135)/ (143)
=0.9441

Therefore**,** Linear Regression Accuracy was=**94.41%**

4.4 Confusion Matrix for Decision Tree Algorithm

		Predicted Cancer Status		
Actual Cancer Status	Total = 143	Benign	Malignant	
	Benign	TN = 84	FP = 6	90
	Malignant	FN = 1	TP = 52	53
		85	58	**Total = 143**

Table 3 above apparently displays that the Decision Tree algorithm correctly classified 136 patients' datasets as benign or malignant and 7 patients' datasets incorrectly as benign or malignant. As such, the accuracy of the Decision Tree algorithm was 95.10%.

Accuracy= (TN+TP)/ (TN+TP+FN+FP)
= (84+52)/ (84+52+1+6)
= (136)/ (143) =0.9510

Therefore, the accuracy for the Decision Tree Algorithm is**=95.10%**

4.5 Confusion for the Random Forest Algorithm

As regards the Random Forest algorithm, it correctly classified 138 patients' datasets as malignant or benign and 5 patients' datasets incorrectly as malignant or benign. As such, the accuracy of the Random Forest model was 96.50%.

Accuracy= (TN+TP)/ (TN+TP+FN+FP)

= (87+51)/ (87+5+2+3)

= (138) / (143) =0.9650

Therefore, Random Forest Accuracy=**96.50%**

Considering everything, the results demonstrated that, a random forest maintained higher classification performance by providing relatively better results. This outcome is premised on the fact that linear regression and decision tree models require more significant quantities of training data to maintain higher levels of accuracy. The highest proper accuracy rates were obtained by the random forest classifiers that the available data had trained. These show the effectiveness of such a classifier in terms of detecting breast cancer malignancies.

4.6 How to Use the Proposed Algorithms

Step 1: Define Healthcare goals- Establish what needs to be predicted by the intended model such as the root cause of the disease, disease risk, treatment response, and how it will help patients.

Step 2: Gather Electronic Health Record- Retrieve patient demographics, biopsy and mammography reports, risk factors, as well as family history from Electronic Health Records**.**

Step 3: Pre-process the data- Remove missing data, normalize numeric variables, as well as encode categories.

Step 4: Define Target Attributes- Clearly describe the targeted attributes (presence of breast cancer as well predictor attributes (age, density, radius, texture, or compactness)

Step 5: Split Data-Split data into train and test data keeping a profile and record of both datasets.

Step 6: Train the Model- Train the random forest algorithm on mammography images and related data.

Step 7: Consolidate the Random Forest Algorithm with the clinical workflow-Tailor APIs to consolidate the algorithm into the Electronic Health Records.

Step 8: Continuously Monitor and Evaluate the Model: Monitor the model for errors, biases, or unexpected outcomes and update the algorithm respectively.

5. Conclusion

The chief objective of this research paper was to explore machine learning algorithms used in the early detection of breast cancer in the USA. Three machine learning algorithms were trained with the training dataset, most notably, the Decision Tree algorithm, and the Random Forest. Subsequently, the performance of these models was compared to obtain the best model. Considering everything, the results demonstrated that, a random forest maintained higher classification performance by providing relatively better results. This outcome is premised on the fact that linear regression and decision tree models require more significant quantities of training data to maintain higher levels of accuracy. The highest proper accuracy rates were obtained by the random forest classifiers that the available data had trained. These show the effectiveness of such a classifier in terms of detecting breast cancer malignancies. In that respect, Random Forest models can assist in identifying high-risk patients in advance for prompt treatment. In that regard, such detection saves lives and decreases long-term healthcare costs for the US government.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Abunasser, B. S., AL-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2023, May). Literature review of breast cancer detection using machine learning algorithms. In AIP Conference Proceedings (2808, 1). AIP Publishing.

[2] Adam, W., & Kevin, M. (2024). Scrum Master's Role in Orchestrating Big Data Analytics and Machine Learning Projects for Business Success. *Journal Environmental Sciences And Technology*, *3*(1), 504-514.

[3] Agarap, A. F. M. (2018, February). 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 (5-9).

[4] Al-Mansouri, A. (2024). Anticipating Churn: AI-driven Insights for Sustaining Customer Loyalty in US Business Markets. *Journal of Engineering and Technology*, *6*(1), 1-8.

[5] Allugunti, V. R. (2022). Breast cancer detection is based on thermographic images using machine learning and deep learning algorithms. *International Journal of Engineering in Computer Science*, *4*(1), 49-56.

[6] Basker, N., Theetchenya, S., Vidyabharathi, D., Dhaynithi, J., Mohanraj, G., Marimuthu, M., & Vidhya, G. (2021). Breast cancer detection using machine learning algorithms. Annals of the Romanian Society for Cell Biology, 2551-2562.

[7] Bazazeh, D., & Shubair, R. (2016, December). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th International Conference on Electronic Devices, systems, and Applications (ICEDSA) (1-4). IEEE.

[8] Bhise, S., Gadekar, S., Gaur, A. S., Bepari, S., & Deepmala Kale, D. S. A. (2021). Breast cancer detection using machine learning techniques. Int. J. Eng. Res. Technol, 10(7), 2278-0181.

[9] Chopra, B., & Raja, V. (2024). Towards Improved Privacy in Digital Marketing: A Unified Approach to User Modeling with Deep Learning on a Data Monetization Platform. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, *4*(1), 163-178.

[10] Ding, R., & Hao, T. (2024). Analytics-Driven Transformation: Crafting Business Success Stories with Data. *Journal Environmental Sciences And Technology*, *3*(1), 138-149.

[11] IJSRCSEIT. (2020). Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction - a review. Technoscienceacademy. https://www.academia.edu/44802327/Comparative_Study_of_Machine_Learning_Algorithms_for_Breast_Cancer_Prediction_A_Review?sm=b

[12] IJSRCSEIT. (2021). Detection of breast cancer using machine learning algorithms. Technoscienceacademy. https://www.academia.edu/45327721/Detection_of_Breast_Cancer_Using_Machine_Learning_Algorithms?sm=b

[13] Jack, E., & Logan, J. (2024). Leveraging Machine Learning for Enhanced Big Data Analysis in Business Markets: A Scrum Master's Perspective. *Journal Environmental Sciences And Technology*, *3*(1), 484-495.

[14] Kumar, W. (2024). *Artificial Intelligence: a Comprehensive Exploration of Current Realities and Future Trajectories, Guiding Informed Decision-Making in a Dynamic Technological Landscape* (No. 12259). EasyChair.

[15] Lambert, A., & Smith, J. (2024). *Guardians of the Virtual Gate: AI's Impact on Cyber Defense Strategies* (13311). EasyChair.

[16] Master, T. (2019). Comparative Study of Machine learning Algorithms for breast cancer detection and diagnosis. Lkouniv. https://www.academia.edu/39157528/Comparative_Study_of_Machine_Learning_Algorithms_for_Breast_Cancer_Detection_and_Diagnosis?sm=b

[17] Meenalochini, G., & Ramkumar, S. (2021). Survey of machine learning algorithms for breast cancer detection using mammogram images. *Materials Today: Proceedings*, *37*, 2738-2743.

[18] Mesleh, A. (2021). Breast cancer detection using machine learning algorithms. Al-balqa. https://www.academia.edu/61970439/Breast_Cancer_Detection_Using_Machine_Learning_Algorithms?sm=b

[19] Moreno, M., & Hernández, W. G. (2024). The Business Intelligence Blueprint: Integrating Analytics, Big Data, and Project Management for Organizational Triumph. *Journal Environmental Sciences And Technology*, *3*(1), 161-175.

[20] Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. In *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5* (pp. 108-117). Springer Singapore.

[21] Preston, R., & Smith, J. (2024). *Fortifying Cybersecurity: Exploring AI's Role as Guardians of the Virtual Gate* (13310). EasyChair.

[22] Smith, J., & Anderson, J. (2024). *AI's Watchful Eye: Strengthening Cyber Defense as Guardians of the Virtual Gate* (13302). EasyChair.

[23] Sindhwani, N., Rana, A., & Chaudhary, A. (2021, September). Breast cancer detection using machine learning algorithms. In *2021 9th International Conference on Reliability, Infocom technologies and optimization (trends and future directions)(ICRITO)* (1-5). IEEE.