JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Machine Learning-Based Rational Breast Cancer Diagnosis

Khadeeja Naqvi*[1], Divyanshi Gautam [2], Ashish Kumar Srivastava[3], Prof. (Dr.) Syed Qamar Abbas[4], Dr. Nikhat Akhtar[5]

[1*]Scholar (B.Tech Final Year) Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow, India
[2]Scholar (B.Tech Final Year) Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow, India
[3]Assistant Professor, Department of Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow
[4]Director General, Ambalika Institute of Management & Technology, Lucknow, India
[5]Associate Professor, Department of Computer Science & Engineering, Ambalika Institute of Management & Technology, Lucknow

Abstract: One of the most hazardous illnesses for people is cancer, yet there is currently no long-term treatment available. The most frequent cause of cancer-related mortality is breast cancer. Finding cancer in its early stages is crucial. While rates are rising in practically every location worldwide, they are greater among women in more developed areas. But while it's still in its early stages, the cancer can still be cured. The prognosis and recovery of breast cancer patients are enhanced by early identification and rapid, efficient therapy. When identifying tumors, there is a significant chance of ambiguity and inaccurate detection, which has to be addressed. If patients are appropriately categorized, unnecessary treatments can be avoided. Medical imaging research now heavily relies on machine learning (ML). Data classification techniques based on machine learning are efficient. Particularly in the realm of medicine, where those techniques are frequently utilized in diagnosis and analysis for decision-making. Based on the features supplied by the data, we employ a variety of machine learning techniques to predict whether a tumor is benign or malignant in this situation. In this article, we are presented a system that can identify breast cancer and discussed how machine learning (ML) algorithms might enhance breast cancer early detection and diagnosis.

Keywords: Machine Learning, Breast Cancer, Data Exploration, Classification, Convolutional Neural Network (CNN), UCI Machine Learning Repository.

1. Introduction

One of the leading causes of mortality for women worldwide is breast cancer. The World Health Organization (WHO) reports that recently, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally [1]. The most typical cancer in women is breast cancer, according to the Centers for Disease Control and Prevention (CDC) Trusted Source. The large variations in breast cancer survival rates are caused by several variables. This type of cancer [2] women has and the stage of the disease when they receive a diagnosis is two of the most crucial variables. Breast cancer is a kind of cancer that starts in breast tissue. Usually, breast cancer develops in the ducts or lobules of the breast. Additionally, cancer can develop in your breast's fat tissue or fibrous connective tissue [3]. In addition to often invading healthy breast tissue, unchecked cancer cells can also go to the lymph nodes beneath the arms. According to medical professionals, breast cancer was caused by breast cells that grew abnormally and then spread to the lymph nodes or other regions of the body. In order to prevent the effects of the following phase, it is [4] crucial to identify and stop the proliferation of these undesirable cells as soon as feasible. As a result, there is a lot of study on the proper diagnosis of breast cancer and the classification of individuals into benign or malignant categories.

Machine learning (ML) [5] is widely acknowledged as the preferred approach in breast cancer pattern classification and forecast modeling due to its distinct benefits in essential features discovery from complicated breast cancer datasets. The majority of breast cancer cases, nevertheless, cannot be traced back to a single factor [6]. Discuss your unique risk with your doctor. In the age women become older, their chances of developing breast cancer rise. Women over the age of 50 are the ones who develop breast cancer at a rate of about 80%. The personal experience with breast cancer a woman who has already had breast cancer in one breast is more likely to have it in the other breast. The history of breast cancer in the family, if a woman's mother, sister, or daughter had breast cancer, particularly when she was young, her chance of developing breast cancer is increased (before 40). Another risk factor is having relatives who have breast cancer. Genetic influences, while it is currently difficult for doctors to determine whether a tumors is dangerous or not by looking at merely x-ray pictures, building a machine learning model according to the identification of the tumors [7] can be very helpful. Machine learning is becoming more and more in

demand, eventually becoming a service. Unfortunately, there are still significant obstacles to entry and specialised skills in the field of machine learning. It takes a certain set of abilities and knowledge to create an efficient machine learning model that includes the pre-processing, feature-selection [8], and classification phases [9].

The Pre-processing, feature extraction, and classification are the three basic steps of the many data mining [10] and machine learning algorithms that have been developed in the previous several decades for breast cancer detection and classification [11]. The pre-processing of mammography films aids in improving the visibility of peripheral regions and intensity distribution, which facilitates interpretation and analysis [12]. Several approaches have been described to help with this procedure. Because it aids in the differentiation between benign and malignant tumours, feature extraction is a crucial step in the identification of breast cancer. Following extraction, segmentation is used to extract picture attributes such smoothness, coarseness, depth, and regularity [13]. However, digital pathology (DP) is a technique for digitising histology slides in order to provide high-resolution photographs. Through the use of image analysis tools, these digital pictures are employed for detection, segmentation, and classification. The possibility that CNN [14] presents to research on medical imaging is not limited to deep CNN for extraction of the imaging [15] feature. Additional stages are necessary in deep learning (DL) employing CNNs, such as digital staining, to grasp patterns for image categorization [16]. In fact, the application of CNN for synthetic picture rendering is a second area that might aid in medical research.

2. Related Work

This section discusses some of the related research on machine learning-based breast cancer diagnosis that has been conducted in the past. By using machine learning techniques like the Convolutional Neural Network (CNN) method for breast image classification, conventional Neural Network (NN), Random Forest (RF) algorithm, Support Vector Machines (SVM) [17], and Bayesian methods, Abdullah-Al Nahid and Yinan Kong [18] presented a novel method to detect breast cancer by image classification. Since Convolutional Neural Network (CNN) approaches often extract the features globally using kernels and these Global Features have been utilized for image classification, the CNN method showed to be the best for the diagnosis of breast cancer.

The dataset from William H. Walberg of the University of Wisconsin Hospital was used by Muhammet Fatih Ak [19]. This dataset was subjected to data visualization [20] and machine learning methods as logistic regression, k-nearest neighbors, support vector machine, naive Bayes, decision tree, random forest, and rotation forest. These machine learning methods and visualization were implemented using R, Minitab, and Python. All the approaches were compared in a comparative study. The best classification accuracy (98.1%) was obtained using the logistic regression model with all characteristics included, and the suggested method demonstrated improved accuracy results. Three machine learning techniques Support Vector Machine (SVM), Random Forest (RF), and Bayesian Networks (BN) [21] were compared by Dana Bazazeh [22] and Raed Shubair. As a training set, the original Wisconsin breast cancer data set was employed. The performance of classification is proven to vary depending on the approach chosen through simulation results. Results revealed that SVMs perform better in terms of precision, sensitivity, and accuracy. However, RFs are more likely to categories tumors accurately. The authors in [23] provided a summary of the many methods for classifying breast cancer using histopathological image analysis (HIA) based on several ANN designs [24]. The applicable dataset was used to group the authors' research. They put it in chronological sequence, going up. This study discovered that ANNs were initially used to HIA circa 2012. The most often used algorithms were ANNs and PNNs. The majority of the work in feature extraction, however, utilized textural and morphological traits. It was evident that Deep CNNs were very useful for diagnosing and early detection of breast cancer, resulting in more successful therapy. Numerous methods were used for non-communicable disease (NCD) prediction.

The efficiency of NNs in the categorization of cancer diagnoses, particularly in the early stages, was shown by the authors in [25]. The majority of NNs have demonstrated potential in identifying malignancy cells, according to their study. To preprocess the pictures, however, the imaging technique needs powerful computing power [26]. In [27] examined several data mining, deep learning, and machine learning techniques relevant to breast prediction and diagnosis. The authors noted that only a small number of studies utilized genetics, whereas the majority of the papers made use of imaging. SVM, decision trees, and random forests were the three major algorithms utilised in the genetic breast cancer prediction process. However, several algorithms, like CNNs and Nave Bayes, were applied in imaging approaches. The authors of [28] reviewed current research using several imaging modalities to detect breast cancer using deep learning [5]. These studies were arranged according to the dataset, architecture, application, and assessment criteria. They concentrated on three breast imaging modalities' deep learning frameworks (ultrasound, mammography and MRI). They made an effort in their study to present cutting-edge [29] discoveries on breast cancer imaging using DLR-based CAD systems. They used CNN classification and private datasets in their research.

3. Objectives

One of the most common cancers to be discovered worldwide, including in India, is breast cancer. Despite the excellent survival rate, 97% of women can survive for more than 5 years with early diagnosis. According to statistics, the number of deaths caused by this illness has dramatically grown in recent years. The goal of this research is to determine which characteristics are most useful [30] in predicting whether a malignancy is malignant or benign as well as to identify broad trends that might help us choose the right model and hyper parameters. Identifying whether a breast cancer is benign or aggressive is the objective. In order to do this, I fitted a function that can predict the discrete class of fresh input using machine learning classification algorithms. Identifying whether a breast cancer is benign or aggressive is the objective.

4. Prerequisite

This section is important when starting a project since it explains which section uses or follows which piece of technology. The development team should be aware of all the project's features and applications before defining the hardware and software requirements.

.

4.1 UCI Machine Learning Repository

In this project, we will analyses data to find breast cancer using data mining and machine learning algorithms. Women all throughout the world commonly get breast cancer (BC). By encouraging patients to receive therapeutic therapy, early diagnosis of BC can significantly increase prognosis and survival prospects. For the breast cancer dataset, we'll utilize the UCI Machine Learning Repository. We are browsing a database of machine learning issues for free at the UCI Machine Learning Repository [31]. The University of California, Irvine's Center for Machine Learning and Intelligent Systems is responsible for hosting and maintaining it. As a PhD student at UC Irvine, David Aha first developed it. It has been the go-to resource for machine learning practitioners and researchers who need a dataset for more than 25 years. Every dataset has a webpage with all the information that is currently available about it, including any pertinent research articles. The actual datasets are available for download as ASCII files, frequently in the practical CSV format. The dataset used in this article was developed by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, Wisconsin, in the United States, and is openly accessible. Dr. Wolberg used fluid samples obtained from patients with solid breast masses and Xcyt, an easy-to-use graphical computer tool that can analyses cytological characteristics based on a digital scan, to construct the dataset.

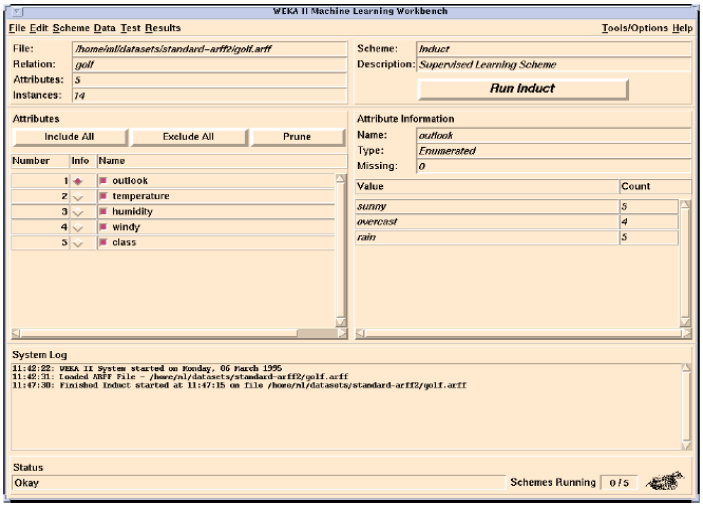

Figure 1 The WEKA Main Screen

4.2 Waikato Environment for Knowledge Analysis (WEKA)

The Waikato Environment for Knowledge Analysis examines us (WEKA). The three most well-liked data mining algorithms Naive Bayes, RBF Network, and J48 were utilised to create the prediction models. A workbench called WEKA is intended to help in the application of machine learning to real-world data sets. For data mining jobs, WEKA is a set of machine learning algorithms [32]. We are either invoking the algorithms directly from your own Java code or apply them directly to a dataset. It has tools for mining association rules, grouping, regression, classification, and visualization of data. The modules will be merged in a way that results in the required output when the user interacts with WEKA. Sun Microsystems UNIX workstations running Solaris 2 are used to run the WEKA workbench, which was created using the TCL/TK programming language and X window toolkit.

4.3 Data Mining and Machine Learning

The phrase "data mining" is misleading because the objective is not the extraction (mining) [33] of data itself but rather the extraction of patterns and information from vast volumes of data. It is also a buzzword that is frequently used to refer to any kind of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics), as well as any use of computer decision support systems, such as business intelligence and artificial intelligence [34] (such as machine learning). The following machine learning algorithms are employed in this research.

4.3.1 Decision Tree Algorithms

Successful machine learning classification approaches include decision tree algorithms. They are supervised learning techniques that make use of gathered and edited information to enhance outcomes. Additionally, decision tree algorithms are frequently employed for categorization in a wide range of studies, for instance, in the field of medicine and health problems. Decision tree algorithms come in numerous varieties, including ID3 and C4.5 [29]. The most often used decision tree algorithm, nevertheless, is J48. J48 is an extension of ID3 and the implementation of a better version of C4.5.

4.3.2 K-nearest-Neighbors (KNN) Algorithm

It is a straightforward approach for supervised learning in pattern recognition. Due to its ease of use and effectiveness in the field of machine learning, it is one of the most often used neighborhood classifiers. KNN method searches the pattern space for the k training tuples that are most similar to the unknown tuples. It then saves all cases and categorizes new instances based on similarity measurements. The appropriate number of neighbors (k) relies on performance and varies from one data sample to another.

4.3.3 Support Vector Machine (SVM)

It is a supervised learning technique for categorizing both linear and nonlinear data that is developed from statistical learning theory. By increasing the margin of hyper plane splitting, SVM [35] divides data into two classes over a hyper plane while avoiding over-fitting the data.

4.3.4 Naïve Bayes (NB)

A probabilistic classifier, it uses strong (naive) independent assumptions to apply Bayes' theorem [36] to one of the most effective classification methods. Given the class variable, it is assumed that the feature's value is unrelated to the values of

any other features. Based on the greatest likelihood. It determines whether the provided tuple belongs to a specific class.

4.3.5 Logistic Regression

The logistic model (also known as the genuine model) in statistics is used to estimate the likelihood that a certain class or event, such as pass/fail, win/lose, alive/dead, or healthy/ill, will occur. This may be expanded to simulate a variety of event classes, such identifying the presence of a cat, dog, lion, etc. in a picture. Each object in the image that is detected would be given a probability between 0 and 1, with the aggregate equaling 1. Early in the 20th century, the biological sciences began to employ logistic regression. Then, it was put to many different social science uses. When the dependent variable (target) is categorical, logistic regression is utilized.

5. Proposed System

We will be able to differentiate between malignant and benign tumors more quickly with the suggested technique. Despite being a sophisticated and difficult classifier, CNN can automatically extract important characteristics without the need for preprocessing. It is more effective since it filters the crucial variables and is versatile enough to perform incredibly well with picture data. In order to handle data having a grid-like architecture, CNNs [37] is a sort of deep learning method. CNNs are a subset of deep learning algorithms that are employed to interpret spatially or temporally related data. CNNs are comparable to other neural networks, but since they employ the convolutional layers seen in figure 2, they have an additional level of complexity.

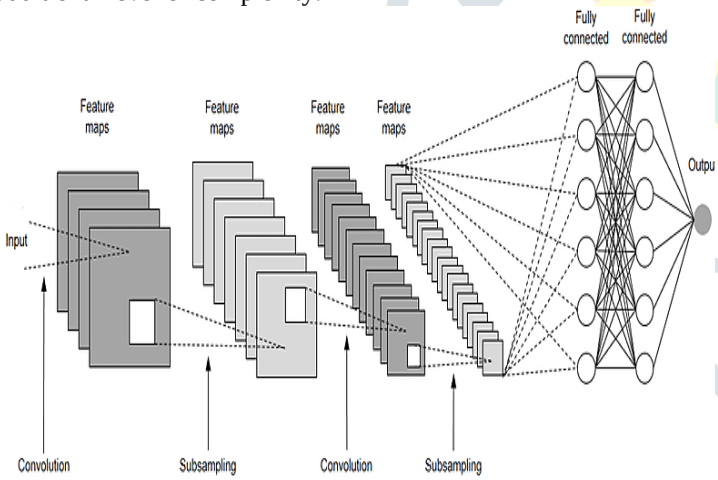

Figure 2 The Convolutional Neural Network (CNN)

An integral part of convolutional neural networks are convolutional layers (CNNs) [38]. Our project's main goal is to employ a convolution neural network using Keras as the back end to distinguish between malignant and benign tumors, and then we'll study the results to determine how the model may be beneficial in real-world scenarios.

Step 0 — Data Preparation

For the breast cancer dataset, we'll utilize the UCI Machine Learning Repository. After computing ten features from each sample cell using a curve-fitting technique, the computer calculates each feature's mean value, extreme value, and standard error for the picture before delivering a 30 real-valued vector of attribute information.

- ID number
- Diagnosis (M = malignant, B = benign) 3–32) Ten real-valued features are computed for each cell nucleus: 2. 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 (perimeter² / area — 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- 10. Symmetry
- Fractal dimension (“coastline approximation” -1) the mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius.

Step 1 — Data Exploration

To work with this dataset, we are use Spyder. We'll start by adding the required libraries before adding our dataset to Spyder.

```
#importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd#importing our cancer dataset
dataset = pd.read_csv('cancer.csv')
X = dataset.iloc[:, 1:31].values
Y = dataset.iloc[:, 31].values
```

Figure 3 Pandas Data Set

We can examine the data set using the pandas’ head() method shown in figure 3.

	id	radius_mean	...	fractal_dimension_worst	diagnosis
0	842302	17.99	...	0.11890	M
1	842517	20.57	...	0.08902	M
2	84300903	19.69	...	0.08758	M
3	84348301	11.42	...	0.17300	M
4	84358402	20.29	...	0.07678	M

Top 5 data of our dataset

We can find the dimensions of the data set using the panda dataset 'shape' attribute.

print("Cancer data set dimensions : {} ".format (dataset.shape))Cancer data set dimensions : (569, 32)

The data set has 569 rows and 32 columns, as we can see. The column labeled "Diagnosis" will tell us if the cancer is M = malignant or B = benign. Malignant cancer is indicated by a 1 and benign cancer by a 0. We can see that 357 of the 569 individuals are classified as B (benign), and 212 as M. (malignant). Data visualization [39] is a crucial component of data science. Understanding data and explaining it to someone else both assist. As seen in figure 4, Python includes a number of useful visualization packages, including Matplotlib, Seaborn, and others. To determine the data distribution of the features, we will utilize pandas' visualization, which is built on top of matplotlib, in this work.

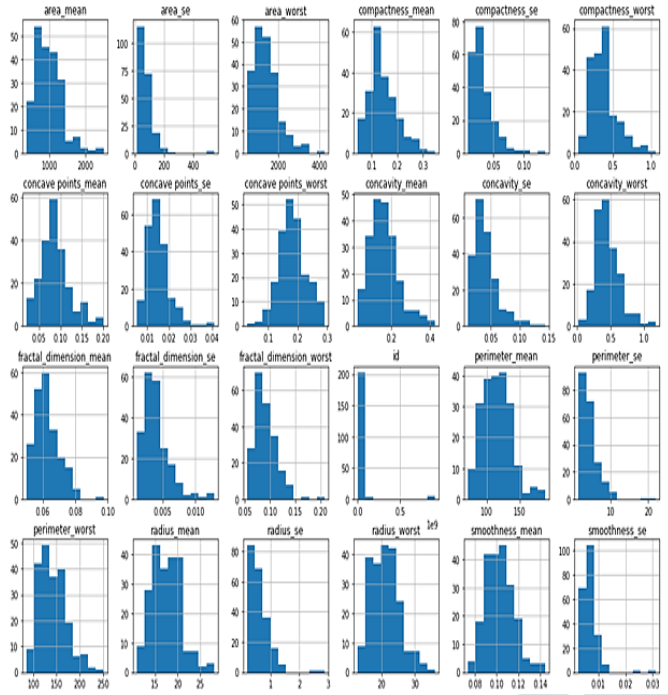

Figure 4 The Visualization of Dataset

We can find any missing or null data points of the data set (if there is any) using the following pandas function shown in figure 5.

```
id                         0
radius_mean                0
texture_mean               0
perimeter_mean             0
area_mean                  0
smoothness_mean            0
compactness_mean           0
concavity_mean             0
concave points_mean        0
symmetry_mean              0
fractal_dimension_mean     0
radius_se                  0
texture_se                 0
perimeter_se               0
area_se                    0
smoothness_se              0
compactness_se             0
concavity_se               0
concave points_se          0
symmetry_se                0
fractal_dimension_se       0
radius_worst               0
texture_worst              0
perimeter_worst            0
area_worst                 0
smoothness_worst           0
compactness_worst          0
concavity_worst            0
concave points_worst       0
symmetry_worst             0
fractal_dimension_worst    0
diagnosis                  0
dtype: int64
```

Figure 5 The Missing or Null Data points

Step 2 — Categorical Data

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. For example, users are typically described by country, gender, age group etc. We are use Label Encoder to label the categorical data [40]. Label Encoder is the part of SciKit Learn library in Python and used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

#Encoding categorical data values
from sklearn.preprocessing import LabelEncoder
labelencoder_Y = LabelEncoder()
Y = labelencoder_Y.fit_transform(Y)

The data we use is usually split into training data and test data shown in figure 6. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. We are using SciKit-Learn library in Python [41] using the train_test_split method.

 # Splitting the dataset into the Training set and Test setfrom sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25, random_state = 0)

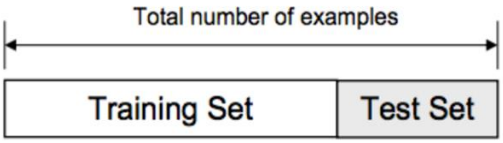

Figure 6 Training and test set

Step 3 — Feature Scaling

Our dataset typically includes features with a wide range of magnitudes, units, and ranges. However, as Eucledian distance between two data points is used in the majority of machine learning methods, All characteristics must be brought to the same magnitude level. Scaling can be used to accomplish this.

This implies that you're altering your data to make it compatible with a certain scale, such as 0-100 or 0-1. We are using of the SciKit-Learn library's StandardScaler function.

```
#Feature Scalingfrom sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Step 4 — Model Selection

The most thrilling part of applying machine learning to any dataset is this stage. It is sometimes referred to as algorithm selection for optimum outcome prediction. To analyses massive data sets, data scientists often utilize a variety of machine learning methods. However, on a broad scale, all of those various algorithms may be divided into two categories: supervised learning and unsupervised learning. I'll only offer a quick review of these two styles of learning to avoid spending too much time. In a supervised learning system, both the input data and the intended output data are given. In order to establish a learning foundation for future data processing, input and output data are labeled for categorization. Regression and classification problems are within the category of supervised learning issues. When the output variable is a real or continuous value, like "salary" or "weight," a regression problem exists. When the output variable is a category, such as filtering emails as "spam" or "not spam," there is a categorization difficulty. Unsupervised learning is when an algorithm uses data that has neither been classed nor labelled and is allowed to act on the data without supervision. Our dataset's outcome variable, or dependent variable, Y, only has two possible sets of values: M (Malign) or B. (Benign). So, we'll apply the supervised learning classification technique. Let's begin using the algorithms. To import all of the classification algorithm methods, we are using the Sklearn package. In order to employ the Logistic Regression algorithm, we use the Logistic Regression technique of model selection.

```
#Using Logistic Regression Algorithm to the Training Setfrom sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, Y_train)#Using KNeighborsClassifier Method of neighbors class to use Nearest Neighbor algorithmfrom sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
classifier.fit(X_train, Y_train)
#Using SVC method of svm class to use Support Vector Machine Algorithm
from sklearn.svm import SVC
classifier = SVC(kernel = 'linear', random_state = 0)
classifier.fit(X_train, Y_train)
#Using SVC method of svm class to use Kernel SVM Algorithm
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)
classifier.fit(X_train, Y_train)
#Using GaussianNB method of naïve_bayes class to use Naïve Bayes Algorithm
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, Y_train)
#Using DecisionTreeClassifier of tree class to use Decision Tree Algorithm
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, Y_train)
#Using RandomForestClassifier method of ensemble class to use Random Forest Classification algorithm
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
classifier.fit(X_train, Y_train)
```

We forecast the outcomes of the test set and evaluate the accuracy of each of our models. We must import the metrics class's confusion_matrix function in order to assess the correctness. The confusion matrix is a technique of keeping track of how many anticipated classes ended up in the incorrect classification bin based on the actual classes.

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred)
```

To determine the correctness of our models, we employ the classification accuracy approach. When we use the term accuracy, we often imply classification accuracy. It measures the proportion of accurate predictions to all input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

The confusion matrix object must be checked to determine the proper prediction, and the predicted outcomes must be added diagonally to get the number of accurate predictions, which must then be divided by the overall number of predictions, as illustrated in figure 7.

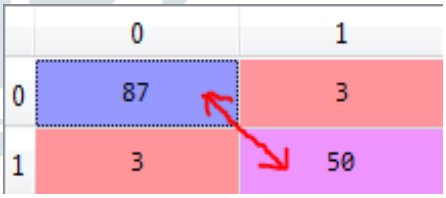

Figure 7 The Confusion Matrix

6. Result and Analysis

Using the head() function of the pandas package, we analyses the data collection. The first 7 rows of data in df.head(7) By utilising the panda [42] dataset's "shape" feature, we can determine the dimensions of the data set. df.shape (569, 33) we are see that there are 569 rows and 33 columns in the data set. The column labeled "Diagnosis" will tell us if the cancer is M = malignant or B = benign. Malignant cancer is indicated by a 1 and benign cancer by a 0. We can see that 357 of the 569 people have the label "B" (benign), while 212 have the label "M." (malignant). 33 characteristics from the 569 patients are represented by a patient in each row. We must eliminate the final column, Unnamed 32, since it contains NaN values. Therefore, we count how many columns are empty and remove the columns with empty values. 5 We eliminate column Unnamed: 32 because there are 569 missing

data. Therefore, the data's new form is (569, 32), which denotes the data's 569 rows and 32 columns. The number of malignant (M) or benign (B) cells which are hazardous or not can now be seen and shown on a graph. As we can see, there are six different data types for the columns. The id column, which serves as the patient's identification number and is of the integer data type, cannot be utilized as a feature to forecast the tumors. The values of categorical data are then encoded (categorical data is converted from strings to integers).

Here, the number 1 denotes malignant (M) cells, which are hazardous, while the value 0 denotes benign (B) cells, which are not harmful. We can now see a relationship between the various qualities. How strongly one column effects all the other columns in this heat map (e.g radius means has 32 percent influence on texture mean). Testing and Training. The datasets were then divided into independent (X) and dependent (Y) datasets. Y = df.iloc[:, 1].values where X = df.iloc[:, 2:31].values. They have an array type. The independent dataset (X) contains the attributes that are used to predict the result, whereas the dependent data set (Y) contains the diagnosis of the patient's malignancy. Now, we divide the dataset into 75 percent training data and 25 percent testing data, and we apply several machine learning [43] models to the training set, including decision trees, logistic regression, and random forest classifiers. The correctness of the training data is now printed.

As a result, we can see that the decision tree classifier, with a 100% accuracy rating, is the best model. We will now project the outcomes of the test set and evaluate the precision of each of our models: We must import the confusion matrix function [44] from the metrics class in order to verify correctness. The confusion matrix is a tool for keeping track of misclassifications, or the number of anticipated classes that were incorrectly classified when compared to the actual classes. 7 Where TP is true positive, the matrices in this case are of the type [TP FP] [FN TP]. When the model accurately predicts the positive class, the result is a genuine positive. True Negative (TN) and result where the model properly predicted the negative class is referred to as a true negative.
False Negative (FN). A false negative is a result that the model predicted wrongly. We may use Model 5, or Random Forest Classifier, to forecast whether a patient will get cancer or not based on the test results since it has a 96.5 percent accuracy rate. Model prediction: pred = model [5]. Print predict(X test) (pred).

So, we have the forecasts printed here. The first data displays the actual outcome for each cancer patient, while the second data represents the model's forecast. The model's accuracy is 96.5 percent, so while there are occasional incorrect predictions, for the most part it is effective in determining whether a tumors is malignant (M) (dangerous) or benign (B) (not harmful) based on the attributes offered in the data and the training supplied. The Kaggle website was used to get the Breast Cancer dataset. The datasets will be examined by the algorithm based on several criteria (such as area, smoothness, concavity).

6.1 Attributes

- **diagnosis:** The diagnosis of breast tissues (1=malignant,0=benign)
- **mean_radius:** mean of distances from center to points on the peremeter
- **mean_texture:** standard deviation of gray-scale
- **mean_perimenter:** mean size of the core tumor
- **mean_area mean_smoothness:** mean of local variation in radius length.

6.2 Classification Report

The accuracy of predictions made by a classification algorithm is evaluated using a classification report. The primary classification metrics of accuracy, recall, and f1-score are displayed in the report on a per-class basis. True and false positives and false negatives are calculated by the metrics. Positive and negative names for the anticipated classes are used here. There are four techniques to determine if the forecasts were accurate or not.

- **TP/True Positive:** when a case is positive and predicted positive.
- **TN/True Negative:** when a case is negative and predictive negative.
- **FN/False Negative:** when a case is positive and predicted negative.
- **FP/False Positive:** when a case is negative and predicted positive
-

```
dataframe = pd.DataFrame(Y)
#Encoding categorical data values
from sklearn.preprocessing import LabelEncoder
labelencoder_Y = LabelEncoder()
Y = labelencoder_Y.fit_transform(Y)


# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.25, random_state = 0)


#Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

#Fitting the Logistic Regression Algorithm to the Training Set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, Y_train)
#95.8 Acuracy

```

```
#Fitting K-NN Algorithm
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
classifier.fit(X_train, Y_train)
#95.1 Acuracy

#Fitting SVM
from sklearn.svm import SVC
classifier = SVC(kernel = 'linear', random_state = 0)
classifier.fit(X_train, Y_train)
#97.2 Acuracy

#Fitting K-SVM
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)
classifier.fit(X_train, Y_train)
#96.5 Acuracy

#Fitting Naive_Bayes
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, Y_train)
#91.6 Acuracy

#Fitting Decision Tree Algorithm
from sklearn.tree import DecisionTreeClassifier
```

```
#Fitting Naive_Bayes
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, Y_train)
#91.6 Acuracy

#Fitting Decision Tree Algorithm
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, Y_train)
#95.8 Acuracy

#Fitting Random Forest Classification Algorithm
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
classifier.fit(X_train, Y_train)
#98.6 Acuracy

#predicting the Test set results
Y_pred = classifier.predict(X_test)

#Creating the confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred)
c = print(cm[0, 0] + cm[1, 1])
```

Following the application of several classification models, we obtained the accuracies displayed in figure 8 using various models.

- Logistic Regression — 95.8%
- Nearest Neighbor — 95.1%
- Support Vector Machines — 97.2%
- Kernel SVM — 96.5%
- Naive Bayes — 91.6%
- Decision Tree Algorithm — 95.8%
- Random Forest Classification — 98.6%

Having completed the construction of our classification model, it is clear that the Random Forest Classification algorithm [45] produces the best outcomes for the data we have available. It doesn't, however, apply to every dataset. We must always first assess our dataset before applying our machine learning model to the model we want to use.

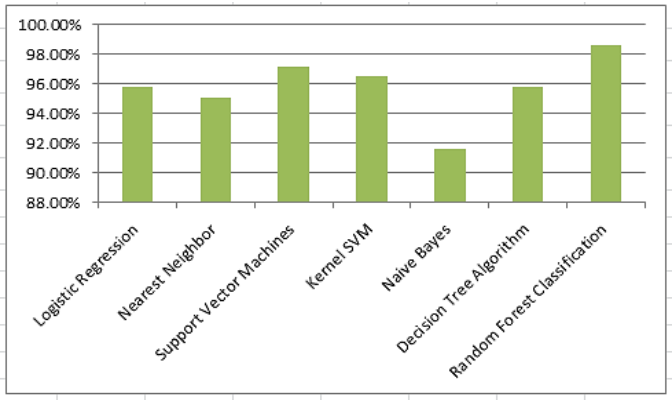

Figure 8 The Accuracies with Different Models

7. Conclusion

This work makes an effort to use a machine learning algorithm to address the issue of automated breast cancer diagnosis. This study investigated various machine learning methods for detecting breast cancer. The breast cancer dataset was used to execute the several different investigations. In this Python project, we developed graphs and results for the same breast cancer tumors predictor that we learnt to build using the dataset. A good dataset has been shown to offer improved accuracy. The creation of prediction systems will result from the selection of suitable algorithms with a strong home dataset. When a patient is diagnosed with breast cancer, these systems can help determine the best course of therapy. Based on the stage of a patient's breast cancer, there are a variety of therapies available; data mining and machine learning may be a big help in selecting the course of therapy to be taken by extracting knowledge from such appropriate databases. It is a difficult challenge to automate the identification of breast cancer to improve patient care. Finally, the suggested model appears to be ideally suited for automated breast cancer detection on the one side and control parameter setting of machine learning algorithms on the other.

8. Work in the Future

One of the top causes of mortality for women is breast cancer. The most serious problem for women is breast cancer. Breast cancer has surpassed lung cancer as the most frequent cancer in women diagnosed globally, according to data provided by the International Agency for Research on Cancer (IARC) in December 2021. Future research can focus on transforming the selected strategy into a potentially useful technique for giving clinicians a prompt second opinion when detecting breast cancer. In the future, we would like to expand the dataset and evaluate the algorithm's effectiveness and scalability.

References

[1] Tamar Kakiashvili, Waldemar W. Koczkodaj," Assessing the properties of the World Health Organization's Quality of Life Index", International Multiconference on Computer Science and Information Technology, IEEE, Poland, 2008

[2] M. Akram, M. Iqbal, M. Daniyal and A. U. Khan, "Awareness and current knowledge of breast cancer", Biological research, vol. 50, no. 1, pp. 33, 2017

[3] S.-J. Han et al., "Prognostic significance of interactions between ER alpha and ER beta and lymph node status in breast cancer cases", Asian Pacific Journal of Cancer Prevention, vol. 14, no. 10, pp. 6081-6084, 2013

[4] H. Safizade, N. Amirzadeh and P. Mangolian Shahrbabaki, "Motivational Factors for Breast Cancer Screening Behaviors in Iranian Women: A Qualitative Study", Asian Pacific Journal of Cancer Prevention, vol. 21, no. 10, pp. 3109-3114, 2020

[5] Yusuf Perwej, "An Evaluation of Deep Learning Miniature Concerning in Soft Computing", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Volume 4, Issue 2, Pages 10 - 16, 2015, DOI: 10.17148/IJARCCE.2015.4203

[6] M. S. Moran et al., "Society of Surgical Oncology–American Society for Radiation Oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in stages I and II invasive breast cancer", Annals of surgical oncology, vol. 21, no. 3, pp. 704-716, 2014

[7] F. F. Ting and K. S. Sim, "Self-regulated Multilayer Perceptron Neural Network for Breast Cancer Classification", International Conference on Robotics Automotion and Sciences (ICORAS), 2017

[8] Yusuf Perwej, Shaikh Abdul Hannan, Nikhat Akhtar, "The State-of-the-Art Handwritten Recognition of Arabic Script Using Simplified Fuzzy ARTMAP and Hidden Markov Models", International Journal of Computer Science and Telecommunications (IJCST),,Sysbase Solution (Ltd), UK, London, Volume, Issue 8, Pages 26 - 32, 2014

[9] Mohamed A. Berbar, "Hybrid methods for feature extraction for breast masses classification", Egyptian Informatics Journal, 2017

[10] Yusuf Perwej, Mohammed Y. Alzahrani, F. A. Mazarbhuiya, Md. Husamuddin, "The State of the Art Cardiac Illness Prediction Using Novel Data Mining Technique", International Journal of Engineering Sciences & Research Technology (IJESRT), ISSN: 2277-9655, Volume 7, Issue 2, Pages 725-739, 2018, DOI: 10.5281/zenodo.1184068

[11] G. Valvano, G. Santini, N. Martini et al., "Convolutional neural networks for the segmentation of microcalcification in mammography imaging," Journal of Healthcare Engineering, vol. 2019, Article ID 9360941, 9 pages, 2019.

[12] A. P. Charate and S. B. Jamge, "*e preprocessing methods of mammogram images for breast cancer detection," International Journal on Recent and Innovation Trends in Computing and Commu., vol. 5, no. 1, pp. 261–264, 2017

[13] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," IEEE Transactions on Image Processing, vol. 9, no. 4, pp. 561–576, 2000.

[14] B. Sahiner et al., "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images", IEEE Trans. Med. Imaging, vol. 15, no. 5, pp. 598-610, 1996

[15] Yusuf Perwej , Firoj Parwej, Asif Perwej, "Copyright Protection of Digital Images Using Robust Watermarking Based on Joint DLT and DWT ", International Journal of Scientific & Engineering Research (IJSER), France, ISSN 2229-5518, Volume 3, Issue 6, Pages 1- 9, 2012

[16] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," IEEE Access, vol. 7, pp. 18940–18950, 2019

[17] Yusuf Perwej, Nikhat Akhtar, Firoj Parwej, "The Kingdom of Saudi Arabia Vehicle License Plate Recognition using Learning Vector Quantization Artificial Neural Network", International Journal of Computer Applications (IJCA), USA, ISSN 0975 – 8887, Volume 98, No.11, Pages 32 – 38, 2014, DOI: 10.5120/17230-7556

[18] Abdullah-Al Nahid and Yinan Kong Involvement of Machine Learning for Breast Cancer Image Classification: Asurvey, 2017

[19] Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications", 2020

[20] Nikhat Akhtar, Nazia Tabassum, Dr. Asif Perwej, Dr. Yusuf Perwej," Data Analytics and Visualization Using Tableau Utilitarian for COVID-19 (Coronavirus)", Global Journal of Engineering and Technology Advances (GJETA), ISSN : 2582-5003, Volume 3, Issue 2, Pages 28-50, 2020, DOI: 10.30574/gjeta.2020.3.2.0029

[21] Nikhat Akhtar, Devendera Agarwal, "An Efficient Mining for Recommendation System for Academics", International Journal of Recent Technology and Engineering (IJRTE), ISSN 2277-3878 (online), Volume-8, Issue-5, Pages 1619-1626, 2020, DOI: 10.35940/ijrte.E5924.018520

[22] Dana Bazazeh and Raed Shubair "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis", 2016

[23] Zhou X, Li C, Rahaman MM, Yao Y, Ai S, Sun C, et al. A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks. IEEE Access 2020;8:90931–56

[24] Yusuf Perwej, "The Bidirectional Long-Short-Term Memory Neural Network based Word Retrieval for Arabic Documents", Transactions on Machine Learning and Artificial Intelligence (TMLAI), Society for Science and Education, United Kingdom (UK), ISSN 2054-7390, Volume 3, Issue 1, Pages 16 - 27, 2015, DOI: 10.14738/tmlai.31.863

[25] Mahmood M, Al-Khateeb B, Alwash WM. A review on neural networks approach on classifying cancers. IAES Int J Artif Intell 2020;9:317–26

[26] Yusuf Perwej, Asif Perwej, Firoj Parwej, "An Adaptive Watermarking Technique for the copyright of digital images and Digital Image Protection", International Journal of Multimedia & Its Applications (IJMA), Volume 4, No.2, Pages 21- 38, 2012 , DOI: 10.5121/ijma.2012.4202

[27] Fatima N, Liu L, Hong S, Ahmed H. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access 2020, 8, 150360–76

[28] Pang T, Wong JHD, Ng WL, Chan CS. Deep learning radiomics in breast cancer with different modalities: Overview and future. Expert Syst Appl 2020;158:113501.

[29] Yusuf Perwej, Firoj, Nikhat Akhtar, "An Intelligent Cardiac Ailment Prediction Using Efficient ROCK Algorithm and K- Means & C4.5 Algorithm" , European Journal of Engineering Research and Science, Belgium, Vol. 3, No. 12, Pages 126 – 134, 2018, DOI: 10.24018/ejers.2018.3.12.989

[30] O. Golubnitschaja, M. Debald, K. Yeghiazaryan, W. Kuhn, M. Pešta, V. Costigliola, et al., "Breast cancer epidemic in the early twenty-first century: evaluation of risk factors cumulative questionnaires and recommendations for preventive measures", Tumor Biology, vol. 37, no. 10, pp. 12941-12957, 2016

[31] Sun Chang, Yue Shihong," Clustering Characteristics of UCI Dataset", 39th Chinese Control Conference (CCC),IEEE, Accession Number: 19948911 , China,2020

[32] H Malik, S Mishra and AP Mittal, "Selection of Most Relevant Input Parameters Using Waikato Environment for Knowledge Analysis for Gene Expression Programming Based Power Transformer Fault Diagnosis", Electric Power Components and Sys, vol. 42, no. 16, pp. 1849-1861, 2014

[33] anal Bani Issa, Omar Darwish, Doaa Habeeb Allah, Farah Shatnawi, Dirar Darweesh, Yahya M. Tashtoush, "Analysis of Jordanian University Students Problems Using Data Mining System", 2022 13th International Conference on Information and Communication Systems (ICICS), pp.220-225, 2022

[34] Asif Perwej, Prof. K. P. Yadav, Prof. Vishal Sood, Yusuf Perwej, "An Evolutionary Approach to Bombay Stock Exchange Prediction with Deep Learning Technique", IOSR Journal of Business and Management (IOSR-JBM), USA, Volume 20, Issue 12, Ver. V, Pages 63-79, 2018, DOI: 10.9790/487X-2012056379

[35] Asif Perwej, Yusuf Perwej, Nikhat Akhtar, and Firoj Parwej, "A FLANN and RBF with PSO Viewpoint to Identify a Model for Competent Forecasting Bombay Stock Exchange", International Journal of Advanced Computer Technology, 4 (1), Volume-IV, Issue-I, Pages 1454-1461, 2015, DOI: 10.6084/ijact.v4i1.60

[36] P. Ruiz, J. Mateos, G. Valls, R. Molina and A.K. Katsaggelos, "Bayesian Active Remote Sensing Image Classification", IEEE Transaction on Geoscience and remote sensing, 2012

[37] O. Abdel-hamid, L. Deng and D. Yu, Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition, pp. 3366-3370, 2013

[38] N. Kwak, Introduction to Convolutional Neural Networks (CNNs), 2016

[39] Nikhat Akhtar, Nazia Tabassum, Asif Perwej, Yusuf Perwej," Data Analytics and Visualization Using Tableau Utilitarian for COVID-19 (Coronavirus)", Global Journal of Engineering and Technology Advances, Volume 3, Issue 2, Pages 28-50, 2020, DOI: 10.30574/gjeta.2020.3.2.0029

[40] D. Gibson, J. Kleinberg and P. Raghavan, "Clustering categorical data: An approach based on dynamical systems", Proceedings of the 24th VLDB Conference, 1998

[41] V.L. Ceder, K. McDonald and D.D, Harms, The quick Python book, Manning, pp. 335, 2010

[42] R. Ding, L. Wang, Q. Zhang, Z. Niu, N. Zheng and G. Hud, "Fine-grained giant panda identification", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 2108-2112, May 2020

[43] Yusuf Perwej, Dr. Ashish Chaturvedi, "Machine Recognition of Hand Written Characters using Neural Networks", International Journal of Computer Applications (IJCA), USA, ISSN 0975 – 8887, Volume 14, No. 2, Pages 6-9, 2011, DOI: 10.5120/1819-2380

[44] M. Ohsaki, K. Matsuda, P. Wang, S. Katagiri and H. Watanabe, "Formulation of the kernel logistic regression based on the confusion matrix", Proc. IEEE Congr. Evol. Comput, pp. 2327-2334, 2015

[45] John Halloran, "Classification: Naive Bayes vs logistic regression", Technical report, 2009