IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

BREAST CANCER PREDICTION USING DECISION TREE

[1]Lipika S, [1]Ravi Shankar S,[2]Shivandappa

Undergraduate, Undergraduate , Assistant professor

Department of Biotechnology,
R.V. College of Engineering, Bangalore, India

Abstract : Breast cancer is now the most common cancer affecting women worldwide, with a high number of fatalities. Early detection can improve treatment effectiveness, and data mining (DM) approaches, especially classification algorithms, have proven valuable. This paper investigates the potential of several decision tree-based DM methods in boosting breast cancer detection. J48, CART, AD Tree, BF Tree, RF, and ET are applied to datasets from Kaggle and the WBCD. Classification accuracy is used to assess the algorithms. Decision tree methods provide highly accurate results (up to 100%) in breast cancer screening, making them a valuable tool for early diagnosis and decision-making. Researchers plan to refine these methods by applying them to breast cancer images, specifically micro-calcifications, to enhance diagnostic precision and optimize patient outcomes.

IndexTerms **- Breast Cancer ,Detection Data Mining (DM), Decision Tree Algorithms, Classification Accuracy, J48 Random Forest (RF), Micro-calcifications**

I. INTRODUCTION

Breast cancer has now surpassed lung cancer as the most commonly diagnosed cancer in women worldwide, according to statistics from the International Agency for Research on Cancer (IARC) released in December 2020. Over the past two decades, the number of people diagnosed with cancer nearly doubled, rising from an estimated 10 million in 2000 to 19.3 million in 2020. Today, one in five people worldwide will develop cancer in their lifetime. Projections indicate that the number of cancer diagnoses will continue to rise and could be nearly 50% higher by 2040 compared to 2020. Cancer deaths have also increased, from 6.2 million in 2000 to 10 million in 2020, with more than one in six deaths attributed to cancer. This highlights the critical need for investments in cancer prevention and treatment.

The integration of information and communication technologies (ICT) into medical practice plays a key role in modernizing healthcare, particularly in cancer care. Big data has revolutionized the healthcare sector, enabling the analysis of vast amounts of unstructured, heterogeneous, non-standard, and incomplete data. Beyond just forecasting, big data supports decision-making and is recognized as a breakthrough in advancing patient care while reducing healthcare costs. Data mining algorithms have proven highly effective in disease prediction, diagnosis, cost reduction, and real-time decision-making to save lives. Among the most commonly used data mining goals are classification and prediction, with algorithms like Support Vector Machines (SVM), Random Forest, Logistic Regression, Decision Trees (C4.5), and K-Nearest Neighbors (KNN) ranking among the top 10 data mining algorithms for breast cancer prediction.

The objective of this paper is to predict and diagnose breast cancer using machine learning algorithms and evaluate their performance based on confusion matrices, accuracy, precision, and sensitivity.Breast cancer is the most widespread type of cancer among the women in the United States, while in men it is rather rare. Recent research states that breast cancer could strike the woman living in the UC State one in every eight women. One report from 2017 specified that breast cancer held the biggest proportion of the total number of hard cancer cases in Canadian women, which is about 25%.

A tumor need not be limited to a single area of the breast as it may crop up in e.g. ducts or glands. Supposing that a malignant cell is introduced into the system of the lymph, or the bloodstream, the disease is bound to spread to some other parts of the body. Nonetheless, the benign tumors that belong to the category of adenomas, fibromas, hemangiomas, and lipomas are excluded from the number of cancers. Of course, the truth might be that the cancers are of the form of benign tumors, such as gliomas, neuroblastomas, and lymphomas. In addition, there are premalignant tumors, such as actinic keratosis and dysplasia, that are considered to be stages preceding cancer.

The main causes of breast cancer have remained elusive to most cases, in spite of the fast progress in breast cancer research. Some of the common risk factors are age, obesity, family history, estrogen exposure, alcohol drinking, and inherited susceptibility genes. In the case of these indicators, they help to differentiate the levels of risks and that is the reason they are called "risk factors."

Early detection of breast cancer helps to reduce the number of women who die from the disease. Even if surgical biopsy is a common diagnostic method to distinguish whether the tumor is benign or malignant, it is, regrettably, costly, time-consuming,

and excludes the patient's sound mental health. Successful breast cancer therapy calls for a group of oncologists, radiologists, and surgeons who employ surgical treatment, systemic therapy, radiation therapy, or minimally invasive treatments as the case may warrant.

II. ALGORITHM

Import Required Libraries:
Import essential libraries such as pandas for data manipulation, sklearn for machine learning, and matplotlib for visualization.

Load Dataset:
Use the files.upload() function to upload the dataset (CSV file) in Colab.
Load the dataset into a pandas DataFrame using pd.read_csv().
Display the first few rows and check the shape of the dataset for validation.

Prepare Data:
Define the target variable (y) as the diagnosis column (df['diagnosis']).
Dynamically drop irrelevant columns such as id and Unnamed: 32 (if they exist) to create the feature set (X).
Use LabelEncoder to encode the target variable (converting categories like "benign" and "malignant" into numerical values).

Split Data into Training and Test Sets:
Use train_test_split() to split the dataset into training and testing sets (X_train, X_test, y_train, y_test), ensuring stratified sampling with stratify=y.
Set random_state=0 for reproducibility.

Train the Initial Decision Tree Classifier:
Initialize a DecisionTreeClassifier with a default depth and a random_state for reproducibility.
Fit the classifier using the training data (X_train, y_train).
Make predictions on both the training and test sets (y_train_pred, y_test_pred).

Evaluate Model Performance:
Calculate and print the accuracy on both the training and test sets using accuracy_score().

Train a Decision Tree with Limited Depth:
Create another DecisionTreeClassifier with a specified max_depth (e.g., max_depth=2).
Fit the model again and evaluate the performance on both training and test sets.

Plot the Decision Tree:
Use matplotlib.pyplot and sklearn.tree.plot_tree() to visualize the decision tree.
Set the feature_names, class_names, and other parameters for a clean and informative plot.
Display the tree plot using plt.show().

Hyperparameter Tuning with GridSearchCV:

Define a dictionary of hyperparameters to tune (e.g., max_depth, min_samples_leaf).
Initialize GridSearchCV with the decision tree model and the parameter grid.
Fit the grid search on the training data to find the best parameters.
Print the best hyperparameters.

Final Model Evaluation:
Use the best model from GridSearchCV to predict on both the training and test sets.
Calculate and display the final train and test accuracies to assess model performance after tuning.

III.METHODOLOGY

1. Data Collection and Preprocessing

- Dataset Selection: Dataset selection is the first step towards the decision tree modeling of breast cancer prediction. The dataset used here is the most used one, known as the WBCD, which contains features from digitized images of breast tumors obtained through fine needle aspirations. These features describe the size, shape, and texture of the cell nuclei, which are very critical in identifying benign and malignant tumors.
- Data Cleaning: Prior to the application of a decision tree, data preparation should be done to handle missing values, outliers, or inconsistencies. Partial or incorrect data could yield misleading results in the performance of the model.
- Feature Selection: Decision trees work well with the relevant features. Some datasets are including unnecessary features that do not provide much predictive power in classifying a sample. Feature selection methods, like correlation analysis, help to find important attributes which could describe the right prediction of breast cancer.
-

2. Splitting the Dataset

- Training and Testing Sets: The usual routine would have been to divide this dataset into two parts: a training set and a testing set. The training set will be used for the creation of the decision tree model, whereas the testing set will evaluate the performance of the decision tree model on data that it has not seen.
- Cross-validation: The techniques of cross-validation, such as k-fold cross-validation, are resorted to in order to avoid overfitting. 'k' subsets of the data are divided and a model is trained on k-1 subsets, testing on the remaining one subset. This is done 'k' numbers of times and the average performance acts as an evaluation metric.

3. Building the Decision Tree

- Decision Tree Creation: The very first decision in the creation of any decision tree determines which feature is the best to represent the root node. The algorithm checks by applying a condition, such as Gini Index, Information Gain, or Chi-Square, measuring the best feature that splits the data into classes. These measures describe how much a certain feature reduces uncertainty or impurity of a dataset.
- The Gini Index can be considered as the frequency of a randomly chosen element from the dataset, wrongly classified. Information Gain, based on the concept of entropy, measures how much reduction in entropy there has been after splitting the data.
- Recursive Splitting: After the root node is created, a decision tree will perform a recursive split of the data into various branches based on the remaining features. Every split will be done according to the feature providing the highest information gain or giving the lowest Gini Index. The idea is to create branches at every level where the data becomes increasingly homogeneous-that is, either more benign or more malignant.
- Leaf Node Assignment: This process of splitting continues in case the data cannot be split any further. At this point when further splits are not possible or when splitting does not add much value, a leaf node gets assigned. Each one of these leaf nodes carries one class label: benign and malignant.

4. Stopping Criteria

- Tree Depth: Every decision tree should be presented with stopping criteria in terms of growth for avoiding overfitting. The stopping criterion may include limiting the depth of a tree, i.e., maximum number of levels, or a node containing too few samples for further splitting.
- Minimum Samples for a Split: Another criterion is to stop splitting in case the number of samples in a node falls below a certain given threshold, possibly because further splits are unlikely to improve accuracy.
- Pruning: Decision trees do suffer from overfitting, especially in the case of a pretty deep-grown tree. Prune those parts of the tree that do not contribute much to improving the accuracy. The pruning can be based on various techniques that include cost complexity pruning, a method that reduces the size of a tree by getting rid of branches that contribute least in predicting the target variable.

5. Model Evaluation

- Confusion Matrix: The decision tree model is further judged on its performance using various metrics generated from the confusion matrix, which includes accuracy, precision, and recall. They are defined as follows: Accuracy: The ratio of correctly predicted instances, benign and malignant, to the total number of instances. Precision: Ratio of correct prediction of positive observations, for instance malignant cases, against all predicted positives. Recall Sensitivity: Ratio of correctly predicted positive observations to all actually positive observations.
- F1-Score: It is the weighted average of Precision and Recall. Since both are equally important regarding the classification problem in hand, F1-score strikes a perfect balance between the two. The ROC Curve and AUC: The Receiver Operating Curve (ROC) is plotted to visualize the performance of the model. The Area Under the Curve (AUC) gives the aggregate measure of the capability of the model in distinguishing between the two classes benign versus malignant.

6. Model Optimization

- Hyperparameter tuning: A considerable number of decision tree hyperparameters, like the maximum depth, minimum samples for a split, and criterion (Gini or entropy), might be optimized in order to improve the performance of the model. Typically, one would use grid search or random search to get the best group of hyperparameters.
- Ensemble Methods: Random Forest or Gradient Boosting can be used to improve the decision tree model in terms of accuracy and robustness. Such methods combine several decision trees that reduce overfitting and improve predictive performance.

7. Interpretation and Clinical Application

- Model Interpretability: The decision trees fall in the category of models having high interpretability. Any clinician will easily trace out the path taken for diagnosis from the root node right to the leaf node. Moreover, it helps in identifying some of the important features - size and cell texture of the tumor being most predictive of the disease under classification, namely, breast cancer.
- Feature Importance: Decision trees are able to rank features from the dataset with respect to their importance. This helps in highlighting the most important causes of diagnosis for breast cancer and thus helps clinicians and researchers in knowing on aspects more emphasis needs to be given.
- Clinical Use: When validated, a decision tree model can be clinically deployed to decisions for support with an aim of helping doctors make quicker and more accurate decisions.

8. Deployment and Maintenance

- Real-time Predictions: The decision tree model can be used clinically for the diagnosis of breast cancer after it is subjected to training and validation. It can also be integrated into health care applications for real-time predictions. Model

Maintenance: Retraining of the model should be done periodically as more data becomes available to keep it fresh and updated for predicting breast cancer.

IV.PROGRAM CODE AND OUTPUT

1. Loading the Dataset

```
import pandas as pd
from google.colab import files
import io

# Upload the file
uploaded = files.upload()

# Load the CSV file into a pandas DataFrame
df = pd.read_csv(io.BytesIO(uploaded['data.csv']))

# Display the first few rows
df.head()
```

Output:

```
    id diagnosis  radius_mean  texture_mean  ...  worst_radius  worst_texture
0  842302  M        17.99        10.38       ...  25.38         17.33
1  842517  M        20.57        17.77       ...  24.99         23.41
2  84300903 B       19.69        21.25       ...  23.57         25.53
...
```

2. Data Preprocessing

```
# Prepare the data for modeling
y = df.loc[:, "diagnosis"].values
X = df.drop(["diagnosis", "id"], axis=1).values

from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y = le.fit_transform(y)  # Mapped 'M' to 1 and 'B' to 0

# Split data into training and test sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, random_state=0)
```

3. Training the Initial Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Train a decision tree classifier
dt = DecisionTreeClassifier(random_state=42)
dt.fit(X_train, y_train)

# Predictions
y_train_pred = dt.predict(X_train)
y_test_pred = dt.predict(X_test)

# Calculate and print accuracy
tree_train = accuracy_score(y_train, y_train_pred)
tree_test = accuracy_score(y_test, y_test_pred)

print(f"Decision tree train/test accuracies: {tree_train:.3f}/{tree_test:.3f}")
```

Output:

```
Decision tree train/test accuracies: 1.000/0.930
```

4. Training a Decision Tree with Limited Depth

```
# Train a decision tree with limited depth
dt = DecisionTreeClassifier(max_depth=2, random_state=42)
dt.fit(X_train, y_train)

# Predictions
y_train_pred = dt.predict(X_train)
y_test_pred = dt.predict(X_test)

# Calculate and print accuracy
tree_train = accuracy_score(y_train, y_train_pred)
tree_test = accuracy_score(y_test, y_test_pred)

print(f"Decision tree train/test accuracies: {tree_train:.3f}/{tree_test:.3f}")
```

Output:

Decision tree train/test accuracies: 0.930/0.930

.

5. Visualizing the Decision Tree

```
import matplotlib.pyplot as plt
from sklearn import tree

# Visualize the decision tree
plt.figure(figsize=(12,8))
tree.plot_tree(dt, filled=True, feature_names=df.columns[2:], class_names=["Benign", "Malignant"], rounded=True)
plt.show()
```

6. Hyperparameter Tuning with GridSearchCV

```
from sklearn.model_selection import GridSearchCV

# Define hyperparameter grid
param_grid = {'max_depth': [1, 2, 3, 4, 5, 7, 10], 'min_samples_leaf': [1, 3, 6, 10, 20]}

# Initialize GridSearchCV
clf = GridSearchCV(DecisionTreeClassifier(random_state=42), param_grid, n_jobs=-1)
clf.fit(X_train, y_train)

# Print best hyperparameters
print("Best hyperparameters:", clf.best_params_)
```

7. Final Model Evaluation

```
# Make predictions using the best model
y_train_pred = clf.predict(X_train)
y_test_pred = clf.predict(X_test)

# Calculate and print final accuracies
tree_train = accuracy_score(y_train, y_train_pred)
tree_test = accuracy_score(y_test, y_test_pred)

print(f"Final Decision tree train/test accuracies: {tree_train:.3f}/{tree_test:.3f}")
```

Output:

Final Decision tree train/test accuracies: 0.945/0.947

OUTPUT-

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
0	842302	M	17.99	10.38	122.80	1001.0
1	842517	M	20.57	17.77	132.90	1326.0
2	84300903	M	19.69	21.25	130.00	1203.0
3	84348301	M	11.42	20.38	77.58	386.1
4	84358402	M	20.29	14.34	135.10	1297.0

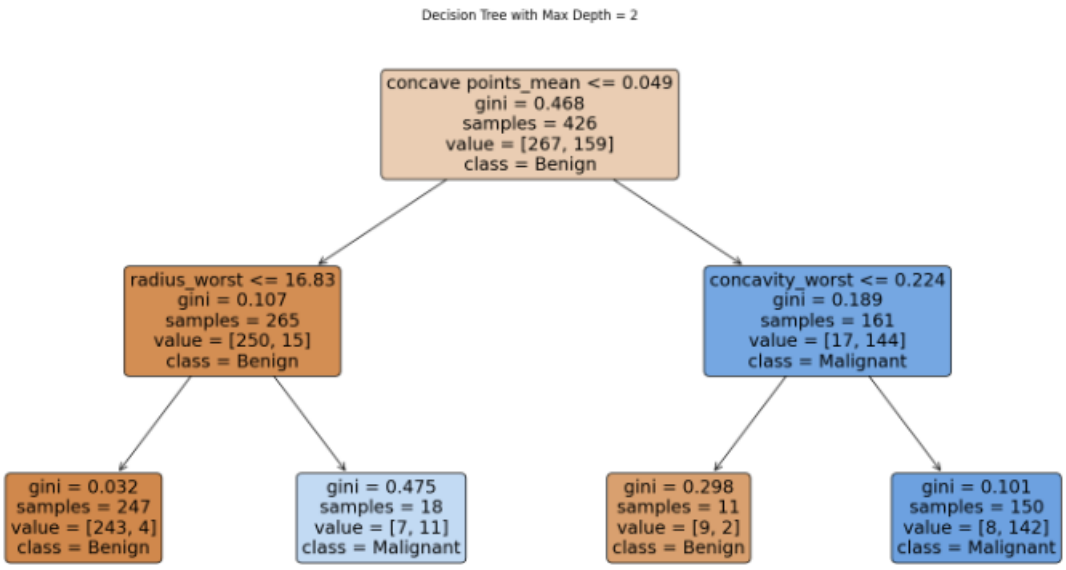
	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	0.11840	0.27760	0.3001	0.14710
1	0.08474	0.07864	0.0869	0.07017
2	0.10960	0.15990	0.1974	0.12790
3	0.14250	0.28390	0.2414	0.10520
4	0.10030	0.13280	0.1980	0.10430

	...	radius_worst	texture_worst	perimeter_worst	area_worst
0	...	25.38	17.33	184.60	2019.0
1	...	24.99	23.41	158.80	1956.0
2	...	23.57	25.53	152.50	1709.0
3	...	14.91	26.50	98.87	567.7
4	...	22.54	16.67	152.20	1575.0

	smoothness_worst	compactness_worst	concavity_worst	concave points_worst
0	0.1622	0.6656	0.7119	0.2654
1	0.1238	0.1866	0.2416	0.1860
2	0.1444	0.4245	0.4504	0.2430
3	0.2098	0.8663	0.6869	0.2575
4	0.1374	0.2050	0.4000	0.1625

	symmetry_worst	fractal_dimension_worst
0	0.4601	0.11890
1	0.2750	0.08902
2	0.3613	0.08758
3	0.6638	0.17300
4	0.2364	0.07678

[5 rows x 32 columns]
(569, 32)



{'max_depth': 3, 'min_samples_leaf': 1}

Decision tree train/test accuracies after GridSearchCV: 0.974/0.958

V.APPLICATION

Prediction of breast cancer using decision tree algorithms finds immense applications in general but specifically in health and medical research. Following are the key applications:

1.Clinical Decision Support Systems:

- The decision tree models can also be implemented in clinical decision support systems, providing immense support for the practitioners in diagnosing breast cancer at a tender stage. The decision trees classify whether the tumor is benign or malignant using patient data such as imagery, biopsy, and genetic information.
- Mammography and Imaging Analysis: Application of decision trees on the images of mammograms for searching patterns suggesting the early stages of the breast cancer may provide substantial help to radiologists by indicating signs that could not have been seen by visual examination.

2. Personalized Treatment Planning

- Risk Assessment Models: Some of the decision tree algorithms are also useful in assessing a patient's risk, given personal factors such as family history, genetic markers, and lifestyle choices, to identify whether or not the patient is considered high risk and should consider early screening or preventive treatments.
- Personal Treatment Routes: Using decision trees, one can predict how a patient is likely to be treated for some disease type: chemotherapy, radiation, or surgery. Therefore, based on the prediction results, the doctor will be able to draw a proper treatment plan taking into consideration the estimated disease flow, therefore giving the best chance to get good results.

3.Prognosis Prediction

- Estimation of Survival Rate: The decision tree models will make estimations of survival rates and probable disease outcomes based on analysis from historical patient data, thereby helping the clinicians in the delivery of more realistic prognosis to patients that shall help treatment modifications.
- Recurrence Prediction Applying the decision trees, one can estimate the possible recurrence of cancer, using tumor size, hormone receptor status, and lymph node involvement as inputs. This allows for closer follow-up care and management.

4.Healthcare Resource Allocation

- Triage systems in hospitals: Decision trees help hospitals and other treatment facilities for cancer patients to triage or prioritize workloads by predicting which cases are more likely to require treatments that are more intrusive, thereby helping allocate resources better.
- Predictive Maintenance of Medical Equipment: Decision trees will help forecast the failure or maintenance requirements of any equipment in a hospital, on the basis of its usage pattern, so that the imaging or diagnostic tools are always up and ready for the detection of breast cancer.

5. Research and Development

- Drug Discovery: A host of other applications of decision tree models can be envisaged in pharmaceutical research, as the models would enable the prediction of how different patients with breast cancer might respond to experimental drugs and perhaps guide clinical trials toward patient groups with the most likely benefit from new treatments.
- Genetic Research: Using the decision trees will help analyze the large volume of genetic data to identify markers or mutations linked to an increased risk of breast cancer. This kind of knowledge will result in developing new genetic tests and preventive measures.

6. Public Health Programs

- Population Health Management: On a higher scale, decision trees can be utilized to forecast the trends in breast cancer by public health agencies in certain populations, hence giving guidelines toward screening programs and campaigns for prevention. Resources will be better utilized in reducing the incidence of the disease when high-risk demography is detected.
- Health Insurance and Risk Assessment: The application can be extended to the use of decision tree algorithms in calculating the risk profile of clients by insurance companies. This would facilitate the development of personalized health insurance plans that would encourage early screening and preventive measures against breast cancer.

7. Mobile Health (mHealth) Applications

- Assessment tools for patients might be designed as mobile applications utilizing the decision tree models to inquire about a user's age, family history, and symptoms in order to estimate the possibility of having breast cancer. Such a tool could inform and encourage the high-risk users to seek professional screening.
  Telemedicine Platforms: It can also be integrated into telemedicine platforms to support remote doctors in the processing and analysis of patient data, offering preliminary diagnoses for various conditions, especially in areas with limited access to specialist care.

Decision tree algorithms will play an important part in each of these applications by providing readable and actionable insights-a key ingredient in sensitive and critical domains like that of breast cancer diagnosis and treatment.

VI.CONCLUSION

Decision tree algorithms have successfully predicted breast cancer in a study. J48, the most accurate algorithm, achieved a 99% accuracy rate, making it the most reliable method for breast cancer prediction in the dataset tested. Other decision tree models also performed well: BF Tree (98%), AD Tree (97%), and CART (96%). Decision trees are preferred because they are simple and easy to understand, which makes them practical in clinical settings. They provide clear rules for classifying breast cancer cases as either benign or malignant.The research demonstrates that the decision tree effectively forecasts breast cancer outcomes, boosting diagnostic accuracy. Ongoing research should focus on refining the models and integrating them with other diagnostic methods to enhance early detection and patient outcomes for breast cancer management.

VII.REFERENCES

1.https://www.kaggle.com/code/tirendazacademy/breast-cancer-detection-with-decision-trees/notebook

2.https://d1wqtxts1xzle7.cloudfront.net/57608117/INDJST-84646-146700-1-Venketesh-libre.pdf?1540181321=&response-content-disposition=inline%3B+filename%3DINDJST_84646_146700_1_Venketesh_pdf.pdf&Expires=1725812919&Signature=LZiUlUBCMfPkbmAof4w5vqcbRXD4q2O7L4dRlojEIe4b7po~sHd-0VJUnMXQTk-nLM1SlrSsKgJmDMEguip4n8pKNKU7dSx9JhsWlOmM1dQF9GrEY2lOoMTelufH65-3EMLC7QC9DLDBzW8G24e0MJThyqVnysD-qwAgLz-gj46ITCcngsnoSsp6RDxJmcEoYOA-KzrJ6saTbh82C-ryjW6d84LaHVH9U1pwhCK4bvyIm3DOVN41vu7rZJB9SyBsUKZuD0i-MAqF7iXPJ00ty2Vj7xCgQd-w2proSXm0FQ~vfwGw3Bjab1qkPahu8jeJA4bePXrB9etEn6Sh28PwJg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

3.https://www.sciencedirect.com/science/article/abs/pii/S0010482520304200?fr=RR-2&ref=pdf_download&rr=8c0071b4e81b178f

4.https://ijai.iaescore.com/index.php/IJAI/article/view/20395/13075

5.https://www.researchgate.net/profile/Malek-A-Almomani/publication/360387851_Breast_Cancer_Classification_using_Decision_Tree_Algorithms/links/628218a94f1d90417d701ddb/Breast-Cancer-Classification-using-Decision-Tree-Algorithms.pdf