Breast Cancer Prediction Using Machine Learning Algorithms

Gayathri Keshamoni
dept. of Computer Science
University of Central Missouri
Missouri, USA
gxk24880@ucmo.edu

Laxman Reddy Nalla
dept. of Computer Science
University of Central Missouri
Missouri, USA
lxn20710@ucmo.edu

Uday Kumar Kuppam
dept. of Computer Science
University of Central Missouri
Missouri, USA
uxk97680@ucmo.edu

Vinay Kumar Avirneni
dept. of Computer Science
University of Central Missouri
Missouri, USA
vxa45490@ucmo.edu

Abstract— **Breast cancer is one of the major causes of death in women worldwide. Radiologists and oncologists find it challenging to make an accurate diagnosis of breast cancer due to the complicated nature of the disease's cells (microcalcification and tumors). Breast cancer, along with lung and bronchus cancer, prostate cancer, colon cancer, and pancreatic cancer among others, is the type of cancer that is most known throughout the world. It happens when the size of cells in breast tissue becomes out of control. The organs and tissues in the body are constructed from cells. When the development of new cells is not regulated, a tumor-like mass of tissue will develop. The tumors are classified as benign and malignant tumors. Having said that, early diagnosis of these cancer cells is essential for quick reaction and better chances of cure. To help radiologists identify these cancer cells, a variety of computer-aided diagnosis (CAD) systems have been created and are currently in use. But, unfortunately, early identification of cancer is often difficult because the symptoms of the disease are lacking at the beginning. To make a better understanding of this problem tools are required for healing. Machine learning is a branch of artificial intelligence that enables a machine to develop over time. In this context, machine learning has lately shown promise for the rapid and accurate diagnosis of breast cancer. This paper presents a new machine learning-based framework tool that utilizes the PCA, KNN, NCA, Logistic Regression, Decision Tree, XG Boost, and Support Vector Machine algorithm approaches that are efficiently used for the prediction of breast cancer. For this purpose, cancer patient data were collected from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Pre-processing of data was performed followed by feature extraction of the data set using Principal Component Analysis (PCA). Various Machine Learning techniques were proposed in the paper which helped us achieve an accuracy of 99 % using K-Nearest Neighbors, 93% with PCA, and 88 % using SVM.**

Keywords—Breast Cancer, Dataset, PCA, K-Nearest-Neighbor (KNN), Logistic regression (LR), Decision Tree (DT), SVM, XG Boost, Prediction, NCA.

I. INTRODUCTION

Breast cancer is the second most dangerous cancer after lung cancer, and it is the leading cause of cancer deaths among women [1]. According to published statistics, Breast Cancer has become a major health problem in both developed and developing countries over the past 50 years. Its incidence has increased recently with an estimated 1,152,161 new cases in which 411,093 women die each year [2]. Currently, one in twelve females in Britain between the ages of 1 and 85 years gets breast cancer. With one million new cases of cancers reported in the World, breast cancer is common in females and comprises 18% of all women's cancer. By 2021, it is expected that 85 women per 100,000 would develop breast cancer [12]. Breast cancer accounted for 1.67 million new instances of cancer in 2012, or 25% of all malignancies in women. According to Ferlay et al. [13], there are 883,000 cases in less developed nations and 794,000 in the most developed nations. Data show that between 100,000 people, 145.2 women in Belgium and 66.3 in Poland have breast cancer [14]. One in eight women in the United States and one in 35 women in Asia are diagnosed with breast cancer, respectively. Ten cases per 100,000 people and 7000 new cases are recorded annually in Iran [15]. In Pakistan, the likelihood of developing breast cancer is rising [16]. In South Asian developing nations, breast cancer is primarily seen in densely populated locations [19, 20]. Since its source is still unclear, there are currently no reliable methods of prevention. When cells start to multiply uncontrollably, cancer develops. Breast cancer cells typically develop a tumor, which is frequently detectable on an x-ray or as a lump. The two types of tumors are benign and malignant. Tumors that are malignant are more dangerous than benign ones. Malignant and benign tumors cannot always be distinguished by doctors, and it can take up to two days to classify the tumor cells.

GIT HUB LINK: https://github.com/KUPPAM-700739768/GROUP_PROJECT

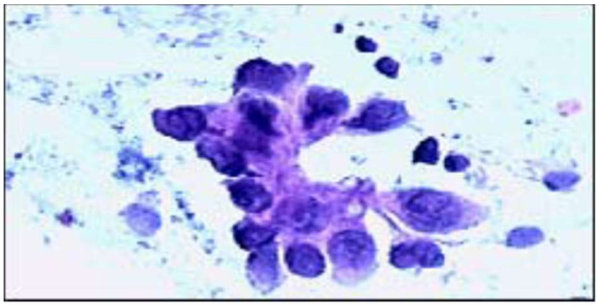


Figure 1: Fine Needle Biopsy of Breast Malignant tumors.

Machine learning techniques are utilized to determine the type of malignant cells efficiently and precisely. Artificial intelligence (AI) systems can automatically learn from their experiences and get better over time thanks to a technique called machine learning [4]. When cells start to multiply uncontrollably, cancer develops. Although it mostly affects women, breast cancer can also affect men. It's crucial to realize that many breast lumps are benign and not cancerous (malignant).

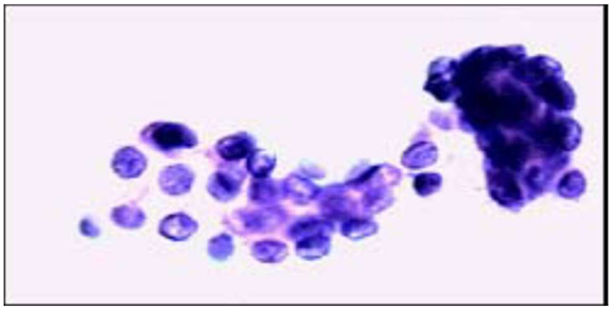


Figure 2: Fine Needle Biopsies Benign breast tumors

Others are less frequent, such as phyllodes tumors and angiosarcoma. Fatigue, headaches, pain, and numbness (peripheral neuropathy), bone loss, and osteoporosis are some of the adverse effects of breast cancer. There are numerous categorization and outcome prediction algorithms for breast cancer. Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), and K-Nearest Neighbors (K-NN) with PCA and NCA are some of the different techniques employed.

A. *K – Nearest Neighbor (KNN)*

KNN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variables for those K instances. For Regression, this might be the mean output variables for those K instances. For regression this might be the mean output variable, in classification, this might be the mode (or most common) class value. To determine which of the K instances in the training dataset are most like a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance. Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j. Euclidean Distance (x, xi) = sqrt (sum ((xj – xij)^2)). The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

B. Decision Trees (DT)

A predictive model called a Decision Tree (DT) describes how the values of an outcome variable can be anticipated based on the values of other variables. A DT output is a decision tree with each end node containing a prediction for the result variable and each fork split into a predictor variable. A binary tree serves as the DT model's representation. A single input variable (x) and a split point on that variable are represented by each root node (assuming the variable is numeric). A prediction is made using an output variable (y) that is present in the tree's leaf nodes.

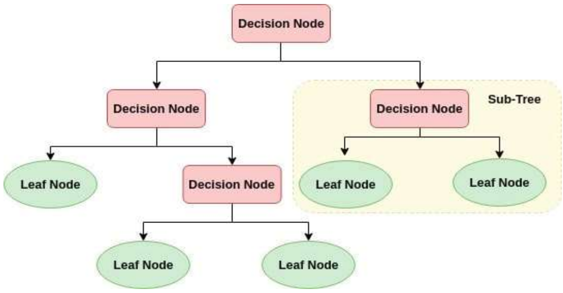


Figure 3: Decision Tree

C. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges.

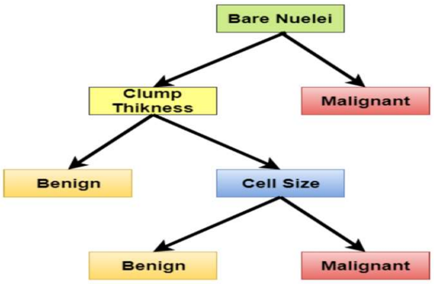
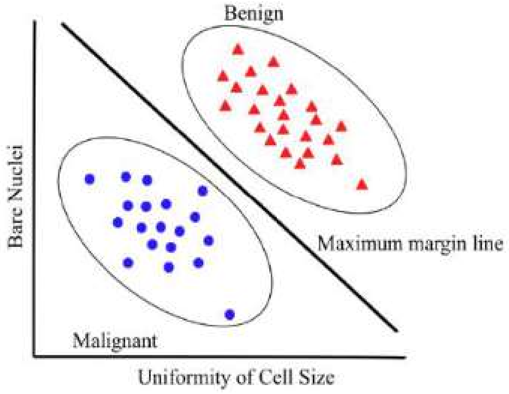


Fig 4: DT structure used in breast cancer classification

Determining a certain indentation point as ‘Normal’ or ‘Cancerous’ based on the mechanical characterization of breast tissue can be defined as a binary classification task. A paradigm of the SVM framework of breast cancer is shown in the figure.



Support vector machine (SVM), known as one of the useful classifiers in the machine learning area, is used for the classification task. For a given training data set of n points, (x1, y1),…, (xn, yn) where xi is a real vector and yi (i = 1,.., n) is either - 1 or +1, which indicates the class of the tissue as ‘Normal’ and ‘Cancerous’ respectively, an SVM algorithm separates the given training data into two classes by finding an optimal hyperplane which maximizes the hard-margin, 2 ||w||, where w is the normal vector to the hyperplane and b is the bias, as shown in Fig. 5. To extend SVM for cases that the given data are not linearly separable, soft-margin SVM is introduced with a slack variable, ξi, which accounts for an upper bound on the training classification errors. To get an optimal hyperplane from linearly non-separable data, the total penalties for misclassified data should be minimized with respect to the optimal hyperplane. Hence, the soft-margin SVM algorithm can be expressed as [28]:

$$\min_{\mathbf{w},b,C}\left[\frac{1}{2}||\mathbf{w}||^2+C\sum_{i=1}^{n}\xi_i\right]$$
$$\forall(\mathbf{x}_i,y_i)\quad s.t.\quad y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1-\xi_i,\quad \xi_i\geq 0$$

Fig: 5

where c is the regularization parameter that decides how much of an increase in geometric margin is worth sacrificing to reduce the number of incorrect classifications.

D. Logistic Regression (LR)

A statistical analysis method called logistic regression uses previous observations from a data set to predict a binary outcome, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable. For instance, logistic regression could be used to forecast whether a candidate for office will win or lose, or if a high school student will be accepted into a particular institution or not. These simple choices between two options allow for binary outcomes. Multiple criteria for input can be considered by a Logistic regression model.

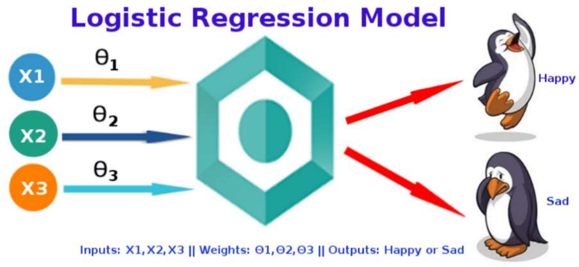

Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1,X2,X3 || Weights: Θ1,Θ2,Θ3 || Outputs: Happy or Sad

Fig 5: Logistic Regression Model

II. Motivation

The most common disease affecting women worldwide is breast cancer. 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the U.S. during 2016 and 40,450 women’s death is estimated. The development of Breast Cancer and its prediction fascinated me. The UCI Wisconsin Machine Learning Repository Breast Cancer Dataset attracted a large number of patients with multivariate attributes that were taken as a sample set.

Today's healthcare sector can benefit from the delivery of medical big data through machine learning knowledge. Machine learning techniques can describe cases from an objective perspective and generate predictions about diagnostic outcomes from combinations of related cases pathological factors. The introduction of machine learning to medical diagnostic methods and diagnostic accuracy will greatly change medical care in the future, and it is a direction that cannot be avoided. In 2019, an estimated 268,600 cases of invasive breast cancer were diagnosed in women and approximately 2,670 cases in men. Additionally, an estimated 48,100 cases of her DCIS have been diagnosed in women. About 41,760 women and 500 men are expected to die from breast cancer in 2019. Of interest is the development of breast cancer and its prognosis. Breast cancer has become the most common type of cancer in humans. Teams of doctors Cancer is divided into four stages, stages 1-4. Cancer diseases are more likely to be cured if they are detected early. There are new technologies that can identify the type and stage of cancer. However, false positives and true negatives can exist. This FP and TN can be deadly. Machine learning classification algorithms using this technology can be used to score FP and TN. This machine learning project proposes a comparison with ML Algorithms. A classification algorithm that can accurately predict cancer types.

III. MAIN CONTRIBUTION & OBJECTIVES

- Our objective is to compare machine learning algorithms to predict and diagnose breast cancer.
- To find the most effective one based on the performance of each classifier in terms of confusion matrix, precision, accuracy, and sensitivity.
- To analyze data from clinical trials to find previously unknown side effects.
- Machine learning algorithms can be used in medical imaging (such as X-rays or MRI scans) using patterns.
- Recognition to look for patterns that indicate a particular disease. This could potentially help doctors make quicker, more accurate diagnoses.
- Pattern recognition to seek diseases that might be indicated by certain patterns. This might enable medical professionals to diagnose patients more quickly and effectively.
- The goal of this analysis is to determine which characteristics are most useful in predicting either malignant or benign cancer as well as to identify broad trends that may help us choose the right model and hyperparameters.
- The remainder of this paper is divided into sections focused on the introduction of the method, the results of previous studies on breast cancer diagnosis, the description of the proposed methodology, and the detailed presentation and explanation of the experimental results.

IV. RELATED WORK

A substantial amount of work is already done in the determination of breast cancer by employing different methods. A new technique in support of breast cancer detection was introduced by the researcher. Bayesian Networks and Support Vector Machines are put forward by them in support of breast cancer detection. Based on the quantity chosen method, classifiers of Machine Learning were compared in support of Breast Cancer treatment. A DNA alteration or mutation is one of the causes of breast cancer. When cells start to multiply uncontrollably, cancer develops. Breast cancer cells frequently cluster together to create a lump or an x-ray-visible tumor. The two most prevalent kinds of breast cancer are invasive carcinoma and ductal carcinoma in situ (DCIS). Some are less frequent, such as phyllodes tumors and angiosarcoma. Logistic regression was applied by Wang, D.; Zhang, and Y.-H Huang (2018) et al. [1] and resulted in an accuracy of 96.4%. With an accuracy of 96.85%, Akbugday et al. [2] completed classification on the Breast Cancer Dataset using KNN and SVM. Random Forest was used by KAYA KELES et al. [3] to reach an accuracy of 92.2% in their study titled "Breast Cancer Prediction and Detection Using Data Mining." To determine the best classifier for breast cancer datasets, Vikas Chaurasia and Saurabh Pal et al. [4] examine the performance criteria of supervised learning classifiers including Naive Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48), and basic CART. Using AdaBoost, Dalen, D. Walker, and G. Kadam et al. [5] outperformed Random Forest in accuracy by 97.5%. With ensemble approaches and neural networks, Kavitha et al [6].'s accuracy was 96.3% lower than that of earlier investigations. Sinthia et al. [7] claim that the backpropagation method was used with 94.2% accuracy. The experimental outcome demonstrates that the SVM-RBF kernel outperforms other classifiers in terms of accuracy; in the Wisconsin Breast Cancer (original) datasets, it achieves a score of 96.84%. We have employed SVM, KNN, Random Forest, Naive Bayes, and ANN as classification techniques. Three key areas are the focus of cancer development prediction and prognosis: risk assessment or cancer susceptibility prediction, cancer relapse prediction, and cancer survival rate prediction. The likelihood of developing a specific cancer is predicted in the first domain before the patient is diagnosed. The third case aims to predict several potential parameters characterizing cancer development and treatment after the diagnosis of the disease: survival time, life expectancy, progression, drug sensitivity, etc. The second issue relates to the prediction of cancer recurrence in terms of diagnostics and treatment. The effectiveness of medical care and the accuracy of the diagnosis has a significant impact on the cancer survival rate and relapse rate. Data pre-processing, as we are aware, is a data mining technique used to filter data into a format that may be utilized. since the dataset from the real world is generally always available in numerous formats. It must be filtered in a comprehensible way because it is not available in accordance with our needs. Pre-processing data is a tried-and-true way to fix these problems. Data pretreatment involves converting the dataset into a format that may be utilized for preprocessing.

The Summary List of existing works on the specified domain is as follows:

A. Wang, D.;Zhang and Y.-H Huang(2018)[1]

The Logistic Regression method was proposed by the author and attained an accuracy of 96.4 %. The dataset used here is electronic health records, and with the help of the tool WEKA, it resulted in an error rate of 0.33%. The points we observed here are 5- year survivability predictions using logistic regression.

B. Akbugday(2018)[2]

The techniques used here are KNN SVM and Naïve Bayes methods by the author and attained an accuracy of 96.85 %, 95.99, and 96.85 % respectively. The dataset used here is the Breast Cancer Wisconsin dataset, and with the help of the tool WEKA, it resulted in an error rate of 0.66%. The points we observed here are the optimal k-value for a K-NN classifier, g K-NN is a lightweight, lazy learning algorithm with very short build times.

C. V Chauriya & S Paul [4]

The techniques used here are the Statistical Feature Selection method by the author which attained an accuracy of 92.3%. The dataset used here is the Breast Cancer Wisconsin dataset, and with the help of the tool WEKA, it resulted in an error rate of 0.3%. The points we observed here are the Patient features sorted out from data materials that are statistically tested based on the type of individual feature. Then 51 attributes or features are selected, and a feature's importance score is calculated. XGBoost algorithm is done by repeating 10-fold cross-validation.

D. Keles, M. Kaya, [3]

The techniques used here are KNN, SVM, Decision Tree, and Naïve Bayes methods by the author and attained an accuracy of up to 96.91%. The dataset used here is the Wisconsin Diagnostic Breast Cancer dataset, and with the help of the tool Python, it resulted in an error rate of 0.33 The points we observed here are the SVM map of the input vector into a feature space of higher dimensionality and identify the hyperplane that separates the data points into two classes. The marginal distance between the decision hyperplane and the instances that are closest to the boundary is maximized.

E. KELES et al., (2019) [3]

The techniques used here are Random Forest methods by the author and attained an accuracy of 92.2%. The dataset used here is the Breast Cancer Wisconsin dataset and with the help of the tool WEKA. The points we observed here are each dataset is generated with displacement from the original dataset. Then trees are developed using a random selection feature but are not pruned.

F. Chauraisa et. al [4]

The techniques used here are Naïve Bayes, Decision Tree, and Bagging Algorithm methods by the author and attained an accuracy of 96.5%. The dataset used here is the UC Irvine machine learning repository and with the help of the tool WEKA. The points we observed here are that the Decision Tree (DT) is the best predictor on the holdout sample (this prediction accuracy is better than any reported in the literature).

G. Delen at al. [5]

The techniques used here are AdaBoost methods by the author and attained an accuracy of 97.5 %. The dataset used here is Cancer Society and with the help of the tool WEKA. The observed advantages are the Low in error rate, performing well in the low noise data set. The advantage of this algorithm is that it requires fewer input parameters and needs little prior knowledge about the weak learner.

H. Sinthia et al. [7]

The techniques used here are KNN SVM and Naïve Bayes methods by the author and attained an accuracy of 94.2%. The dataset used here is the Wisconsin Diagnosis Breast Cancer BCI dataset, and with the help of the tool CAD System, it resulted in an error rate of 0.66%. The points we observed here are the Logistic Regression and the Backpropagation neural Network.

I. Chaurasia et. at [8]

The techniques used here are SVM methods by the author and attained an accuracy of 97.13 %. The dataset used here is the Breast Cancer Wisconsin dataset and with the help of the tool WEKA. The points we observed here are the optimal k-value for an It gives the most optimal hyperplane to distinguish between two classes.

J. Khourdifi et al. [10]

The techniques used here are Fast Correlation Based Filter with SVM, Random Forest, Naïve Bayes, KNN, and MLP methods by the author and attained an accuracy of 96.1 %. The dataset used here is the Breast Cancer Wisconsin dataset, and with the help of the tool WEKA, it resulted in an error rate of 0.0404%. The points we observed here are that Attributes are reduced by deleting irrelevant and redundant attributes which have no meaning in the classification task technique.

K. Khuriwal et.al [9]

The techniques used here are Naïve Bayes and SVM, methods by the author and attained an accuracy of 74.44 %. The dataset used here is Haberman's Survival dataset and with the help of the tool WEKA. The points we observed here help in marginalizing the hyperparameters and differentiating classes.

L. Mohana,et. al [11]

The techniques used here are the Decision Tree method by the author which attained an accuracy of 96.33. The dataset used here is the Breast Cancer Wisconsin dataset and with the help of the tool WEKA. The observed points here help in Splitting and choosing the best attributes.

M. Shravya at al. [12]

The techniques used here are the SVM method by the author and attained an accuracy of 92.7 % respectively. The dataset used here is the Breast UCI repository and with the help of the tool Spyder. The points we observed here are the Hyperplane separates two classes which helps in higher accuracy.

N. Wang et. al [13]

The techniques used here are PCA methods by the author and Eight PCs are chosen based on the screen plot, which explains 92.6% of the total correlation. And ten PCs are selected based on a 95% correlation. The dataset used here is Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995), and with the help of the tool WEKA, it resulted in an error rate of 0.66%. The points we observed here are dimension reduction Wang and Yoon technique, which manifests some advantages in terms of prediction accuracy and efficiency.

O. Bellaachia et. al [14]

The techniques used here are Naïve Bayes methods by the author and attained an accuracy of 96.3%. The dataset used here is the SEER Public-Use Data and with the help of the tool WEKA. The points we observed here are the Gives a probabilistic model for classification Helping in classification.

P. AMRANE et al. (2018) [17]

The techniques used here are KNN and Naïve Bayes methods by the author and attained an accuracy of 0.975109 % and KNN 0.961932 % respectively. The dataset used here is the Breast Cancer Wisconsin dataset, and with the help of the tool WEKA, it resulted in an error rate of 0.66%. The points we observed here are KNN classifiers are ranked first in terms of accuracy and duration.

Q. Khuriwal and Mishra (2018) [16]

The techniques used here are the Ensemble Voting Method and the Logistic Regression methods by the author which attained an accuracy of 98.50%. The dataset used here is the Wisconsin Diagnosis Breast Cancer dataset. UCI open the database, and with the help of the tool DWT, it resulted in an error rate of 0.99%. The points we observed here are useful for predicting the class of a binomial target feature.

R. Kibeom et. al [19]

The techniques used here are Bagging and AdaBoost Decision trees methods by the author and attained a Single C4.5 – 95.6%, Bagging C4.5 – 93.29%, and AdaBoost C4.5 – 92.62%. The dataset used here is Gene Expression Dataset Collection, and with the help of the tool WEKA, it resulted in a Sensitivity of 56% and 72%. The observed points that are Ensemble Method help to combine multiple learners.

S. Al-hadidi et al. [18]

The techniques used here are Logistic Regression and Backpropagation neural Network methods by the author and attained an accuracy of Greater than 93.7%. The dataset used here is the General Sample and with the help of the tool MATL AB, it resulted in an error rate of Less than 0.07. The points we observed here are BPNN is easy to implement and has been used widely for classification purposes. LR needs a hypothesis and a cost function that optimizes performance.

V. PROPOSED FRAMEWORK

Proposed Methodology

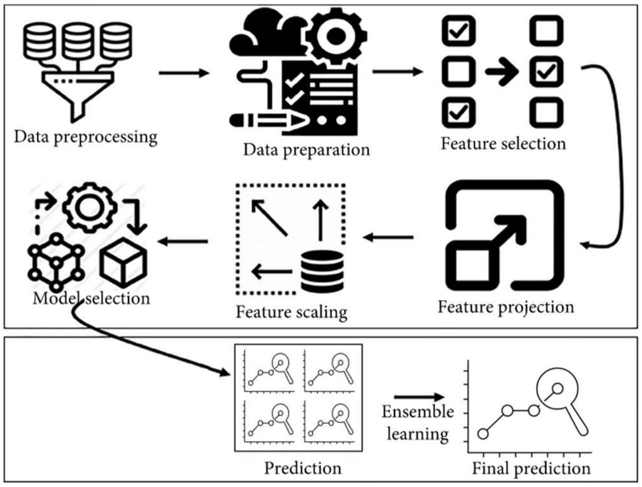


Fig 7: Various Phases of Machine Learning.

Phase 1: Pre-Processing Data

In the initial stage, we gather the data that we are interested in gathering to apply classification and regression methods before pre-processing. A data mining approach called data pre-processing entails putting raw data into a comprehensible format. Often, real-world data is insufficient, unreliable, and lacking likely to be filled with mistakes. Pre-processing data is a tried-and-true way to fix these problems. Raw data is prepared for subsequent processing by data pre-processing. We pre-processed the UCI dataset by using the standardization approach. This stage is highly crucial since the quality and quantity of data that you acquire will directly affect how good your prediction model may be. In this instance, we gather samples of both benign and malignant breast cancer. This will be our training data.

Phase 2 – Data Preparation

The process of loading our data into the proper location and getting it ready for use in our machine-learning training is known as data preparation. All our data will be assembled initially, and the ordering will be randomized after that.

Phase 3 – Feature Selection

The process of choosing a subset of pertinent characteristics to be used in the creation of a model is known as feature selection in machine learning and statistics, as well as the variable selection and attribute selection.

Selection of the Data File and Feature Wisconsin Breast Cancer (Diagnostic) - Data We chose roughly 8–9 parameters from the available 31 in the Kaggle library. Breast cancer diagnosis, whether it is malignant or benign, is our target parameter. For Feature Selection, the Wrapper Method was utilized. The important features found by the study are: Concave points worst, Texture worst, Area se, Area worst, Texture mean, Smoothness worst, Radius mean, Smoothness mean, and Symmetry means.

For Feature Selection, the Wrapper Method was utilized. The study's key findings include the following: concave points are the worst 2. Worst area 3. The texture is poor 4. Area se 5. The texture is poor 6. Smoothness is the worst Smoothness means 7. Radius means 8. 9. The term symmetry

Attribute Information:

ID number 2) Diagnosis (M = malignant, B = benign) 3–32).

Phase 4 – Feature Projection

Data from a high-dimensional space is transformed into a lower-dimensional space (with fewer properties) using feature projection. Depending on the nature of correlations between the features in the dataset, both linear and nonlinear reduction strategies can be applied.

Phase 5 – Feature Scaling

The Dataset will often include features with a wide range of magnitudes, units, and ranges. But since most machine learning algorithms compute the Euclidean distance between two data points.

All characteristics must be brought to the same magnitude level. Scaling can be used to accomplish this.

Phase 6 – Model Selection

The process of supervised learning involves labeling the input and output of the data before the machine is trained on it. The model may be trained using historical data, and it can analyze new data to make predictions about the future. They are divided into strategies for regression and classification. When the outcome is a regression, there is a genuine or consistent value, such as "wage" or "weight." When the outcome is a category like screening emails as "spam" or "not spam," there is a classification difficulty. Unsupervised learning is the process of giving the computer information that has not been categorized or labeled and letting the algorithm analyze the data without being provided any instructions. In an unsupervised learning algorithm, the machine is trained from the data which is not labeled or classified making the algorithm work without proper instructions. The outcome variable, or dependent variable, in our dataset, Y, only has two possible sets of values: M (Malign) or B. (Benign). Therefore, the supervised learning Classification method is used on it. Three main categories of machine learning classification algorithms have been selected. We can employ a straightforward tiny linear model.

Phase 7 – Prediction

Data is used by machine learning to provide answers to queries. Therefore, the prediction or inference step is when we get to provide answers to various queries. This is the point of all this work, where the value of machine learning is real.

METHODS USED

Logistic Regression

The field of machine learning predates the introduction of logistic regression, which was developed in 1958 by statistician DR Cox. It is a method of supervised machine learning that is used for categorization tasks (for predictions based on training data). Similar to linear regression, logistic regression uses an equation, but the result is a categorical variable as opposed to a value in other regression models. The independent variables can be used to forecast binary outcomes.

The general workflow is:

(1) Get a dataset

(2) Train a classifier

(3) Make a prediction using the classifier

K-Nearest Neighbor (k-NN)

As the data provided to K-Nearest Neighbor is labeled, it is a supervised machine learning algorithm. It is a nonparametric method since the closest training data points are used to classify test data points rather than taking the dataset's dimensions (or parameters) into account.

The general workflow is:

(1) The dataset should be entered and divided into a training and testing set.

(2) Pick an instance from the testing sets and calculate its distance with the training set.

(3) Rank the distances from the closest point.

(4) The class instance is the most common class of the first training instances (k=2).

Support Vector Machine

Support A training technique for learning classification and regression rules from data, Vector Machine is a supervised machine learning system that performs well in pattern recognition applications. When there are many features and instances, SVM is employed most precisely. The SVM algorithm creates a binary classifier. Each data point in an n-dimensional space in an SVM model, where n is the number of features, is represented as the value of a coordinate in the n-dimensional space.

A support vector machine algorithm model operates as follows:

(1) It starts by identifying boundaries or lines that classify the training dataset appropriately.

(2) The line or boundary with the greatest distance from the nearest data points is then selected from among those available.

XGBoost

A high-scalability DT ensemble built on gradient boosting is called XGBoost. Similar to gradient boosting, XGBoost minimizes a loss function to create an expansion of the objective function that is addictive. The complexity of the trees is managed using a variation of the loss function as indicated in equations [2] because XGBoost only employs DTs as base classifiers.

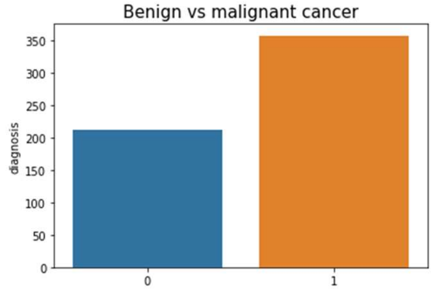
$$L_{\text{xgb}} = \sum_{i=1}^{N} L\left(y_i, F\left(x_i\right)\right) + \sum_{m=1}^{M} \Omega\left(h_m\right),$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2,$$

Where T is the number of leaves on the tree and W denotes the leaf output scores. This loss function can be incorporated into decision trees' split criterion, resulting in a prepruning approach. Trees with higher values are simpler. Determines how much loss reduction gain is required to split an internal node. Shrinkage is an additional regularization parameter in XGBoost that reduces the additive expansion step size. Finally, other tactics such as tree depth can be used to limit the complexity of the trees. The models are trained faster and need less storage space as a side effect of reducing tree complexity.

VI. DATA DESCRIPTION

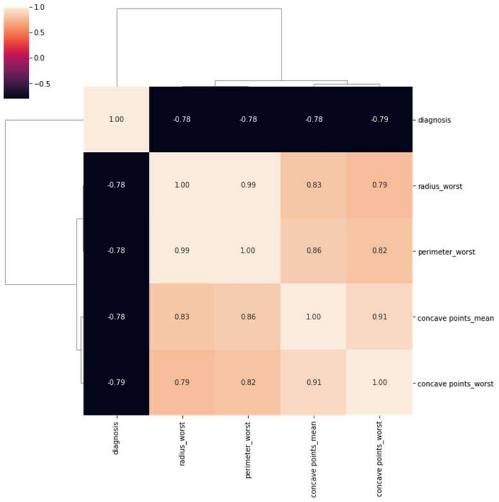
The Wisconsin Breast Cancer (Diagnostic) dataset is utilized in this investigation. The dataset and (CSV) file can both be found in Ref [15] and the UCI machine learning repository, respectively. The (CSV) file was then translated by MATLAB into (MAT) format. The dataset's characterization, in brief, is given below. The dataset includes 569 patterns, 357 of which are benign and 212 of which are malignant, divided into three classifications (ID number, benign, and malignant), with 32 columns for characteristics. A digital image of a fine needle aspirate (FNA) for the breast mass is used to record the features. The dataset does not have any missing attribute values and is coded with four considerable digits. A simulation environment (MATLAB 2015a) was utilized for this study. All the experiments of the used classifiers were conducted using the machine learning toolbox (classification learner) which contains a collection of machine learning algorithms.



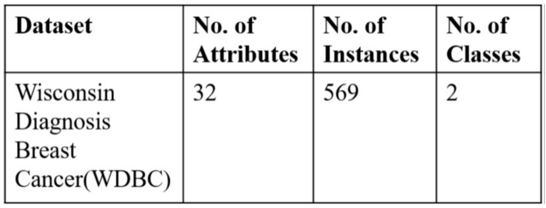
Count of Benign and malignant cancer

Dataset Details:

In this study, we used the dataset provided by researchers at the University of Wisconsin. Dr. Wolberg, at the University of Wisconsin Hospital, first created the group of images using Fine-Needle Aspiration (FNA) biopsies of the breast. Image processing was then applied on the set of images to come up with the WDBC dataset. The dataset was obtained from the University of California Irvine (UCI) Machine Learning Repository [2]. The features in this dataset were computed from digitized FNA samples. A portion of the well-differentiated cell was scanned using a digital camera.



Correlation of features with threshold > 0.75



Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Diagnosis Breast Cancer(WDBC)	32	569	2

Fig: Description of WDBC Dataset

VII. Results/Experimentation & Comparison Analysis

All tests on the classifiers described in this paper were carried out using libraries from the Anaconda machine learning environment, and the work was implemented on an i3 CPU running at 2.30GHz, with 2 GB of RAM, and 320 GB of external storage. In experimental investigations, training and testing are divided by 70% and 30%, respectively. Machine learning techniques for pre-processing data, classification, regression, clustering and association rules are available in JUPITER. Jupiter's machine learning tools are used to tackle a range of practical issues. The analysis of the data's findings is presented.

We utilize the 10-fold cross-validation test, a method for assessing predictive models that divides the original set into a training sample to train the model and a test set to evaluate it, to apply and test our classifiers. We attempt to visually analyze the data following the application of the pre-processing and preparation methods in order to determine the distribution of values in terms of efficacy and efficiency.

In terms of model-building time, correctly categorized cases erroneously classed instances, and accuracy, we assess each classifier.

Accuracy Comparison with other methods -

METHOD	ACCURACY %	Training Acc
SVM	97.36	98.46
Decision Tree	94.73	99.12
Logistic Regression	96.49	98.46
KNN	97.36	97.58
XG Boost	98.24	99.78

Table 1

In Table, I contrast the precision of our GA with the most accurate findings from prior research on the WBCD data set. Our aim is to demonstrate the confidence levels we offer for the predictions, not to achieve improved accuracy. We can affirm that our GA implementation is accurate enough based on the data in Table I, which provide an average accuracy of 97.20%.

VIII. Conclusion and Future Work

In this part, we can see that SVM builds its model in roughly 0.07 seconds as opposed to k-0.01 NN's seconds. It can be explained by the fact that, in contrast to other classifiers that create models, k-NN is a lazy learner and does not accomplish anything throughout the training process. On the other hand, SVM's accuracy (97.13%) is higher than that of C4.5, Naive Bayes, and k-NN, whose accuracy ranges from 95.12% to 95.28%. It is also clear that, compared to the other classifiers, SVM has the highest value of examples that are successfully classified and the lowest value of instances that are wrongly classified.

Following the development of the projected model, we can now analyze the outcomes in order to gauge the effectiveness of our algorithms. The highest TP value for the benign class was obtained by SVM and C4.5 (97%), although k-NN properly predicts 97% of instances that fall into the malignant class. SVM classifiers have a lower FP rate (0.03 for the benign class and 0.02 for the malignant class), and then the k-NN, C4.5, and NB methods are used. These findings help us to understand why SVM has performed better than other classifiers.

In summary, SVM was able to show its power in terms of effectiveness and efficiency based on accuracy and recall.

Future Work

The results analysis shows that the combination of multidimensional data with various feature selection, classification, and dimensionality reduction techniques can offer advantageous tools for inference in this field. It is necessary to conduct additional studies in this area to improve the performance of classification algorithms and enable them to make predictions on a wider range of factors. To attain high accuracy, we plan to parametrize our categorization systems. We are investigating a variety of datasets and the potential applications of machine learning techniques to further describe breast cancer. We want to maximize accuracy while lowering error rates.

REFERENCES

[1] Wang, D. Zhang and Y. H. Huang "Breast Cancer Prediction Using Machine Learning" (2018), Vol. 66, NO. 7.

[2] B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

[3] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

[4] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", IJCSMC, Vol. 3, Issue. 1, January 2014, pg.10 – 22.

[5] Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

[6] R. K. Kavitha1, D. D. Rangasamy, "Breast Cancer Survivability Using Adaptive Voting Ensemble Machine Learning Algorithm Adaboost and CART Algorithm" Volume 3, Special Issue 1, February 2014

[7] P. Sinthia, R. Devi, S. Gayathri and R. Sivasankari, "Breast Cancer detection using PCPCET and ADEWNN", CIEEE' 17, p.63-65

[8] Vikas Chaurasia and S.Pal, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" (FAMS 2016) 83 (2016) 1064 – 1069

[9] N. Khuriwal, N. Mishra. "A Review on Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques", (2018), Vol. 1, No. 1.

[10] Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-6

[11] R. M. Mohana, R. Delshi Howsalya Devi, Anita Bai, "Lung Cancer Detection using Nearest Neighbour Classifier", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019

[12] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.

[13] Haifeng Wang and Sang Won Yoon, "Breast Cancer Prediction Using Data Mining Method", Proceedings of the 2015 Industrial and Systems Engineering Research Conference,

[14] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"

[15] Juhyeon Kim, Hyunjung Shin, Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data, Journal of the American Medical Informatics Association, Volume 20, Issue 4, July 2013, Pages 613–618.

[16] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-5.

[17] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4.

[18] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39.

[19] Kibeom Jang, Minsoon Kim, Candace A Gilbert, Fiona Simpkins, Tan A Ince, Joyce M Slingerland "WEGFA activates an epigenetic pathway regulating ovarian cancer-initiating cells" Embo Molecular Medicines Volume 9 Issue 3 (2017)

[20] SA Medjahed, TA Saadi, A Benyettou "Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules" International Journal of Computer Applications 62 (1), 2013

[21] Abien Fred Agarap. 2017. A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data. arXiv preprint arXiv:1709.03082 (2017).

[22] Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and other kernel- based learning methods.Ai Magazine 2000, 22, 190.