**BREAST CANCER PREDICTION USING MACHINE LEARNING ALGORITHMS**

**Submitted By**

Laxman Reddy Nalla, Vinay Kumar Avirneni,  
USER ID LXN20710, USER ID VXA45490,   
Department of Computer Science Department of Computer Science

Uday Kumar Kuppam, Gayathri Keshamoni,  
USER ID UXK97680 USER ID GXK24880  
Department of Computer Science, Department of Computer Science,

**Submitted To**

Khan, Muhammad Zubair,  
Course CS 5710  
Department of Computer Science

**UNIVERSITY OF CENTRAL MISSOURI  
Lee Summit's, Missouri**

**Introduction:**

Around the world, Breast cancer is the most widely recognized type of cancer alongside lung and bronchus cancer, prostate cancer, colon cancer, and pancreatic cancer among others. Breast cancer might be a prevalent reason for death, and it's the main kind of malignant growth that is boundless among ladies the around the world. Breast Cancer causes are multifactorial and include family ancestry, weight hormones, radiation treatment, and even reproductive factors. As indicated by the report of the world health organization every year, 2.1 million ladies are recently affected by breast cancer, which furthermore causes the highest number of cancer-related deaths among ladies [1]. In 2018, it is assessed that 627,000 ladies died from breast cancer - that is roughly 15% of all cancer deaths among ladies. While breast cancer growth rates are higher among ladies in extra-developed areas, rates are expanding in about each locale internationally. Many imaging techniques are developed for early identification and treatment of breast cancer and to scale back the amount of death and lots of aided breast cancer diagnosis methods are wont to increase the symptomatic precision. Machine Learning algorithms are widely utilized in intelligent human services frameworks. particularly for breast cancer diagnosis and guess. There are many machine learning classifications and algorithms for the prediction of breast cancer outcomes but during this paper, we are comparing various sorts of classification algorithms like k Nearest Neighbors, Support Vector Machine, Logistic Regression, and Gaussian Naive Bayes. And furthermore, assess and compare the performance of the varied classifiers as far as accuracy, precision, recall, f1-Score, and Jaccard index. The outcomes obtained during this paper provide a summary of the condition of modern Machine Learning strategies for breast cancer detection.

**Motivation:**

* Breast Cancer has become the most common type of cancer in humans. The doctors classified cancer into 4 stages 1 to 4. If cancers are recognized in the early stages, there is a high possibility of recovery from it.
* There are new technologies where we can detect the type of cancer and its stage of it, but there are likely to be false positives and true negative results. This FP and TN can end up with the cost of human life. We can assess the FP, and TN using Machine Learning Classification algorithms with the technology.
* In this Machine Learning Project, we are proposing a comparison between ML classification algorithms that can accurately predict the type of cancer.

**Significance:**

* It is very important to get high prediction accuracy since a false prediction can result in the cost of human life.
* Objectives:
* Compare different classification algorithms which give high accuracy for the breast cancer prediction problem.
* Implementation of different ML classification algorithms against breast cancer dataset.

**Objectives:**

* Compare different classification algorithms which give high accuracy for the breast cancer prediction problem.
* Implementation of different ML classification algorithms against breast cancer datasets.

**Increment:**

**Dataset:**[Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle](https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)

**Dataset Description:**

* Breast cancer is the most common cancer among women in the world. It accounts for 25% of all cancer cases and affected over 2.1 million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area.
* The key challenge against its detection is how to classify tumors into malignant (cancerous) or benign(non-cancerous). We ask you to complete the analysis of classifying these tumors using machine learning (with SVMs) and the Breast Cancer Wisconsin (Diagnostic) Dataset.

**Detailed Description of Features:**

1) ID number  
2) Diagnosis (M = malignant, B = benign)  
3) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from the center to points on the perimeter)  
b) texture (standard deviation of gray-scale values)  
c) perimeter  
d) area  
e) smoothness (local variation in radius lengths)  
f) compactness (perimeter^2 / area - 1.0)  
g) concavity (severity of concave portions of the contour)  
h) concave points (number of concave portions of the contour)  
i) symmetry  
j) fractal dimension ("coastline approximation" - 1)

**Features analysis:**Dataset statical information:

A screenshot of a computer

Description automatically generated

Type of cancers using count plot:  
Benign: 357  
Malignant: 212

Chart, bar chart

Description automatically generated

**Count Plot of Type of Cancers**

Chart, histogram

Description automatically generated

**Area mean of breast cancer tumour**

Chart, histogram

Description automatically generated

**Radius of Cancer tumour**

Correlation matrix:

The correlation matrix for the will shows the correlated features in dataset. The breast cancer dataset consists of 32 features, so it is difficult to see the correlation matrix so took the correlation matrix of features which has correlation thresholds above 0.75 with respect to Target.

Text

Description automatically generated

Graphical user interface, application, Teams

Description automatically generated

Fig: Correlation matrix of features

A picture containing text, colorful

Description automatically generated

There is a high correlation of 0.75 and above between **radius worst, perimeter worst, concave points mean, and concave points worst** these features so we removed some features such as **Radius worst, perimeter worst, and concave points** mean these are highly related features. By removing we can reduce features in data without losing important features.

**Implementation of Models:**

For this breast cancer data set we thought of using multiple machine-learning algorithms such as

1. PCA-KNN with n\_negibours as 2.
2. NCA-KNN with n\_negibours as 2
3. Logistic Regression
4. Decision Trees
5. SVM
6. XG Boost

**KNN with PCA:**

The k-nearest neighbor’s algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

Principal Component Analysis or PCA is a widely used technique for the dimensionality reduction of a large data set. Reducing the number of components or features costs some accuracy and on the other hand, it makes the large data set simpler and easy to explore and visualize. Also, it reduces the computational complexity of the model which makes machine learning algorithms run faster.

Text

Description automatically generated

This is the sample implementation of the KNN algorithm using 2 neighbors because we are classifying data between whether the cancer is benign or malignant.

Text

Description automatically generated

By applying PCA with the KNN algorithm we got the

Accuracy: 0.9321794871794872

Test Scores: 0.947368421052631

Train Score: 0.9346733668341709

**KNN with NCA:**

Neighborhood component analysis (NCA) is a non-parametric method for selecting features with the goal of maximizing the prediction accuracy of regression and classification algorithms. It learns linear transformation in a supervised fashion to improve the classification accuracy of a stochastic nearest neighbor’s rule in the transformed space.

Chart, scatter chart

Description automatically generatedAs you can see above plt nca gives better solution than pca (dots are seperated).

Text

Description automatically generated

By Applying NCA with KNN we the

Accuracy: 0.99

Test Score: 0.99

Train Accuracy: 0.99

**SVM:**

A Support Vector Machine (SVM) is a binary linear classification whose decision boundary is explicitly constructed to minimize generalization error. It is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression, and even outlier detection.

SVM is well suited for the classification of complex but small or medium-sized datasets.

Graphical user interface, text, application

Description automatically generated

A picture containing graphical user interface

Description automatically generated

By implementing SVM we got

Accuracy: 0.8829787234042553

These are some of the advantages of SVM:

Effective in high-dimensional spaces.

Still effective in cases where the number of dimensions is greater than the number of samples.

Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Versatile: different [Kernel](http://scikit-learn.org/stable/modules/svm.html#svm-kernels) functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

**WORK COMPLETED:**

As we proposed we decided to implement multiple machine learning algorithms we gathered the dataset from Kaggle and carried out some Exploratory data analysis on the dataset. After implementing EDA, we started implementing the KNN algorithm in which we implemented KNN with PCA (Principal component analysis) and KNN with NCA (Neighborhood component analysis). We have also implemented the SVM (Support vector machines) algorithm on the breast cancer dataset.

Responsibilities:

As a group we divided the tasks equally among us:

For each part of the project, we divided the work as follows:

* **Uday and Gayathri** have invested their time into referring to different papers and articles to assess and decide on the machine learning algorithms.
* **Laxma Reddy and Vinay** have handled the dataset collection, preprocessing, EDA, and implementation of algorithms and ML models.

Contribution:

* We four members worked together on this use case by discussing all the aspects that needs to be done.
* Overall, each person in the group contributed equally to the project i.e 25% of work per person.

**WORK TO BE COMPLETED:**

* At the end of part two of the project we left a few other Machine Learning Algorithms and the performance tuning part.

**References**

 [1] 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).

[2] C. Cortes and V. Vapnik, (1995), “Support-vector Networks.Machine Learning 20.3”, (1995), 273– 297. <https://doi.org/10.1007/BF00994018>.

[3] Gönen, M.; Alpaydın, E. Multiple kernel learning algorithms. J. Mach. Learn. Res. 2011, 12, 2211– 2268.

[4] 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+.

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

[6] Y. Sun, J. Liu, K. Yu, M. Alazab, K. Lin, “PMRSS: Privacy-preserving Medical Record Searching Scheme for Intelligent Diagnosis in IoT Healthcare”, IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3070544