IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

COMPARATIVE ANALYSIS OF BREAST CANCER PREDICTION BY USING MACHINE LEARNING ALGORITHMS

Uma Sneka[1] and Nancy Jasmine Goldena[2]

[1]PG Student, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India.

[2]Associate Professor, Department of Computer Applications & Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India.

Abstract:

Breast cancer is a type of tumour that occurs in the tissues of the breast. It is most common type of cancer found in women around the world and it is among the leading causes of deaths in women. This work presents the comparative analysis of machine learning, deep learning and data mining techniques being used for the prediction of breast cancer. Two algorithm KMEANS, PCA (Principal Component Analysis) which predict the breast cancer outcome have been compared in the paper using Kaggle machine learning repository dataset. The datasets consist 598 rows and 30 columns. All experiments are executed within a simulation environment and conducted in R studio platform. Many researchers have put their efforts on breast cancer diagnoses and prognoses, every technique has different accuracy rate and it varies for different situations, tools and datasets being used. Our main focus is to comparatively analyse different existing Machine Learning (ML) and Data Mining (DM) techniques in order to find out the most appropriate method that will support the large dataset with good accuracy of prediction. The main purpose of this research is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction and this project provides the all-necessary information to the beginners who want to analyse the machine learning.

Keywords: Breast Cancer, Machine Learning, Classification, Prediction, Comparative, Performance Evaluation.

Introduction:

Breast cancer is a condition in which the breast's cells multiply out of control. Breast cancer comes in several forms, which breast cells develop into cancer determine the type of breast cancer. Different areas of the breast might give rise to breast cancer. There are three basic components of a breast: connective tissue, ducts, and lobules. The glands that generate milk are called lobules. Milk travels through tubes called ducts to the nipple. The connective tissue, which is made up of fatty and fibrous tissue, envelops and holds everything in place. The ducts or lobules are where most breast cancers start. Blood and lymph vessels are two ways that breast cancer can travel outside of the breast.

Motivation of the work:

Mammograms are for many women the best technique to detect breast cancer at an early stage, when it is less difficult to cure and before it becomes large enough to feel or produce symptoms. Regular mammograms can reduce the risk of breast cancer-related death. For most women who are of screening age right now, a mammography is the best method of detecting breast cancer. False positive test results are one example of firearms; another is when a clinician observes something that appears to be cancer but is not. This may result in additional testing, which can be costly, intrusive, time-consuming, and anxiety-inducing. Additionally, when doctors over diagnose a malignancy that would not have developed into symptoms as a result of tests.

Motivation of the Work:

Treatments like radiation therapy or surgery that are advised for breast cancer can be overtreated. These may have unintended and undesirable side effects. Radiation exposure from the mammography examination itself and pain experienced during treatments are two other potential side effects of breast cancer screening. This requires human labour and takes time to complete. The pathologist's equipment and level of experience also play a role in making the decision. Therefore, to speed up the process and automatically recognise and locate tumour tissue cells, machine learning could be extremely helpful. One might construct a pipeline using vast amounts of tissue image data from multiple hospitals, which were assessed by various experts and encouraged additional research, in order to fully realise the potential.

Dataset Description:

The dataset is made up of features that were taken from a digitised picture of a breast mass that was sampled with a Fine Needle Aspiration (FNA). They define the characteristics of the cell nuclei seen in the images. The mean, standard error, and worst—the average of the three main aspects the additional three pieces of information for each feature. As a result, a patient ID, a response variable, and tumour features that were taken from a FNA images are included in the dataset. There are 598 rows and 30 columns in the breast cancer dataset. The number of women affected by breast cancer and the number of deaths it caused were discovered through analysis of the disease and those who recovered from it. The two techniques we used are PCA and K-MEANS.

Architectural Design:

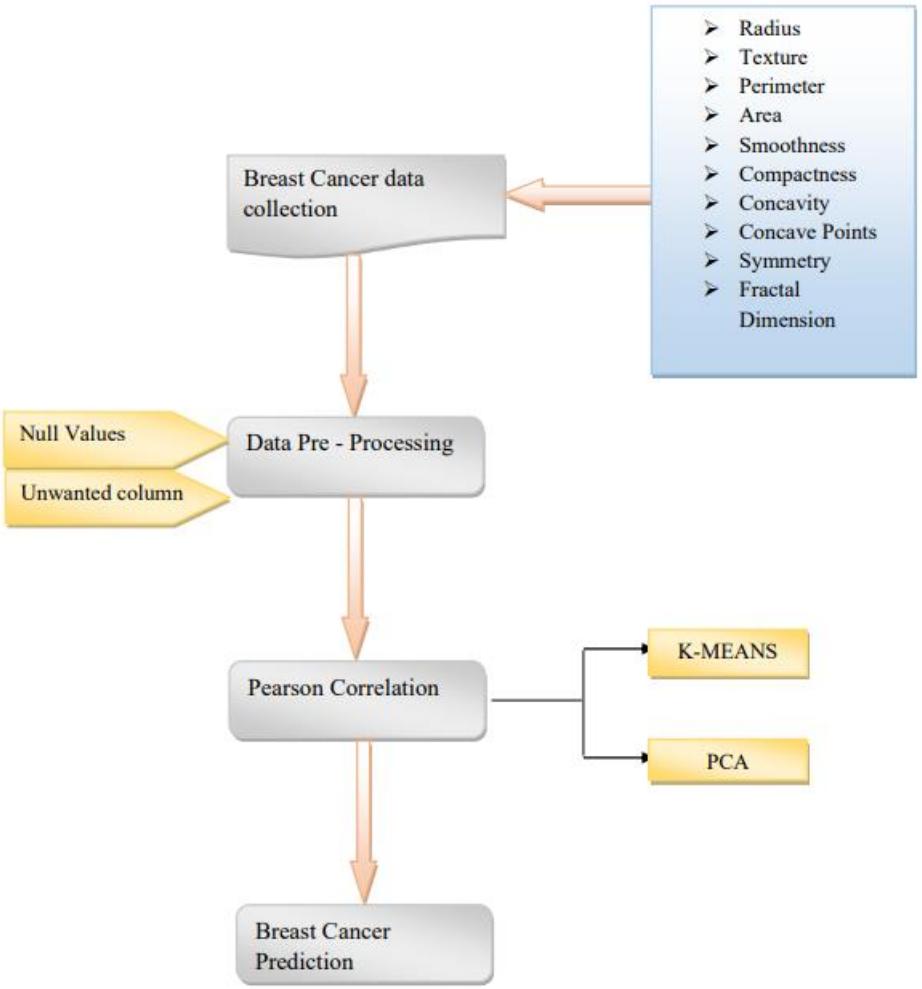

Figure 1: Outline of the Work

Methodology:

Data pre-processing is the process of transforming raw data into something that can be used by a machine learning model. In order to build a machine learning model, it is the first and most important stage. The obtained data may contain inconsistent data due to missing values. Pre-processing must be carried out to improve the model's efficiency in order to obtain the desired results on the final proposed systems. Additionally, it is the fundamental and important component of any classifications or predictions model. The pre-handling of data is also another name for this procedure. Both the method and interpretation are made simpler. To increase the precision of the prediction score, several techniques are used to remove duplicates or inconsistencies in the data.

Feature Selection using Pearson correlation:

The strength of the correlations between the various variables is evaluated by the Pearson correlation, often known as the Pearson R statistical test. As a result, it is always a good idea for the person conducting the analysis to calculate the correlation coefficient value in order to determine how strong the association between the two variables is. The link between the variables is positively correlated and both values decline or increase together if the value is in the positive range. The relationship between the variables is said to be negatively correlated if the value is in the negative range, in which case both values will move in the opposite direction.

K-MEANS Clustering:

K-Means Clustering is a type of unsupervised learning algorithm that is used to address clustering issues in data science or machine learning. We will discover what the K-means clustering method is, how it functions, and how to implement it in RStudio. It gives us the ability to divide the data into various groups and offers an efficient method for automatically identifying the groups in the unlabelled dataset without the need for any training. Each cluster has a centroid assigned to it because the algorithm is centroid-based. This algorithm's primary goal is to reduce the total distances between each data point and its corresponding clusters.

Principal Component Analysis:

An unsupervised learning approach called principal component analysis is used in machine learning to reduce dimensionality. With the use of orthogonal transformation, it is a statistical process that transforms the observations of correlated features into a set of linearly uncorrelated data. The Principal Components Analysis (PCA) is the name given to such newly converted features. One of the widely used tools for exploratory data analysis and predictive modelling is this one. It is a method for identifying significant patterns in the provided dataset by lowering the variances. An unsupervised learning approach called principal component analysis is used in machine learning to reduce dimensionality. It is a statistical procedure that transforms a set of linearly uncorrelated features from an observation of correlated data.

Performance Evaluation:

The performance evaluation metrics are used to calculate the performance of your trained machine learning models. This helps in finding how better your machine learning model can perform on a dataset that it has never seen before. Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Performance evaluation result is shown in **Table 1.**

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Table 1: Performance Evaluation Result

S. No	Classification	Accuracy
1	PCA	80%
2	K-MEANS	76%

Conclusion:

Breast cancer is the leading cause of cancer-related death in women worldwide. The implication of tumour heterogeneity in breast cancer progression and relapse has been well documented, and advocates for the development of tailored treatments. The implementation of personalised medicine requires the discovery of new biomarkers, based on omics data analysis along with clinical information. We have proposed a prediction model, which is specifically designed for prediction of Breast Cancer using Machine learning algorithms PCA and K- Means algorithms. The model predicts the type of tumour, the tumour can be benign (noncancerous) or malignant (cancerous). The model uses supervised learning which is a machine learning concept where we provide dependent and independent columns to machine. It uses classification technique

which predicts the type of tumour. The goal of the research is to classify the patients in Malignant and Benign types of tumours by classification techniques hence achieving higher accuracy.

Future Enhancement:

This study still holds a scope for further research and improvement including other machine learning algorithms to predict breast cancer or any other disease. This dataset only has two classes. In future other types and stages of cancer also be predicted. This is a binary classification prediction model, in future multiclass classification can be performed. In future better normalization technique, feature extraction, and feature engineering. Other ML algorithm can be taken and may increase the accuracy. In future, better performance evaluation metrics can be added for effective result.

References:

1. 'WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosisscreening/breast-cancer/en/ (accessed Feb. 18, 2020).
2. S. Nayak, D. Gope "Comparison of supervised learning algorithms for RF-based breast cancer detection,", 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona (2017)
3. A.H. Osman "An Enhanced Breast Cancer Diagnosis Scheme based on Two-StepSVM Technique," Int. J. Adv. Comput. Sci. Appl., 8 (2017), pp. 158-165
4. "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set."
5. Golatkar, A., Anand, D. &Sethi, A. Classification of breast cancer histology using deep learning. In International Conference Image Analysis and Recognition, 837–844 (Springer, 2018).
6. Albarqouni, S. et al. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE transactions on medical imaging 35, 1313–1321 (2016).
7. 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+.
8. M. M. Y. Al-Hashimi and X. J. Wang, "Breast cancer in Iraq, incidence trends from 2000-2009," Asian Pacific Journal of Cancer Prevention, vol. 15, no. 1, pp. 281– 286, 2014.
9. B. M. Gayathri, C. P. Sumathi, and T. Santhanam, "Breast cancer diagnosis using machine learning algorithms– A survey," International Journal of Distributed and Parallel Systems (IJDPS), vol. 4, no. 3, 2013.
10. I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers," Applied Intelligence, vol. 30, pp. 24–36, 2009.