Comparative Study On Breast Cancer Detection Using Machine Learning Algorithms

Utkarsh Gupta
Information Technology
Dehradun Institute of Technology
Dehradun, India
1000011799@dit.edu.in

Mohd. Faisal
Information Technology
Dehradun Institute of Technology
Dehradun, India
1000010865@dit.edu.in

Mrs. Moumita Ghosh
Assistant Professor
Dehradun Institute of Technology
Dehradun, India
moumita.ghosh@dituniversity.edu.in

Abstract*— This paper presents a machine learning algorithm-based system that detects breast cancers (BCs). BC detection uses logistic regression (LR), decision trees classification (DTC), naïve Bayes (NB), support vector machine (SVM), Kernel support vector machine (KSVM) and K-Nearest neighbor (KNN) classifiers, these machine learning (ML) based algorithms are trained to predicting BCs (malignant or benign) on BC Wisconsin dataset from the UCI repository, during which attribute clump thickness is used as evaluation class. The efficacy of these ML algorithms is evaluated in terms of accuracy and F-measure; decision trees classification outperformed the other classifiers and achieved 99.3% accuracy and 99% F1-Socre.

Keywords*— Breast Cancer, Machine Learning, Cancer Detection

I. INTRODUCTION

BC [1] It builds within the breasts causing some cells to grow abnormally, especially those with milk ducts. The second major cause of death for women is BC (after lung cancer). 246,660 of women's new cases of invasive BC are expected to be found in the US during 2016 and 40,450 of women's death is estimated. BC is a type of cancer that starts inside the breast. Cancer begins when cells begin to grow out of control. BC cells usually form a tumour that can often be recognize on an x-ray or felt as a lump. BC can spread when the cancer cells get into the blood or lymph system and are transferred to other parts of the body. The cause of BC includes changes and mutations in DNA. There are many different types of BC and most common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes and angiosarcoma are rare. There are many algorithms for classification of BC outcomes. The side effects of BC are – Fatigue, Headaches, Pain, and peripheral neuropathy, Bone loss and osteoporosis. There are many algorithms for classification and predicting BC results. The current paper provides a comparison between the performance of six classifiers: logistic regression (LR), decision trees classification (DTC), naïve Bayes (NB), support vector machine (SVM), Kernel support vector machine (KSVM) and K-Nearest neighbor (KNN) which are among the foremost influential data mining algorithms. It can be diagnosed early during a screening examination by portable cancer diagnostic tool or through mammography. Cancerous breast tissues change with the development of the disease, which might be directly connected to cancer staging. The stage of BC (I–IV) describes how far a patient's cancer has proliferated. Statistical indicators like as tumour size, bare nuclei, mitoses etc are accustomed to determining stages. To prevent cancer from spreading, patients must undergo BC surgery, chemotherapy, radiotherapy and endocrine. The aim of this research is to spot and classify Malignant and Benign patients and intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and the way further Machine Learning algorithms will be used to characterize BC. We want to scale back the error rates with maximum accuracy in an efficient way.

The following paper is organized as: section II presents the recent related work, section III presents the implemented Machine Learning algorithms, section IV presents results and related discussions, and finally, conclusion in section V.

II. LITERATIRE REVIEW

In [2], comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis has been done in 2016 using algorithms like SVM, RF, NB and as a result SVM was best amongst with 97.1% accuracy. In [3], some ML algorithms were used like- SVM-Poly, SVM-RBF, KNN, PNN which was tested in 2010 and SVM-RBF came up with highest accuracy of 98.8%. In [4], prediction of breast cancer was done in 2017 using SVM and KNN as a result SVM was more accurate than KNN with 98.5% accuracy. In [5], performance was evaluated of ML techniques in the classification of Wisconsin Breast cancer in the year 2018 using algorithms like LR, CUBIC, QUADRATIC and QUADRATIC was most accurate with an accuracy of 98.1%. In [6], ensemble learning methods are used to predict BCs such AdaBoost, Random Forest, and XGBoost, their results write down that random forest achieves 97% accuracy. In [7], Comparative study was done of ML algorithms for breast cancer prediction in 2020 using algorithms like LR and DTC and as a result DTC was more accurate than LR with 95.1% accuracy. In [8], mortality aspect of BCs in India is investigated with respect to many risk parameters such as the characteristics of demography, lifestyle, water intake, etc., and an ensemble algorithm (Bagoost) is used to predict BCs for Malwa women in India and achieved an accuracy of 98.21%. In [9], an ensemble algorithm that combines multiple

classification methods to classify benign and malignant tumor on UC Irvine ML repository, and their stacked ensemble classifier achieved an accuracy of 97.20%.

III. METHODOLOGY

With ML, computer systems can learn and behave like humans, gradually improving performance on specific tasks. ML algorithms are used to address many applications such as classification, feature selection, and clustering. Learning the function "f" to associate each attribute set "x" with one of the specified class labels "y" is a matter of classification. The classification model is another name for the objective function. A classification algorithm (also known as a classifier) is a method of creating a classification model from a dataset. Each method uses a learning mechanism to name the best model for the training attributes and associated class labels. Such a learning mechanism must match both the input data and the correct corresponding class specification in the test dataset. Therefore, the goal of learning is generalization. H. Processing of any future dataset. Regarding the application of the ML algorithm in the classification of cancer, the ML algorithm effectively distinguishes between benign and malignant and aids the doctor's diagnosis. In addition, for ML classifiers, it is important to find a subset of features. There are many ML techniques commonly used in BC classification, history monitoring, treatment, and prediction, such as neural networks and traditional neural networks. The ML algorithm used in this study is briefly described below.:

A. K-Nearest Neighbour

K-nearest neighbour (KNN) algorithm is a form of supervised Machine Learning algorithms which may be used for both classification as well as regression predictive problems. However, it's mainly used for classification predictive problems in industry. The subsequent two properties would define KNN well –

Lazy learning algorithm − KNN could be a lazy learning algorithm because it doesn't have a specialized training state and uses all the information for training while classification.

Non-parametric learning algorithm − KNN is additionally a non-parametric learning algorithm because it doesn't deduce anything about the underlying data.

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of recent datapoints which further implies that the new datum is going to be assigned a value based on how closely it matches the points within the training data set. We can know how its working with the help of following steps –

Step 1 - For implementing any algorithm, we want dataset. So, during the primary step of KNN, we must load the training as well as test data.

Step 2 - Next, we need to decide the value of K i.e., the nearest data points. K are often any integer.

Step 3 - For each point in the test data do the following –

1. Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. Mostly we use Euclidean method to calculate distance.
2. Now, based on the distance value, sort them in ascending order.
3. Next, it will choose the highest K rows from the sorted array.
4. Now, it will assign a category to the test point based on most frequent class of these rows.

Step 4 - END.

B. Support Vector Machine

Use The Support vector machine is a supervised learning system and used for classification and regression problems. Support vector machine is extremely favored by many as it produces remarkable correctness with low computation power. It is widely used in classification problems. We've got three styles of learning supervised, unsupervised, and reinforcement learning. A support vector machine could be a selective classifier officially defined by separating the hyperplane.

With the provision of labelled training data, the algorithm releases best hyperplane that classified new examples or models. In two-dimensional space, this hyperplane may be a line dividing a plane into two parts where each class lies on either side. The purpose of the support vector machine algorithm is to detect a hyperplane in an N-dimensional space that individually classifies the data points.

Identify applicable funding agency here. If none, delete this text box.

C. Kernel Support Vector Machine

The SVM algorithms use a collection of mathematical functions that are interpret as the kernel. The function of kernel is to take data as input and transform it into the desired required form. Different SVM algorithms use distinct types of kernel functions. These functions may be several types. For instance, linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

Introduce Kernel functions for sequence data, graphs, text, images, as well as vectors. The foremost used type of kernel function is RBF. Because it has localized and finite response along the whole x-axis.

The kernel functions return the inner product between two points in an exceedingly suitable feature space. Thus, by defining a notion of resemblance, with little computational cost even in very high-dimensional spaces.

D. Logistic Regression

Logistic regression is one of the most popular algorithms used in Machine Learning, which falls under Supervised Learning technique. It's used for predicting

the variables that are categorical dependent by using a given set of independent variables. It provides possible values between 0 and 1 and is divided into the following three types:

- Binomial: there are only two values for the dependent variables, like 0 or 1.
- Multinomial: there are often 3 or more possible dependent variables which are unordered, like “triangle", "square", or "circle"
- Ordinal: there may be 3 or more possible dependent variables which are ordered, such as "slow", "Medium", or "fast".

It uses function like sigmoid activation to predict the odds of a binary event occurring. The activation function aka logistic function generates an "S" shaped curve which will be accustomed to convert any real-valued integer to a number between 0 and 1. The output of this activation function becomes 1 when the curve reaches positive infinity, and 0 when the curve reaches negative infinity. If the output of the sigmoid is larger than 0.5, the output is 1 or True, otherwise, the output becomes 0 or False.

E. Decision Tree

Decision trees are supervised learning techniques used to classify data. The use of decision trees aims to build models capable of predicting target classes represented by the observed training data, which is achieved by learning simple decision rules. In a decision tree, branches stand for observations about an element, and leaves are conclusions about the element's target value. A decision tree consists of root, inner, and leaf nodes. Each leaf node in the decision tree is assigned a class label. Attribute validation conditions on non-terminal nodes are used to distinguish items with different attributes. Here are the basic steps of a decision tree classifier:

1. Run the decision tree on the root node (e.g., S) that holds the entire data set.
2. Get the best properties from the specified data set.
3. Divide tree S into subsets that may hold candidate values for the best properties. These best attributes should build decision trees.
4. Create a new decision tree using a subset of the data set created in the earlier step.

   Continue until you get a leaf node from which the node can no longer be classified.

F. Naïve Bayes

Naive Bayes - A probabilistic classifier based on Bayes' rule given evidence E and hypothesis:

$$P(H|E) = P(H)\, P(E|H) / P(E) \qquad (1)$$

where P(H|E) is the belief about the hypothesis after receiving E (called the posterior probability), P(H) is the belief about H before observing E (called the prior probability), and P(E|H) is the probability of H (L(H|E)) if E measures how well H explains E and P(E) is a regularization constant that guarantees P(H|E) = 1 for all hypotheses.

IV. RESULTS AND DISCUSSIONS

A. Dataset

In this paper, the Wisconsin Diagnostic Breast Cancer (WDBC) is used, it has 570 instances: Benign: 357, Malignant: 212), two classes: 37.19% malignant and 62.63% benign) and 32 integer-valued attributes (see Fig. 2). The five classifiers are trained to predicting cancers in the BC Wisconsin dataset, in which attribute clump thickness is used as evaluation class, in our experiments, ten attributes are taken from the dataset, these attributes shown in Table 1 noting that class comes with 2 and 4 numeric values to represent benign and malignant tumors respectively, however, their scale varies from 1 to 10.

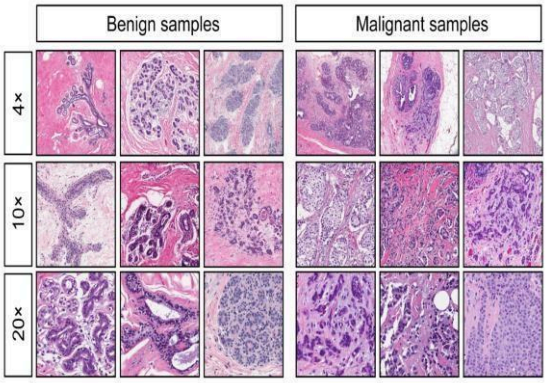

Fig. 2 Benign and malignant Samples for BCs.

B. Performance Measures

F-measure [25] is the weighted mean of precision (Pr) and recall (Re), where Pr is the ratio of true positives over all positives, on the other hand, re is the ration of true positives over all samples that must be identified as positive. Accuracy [25] is-

Number of correct predictions / total number of predictions.

TABLE I. ATTRIBUTES RETRIEVED BY CLASSIFIERS

Clump thickness
Uniformity of cell size
Uniformity of cell shape
Marginal adhesion
Single epithelial cell size
Bare nuclei
Bland chromatin
Normal nucleoli

Mitoses
Class

C. Results

In this paper, experiment runs are executed on an Intel Pentium (R) HP Pavilion notebook with 2.60 GHz i3- CPU and 4 GB RAM, running 64-bit Windows 10, the classification algorithms are implemented using scikit-learn ML library. Accuracy and f-measure values of the six classifiers are shown in Table II. It is obvious that decision tree classification achieved highest classification results, even slightly better than SVM and KSVM algorithms.

TABLE II. RESULTS ACHIEVED BY CLASSIFIERS

Classifier Type	**Performance**	
	Accuracy	***F-measure***
LR	97.9%	97.1%
KNN	97.3%	96.2%
SVM	98.8%	98.4%
KSVM	98.8%	98.4%
NB	96.4%	95.1%
DTC	99.3%	99%

D. CONCLUSION

BCs are among the common malignant tumors of many women all over the globe, and it is most spread in elderly women, recently, it is also spread in youngers. As a result, it is always suggested to do more research that may help in detecting BCs early, moreover, it is obvious that many of the characteristics of BC are still to be explored. In this paper, Six ML algorithm are used to detect BCs on the BCs Wisconsin dataset, results of Decision trees are the best as it achieved highest F-measure and accuracy scores, moreover, Decision trees achieved the testing process in less than one second, this indicates that Decision trees with this 99.3% accuracy will be able to help doctors to automatically identify whether their patients have malignant or benign tumors with high confidence.

REFERENCES

[1] A. Alzu'bi, H. Najadat, W. Doulat, O. Al-Shari, and L. Zhou, "Predicting the recurrence of breast cancer using machine learning algorithms," Multimedia Tools and Applications, vol. 80, no. 9, pp. 13787–13800, 2021.

[2] S. Laghmati, B. Cherradi, A. Tmiri, O. Daanouni, and S. Hamida, "Classification of patients with breast cancer using neighbourhood component analysis and supervised machine learning techniques," in 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet). IEEE, 2020, pp. 1–6.

[3] P. Gupta and S. Garg, "Breast cancer prediction using varying parameters of machine learning models," Procedia Computer Science, vol. 171, pp. 593–601, 2020.

[4] P. Gupta and S. Garg, "Breast cancer prediction using varying parameters of machine learning models," Procedia Computer Science, vol. 171, pp. 593–601, 2020.

[5] G. Battineni, N. Chintalapudi, and F. Amenta, "Performance analysis of different machine learning algorithms in breast cancer predictions," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 6, no. 23, 2020.

[6] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in 2016 5th international conference on electronic devices, systems and applications (ICEDSA). IEEE, 2016, pp. 1–4.

[7] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," Clinical Epidemiology and Global Health, vol. 7, no. 3, pp. 293–299, 2019.

[8] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT). IEEE, 2018, pp. 1–4.

[9] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," in Healthcare, vol. 8, no. 2. Multidisciplinary Digital Publishing Institute, 2020, p. 111.