

**PREDICTION OF CANCER IN BREAST USING MACHINE
LEARNING TECHNIQUES: A COMPARATIVE STUDY OF CLASSIFICATION
ALGORITHMS**

Vasu Bansal

Dept. of CSE, KIET Group of Institutions, Ghaziabad, Delhi-NCR, India
vasubansal3926@gmail.com

Shivangi Gaur

Dept. of CSE, KIET Group of Institutions, Ghaziabad , Delhi-NCR, India
shivangigaur1002@gmail.com

Seema Maitrey*

Dept. of CSE, KIET Group of Institutions, Ghaziabad , Delhi-NCR, India
seema.maitrey@kiet.edu

Deepti Seth

Dept. of Applied Science, KIET Group of Institutions, Ghaziabad , Delhi-NCR, India
deepti.seth@kiet.edu

Akanksha

Dept. of CS, KIET Group of Institutions, Ghaziabad , Delhi-NCR, India
akanksha.cs@kiet.edu

Abstract—Cancer in breast is a frequently occurring type of cancer that impacts individuals globally and can be fatal if not detected and treated in a timely manner. Depending on the tools, datasets, and conditions, researchers have investigated a variety of strategies, including ML techniques , DL methods, and data mining techniques, to predict breast cancer with variable degrees of accuracy. The main aim of this study is to analyze and distinguish the existing machine learning and data mining techniques to determine the most effective way for correctly predict breast cancer utilizing large datasets. The main objective is to give newcomers helpful knowledge on understanding machine learning algorithms and laying the groundwork for deep learning.

This study looked at several studies that show the usefulness of DL and ML algorithms in detecting breast cancer across multiple datasets. SVM, KNN, RF, ANN, and CNN were discovered to have the highest accuracy rates among the algorithms. Future study will focus that allows users to calculate their risk of developing breast cancer.

Keywords-- Breast cancer, prediction, diagnosis, deep learning algorithm, machine learning algorithm.

I. INTRODUCTION

A lot of people die each year from the dangerous and widespread disease known as breast cancer. According to the World Health Organization, 2.9 million women could pass away from breast cancer worldwide. After skin cancer, the second most dangerous cancer in women in the US is breast cancer. In the U. S. A. in 2021, the Society of American Cancer predicts that there will be 48,530 new instances of non-invasive (in situ) breast cancer and approximately 284,200 new cases of invasive breast cancer. However, with an anticipated 685,000 deaths from breast cancer in 2020, it will continue to be a leading cause of death for women worldwide.[1] The World Health Organization (WHO) states that the leading cause of death for women is breast cancer (BC) [2].

A. The indications and manifestations of breast cancer include:-

Breast cancer symptoms can vary among individuals.

While some may not experience any symptoms, others may notice a new lump in the breast or underarm. Additional indications of breast cancer may involve enlargement or hardening of a specific breast area, puckering or inflammation of the breast skin, or the appearance of reddened or scaly skin surrounding the breast or nipple, nipple soreness or pulling inwards, discharge from the nipple that is not breast milk, any changes in the size or shape of the breast, and breast pain in any area.

Machine learning [3] and data mining methods are methods are utilized to predict breast cancer, and Identifying the appropriate algorithm for the job is the obstacle to overcome. Breast cancer emerges from malignant tumors that arise when cellular growth becomes unmanageable, resulting in a typical multiplication of adipose tissue. There are different types of breast cancer, such as DCIS, IDC, MTBC, LBC, MBC, and IBC.

These categories of breast cancer happen when cancerous cells and tissues metastasize to other parts of the body.

DCIS is a form of breast cancer that is sometimes referred to as non-invasive cancer.

***Corresponding author: seema.maitrey@kiet.edu**

It arises when abnormal cells grow beyond the breast[4].

The second type of breast cancer is known as infiltrative ductal carcinoma (IDC) [5] or invasive ductal carcinoma, depending on the classification. IDC is more commonly found in men, and it happens when abnormal breast cells spread throughout the entire breast tissue[6].

Mixed Tumours Breast Cancer (MTBC) [7], also called invasive mammary breast cancer, is the third category of breast cancer. This type of cancer develops when abnormal cells in the ducts and lobules of the breast grow uncontrollably.

Lobular breast Cancer (LBC) is the fourth type of breast cancer and it originates in the lobules of the breast. This cancer increases the risk of developing more invasive types of breast cancer[8].

Colloid breast cancer, also known as mucinousbreast.is the fifth kind of breast cancer. It develops from invasive ductal cells, and it takes place when anomalous tissue surrounds the duct[9].

Inflammatory Breast Cancer (IBC) is the last form which leads to breast inflammation and redness. IBC is a rapidly growing cancer that occurs when cancer cells block the lymph vessels[10].

Finding Data mining involves the retrieval of significant details from an extensive collection of data. Various techniques, such as machine learning, statistics, databases, fuzzy sets, data warehouses, and neural networks, can be applied to analyze and predict the prognosis of various types of cancer, including prostate cancer, lung cancer, and leukemia [11]. Unlike the conventional method of cancer diagnosis, which involves three tests --, radiological imaging, clinical examination and pathology examination [12] -- modern machine learning approaches and algorithms focus on creating models to forecast expected outcomes using unknown data. Machine learning is based on three primary methodologies: preprocessing, classification, feature selection or extraction[13]. Feature extraction is particularly important, as it can help to find between benign and malignant tumors. The primary goal of creating a model is to forecast and obtain accurate results during both the training and testing phases, using data that was previously unknown [14].

II. LITERATURE REVIEW

This section discusses about few of the previous work on breast cancer treatment ways done by researchers using various ML methods. Table 1 below show some of the things previously done on breast cancer diagnosis by researchers using different methods, including ML, deep learning, which are discussed.

Table 1: Literature Review

Reference	Study	Published Year
[3]	Proposed BCAD Framework, Multilayer Perceptron Model, 99.12% Accuracy.	2022
[4]	Analysis of MRI imaging and gene sequencing methods applied to various data sources.	2022
[5]	When compared to deep learning models, the GB-based model offers more accuracy along with faster classification and requires less computer power.	2020
[6]	This study introduced an uncomplicated and efficient technique for categorizing HE stained histological images of breast cancer, even with limited training data. (328 samples).	2022
[7]	The SVM, NN, RF, DT and LR classification models are used. Random forest is most effective method.	2018
[8]	Random forest produced the best accuracy (82.7%), whereas decision trees produced the lowest accuracy (79.8%).	2019
[9]	Within this research, Random Forest classifier showed the best results in terms of precision and execution time.	2020
[10]	The paper comes to the conclusion that deep learning models perform more accurately than machine learning algorithms.	2020

[11]	Researchers can use several data augmentation strategies to address the problem of the small amount of available dataset. Researchers should take into account the issue of the disparity between positive and negative data since it can result in bias towards either a positive or negative prediction. For accurate breast cancer diagnosis and prognosis, an essential problem with an uneven number of breast cancer photos against affected patches needs to be resolved..	2020
[12]	SVM shows higher accuracy(97.13%) compared to C4.5, Naïve Bayes, and k-NN.	2020
[13]	SVM, KNN, and LR provide the most accurate results among the three methods implemented.	2019
[14]	KNN has the highest accuracy among Support Vector Machine, Naive Bayes, and Logistic Regression.	2020
[15]	SVM has the best performance with a 97.07% perfect prediction accuracy when compared to NB, KNN, RF, DT, and LR.	2020
[16]	The highest accuracy is achieved by Adam Gradient Learning, which combines the advantages of AdaGrad and RMSProp.	2020
[20]	KNN classifier performs best when used with the most predictive variables.	2018
[21]	In a study using Wisconsin Breast Cancer (original) datasets, SVM, NB, KNN, and C4.5 algorithms were employed, and SVM achieved the best performance in terms of precision and low error rate in 2016	2016

II. METHODOLOGY

We must first gather the data and execute preprocessing in order to create machine learning methods for breast cancer prediction. The integral components of the pre-processing phase are data cleaning, attribute selection, target role definition, and feature extraction.

Properly processed data is vital for creating a machine learning model that can accurately predict breast cancer. To evaluate the performance of the algorithms, a new set of labelled data is used. A common method is to divide the labelled data into two parts using the Train_test_split approach. The training data consists of three-fourths of the total data, while the remaining one-fourth is reserved for testing the model's accuracy. This is illustrated in the table 2.

Based on the findings, the algorithm that predicts the presence of breast cancer with the highest degree of accuracy is chosen. The algorithms that offer the highest level of accuracy and dependability in predicting the existence of breast cancer are identified after models have been examined. The figure below depicts the overall steps required in early prediction of breast cancer along with the algorithms used:

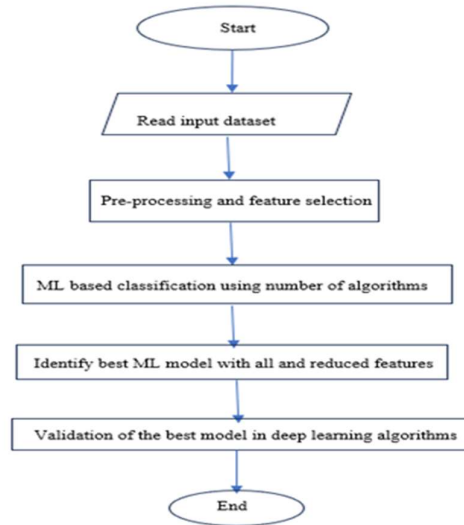


Figure 1: Early prediction of breast cancer using ML algorithms

Dataset in Use

Various datasets are used for breast cancer prediction, such as Coimbra Breast Cancer (CBC), Wisconsin (Prognostic) Breast Cancer (WPBC), Wisconsin (Diagnostic) Breast Cancer, Wisconsin Original Breast Cancer (WOBC), and Breast Tissue Dataset (BTD). The study in question utilized the WBCD, which contains information about both malignant and benign breast cancer. The dataset includes thirty features that are obtained from fine needle aspiration (FNA) of the breast mass. Unlike typical cancer datasets that consist of images, the WBCD consists of feature vectors that describe the cell nuclei in the image. These vectors consist of ten real-valued features that characterize the cell nuclei

Table 2: Features of training data

Features		
Radius_mean	Texture	Radius worst
Perimeter_mean	Smooth_mean	Texture worst
Concavity_mean	Concave_pts.	Perimeter worst
Fractal_dim.	SE_Compact	Area worst

III. MACHINE LEARNING ALGORITHMS

Our research involves carrying out predictive analysis of data using machine learning algorithms. The algorithms that are used in this paper are:

A. Support Vector Machine

It is a classification algorithm that partitions the dataset into different classes using the nearest points of data to identify the maximum margin hyperplane (MMH). SVMs have been widely used in breast cancer classification tasks. Here's an elaboration on how SVMs are applied in breast cancer:

1. **Dataset:** To train an SVM model for breast cancer classification, a dataset containing features extracted from breast cancer samples is required. These features can include various characteristics of the tumors, such as their size, shape, texture, and other clinical attributes.
2. **Data Preprocessing:** Before training an SVM model, it's important to preprocess the dataset. This typically involves steps such as removing missing values, normalizing or standardizing the features, and balancing the class distribution if there is a class imbalance issue.
3. **Feature Selection:** Feature selection techniques may be employed to identify the most informative and relevant features for the classification task. This helps to reduce the dimensionality of the dataset and improve the efficiency and performance of the SVM model.
4. **Training:** Once the dataset is preprocessed, it is split into a training set and a testing/validation set. The training set is used to train the SVM model. During training, the SVM algorithm tries to find the optimal hyperplane that maximizes the margin between different classes while minimizing the classification errors.
5. **Kernel Selection:** In breast cancer classification, SVMs commonly use different kernel functions to handle non-linearly separable data. The choice of the kernel function depends on the characteristics of the dataset. Some popular kernel functions used in breast cancer classification include the linear kernel, polynomial kernel, Gaussian (RBF) kernel, and sigmoid kernel.
6. **Model Evaluation:** After training, the SVM model is evaluated using the testing/validation set. The accuracy, precision, recall, F1-score, and other metrics are commonly used to assess the performance of the model. Cross-validation techniques may also be employed to obtain a more reliable estimate of the model's performance.
7. **Hyperparameter Tuning:** SVMs have hyperparameters that need to be tuned to optimize their performance. The C parameter controls the trade-off between maximizing the margin and minimizing the classification errors. The kernel-specific parameters, such as the degree of the polynomial kernel or the width of the Gaussian kernel, also need to be tuned. Grid search, random search, or other optimization techniques can be used to find the optimal hyperparameters.
8. **Prediction:** Once the SVM model is trained and evaluated, it can be used to make predictions on new, unseen breast cancer samples. The model assigns a class label to each sample based on its learned decision boundary.
9. **Interpretability:** SVMs offer good interpretability as they provide support vectors that influence the decision boundary. These support vectors represent the most critical samples in determining the classification decision. Examining the support vectors can provide insights into the characteristics of the breast cancer samples.

SVMs can assist in diagnosing breast cancer, predicting its severity, and aiding in treatment planning by providing valuable insights into the characteristics of the tumors.

B. Random forest

Random forests or decision forests, on the other hand, are an ensemble method used for various tasks such as classification and regression. In this method, throughout the training phase, multiple decision trees are developed and outcome is mean class. Random forests are used to overcome the tendency of decision trees to overfit their training data.

It is another popular machine learning algorithm that has been widely applied in breast cancer classification tasks. Here's an elaboration on how Random Forest is used in breast cancer:

1. Dataset: Similar to SVMs, training a Random Forest model for breast cancer classification requires a dataset containing features extracted from breast cancer samples. These features can include tumor characteristics, patient demographics, genetic markers, and other clinical attributes.

2. Data Preprocessing: As with any machine learning algorithm, the dataset needs to be preprocessed before training the Random Forest model. This involves handling missing values, normalizing or standardizing the features, and addressing class imbalances if present.

3. Ensemble of Decision Trees: Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree is trained on a random subset of the original dataset, and the predictions of all the trees are aggregated to obtain the final prediction. This ensemble approach helps to improve the model's performance and robustness.

4. Feature Selection: Feature selection techniques can be applied to identify the most important features for breast cancer classification. Random Forests have an inherent feature importance measure, which is calculated based on how much each feature contributes to reducing the impurity or error in the decision trees. This feature importance information can aid in selecting relevant features and understanding the importance of different factors in breast cancer classification.

5. Training: The Random Forest model is trained by growing multiple decision trees on different subsets of the data. Each decision tree is trained using a process called "bagging" or "bootstrap aggregating," where random samples from the original dataset are drawn with replacement. The decision trees are trained to maximize the information gain or decrease the impurity at each split, typically using criteria like Gini impurity or entropy.

6. Hyperparameter Tuning: Random Forest models have hyperparameters that need to be tuned for optimal performance. These include the number of trees in the forest, the maximum depth of the trees, the number of features considered at each split, and the minimum number of samples required for a node to be split. Techniques such as grid search or random search can be used to find the best combination of hyperparameters.

7. Model Evaluation: After training, the Random Forest model is evaluated on a testing/validation set. Performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) can be used to assess the model's classification performance. Cross-validation techniques can also be applied to obtain a more reliable estimate of the model's performance.

8. Prediction: Once the Random Forest model is trained and evaluated, it can be used to make predictions on new, unseen breast cancer samples. The model combines the predictions from all the decision

trees in the forest to obtain the final prediction for each sample.

9. Interpretability: Random Forest models provide feature importance measures that can help interpret the model's predictions. These measures indicate the relative importance of each

feature in the classification task. By examining the feature importance, researchers and clinicians can gain insights into the factors that contribute most to breast cancer classification. Random Forests have been widely used in breast cancer research and clinical practice due to their ability to handle complex data, handle high-dimensional feature spaces, and provide interpretable feature importance measures. They have demonstrated success in breast cancer diagnosis, prognosis, and treatment planning by leveraging the strengths of ensemble learning and decision trees.

C. K-Nearest Neighbor (K-NN)

To acquire the ability to label new data points, the algorithm requires a vast quantity of labelled data points. When presented with a new data point, the algorithm evaluates its nearest neighbors, or those in close proximity, and seeks their input in predicting its label.

It is another popular machine learning algorithm that can also be applied to breast cancer classification tasks. Here's an elaboration on how K-NN is used in breast cancer:

1. **Dataset:** Like other machine learning algorithms, K-NN requires a dataset containing features extracted from breast cancer samples. These features can include tumor characteristics, patient demographics, genetic markers, and other clinical attributes.

2. **Data Preprocessing:** Prior to training the K-NN model, the dataset needs to be preprocessed. This typically involves handling missing values, normalizing or standardizing the features, and addressing class imbalances if they exist.

3. **Distance Metric:** K-NN is a distance-based algorithm, which means it relies on measuring the distance between data points to make predictions. A distance metric, such as Euclidean distance or Manhattan distance, is used to determine the similarity between samples in the feature space.

4. **Training:** The training phase in K-NN is quite simple. It involves storing the entire training dataset in memory, as the K-NN algorithm doesn't explicitly build a model. During training, K-NN organizes the data points in a space, forming a sort of reference database.

5. **K-Value Selection:** The K in K-NN represents the number of nearest neighbors that are considered when making predictions for a new, unseen data point. The choice of the K value is crucial and can impact the performance of the algorithm. A small K value may lead to overfitting, while a large K value may result in underfitting. It is common to perform hyperparameter tuning to find the optimal K value through techniques like cross-validation.

6. **Prediction:** When a new sample is to be classified, the K-NN algorithm searches for the K nearest neighbors in the feature space. This is done by calculating the distance between the new sample and all the training samples. The class label of the new sample is determined by majority voting among the K nearest neighbors. That is, the most common class label among the K neighbors is assigned to the new sample.

7. Model Evaluation: After making predictions for the test dataset, the performance of the K-NN model is evaluated using various metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Cross-validation techniques can also be applied to obtain a more reliable estimate of the model's performance.

8. Feature Selection: Feature selection techniques can be employed to identify the most relevant features for breast cancer classification using K-NN. Selecting informative features helps improve the efficiency and accuracy of the algorithm.

9. Interpretability: K-NN provides interpretability through its nearest neighbors. The K nearest neighbors for a given sample can be examined to understand the characteristics of similar cases in the dataset. This can provide insights into the factors that contribute to the classification decision.

K-NN is known for its simplicity and intuitive approach. It can handle both binary and multi-class classification problems, making it suitable for breast cancer classification tasks. However, K-NN has some limitations, such as its sensitivity to the choice of distance metric, the need to store the entire training dataset in memory, and its performance degradation with high-dimensional data. Nevertheless, K-NN remains a valuable tool in breast cancer research and can provide accurate predictions when applied appropriately.

D. Logistic regression

It is a widely used and effective modelling technique that extends linear regression. It is commonly used to evaluate the probability of a health condition or disease based on a risk factor. Logistic regression is employed to analyze the connection between independent variables, such as predictor or exposure variables (X_i), and a binary dependent variable, which is also known as an outcome or response variable Y). Logistic regression's primary application is forecasting binary or multiclass dependent variables, and it is employed in both single and multiple logistic regression scenarios.

Logistic regression is a statistical learning algorithm commonly used in breast cancer classification tasks. Here's an elaboration on how logistic regression is applied in breast cancer:

1. Dataset: To train a logistic regression model for breast cancer classification, a dataset containing features extracted from breast cancer samples is required. These features can include tumor characteristics, patient demographics, genetic markers, and other clinical attributes.

2. Data Preprocessing: As with any machine learning algorithm, the dataset needs to be preprocessed before training the logistic regression model. This involves handling missing values, normalizing or standardizing the features, and addressing class imbalances if present.

3. Binary Classification: Logistic regression is typically used for binary classification tasks, where the goal is to predict whether a given sample belongs to a particular class (e.g., malignant or benign). In breast cancer classification, the classes are often labelled as 0 (benign) and 1 (malignant).

4. **Logistic Function:** Logistic regression uses a logistic (or sigmoid) function to model the relationship between the input features and the probability of belonging to a specific class. The logistic function maps the linear combination of the input features to a value between 0 and 1, representing the probability of the positive class.

5. **Model Training:** During the training phase, the logistic regression algorithm estimates the model parameters (coefficients) that best fit the data. This is typically done using optimization techniques such as maximum likelihood estimation or gradient descent. The objective is to minimize the difference between the predicted probabilities and the actual class labels in the training data.

6. **Decision Boundary:** In logistic regression, a decision boundary is established to separate the two classes. The decision boundary represents the threshold probability above which a sample is classified as belonging to the positive class and below which it is classified as belonging to the negative class.

7. **Regularization:** To prevent overfitting, logistic regression often incorporates regularization techniques such as L1 (Lasso) or L2 (Ridge) regularization. Regularization adds a penalty term to the loss function, encouraging smaller coefficient values and reducing model complexity.

8. **Model Evaluation:** After training, the logistic regression model is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Cross-validation techniques can be used to obtain a more reliable estimate of the model's performance.

9. **Feature Selection:** Feature selection techniques can be applied to identify the most relevant features for breast cancer classification using logistic regression. This helps to improve model interpretability and reduce the impact of irrelevant or redundant features.

10. **Interpretability:** Logistic regression offers interpretability by providing the estimated coefficients for each feature. The coefficients indicate the strength and direction of the relationship between the features and the probability of belonging to the positive class. Positive coefficients imply a positive association, while negative coefficients imply a negative association with the positive class.

Logistic regression is widely used in breast cancer research and clinical practice due to its simplicity, interpretability, and ability to provide probabilistic predictions. However, logistic regression assumes a linear relationship between the features and the log-odds of the outcome, which may limit its ability to capture complex non-linear relationships. In such cases, other machine learning algorithms like SVMs or random forests may be more suitable.

E. Decision Tree

It is a versatile predictive modelling technique that can be utilized in a variety of domains. By using an algorithmic approach, it can partition the dataset into multiple segments based on diverse standards.

It is a popular machine learning algorithm used in breast cancer classification tasks. Here's an elaboration on how decision trees are applied in breast cancer:

1. **Dataset:** To train a decision tree model for breast cancer classification, a dataset containing features extracted from breast cancer samples is required. These features can include tumor characteristics, patient demographics, genetic markers, and other clinical attributes.
2. **Data Preprocessing:** As with any machine learning algorithm, the dataset needs to be preprocessed before training the decision tree model. This involves handling missing values, normalizing or standardizing the features, and addressing class imbalances if present.
3. **Tree Structure:** A decision tree is a hierarchical structure consisting of nodes and branches. The nodes represent feature tests or decisions, while the branches represent the possible outcomes of those tests. The tree is constructed by recursively partitioning the data based on the selected features, with each split aiming to maximize the separation of the classes.
4. **Splitting Criteria:** The decision tree algorithm uses various splitting criteria, such as Gini impurity or information gain, to determine the optimal feature and threshold for each split. Gini impurity measures the probability of misclassifying a randomly chosen sample, while information gain measures the reduction in entropy (uncertainty) after a split.
5. **Training:** During the training phase, the decision tree algorithm recursively partitions the dataset based on the selected splitting criteria until a stopping criterion is met. This stopping criterion could be reaching a maximum tree depth, achieving a minimum number of samples per leaf node, or other criteria.
6. **Pruning:** Decision trees are prone to overfitting, where they become too complex and memorize the training data. To mitigate overfitting, pruning techniques can be applied. Pruning involves removing or merging branches/nodes that provide little improvement in classification performance on the validation or testing set.
7. **Model Evaluation:** After training, the decision tree model is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Cross-validation techniques can be used to obtain a more reliable estimate of the model's performance.
8. **Interpretability:** Decision trees offer excellent interpretability as the learned tree structure can be easily visualized. Each decision node represents a feature test, and the path from the root to a leaf node represents the decision rules leading to a classification. The feature importance can be derived from the frequency and position of the features in the tree.

9. Prediction: Once the decision tree model is trained and evaluated, it can be used to make predictions on new, unseen breast cancer samples. The model follows the decision rules learned during training to classify the samples into the corresponding classes based on their feature values.

Decision trees have been widely used in breast cancer classification due to their simplicity, interpretability, and ability to handle both binary and multi-class classification problems. However, decision trees can be prone to high variance and instability, leading to overfitting. To mitigate these issues, ensemble methods like random forests can be utilized, where multiple decision trees are combined to improve performance and reduce overfitting.

The overall process flow is depicted in the figure 3 below:

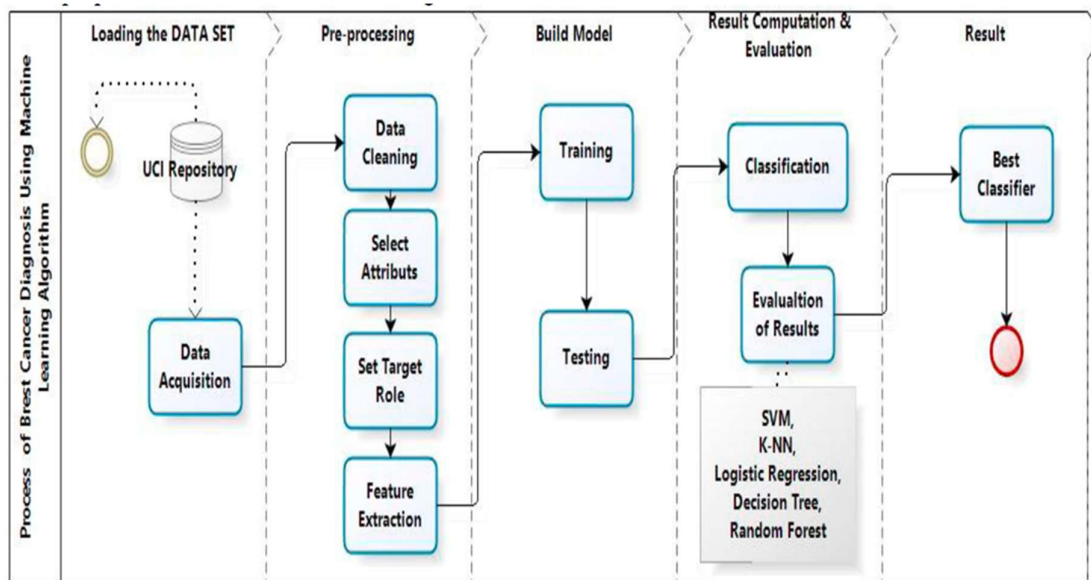


Figure 3: Step-by-step process flow[24]

IV. CONCLUSION AND FUTURE SCOPE

On the Wisconsin dataset, several machine learning techniques were applied to predict breast cancer, including K-NN, SVM, Decision Tree, Nave Bayes Logistic Regression, and Random Forest. The precision attained by ANN, CNN, SVM, Random Forest, and KNN algorithms. These algorithms reached more than 95%, and deep learning algorithms like Convolutional Neural Network (CNN) and Artificial Neural Network (ANN) were used to improve prediction accuracy.

In this study, we looked at several studies that discovered that deep learning algorithms, machine learning approaches, and some proposed algorithms provide improved accuracy for detecting breast cancer on various datasets. However, machine learning algorithms like SVM, KNN, RF, ANN, and CNN provide the highest level of accuracy. In the future, models based on these methodologies will be created in an effort to provide software for the breast cancer prediction so that the user can access themselves.

REFERENCES

- [1] 'WHO |Breast cancer',WHO, <http://www.who.int/cancer/prevention/diagnosis-screening/breastcancer/en/>
- [2] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set."
- [3] S. Aamir, A. Rahim, Z. Aamir, S. F. Abbasi, M. S. Khan, M. Alhaisoni, M. A. Khan, K. Khan, and J. Ahmad "Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques," August 16,2022.
- [4] A.B. Nassif, M.A. Talib, Q. Nasir, Y.Afadar,and O.Elgedy"Breast cancer detection using artificial intelligence techniques: A systematic literature review," March 5,2022.
- [5] H.E. Agouri, M. Azizi, H. E. Attar, M. E. Khannoussi, A. Ibrahimi, R. Kabbaj, H. Kadiri,
- [6] S. BekarSabein, S. EchCharif, C. Mounjid, B. E. Khannoussi, "Assessment of deep learning algorithms to predict histopathological diagnosis of breast cancer: first Moroccan prospective study on a private dataset," Feb 19,2022.
- [7] M.Mangukiya, A.Vaghani, M.Savani, "Breast Cancer Detection with Machine Learning," Feb 2022.
- [8] A. Rasool, C.Bunterngchit, L.Tiejian, Md. R. Islam, Q. Qu, and Q. Jiang, "Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis" March 9,2022.
- [9] H. Zhang, H. Liu, L. Ma, J. Liu, D. Hu "Ultrasound Image Features under Deep Learning in Breast Conservation Surgery for Breast Cancer" Sept 17,2021
- [10] S. Bhise, S. Bepari, S. Gadekar, D. Kale, A. S. Gaur, Dr. S. Aswale, "Breast Cancer Detection using Machine Learning Techniques," July 7,2021.
- [11] S. Aryal, B. Paudel, "Supervised Classification using Gradient Boosting Machine: Wisconsin Breast Cancer Dataset," June 2020.
- [12] M. Srivenkatesh, "Prediction of Breast Cancer Disease using Machine Learning Algorithms," Feb 4,2020.
- [13] N. Fatima, L. Liu, S. Hong, H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," Aug 14,2020.
- [14] R. Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING," May 2020.
- [15]Gaurav Singh, "Breast Cancer Prediction Using Machine Learning," July 30,2020.
- [16] M.Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M.Sazzadur Rahman, Imran Mahmud, R. Akter, "An Analysis on Breast Disease Prediction Using Machine Learning Approaches," Feb 2020.
- [17] P. Gupta, S. Garg, "Breast Cancer Prediction using varying Parameters of Machine Learning Models," June 4,2022.
- [18] M. D. Ganggayah, N.A Talib, Y. C. Har, P. Lio, S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," March 22, 2019.
- [19]Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques," April 2019.
- [20]Yixuan Li, Zixuan Chen, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction," Oct 18,2018.
- [21]Madhu Kumari, Vijendra Singh, "Breast Cancer Prediction System," June 8,2018.

- [22] H. Asri, H. Mossanif, H. Al Moatassime, T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," May 12,2016.
- [23] K. Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Nov 15,2014.
- [24] A. Naji , S. El Filali, K.Aarika, , EL Habib Benlahmar, R. Ait Abdelouhahid, O. Debauche, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis", August 9-12, 2021