Table Of Contents

1. Abstract

Breast cancer is the most common type of cancer in the world to affect women, and it also has the greatest death rate. Early detection of breast cancer greatly lowers the mortality rate, just as early detection of other cancers is crucial. Consequently, early identification of breast cancer greatly raises the likelihood of surviving. Because it can promote prompt clinical therapy, early detection of breast cancer can greatly improve survival rates.

The patient's breast cancer was identified as benign or malignant using machine learning algorithms, and the data quality of the Breast Cancer Wisconsin (Diagnostic) dataset, which contains metric data retrieved from the biopsy piece with various data mining techniques.

The dataset was divided into 80% for the training phase and 20% for the testing phase to implement the ML algorithms. When we compare the developed machine learning algorithms; the Support Vector Machine algorithm showed higher performance than other machine learning algorithms with 96% accuracy. The second most successful model on the test set is the K-Nearest Neighbor and the third approach is the Decision Tree.

2. Introduction

For middle-aged women, breast cancer is the leading cause of cancer-related mortality. The World Health Organisation projects that 1.5 million women will receive a breast cancer diagnosis annually and that 500,000 of those diagnoses will result in breast cancer-related deaths in 2015. So to it is important to find breast cancer early, and there are numerous early detection methods available, including screening, mammography, biopsy, etc.

In many cases such as doctors' negligence or incompetence in addition to a mammography error may also result in a late diagnosis or misdiagnosis, which can be considered a cause of breast cancer death. In the long term, early-stage diagnosis could significantly increase the survival rate of breast cancer, therefore, it is important to improve the accuracy of a breast cancer diagnosis. Machine learning has been applied in medical diagnosis in many papers. To increase the accuracy of breast cancer diagnosis, we aim to use machine learning models and choose the model with higher performance. Breast Cancer Wisconsin is a widely used dataset provided by the UC Irvine machine learning repository. In this project, we will train our models using this dataset.

The input of our algorithm is a set of features calculated from a digitized image of the Fine Needle Aspiration (FNA) of a breast mass from a patient. We will then use three traditional methods including, Nearest Neighbor (k-NN), Support Vector Machines (SVM), and Decision Tree method to predict whether the case is benign or malignant. In addition, the use of machine learning techniques is increasing rapidly, helping medical professionals diagnose diseases. In breast cancer research, machine learning algorithms can be used to detect and predict cancer.

3. Related Work

For the purpose of predicting and diagnosing breast cancer, a multitude of machine learning algorithms are available. Support vector machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5), K-Nearest Neighbours (KNN Network), and others are examples of machine learning algorithms. Many researchers have conducted studies on breast cancer utilizing a variety of datasets, including those from Wisconsin, SEER, mammography pictures, and different hospitals. Authors can finish their research by extracting and choosing different features from these datasets.

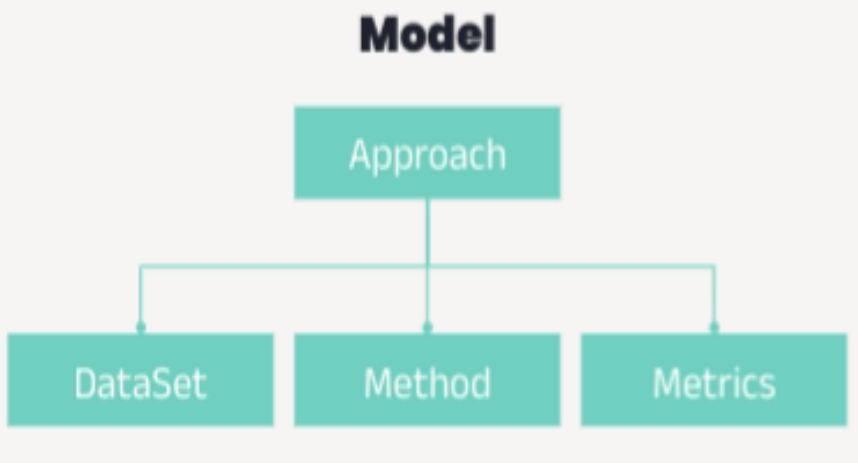
The various authors show how to classify breast cancer using a variety of supervised machine learning algorithms utilizing 3D pictures, and they conclude that SVM performs the best overall. Conversely, we discovered that one of the researchers conducted a comparative analysis of the Relevance vector machine (RVM) and other machine learning techniques for breast cancer detection. RVM offers a low computational cost, and RVM outperforms other machine learning algorithms in terms of breast cancer diagnosis, even when variables are reduced and 97% accuracy is attained.

4. Problem Statement

Developing a predictive model using machine learning techniques to accurately classify breast cancer tumors as benign or malignant based on diagnostic features extracted from breast biopsy samples. The goal is to enhance early detection and diagnosis of breast cancer, thereby improving patient outcomes and reducing mortality rates. The model demonstrates high accuracy, sensitivity, and specificity while being robust and interpretable for clinical decision-making.

5. Our Approach/Model

Our Approach is started with Dataset selection so here as per our selected research paper we used the Breast cancer Wisconsin clinic dataset and we implemented 3 different machine learning algorithms such as Support Vector Machine, K nearest neighbor, and Decision Tree. Lastly, we used a confusion matrix to evaluate the accuracy of each algorithm.



A. Dataset

There are 699 samples in the Breast Cancer Wisconsin dataset, and each sample has 10 features and 1 class information. Below we mentioned all the features with the description.

- Clump thickness: Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers.
- Uniformity of cell size: Cancer cells usually vary in size and these parameters help to determine whether the cells are cancerous or not.
- Uniformity of cell shape: Cancer cells usually vary in size and these parameters help to determine whether the cells are cancerous or not.
- Marginal adhesion: Cancer cells tend to spread away or stay loose. So, loss of adhesion is a sign of malignancy.
- Single epithelial cell size: Epithelial cells that are significantly enlarged is a sign of malignancy.
- Bare nuclei: This is a term used for nuclei that are not surrounded by the cytoplasm as the rest of the cells. Those are typically seen in benign tumors.
- Bland chromatin: It describes the uniform “texture” of the nucleus which is seen in benign cells. In cancer cells, the chromatin is coarser.
- Normal nucleoli: The nucleoli are small in structure and cannot be seen. But in cancer cells, nucleoli are bigger and more prominent.
- Mitoses: it describes is level of division of cancerous cells.
- Class: Benign (non-cancerous) or malignant (cancerous) lump in the breast.

All the features describe the characteristics of cells that are observed in biopsy samples.

Breast tumors can be categorized as malignant (cancerous) or benign (non-cancerous). Generally speaking, benign tumors do not pose a threat to life and do not metastasize to other bodily regions. Depending on their size and symptoms, they might still need medical attention, but they are typically not as dangerous as malignant tumors.

Malignant tumors, on the other hand, are cancerous and have the capacity to metastasize—a term used to describe the spread of a tumor to other parts of the body. Globally, breast cancer is the most prevalent cancer type among women. Effective management of breast cancer and better results depend on early detection and treatment. It's critical to do routine screenings and self-examinations to find any anomalies in the breast tissue.

Features	**Value Range**
Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10

Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10
Class	2: benign 458 (65.5%) 4: malignant 241 (34.5%)

Below, we mentioned few rows of our dataset with all related features and class information.

clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bare_nucleoli	bland_chromatin	normal_nucleoli	mitoses	class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
1	1	1	1	2	10	3	1	1	2
2	1	2	1	2	1	3	1	1	2
2	1	1	1	2	1	1	1	5	2
4	2	1	1	2	1	2	1	1	2

B. Method

The value range of the properties of the dataset is in the range of 1 to 10. The missing data in various properties of 16 samples were filled with an average value of 5 using the missing data filling method, which is a pre-processing method commonly used in data mining.

Input and output values were created by separating the feature and class values in the dataset. 80% of the generated input and output values are divided as training (599 samples) and 20% as test data (140 samples).

Machine learning algorithms like K-NN, Decision Tree, and Support Vector Machine were applied to the prepared data for testing.

C. Evaluation Metrics

The Breast Cancer Wisconsin dataset, which was prepared using data mining methods, was tested with various machine learning algorithms. In this study, accuracy, precision, and recall metrics, are considered as evaluation criteria.

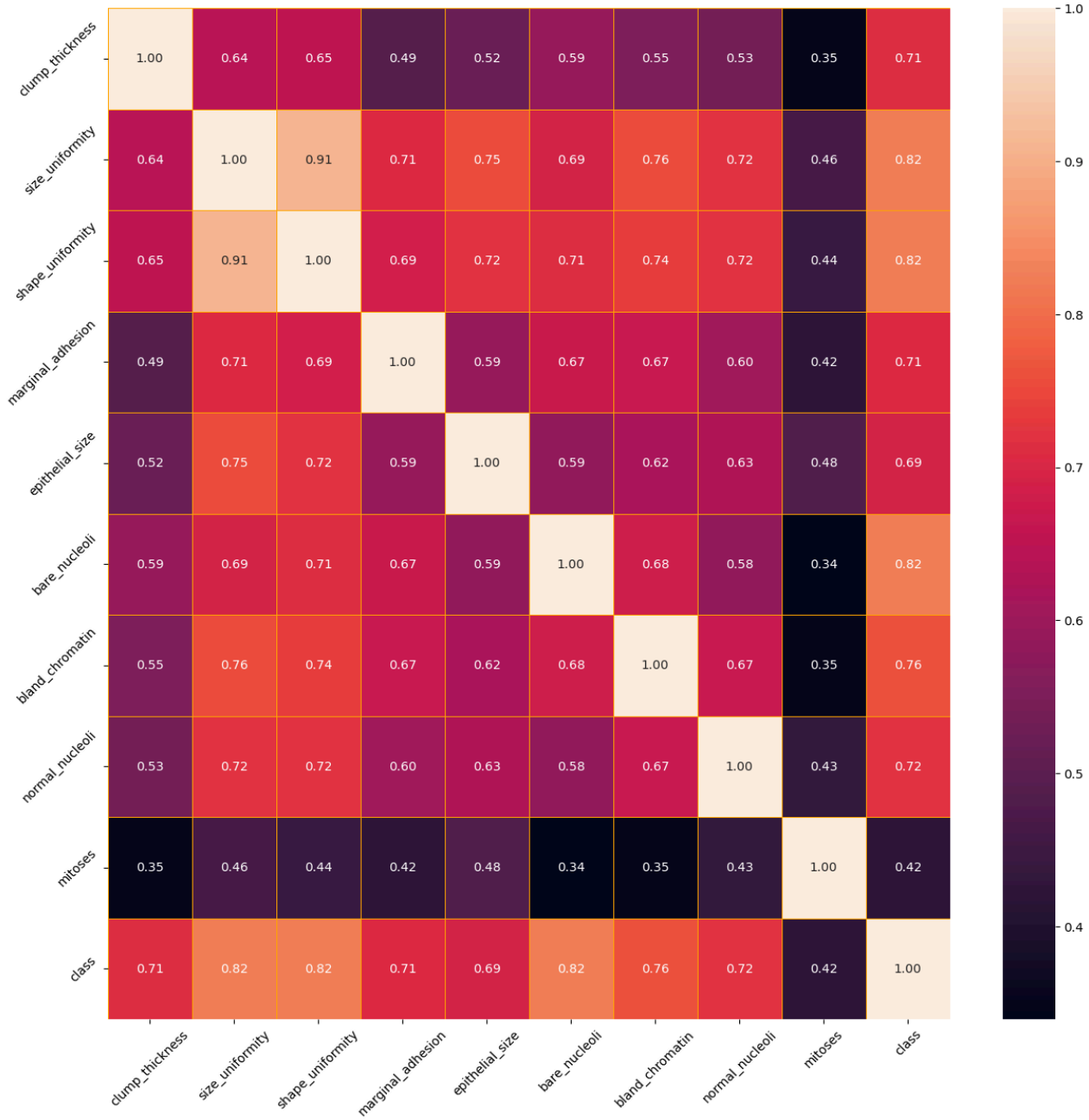
Confusion matrix Actual Values Benign Malign Predicted
Values Benign TP FN Malign FP TN

$$Accuracy = TP+TN / TP+FP+TN+FN$$

$$Precision = TP / TP+FP$$

$$Recall = TP / TP+FN$$

The below figure represents the best-ten correlation results between all variables. The values on the y-axis include diagnosis, radius_mean testure_mean, perimeter_mean, area_mean, smoothness_mean, etc. A correlation coefficient between −1 and 1 indicates no association at all, whereas a positive or negative number indicates a strong relationship between the two variables. It is important to note that the linear connection between variables is all that can be measured using correlation. There is a link of 69% or higher between each of these factors and the outlook for the patient.

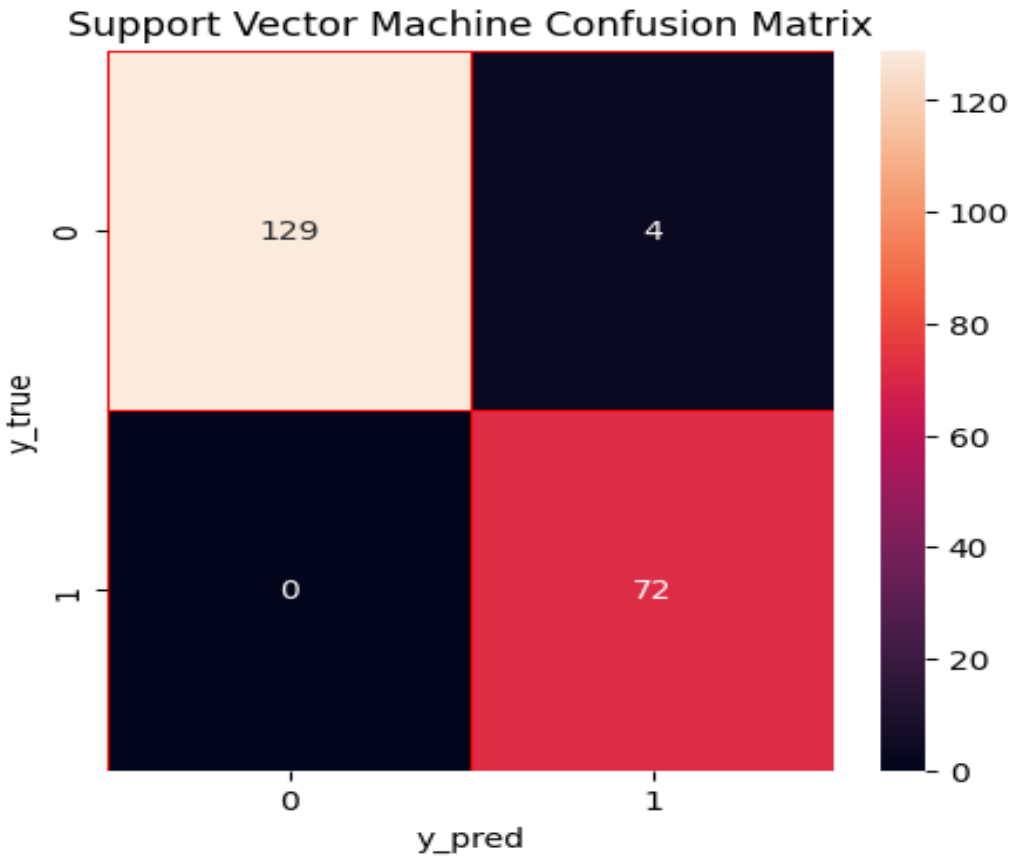


The below figure shows the predicted value of TP, FP, FN, and TN. In this confusion matrix, each entry represents the number of occurrences that share a certain set of labels for both the actual and expected classes. This confusion matrix, in particular, contains:

A large number of true positives and true negatives in the confusion matrix indicates that the model has successfully predicted both groups. The model's success may be seen in its low rate of false positives and false negatives.

Metrics		Description	
		Positive	Negative
Confusion Metrics	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (TP)	True Negative (TN)
		Precision = $\frac{TP}{TP+FP}$	

Confusion matrix for SVM Algorithm



- True Positives (TP): The SVM correctly classified 129 instances as malignant tumors.

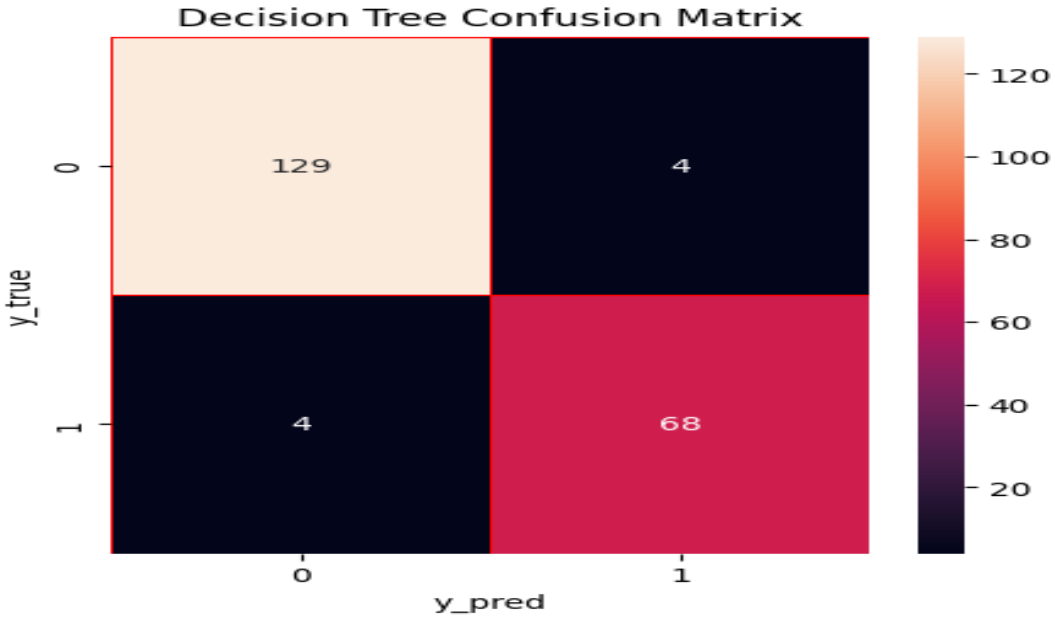
- False Positives (FP): There were 4 instances where benign tumors were incorrectly classified as malignant tumors

- False Negatives (FN): The SVM did not classify malignant tumors as benign tumors.

- True Negatives (TN): The model correctly classified 72 instances as benign tumors.

- SVM Test Accuracy: The SVM model achieved an impressive test accuracy of 98.04%. A high accuracy score suggests that the SVM model performed very well in classifying breast cancer cases.

Confusion Matrix for Decision Tree Algorithm



- True Positives (TP): The Decision Tree correctly classified 129 instances as malignant tumors.

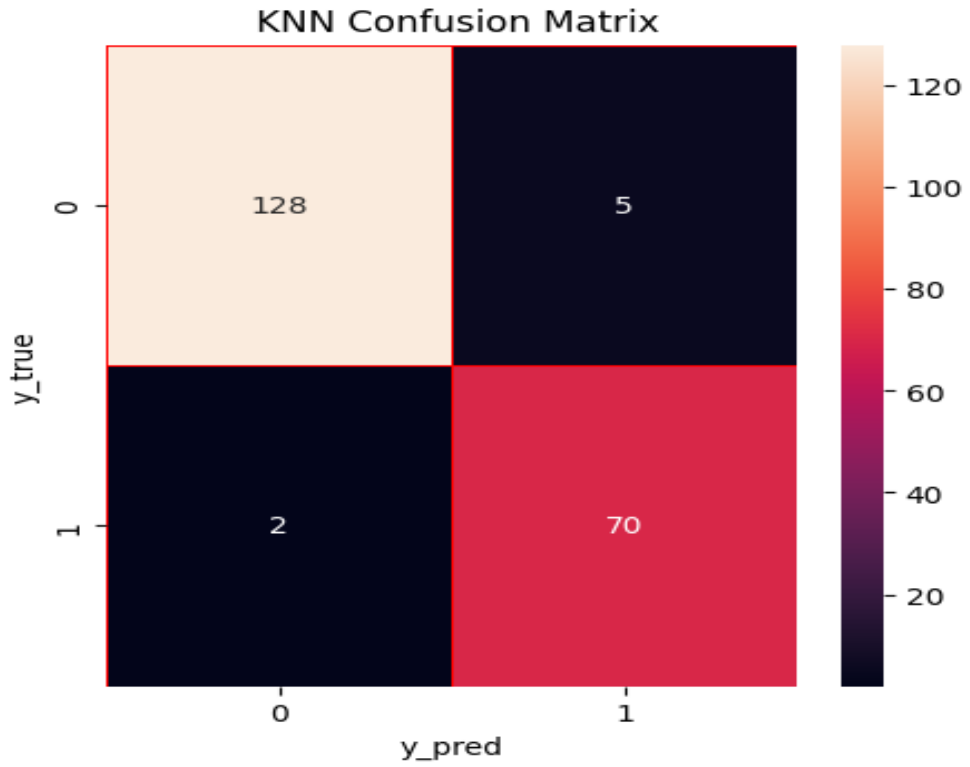
- False Positives (FP): There were 4 instances where benign tumors were incorrectly classified as malignant tumors

- False Negatives (FN): The Decision Tree classifies 4 malignant tumors as benign tumors.

- True Negatives (TN): The model correctly classified 68 instances as benign tumors.

- Accuracy: The Decision Tree model achieved an accuracy of 96.09%. This means that 96.09% of the instances in the test data were correctly classified by the Decision Tree model.

Confusion matrix for K Nearest Neighbour Algorithm



- True Positives (TP): The KNN correctly classified 128 instances as malignant tumors.

- False Positives (FP): There were 5 instances where benign tumors were incorrectly classified as malignant tumors

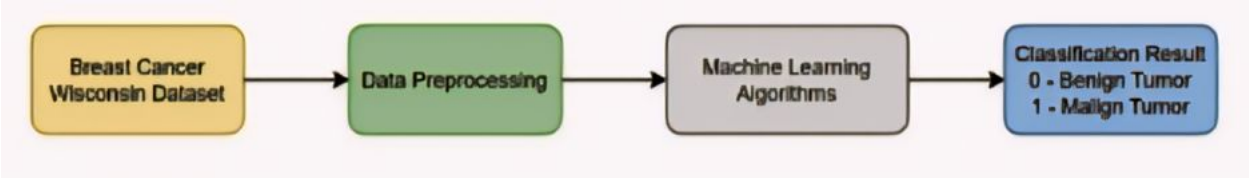
- False Negatives (FN): The KNN classifies 2 malignant tumors as benign tumors.

- True Negatives (TN): The model correctly classified 70 instances as benign tumors.

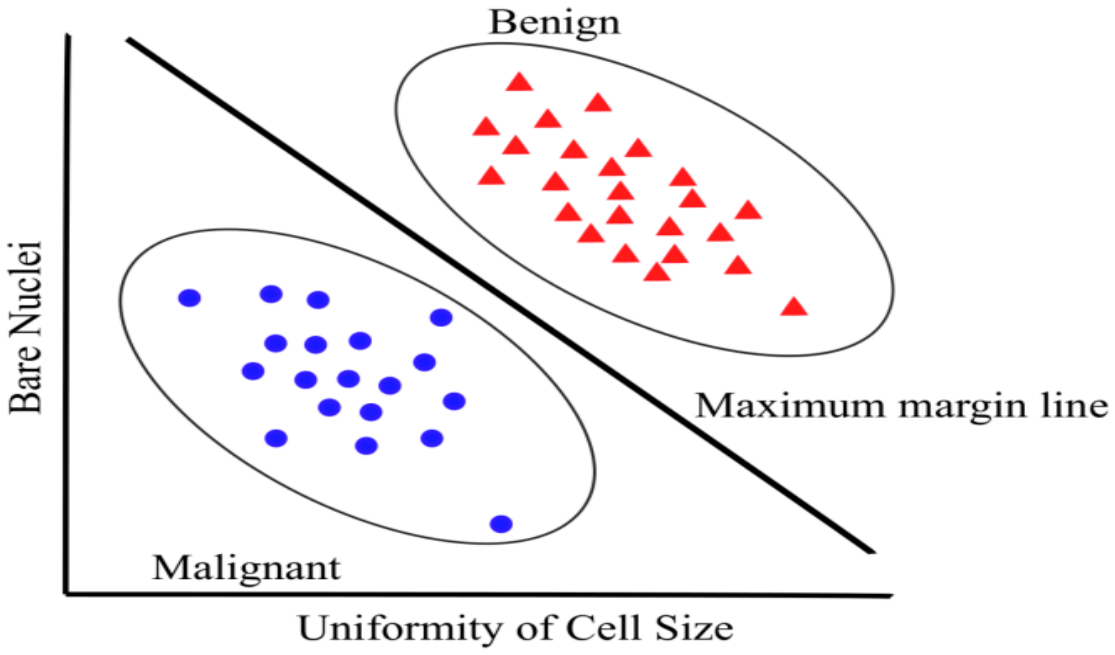
- Accuracy: The KNN model achieved an impressive test accuracy of 96.58%. This metric indicates the proportion of correctly classified instances among all instances when the model was evaluated,

6. Algorithm and Discussion

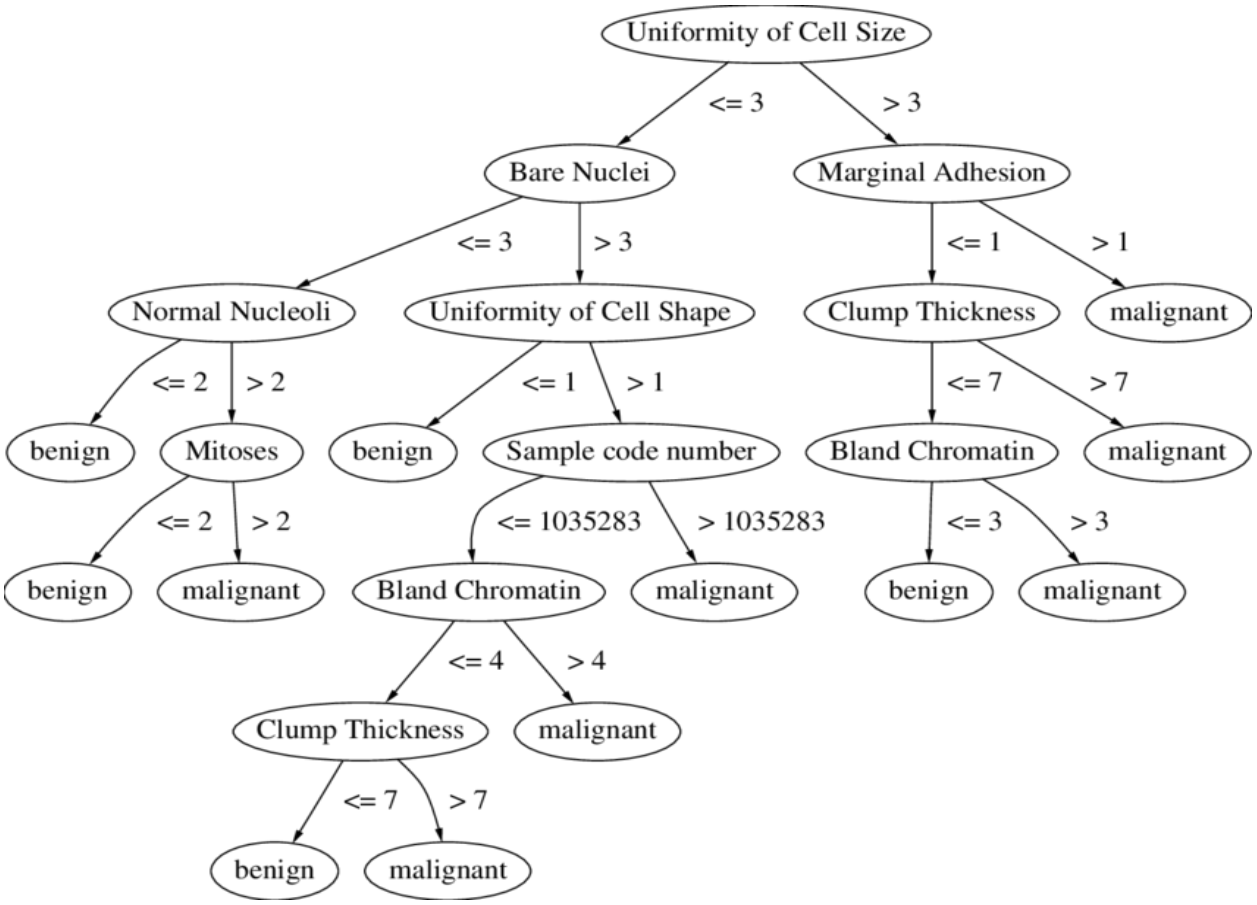
As per our research paper, we implemented our project with the help of machine learning algorithms such as support vector machine, K nearest neighbor, and decision tree. So after applying algorithms to data extracted from the biopsy sample, we identified cancer by 0 as a benign tumor and 1 as a malignant tumor.



The below graph visualization represents cell size uniformity in benign and malignant cells. The x-axis likely represents the uniformity of cell size, while the y-axis represents the number of cells. The graph suggests that benign cells tend to have more uniform cell sizes, while malignant cells have a wider range of cell sizes. This is a common characteristic of cancer cells, which often exhibit uncontrolled growth and division.



Using decision tree we can check for benign and malignant tumor starting from uniformity of cell size if it will be >3 then we need to check for marginal adhesion attribute and if it is less or equal 3 the need to check for bare nuclei. For marginal adhesion if it is greater than 1 then it is sign of malignant and if it is less or equal 1 then check for clum thickness level if it is greate than7 then it is sign of malignant tumor so lie wise based on nine attributes we can determine benign tumor and malignant tumor.



7. Implementation

The implementation performs a machine learning analysis using various classification algorithms on the Breast Cancer Wisconsin dataset.

1. Importing Libraries:

- we Imported necessary libraries like NumPy, pandas, Matplotlib, Seaborn, Plotly, and Scikit-learn for data manipulation, visualization, and machine learning.

2. Mounting Google Drive:

- we mounted Google Drive to access the dataset stored in datset in the Google Drive.

3. Loading and Preprocessing Data:

- Loading the Breast Cancer Wisconsin dataset (CSV file) into a pandas DataFrame.
- Dropping the "id" column as it is not necessary for analysis.
- We handled missing values by removing rows with missing values and converting the

"bare_nucleoli" column to numeric type.

4. Visualization: Heatmap:

- Created a heatmap to visualize the correlation between different features of the dataset.

5. Model Training and Evaluation:

- Splitting the dataset into features (X) and target variable (y).
- Normalizing the features to bring them to a common scale.

- Splitting the data into training and testing sets.
- Training and evaluating three machine learning models: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree.
- Calculating the accuracy of each model on the testing set and appending the scores to a list.

6. Confusion Matrix Visualization:
- Calculating confusion matrices for each model to evaluate performance.
- Visualizing the confusion matrices using heatmaps.

7. Printing Results:
- Printing the accuracy of each model on the testing set.
- Displaying the confusion matrices for each model.

This code demonstrates a standard workflow for performing machine learning classification tasks on a dataset. It includes data loading, preprocessing, model training, evaluation, and result visualization. The choice of algorithms (KNN, SVM, Decision Tree) allows for comparison of different approaches to classification. The visualization of the confusion matrices helps in understanding the performance of each model in terms of true positive, true negative, false positive, and false negative predictions.

- KNN operates on the principle that objects within a dataset are more likely to be similar if they are closer to each other in space.

- Define a distance metric to measure the similarity between data points. Common distance metrics include Euclidean distance, Manhattan distance, and Minkowski distance.

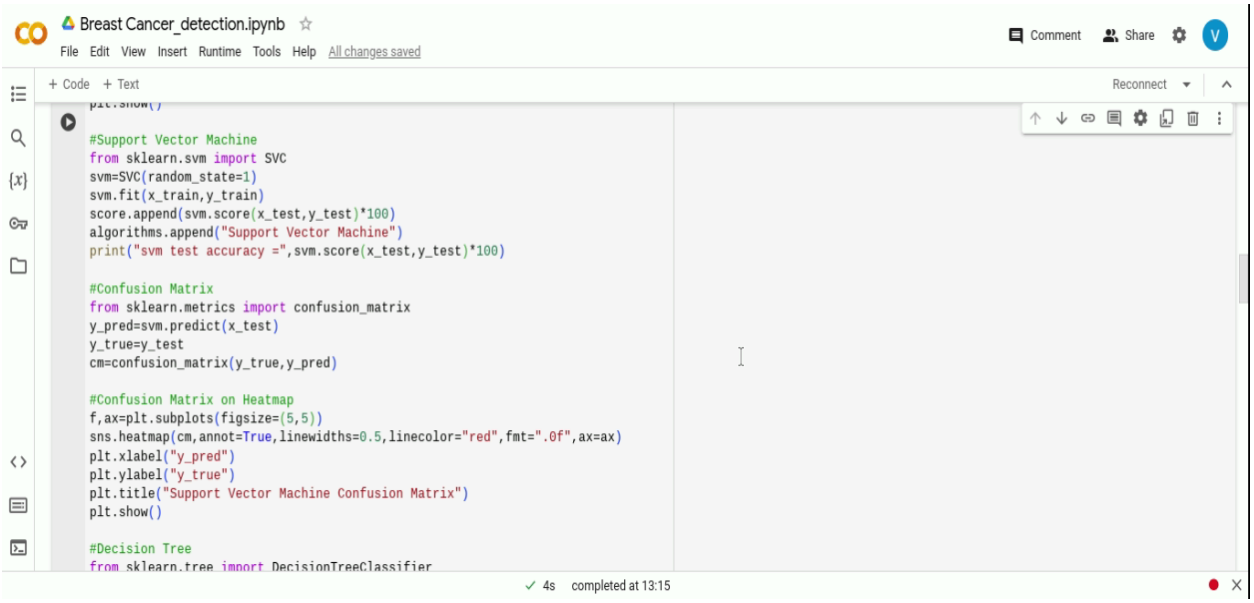
- KNN is non-parametric and instance-based, meaning it doesn't make assumptions about the underlying data distribution and stores all training data for prediction.

- SVM is a powerful supervised learning algorithm used for classification.

- The SVM algorithm aims to find the hyperplane that best separates the data points into different classes (malignant or benign) while maximizing the margin between the classes, which ensures better generalization.

- SVM works by mapping the input data into a higher-dimensional space where it becomes linearly separable.

- SVM can handle high-dimensional data efficiently and is effective in cases where the number of features is greater than the number of samples.



- Decision Tree represents a flowchart-like structure where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents the class label or the predicted value.

- Decision trees are easy to interpret and visualize, making them useful for gaining insights into the data and explaining the model's predictions.

- Decision trees can handle both numerical and categorical data and can be used for both classification and regression tasks.

8. Result and Discussion

After applying machine learning algorithms to the Wisconsin Diagnostic Dataset for Breast Cancer. We used Confusion Matrix, Accuracy, and Precision as performance metrics to assess and contrast the models and determine which algorithm was most effective for the most accurate cancer prediction. The Confusion Matrix is a tool for quantifying the execution of a classification task when the result may belong to two or more different class types. A table with the dimensions "Actual" and "Predicted" along with the columns "True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)" makes up a confusion matrix. The most popular performance indicator for classification algorithms is accuracy. It is defined as the proportion of all predictions made to the number of accurate predictions. The number of accurate documents our ML model returns can be used to determine precision in document retrievals. The quantity of positive results your machine learning model returns can be used to determine sensitivity.

The outcomes of machine learning models that were trained on the Breast Cancer Wisconsin dataset to distinguish between benign and aggressive breast cancer. The algorithm employed indicated that the Decision Tree machine learning algorithm had the lowest classification performance and the Support Vector Machine machine learning algorithm had the highest classification performance for both benign and aggressive breast cancer. A benign tumor is represented by 0 in the confusion matrices below, while a malignant tumor is represented by 1.

9. Conclusion

The literature review has shown that machine learning is widely applied in the medical area, as well as in many other fields, and that it serves as a decision support system for illness identification. It is becoming more and more used, particularly in cancer diagnosis. The most prevalent disease in women is breast cancer, which can be fatal if not caught early. As evidenced by the studies in the literature, it is crucial to precisely and highly performatively detect the diagnosis of breast cancer. The literature was searched for studies using this data set, and a comparison of the various accuracies of machine learning techniques was provided.

The observed reason for this is that variations in the way the data are prepared or pre-processed have an impact on the outcomes. Using the Breast Cancer Wisconsin dataset, data mining, and machine learning algorithms, the patient's breast cancer was diagnosed as benign or malignant in this study. A comparison was made by examining specific metrics in different machine-learning methods. Several classifications on a larger data set are proposed for future research.

10. References

1. Khalid A, Mehmood A, Alabrah A, Alkhamees BF, Amin F, AlSalman H, Choi GS. Breast Cancer Detection and Prevention Using Machine Learning. Diagnostics. 2023; 13(19):3113. https://doi.org/10.3390/diagnostics13193113

2. Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. Prediction of Breast Cancer using Machine Learning Approaches. J Biomed Phys Eng. 2022 Jun 1;12(3):297-308. doi: 10.31661/jbpe.v0i0.2109-1403. PMID: 35698545; PMCID: PMC9175124.

3. Filali, Sanaa & Aarika, Kawtar & Naji, Mohammed & Benlahmar, EL Habib & Ait Abdelouahid, Rachida & Debauche, Olivier. (2021). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science. 191. 487-492. 10.1016/j.procs.2021.07.062.

4. Gopal, V., Al-Turjman, F., Kumar, R., Anand, L., & Rajesh, M. (2021). Feature selection and classification in breast cancer prediction using IoT and machine learning. Measurement (London. Print), 178, 109442. https://doi.org/10.1016/j.measurement.2021.109442

5. Wu J, Hicks C. Breast Cancer Type Classification Using Machine Learning. J Pers Med. 2021 Jan 20;11(2):61. doi: 10.3390/jpm11020061. PMID: 33498339; PMCID: PMC7909418.

6. Rawal, R. (2020, May 1). BREAST CANCER PREDICTION USING MACHINE LEARNING. https://www.jetir.org/view?paper=JETIR2005145

7. Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M., Hasan, M., & Kabir, M. N. (2020). Breast Cancer Prediction: A comparative study using machine learning techniques. SN Computer Science/SN Computer Science, 1(5). https://doi.org/10.1007/s42979-020-00305-w