Breast Cancer Prediction Using Machine Learning

1st Kontham Sai Yashwanth Reddy 2nd Phanindra Dokku 3rd Goutham Parakala

***Abstract*—Around 10% of women worldwide will get breast cancer at some point in their life, making it one of the most feared and frequent diseases among women today [1]. Despite the fact that this sickness is already treatable in practice in many countries, the main issue emerges when the disease cannot be adequately recognized in the very early stages. Machine learning has shown to be crucial in this area for forecasting diseases like cancer. Techniques for classifying data, like data mining, have so far shown to be dependable and effective. These methods have been used to identify diseases, particularly in the field of medicine. In this study, we successfully applied six classification methods to the Wisconsin Breast Cancer (original) datasets, including K-Neighbors, Logistic Regression, Nave Bayes and Support Vector Machine (SVM), Random Forest classifier, both before and after Principal Component analysis. The major goal is to evaluate the accuracy of data classification based on each algorithm's efficiency and effectiveness in terms of accuracy, precision, recall, specificity, and F1 Score.**

***Index Terms*—Decision tree, Machine learning, Support vector machine, Recall, 10-Fold cross-validation**

I. Introduction

The most prevalent kind of cancer in women is breast cancer (BC), which affects 10% of all women at some point in their lives. According to research, the survival rate is 88 percent after five years from diagnosis and 80 percent after ten years from diagnosis in the modern era. Since 1989, there has been significant progress in the early detection of breast cancer, with a 39 percent reduction in mortality. Patients frequently undergo a battery of tests, including but not limited to mammography, ultrasound, and biopsy, to assess their chances of being diagnosed with breast cancer due to the variety of symptoms associated with the disease. The most representative of these procedures, a biopsy, involves the removal of sample cells or tissues for analysis. A pathology laboratory will use a microscope to evaluate the sample of cells after a breast fine needle aspiration (FNA) operation [12]. Microscopical images can be used to calculate numerical properties such as radius, roughness, perimeter, and area. Later, data from the FNA are examined in conjunction with different imaging data to determine the likelihood that the patient has a malignant breast cancer tumor. In this case, an automated method would be quite helpful. It'll probably speed things up and make the doctor's predictions more accurate. Furthermore, if validated by the abundance dataset and the automated

If the system regularly performs well, it may reduce the need for patients to undergo further procedures like mammography, ultrasound, and MRI, which expose patients to a lot of radiation and suffering. All things considered, early prognosis is still a crucial part of the follow-up procedure. Reducing the amount of erroneous positive and false negative judgments can be accomplished via data mining techniques or categorization. As a result, modern techniques like knowledge discovery in databases (KDD) have emerged as the preferred method for medical researchers. The Wisconsin Breast Cancer (original) datasets were used in this study to test six classification models, including Decision Tree, K-Neighbors, Linear Discriminant Analysis (LDA), Logistic Regression, Nave Bayes, and Support Vector Machine (SVM), both before and after applying Principal Component Analysis. The outcomes are then evaluated using a range of performance criteria to compare the algorithms and determine which one is the most effective at predicting cancer.

This type of cancer may be highly prevalent in females and is the second leading cause of cancer death. About one in eight girls in the United States are at risk of acquiring breast cancer during their lifetime. One breast cell dividing uncontrollably causes the start of breast cancer, which manifests as a visible growth known as a tumour. Either a benign or malignant tumor may exist. The right classification in identifying whether the tumor is benign or malignant may save lives. In light of this, the need for precise classification within the clinic may be the cause of some genuine anxiety among specialists and physicians. Since scientists first realized the value of making constrained decisions to cure particular diseases, the importance of artificial intelligence has grown significantly during the previous 25 years. One of the most important diseases in medicine where the classification problem plays a really important part is that of breast cancer diagnosis, where the use of machine learning and data processing as tools for diagnosis becomes incredibly effective. Therefore, using machine learning techniques will make it easier for clinicians to identify breast cancer and classify tumors correctly as benign or malignant. There is no doubt that the evaluation of patient data and the choice of a doctor or specialist are the most crucial elements in the identification, but knowledgeable systems and artificial intelligence techniques like machine learning for classification tasks also significantly help doctors and specialists.

II. Motivation

One of the most prevalent cancers in women and the second leading cause of cancer death in women is breast cancer[1]. Despite the lack of effective treatments, the high incidence and mortality of breast cancer are also largely due to the low diagnostic accuracy. Traditional methods for detecting breast cancer include mammography. Only 78% of breast cancers may be accurately detected by mammography, according to a survey from UCHealth [2]. Therefore, developing an expert system for breast cancer early diagnosis is necessary.

III. Objectives

- To properly clean and prepare the data for training the machine learning model, removing the rows with null or missing values from the data set.
- To put into practice and assess the machine learning classification models for the breast cancer dataset under consideration.

The Machine learning algorithms that we trained for the breast cancer dataset are:

- Logisitc regression
- Support Vector Machine
- K - Nearest Neighbors
- Naive Bayes
- Random forest

IV. Related work

Earlier, research using various data mining approaches was conducted on the categorization and prediction of breast cancer. Information mining frequently employs the techniques of classification and aggregation [1]. The goal of classification, also known as supervised learning in machine learning, is to categorize unknown objects while using the information set to learn about current patterns and classes, and then to predict what will happen in the future. In classification problems, the training set—used to create the classifying structure—and the test set—often used to evaluate the classifier—are typically specified [2].

Furthermore, significant progress has been made in predicting breast cancer survivorship using patient data that has been labeled, unlabeled, and pseudo-labeled. Machine learning algorithms have assisted prognostic studies of breast cancer survivorship because they can forecast a patient's survival based on past patient data. When it comes to forecasting breast cancer, neural networks and related methods play a significant role. Artificial neural networks have gained popularity among researchers over the last few decades and are currently a focus of active study [7–12].

With significant advancements in the categorization and early stage diagnosis of breast cancer, ANNs have achieved a number of achievements [2,7–12]. Input, hidden, and output layers make form the hierarchy of a standard ANN model. The backpropagation artificial neural network (BP-ANN) approach and its derivatives have been the subject of extensive research in the diagnosis of breast cancer [13]–[14]. However, the method has some drawbacks, such as no assurance of global optimum,extensive training times and a large number of tuning parameters. Huang and Babri [15] suggested Single Hidden Layer Neural Networks (SFLN) to address the aforementioned issues using the extreme learning machine (ELM) tree-step learning procedure.For the early prediction of breast cancer, standard [16] and best-parameterized [17] ELM models were suggested. Results indicated that when compared to BP ANN, it generally provided greater accuracy, specificity, and sensitivity.

The majority of currently available works, however, pay little attention to the end users—medical professionals—and applicability issues in actual medical contexts. With all due respect to the previously mentioned related work, this paper compares the effectiveness of the algorithms Decision Tree, K-Neighbors, Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) using Wisconsin Breast Cancer (original) datasets in both diagnosis and analysis to make decisions.

V. Proposed frame work

Our goal is to determine whether a tumor is benign (non-cancerous) or malignant (cancerous). To achieve the most accurate results, we have outlined a straightforward approach. Getting a collection of numerical values for diverse situations was the first goal. In order to train and test six algorithms—Random forest, K-Neighbors, Logistic Regression, Nave Bayes, and Support Vector machine (SVM)—we divided the train-test ration to 70:30 after finalizing our dataset. To lessen the dimensionality of the dataset, feature selection in the form of Principal Component Analysis is performed. The models are retrained using training and testing, and the last step is a comparison of the findings to the earlier results. An overview of the complete thesis is provided in the workflow that follows:-

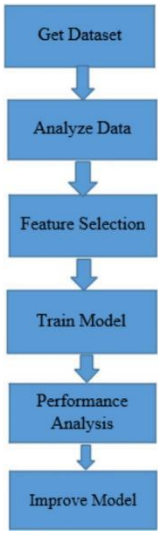
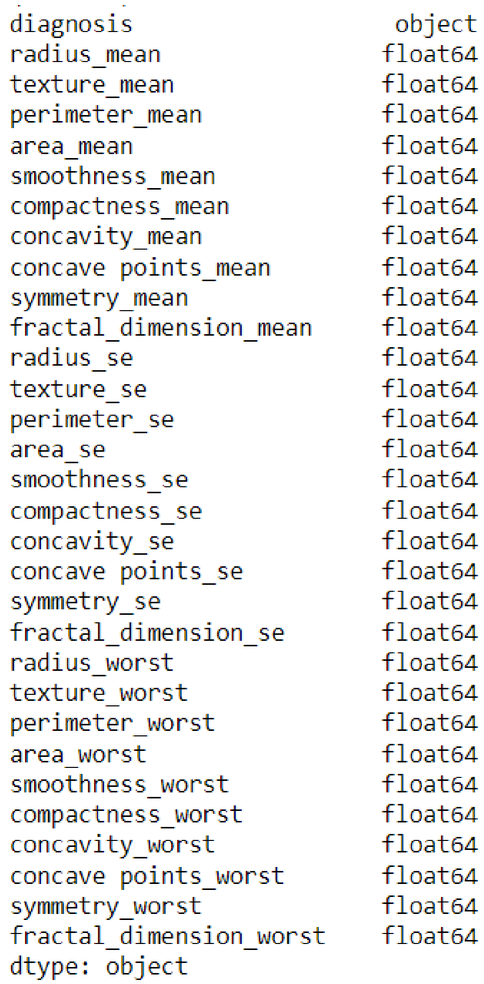


Fig. 1. Process Flow Diagram

A. Dataset Collection

Dr. William H. Wolberg, a physician at the University of California, created the dataset that was used in this study and it is openly accessible.American hospital in Madison, Wisconsin, called Wisconsin Hospital. There are 212 cases of malignant breast cancer and 357 cases of benign breast cancer in the dataset. The dataset has 32 columns, with the ID number as the first column, the diagnosis outcome (benign or malignant), as the second column, and then the mean, standard deviation, and the mean of the worst measurements of ten feature measurements. In the dataset, there were no missing values [18]. The dataset's column names are:



```
diagnosis                  object
radius_mean               float64
texture_mean              float64
perimeter_mean            float64
area_mean                 float64
smoothness_mean           float64
compactness_mean          float64
concavity_mean            float64
concave points_mean       float64
symmetry_mean             float64
fractal_dimension_mean    float64
radius_se                 float64
texture_se                float64
perimeter_se              float64
area_se                   float64
smoothness_se             float64
compactness_se            float64
concavity_se              float64
concave points_se         float64
symmetry_se               float64
fractal_dimension_se      float64
radius_worst              float64
texture_worst             float64
perimeter_worst           float64
area_worst                float64
smoothness_worst          float64
compactness_worst         float64
concavity_worst           float64
concave points_worst      float64
symmetry_worst            float64
fractal_dimension_worst   float64
dtype: object
```

Fig. 2. Features present in the breast cancer dataset

B. Feature selection

The term "feature selection" refers to variable selection or attribute selection. Automatic selection of data properties that are most pertinent to the challenge of predictive modeling. Dimensional reduction and feature selection are entirely independent processes. A dimensional reduction methodology reduces the number of attributes in the dataset by creating new combinations of attributes, whereas feature selection strategies include and exclude attributes that are already present in the data without ever changing them. Both strategies call for reducing the number of attributes in the dataset. Creates a precise forecasting model by ways for choosing features. The best or superior accuracy will be achieved with less data if characteristics are helped in their selection. The feature selection approach can be used to find and eliminate the unnecessary.

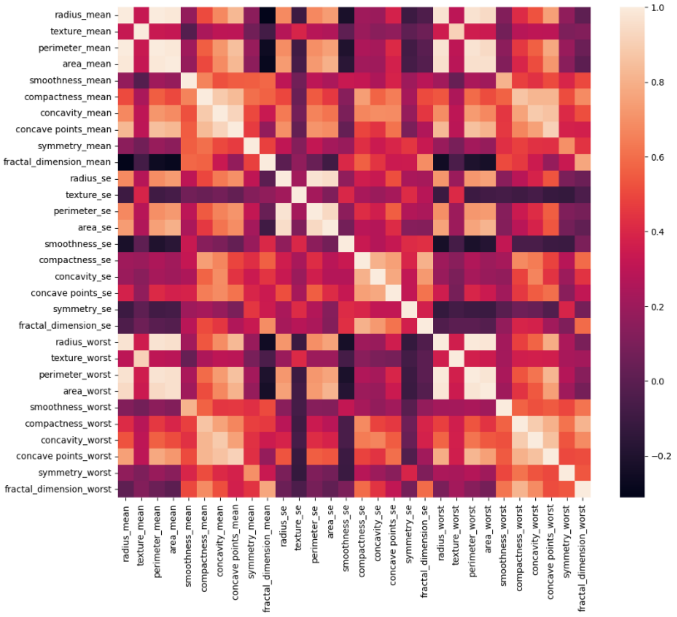
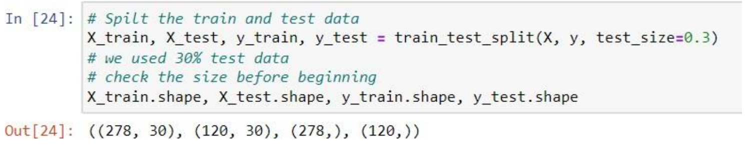


Fig. 3. Co-efficient correlation matrix

C. Division of data for validation and training

Our numerical tabular dataset was divided into two groups in this phase using the Python sklearn tool, with the remaining data serving as validation.



```
In [24]: # Spilt the train and test data
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
         # we used 30% test data
         # check the size before beginning
         X_train.shape, X_test.shape, y_train.shape, y_test.shape

Out[24]: ((278, 30), (120, 30), (278,), (120,))
```

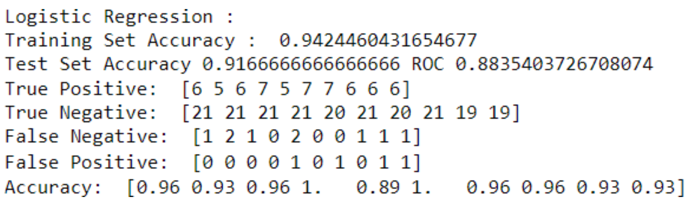
Fig. 4. Test and train split of data

D. Model building

The machine learning techniques in this section were implemented using the Python sklearn package.

1) Logistic Regression: Logistic regression examines the relationship between the independent variables, or the input, and the variable quantity, or the output. By calculating odds while utilizing its underlying supplies. It applies a penalty of 1.2 when regularizing. In order to anticipate a worth, the supply regression formula also employs an equation with independent predictors. The predicted worth is frequently anywhere from a negative to a positive eternity. The supply perform, also known as the sigmoid function, then converts the resulting chances to binary values zero or one.

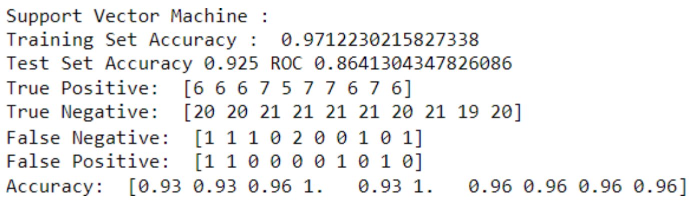
https://github.com/gouthamro45/ML-Project



```
Logistic Regression :
Training Set Accuracy :  0.9424460431654677
Test Set Accuracy 0.9166666666666666 ROC 0.8835403726708074
True Positive:  [6 5 6 7 5 7 7 6 6 6]
True Negative:  [21 21 21 21 20 21 20 21 19 19]
False Negative:  [1 2 1 0 2 0 0 1 1 1]
False Positive:  [0 0 0 0 1 0 1 0 1 1]
Accuracy:  [0.96 0.93 0.96 1.   0.89 1.   0.96 0.96 0.93 0.93]
```

Fig. 5. Logistic Regression performance

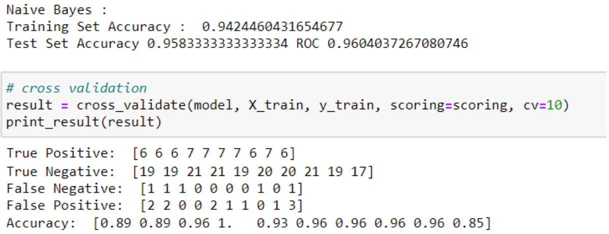
2) Support Vector Machine: A supervised machine learning algorithmic rule called the Support Vector Machine (SVM) may be utilized to solve any classification or regression problems. But it's mostly applied to categorization problems. According to this algorithmic approach, each data point is represented as a point in n-dimensional space, where n is the total number of features, and each feature's value corresponds to a certain coordinate [23]. Then, classification is carried out by locating the hyper-plane that separates the two classes.



```
Support Vector Machine :
Training Set Accuracy :  0.9712230215827338
Test Set Accuracy 0.925 ROC 0.8641304347826086
True Positive:  [6 6 6 7 5 7 7 6 7 6]
True Negative:  [20 20 21 21 21 21 20 21 19 20]
False Negative:  [1 1 1 0 2 0 0 1 0 1]
False Positive:  [1 1 0 0 0 0 1 0 1 0]
Accuracy:  [0.93 0.93 0.96 1.   0.93 1.   0.96 0.96 0.96 0.96]
```

Fig. 6. Support Vector Machine performance

3) Naive Bayes classifier: A group of classification algorithms that are based on the Bayes Theorem are known as naive Bayes classifiers. It's not just one algorithmic rule, but rather a family of algorithms where each algorithm follows a similar guiding concept and each attempt at classifying choices is independent of the others. The Bayes theorem employs contingent probability to compute the likelihood that a future event will occur by iteratively using prior knowledge. The Naive Bayes classifier makes the assumption that the input variables are independent of each alternative and that each alternative can independently affect the likelihood of the target variable.



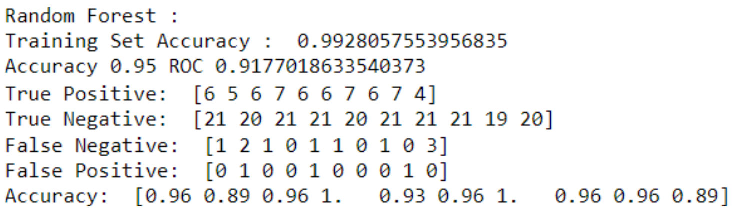
```
Naive Bayes :
Training Set Accuracy :  0.9424460431654677
Test Set Accuracy 0.9583333333333334 ROC 0.9604037267080746
```

```
# cross validation
result = cross_validate(model, X_train, y_train, scoring=scoring, cv=10)
print_result(result)
```

```
True Positive:  [6 6 6 7 7 7 7 6 7 6]
True Negative:  [19 19 21 21 19 20 20 21 19 17]
False Negative:  [1 1 1 0 0 0 0 1 0 1]
False Positive:  [2 2 0 0 2 1 1 0 1 3]
Accuracy:  [0.89 0.89 0.96 1.   0.93 0.96 0.96 0.96 0.96 0.85]
```

Fig. 7. Naive Bayes classifier performance

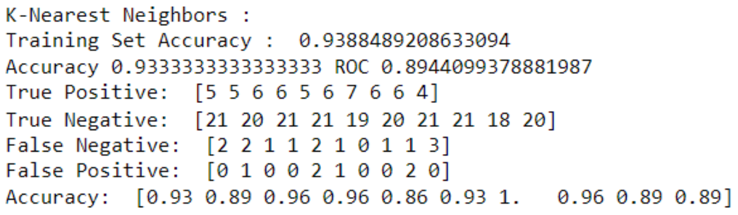
4) Random forest: The ensemble learning method is used for regression in the supervised learning algorithm known as Random Forest Regression. In order to produce predictions that are more accurate than those from a single model, the ensemble learning method integrates predictions from various machine learning algorithms.



```
Random Forest :
Training Set Accuracy :  0.9928057553956835
Accuracy 0.95 ROC 0.9177018633540373
True Positive:  [6 5 6 7 6 6 7 6 7 4]
True Negative:  [21 20 21 21 20 21 21 21 19 20]
False Negative:  [1 2 1 0 1 1 0 1 0 3]
False Positive:  [0 1 0 0 1 0 0 0 1 0]
Accuracy:  [0.96 0.89 0.96 1.   0.93 0.96 1.   0.96 0.96 0.89]
```

Fig. 8. Random forest classifier performance

5) KNN Classifier: The k-nearest neighbors (KNN) is a technique for categorizing data that calculates the likelihood that a data point will belong to one group or another depending on the group to which the data points closest to it belong. An example of a supervised machine learning technique used to address classification and regression issues is the k-nearest neighbor algorithm. However, classification issues are its primary application.



```
K-Nearest Neighbors :
Training Set Accuracy :  0.9388489208633094
Accuracy 0.9333333333333333 ROC 0.8944099378881987
True Positive:  [5 5 6 6 5 6 7 6 6 4]
True Negative:  [21 20 21 21 19 20 21 21 18 20]
False Negative:  [2 2 1 1 2 1 0 1 1 3]
False Positive:  [0 1 0 0 2 1 0 0 2 0]
Accuracy:  [0.93 0.89 0.96 0.96 0.86 0.93 1.   0.96 0.89 0.89]
```

Fig. 9. KNN Classifier performance

VI. Results and Analysis

Following the application and implementation of machine learning models, the next stage is to determine the model's effectiveness, or how well it performed on the datasets. This is done by applying the models to the previously created test dataset. 30% of the dataset for the prediction of breast cancer was included in the test dataset.Breast cancer prediction also underwent 10-fold cross-validation. We tested six machine learning algorithms: support vector machine, Naive Bayes classifier, KNN classifier, logistic regression, and random forest classifier.

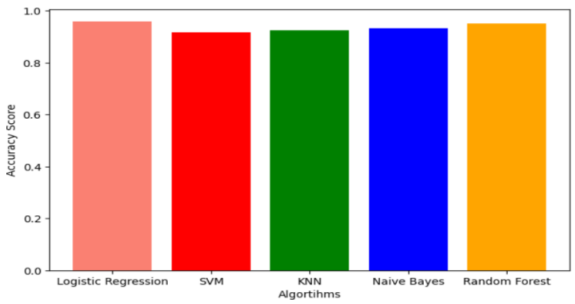


Fig. 10. Accuracy graph for the classifiers

https://github.com/gouthamro45/ML-Project

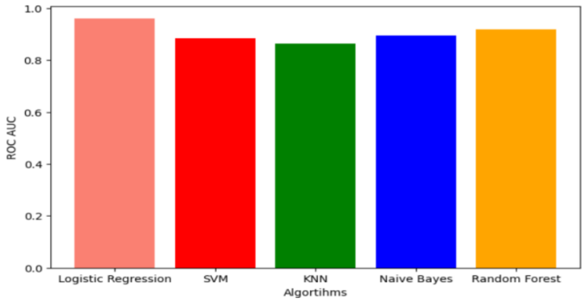


Fig. 11. ROC scores for the different classifiers

CONCLUSION

The accuracy rating for Naive Bayes is 0.964. Logistic regression (0.923) and K-Nearest Neighbors (0.9349) aren't far behind either. SVM has an accuracy rating of 0.917. Out of the six models, Decision Tree performs the worst, scoring 0.834. Much may be learned about the performance of the algorithms by taking into consideration the other performance matrices. The K- Neighbors and Naive Bayes algorithms perform better. Despite reductions in all other performance measures for both the described algorithms, SVM and Logistic Regression both earn a perfect 1.000 for recall, which is crucial for disease prediction. We can conclude that, when the recall score is taken into account, Logistic Regression and Support Vector Analysis outperform other methods for predicting breast cancer.

REFERENCES

[1]] Breast cancer facts and figures 2003-2004 (2003). American Cancer Society

[2] Stages — Mesothelioma — Cancer Research UK Breast cancer survival statistics September 26, 2017

[3] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.

[4] deepsense.ai What is reinforcement learning? The complete guide July 05, 2018

[5] Hacker Noon Absolute Fundamentals of Machine Learning – Hacker Noon January 15, 2018

[6] Furundzic, D.; Djordjevic, M.; Bekic, A.J. Neural networks approach to early breast cancer detection. J. Syst. Archit. 1998, 44, 617–633. [CrossRef]

[7]] Floyd, C.E.; Lo, J.Y.; Yun, A.J.; Sullivan, D.C.; Kornguth, P.J. Prediction of breast cancer malignancy using an artificial neural network. Cancer 1994, 74, 2944–2948. [CrossRef]

[8]H. A. Abbass, “An evolutionary artificial neural networks approach fo

[9]J. Khan, J. S. Wei, M. Ringn ér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.,” Nat. Med., vol. 7, no. 6, pp. 673–9, Jun. 2001.

[10]G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, (Dec. 2006.) vol. 70, no. 1–3, pp. 489–501.

[11]C. P. Utomo, A. Kardiana, and R. Yuliwulandari, “Breast Cancer Diagnosis using Artifi-cial Neural Networks with Extreme Learning Techniques,” Int. J. Adv. Res. Artif. Intell, vol. 3, no. 7,pp. 10–14,

[12]Fine Needle Aspiration Biopsy of the Breast. American Cancer Society.

[13]Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent.” Journal of statistical software 33.1 (2010): 1

Rohith Gandhi. Nearest Neighbor. Understanding Machine Learning (2018)

[14] Machine Learning Mastery Discriminant Analysis for Machine Learning September 22, (2016)

[16]G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, (Dec. 2006.) vol. 70, no. 1–3, pp. 489–501.

[17]C. P. Utomo, A. Kardiana, and R. Yuliwulandari, “Breast Cancer Diagnosis using Artifi-cial Neural Networks with Extreme Learning Techniques,” Int. J. Adv. Res. Artif. Intell, vol. 3, no. 7, pp. 10–14

[18]William H Wolberg, W Nick Street, and Olvi L Mangasarian. (1992). Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository.

[19]H. A. Abbass, “An evolutionary artificial neural networks approach for breast cancer diagnosis.” Artif.Intell. Med., vol. 25, no. 3, pp. 265–81,

[20]J. Khan, J. S. Wei, M. Ringn ér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.,” Nat. Med., vol. 7, no. 6, pp. 673–9

Github Link:

https://github.com/gouthamro45/ML-Project

https://github.com/gouthamro45/ML-Project