Analysis and Prediction of Breast Cancer using Machine Learning Techniques

Shakkeera L, Rahul Raj Pandey, Rahul Bhardwaj, Sidhya Virya Singh, Siddhartha S. Mukherjee

Abstract - Rapid multiplication of cells in the human body leads to cancer. It is the foremost cause of death due to cancer in females, after lung cancer. As the breast cancer is one of the recurrent kinds of cancer, diagnosis of breast cancer recurring is extremelyessential to increase the survival rate of patient suffering from it. Although cancer is avertible and also treatable in primary/early stages yet a vast number of patients are diagnosed with cancer when it is very late. Almost 8% of females are detected with breast cancer. Its characteristics are mutation of genes, constant pain, changes in the size and redness of skin texture of breasts. With the development of technology and machine learning techniques, cancer diagnosis and detection accuracy has greatly improved. This paper presents an outline of evolved machine learning techniques in this medical field by applying machine learning algorithms on breast cancer dataset like Logistic regression, Random Forest, Decision Trees (DT) etc.

Keywords: Classifier, Classification accuracy, Machine Learning, Prediction.

I. INTRODUCTION

Classifying breast cancer can leaddiagnosticians to find anorganized and unbiased prognostic. Normally binary classification that is benign and malignant can be seen frequently. A Classification algorithm, like Decision Treeare broadly used in the world of medicines to categorize theinformation for diagnosis. The process of Feature Selection helps in increasing the overall accuracy of the classifying model since it removesinappropriate attributes.

There are many established ways of detecting and diagnosing cancer but they majorly depend on trained physicians, with the help of medical imaging, wenotice certain indications that commonlyseem to be visible in the advance stages of cancerous cells. Machine learning providesseveral probabilistic and statistical methods that let intelligent systems to learn from past knowledges which repeats to notice and recognize patterns from a dataset. But the limitations are that either they use faulty dataset or they don't wrangle the data correctly or select features properly. The aim of this very project is to guarantee that the benign and malignant classes of breast cancer are predicted and grouped accurately.

Revised Manuscript Received on November 25, 2020.
* Correspondence Author

Shakkeera L*, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India
Rahul Raj Pandey, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India
Rahul Bhardwaj, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India
Sidhya Virya Singh, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India
Siddhartha S. Mukherjee, School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh, India

A. Classification

Theprocess in which objects are recognized, differentiated and understood and are grouped together is called classification.

According*[J. Han et al.,2000]*, classification is a job of data collectionin which objects are segregated into numerouscategories or classes or groups which are pre-defined on the basis features of objects. To do this a data-set is provided as input which is known as training data-set, it consists of numerous specimens each having a number of features. The training set then is used to build a ML system in such a way that new data which is not from the training data-set or from any other source of data can be classified correctly. There are numerous methods for classification, some of them are:

- Decision Tree algorithms
- Support vector machines
- Associative classification
- Bayesian algorithms
- Distance based methods *(like KNN)* and many more.

B. Machine Learning

The learning in which machine can learn by its own by gaining knowledge without being explicitly programmed is called machine learning. It is a subset and very important part of artificial intelligence. In recent days, in fast growing technological world, Machine Learning (ML) becomes popular among many of researchers, industries and government sectors. It is the contrivance of automating and cultivating the learning process of systems/computers based on their capabilities/understanding capabilities without being actually programmed by the human being's assistance. The ML process starts with feeding the raw data as an input to train the machines by building machine learning models or learning tasks. Different ML algorithms or methods uses the input data and targeted output to build ML models or ML tasks. The ML algorithms are based on type of data and kind of task which are trying to automate the machine learning models.

i) Issues in ML

- Nonexistence of proper data.
- Lack of skilled possessions and assets (both man and machines).
- Understanding which processes need mechanization.
- Inadequate availability of infrastructure and implementation.

C. Pre-Processing

Pre-processing refers to the methodsthat we apply our data to transform before passing it in to the algorithm.

The process of preparing the raw data and making it appropriate for a machine learning model is called data pre-processing. It is the initial and key step while creating a machine learning model. It is not necessary that you alwayscome across a clean and formatted data while doing a machine learning project. And it is required to clean data and put in anorganized way while doing any action on our data. So, we use data pre-processing task for this.

A real-world data normally contains missing values, null values/noisesor maybe the data is in an unusable format which cannot be used for training machine learning models and that is why data pre-processing is needed for cleaning the data and making it appropriate for a ML model. Thus, increasing the efficiency and accuracy of a model.

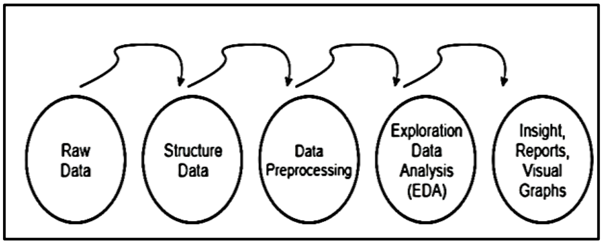

Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

Fig.1Pre-processing

D. Feature Extraction

The process in which we reduce the shape and size of our initial dataset which is raw in nature to a more small and handy set of data is called feature extraction. This process results in usage of computation resources and power unlike large dataset containing numerous attributes. This process also combines many other features into one which increases the quality of our data and also reduces large chunks of data that need to be computed.It also helps in reducing noise/unwanted data for our model building resulting in increase of speed of learning our model.

E. Classification

The process to identify which set of categories are belongs to one group based on the relevant observations is called as classification method.It comes under supervised machine learning. It is very useful when our output is of finite nature and has discrete values. It is of two types-

- **Binary classification:** It predicts one out of two possible outcomes or groups(e.g. is the person eligible tovote or not?)
- **Multi-class classification**: It predicts one out of many possible outcomes or groups (e.g. is this a car, bike or cycle?)

F. Prediction

The method which use data mining methods and probability techniques to get the predicted outcomes from a ML model is called as predictive modeling in ML. Each model is built up by the number of predictors that are highly favourable to determine future decisions. Now a days, predictive modeling methods are used and developed in many fields of business and medical care.

Moreover, Predictive Modeling employs different algorithms and analytics or statistics to estimate the probability of an event using detection theory and largely employed in the field of ML, and AI.

II. RELATED WORK

Ali Al Bataineh in his study on a *Comparative Analysis of Non-Linear ML algorithms* showed comparative study of prediction between five models built on algorithms like Multilayer Perceptron (MLP), KNN, Navie Bayes etc on WDBC dataset. His worked showed that MLP performed best with an accuracy of 99.12% on training dataset.

Another significant work in this field is done by Jian-HuangLai, Chee-Keong Kwoh and their team developed a new ensemble clustering technique and termed is as MDEC. They showed that large random population of varied metrics can be beneficial for clustering using ensemble. The framework they built was combination of three clustering algorithms using consensus functions. They conducted experiments over more than30 real world data-set of high dimensionality. Pragya Chauhan and Amit Swami in their work of prediction of breast cancer using algorithm which wasbased onensemble approach and genetic algorithmsuggested a model where they found that the cancer prediction is a field which is open for research. ML algorithms were used for detection and prognosis of cancer. Linear model, naive bayes, CARTetc. methods were used for prediction. A work was done in prediction of cancer and lymph nodes with tumour marker by Hsiao-Lin Hwa and his team. They used serum samples of femalepatients with and without breast cancer. Then they used logistic regression of univariate and multivariate for evaluation. According to their serum level of TPS had the best value for prediction, with combination of various medical parameter or biomarkers and integrating them with logistic regression raised the model sensitivity and accuracy significantly. A similar research was done in this field by Hoang Pham and David Pham, they proposed a model which was dependent upon logistic and among nine selected biomarkers for better prediction. They also compared their model with several other machine learning model using various training data sets.

III. EXISTING SYSTEM

Many existing models use the concept of mammography image dataset to build models and do predictions. A mammogram is an x-ray picture of the breast. It is also useful if you have a lump or other major sign of cancer. Screening mammography is the type of mammogram that checks patients when they don’t have any symptoms of the breast cancer. It helped in reducing the number of deaths from breast cancer among women of age from 40 to 70. But fewtumors cannot be spottedby a mammogram due to the position of the cancer or the thickness of the chest tissue. About 25 % of cancers in female agingfrom 40 to 49 are not detectable by a screening mammogram, compared to 10% in women older than 50. And that’s the major drawback of using dataset built on mammography image.

There are some other models that do not use image dataset but uses the faulty or improper data to make predictions which often results to be wrong.

A. Issues in existing system

Use of mammography image dataset is surely giving good predictions but not all cancers are identified by it. Other which uses other dataset aren't able to perform well because either the dataset is faulty or they are using less number of attributes in training which are not so important factors in real world in determining BC.

B. Drawbacks in existing system

The major limitations of existing systems are:

- Use of faulty dataset.
- Using less and unimportant features to determine output.
- Tuning model in such a way that

IV. PROPOSED SYSTEM

The proposed system develops a classification and predictive model that will perform accurate classification grouping and prediction of Breast Cancer. This proposed approach will focus on prediction of the disease in the early stage by taking 10 real world value parameters for every cancer cell nucleus. A combination of learning algorithms of classification and ensemble learning are used to implement and develop the proposed model.

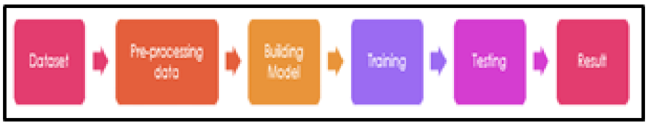


Fig.2Workflow of proposed methodology

A. Advantages of Proposed System

- The accuracy will be more due to use proper dataset consisting of parameters of cell nucleus.
- Proper use of features and normalization of the data.
- Free from unnecessary tunning of model.

B. Proposed System Design

The Fig.3describes the proposed system design for predictive modelling of BC. The proposed system contemplates mechanized diagnosis of Breast Cancer. We will be using different ML classification algorithms like *Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, K- Nearest Neighbour (KNN).* The dataset is retrieved from University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Sample of cancer appears periodically because Dr. Wolberg recorded his all clinical cases.Dataset is accessible on *UCI Machine Learning Repository* as well.

Attributes are calculated from a digitized image of a fine needle aspirate (FNA) of a breast mass which describe characteristics of the cell nuclei present in the image. The features in the dataset describes the properties of the cell nuclei found in the image. The 10 features recorded are divided in to two classes namely Benign and Malignant. There are 569 observations and 32 observations including patient ID, diagnosis and 30 important real-valued features that affects the predictions. The 10 features recorded are –

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Concave points
- Symmetry
- Fractal Dimension

On computing mean, standard error, and "worst" or largest (mean of the three largest values) of these features for each image resulted in 30 features. So, we will be using these features, scale the features and use them to train and test our models to get accurate the predictions and we will compare their performances also

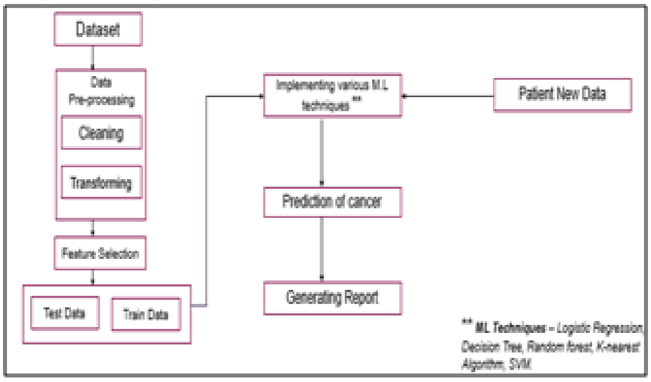


Fig.3Proposed System Design

V. EXPERIMENTAL RESULTS

As said earlier, to increase the accuracy of predictions we will pre-process it. The data *(Wisconsin dataset)* is collected from UCI Machine Learning Repository. Now our collected data is raw and we need to pre-process it to transform numeric values to nominal. Then important features are selected for training and testing and our criteria for feature selection can be anything like choosing only mean value attributes or choosing attributes with high correlation with the predictions. Our next step is to use various classification algorithms to classify our result in to two classes namely: - Benign and Malignant. And finally, we estimate and discuss performance of these models under various cases.

In the experiment we will calculate accuracy score of each model. It compares the true value of diagnosis with the corresponding predicted value of diagnosis for test dataset. After that we will see the classification report of the model which will have highest accuracy.

Now our dataset contains 357 Benign and 212 Malignant values. On dividing our dataset in to 75% training and 25% testing we are having 267 benign and 159 malignant in our training set and 90 benign and 53 malignant in our test data.

A. Correlation Matrix

Correlation is used often used in mathematics to expresses the extent of dependency of two variables. It is also used in data analysis and machine learning to see the dependency among the attributes and with the label.

The following graphs show the correlation between different features in form of a heatmap. Higher the percentage higher the correlation.

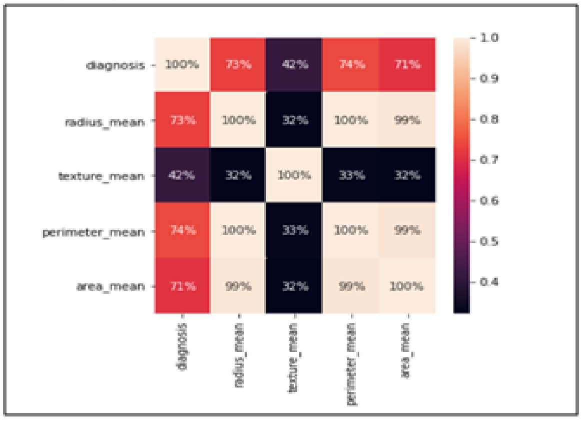


Fig.4 Heatmap of the features

<u>CASE 1</u>.Using all the provided features: -

In this case we used all 32 features in the dataset to train our model, keeping in mind that even a small criterion can be a reason of growth of breast cancer. We trained Decision tree, Random forest, Logistic regression, SVM classifier and XGBoost model for predictions and accuracy of the models are the following: -

Table-I: Showing Accuracy of Models

Accuracy	
Logistic Regression	0.958
Decision Tree	0.944
Random Forest Classifier	0.979
SVM Classifier	0.965
XGBoost	0.965

Clearly, we can see that Random Forest is performing best out of these without specifying its depth. After this SVM and XGBoost are having same accuracy, then Logistic Regression and then at last Decision Tree.

Now the classification report of Random Forest is shown in the following *Table II*.A **classification report** helps us to know the quality of predictions from our model.

Table-II: Classification Report of Random Forest Model

	Precision	Recall	F1-score	Support
Benign	0.98876	0.97778	0.98324	90
Malignant	0.96296	0.98113	0.97196	53
Accuracy			0.97902	143
Macro average	0.97586	0.97945	0.9776	143
Weighted average	0.9792	0.97902	0.97906	143

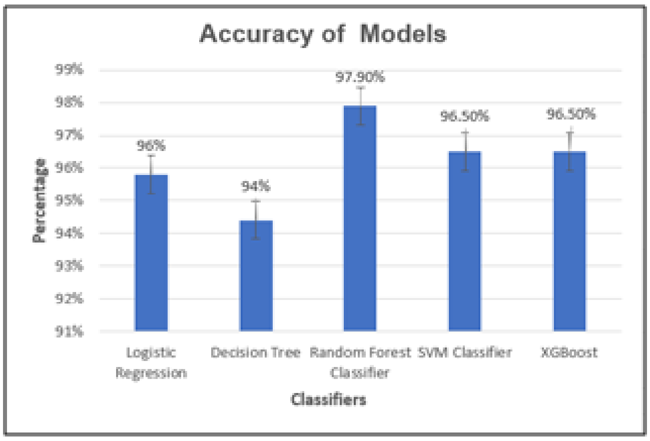


Fig. 5 Graph showing accuracy of models

<u>CASE 2</u>. Selecting few features: -

So now we will be doing feature extraction and take only those features which are having high positive correlation with our diagnosis for training all the models. We will again compare the accuracy of different models which is given below in *Table III.*

The features (along with their correlation) selected are –

1. Concave points worst -0.793566,
2. Perimeter worst - 0.782914
3. Concave points mean - 0.776614
4. Radius worst - 0.776454
5. Perimeter mean - 0.742636
6. Area worst - 0.733825

Table- III: Showing Accuracy of Models after feature selection

Accuracy	
Logistic Regression	0.930
Decision Tree	0.937
Random Forest Classifier	0.930
SVM Classifier	0.930
XGBoost	0.965

As you can see now random forest performed badly compared to the earlier case and also logistic regression and random forest and svm classifier were having same accuracy. In this case also the depth of random forest was not set and we can see as we decreased the number of features its accuracy decreased.

Now as earlier we will see the classification report *Table IV*of the XGBoost classifier as it was the most accurate in this case.

Table- IV: Classification Report of XGBoost Classifier

	Precision	Recall	F1-score	Support
Benign	0.96703	0.97778	0.97778	90
Malignant	0.96154	0.9434	0.9434	53
Accuracy			0.96503	143
Macro average	0.96429	0.96059	0.96238	143
Weighted average	0.965	0.96503	0.96497	143

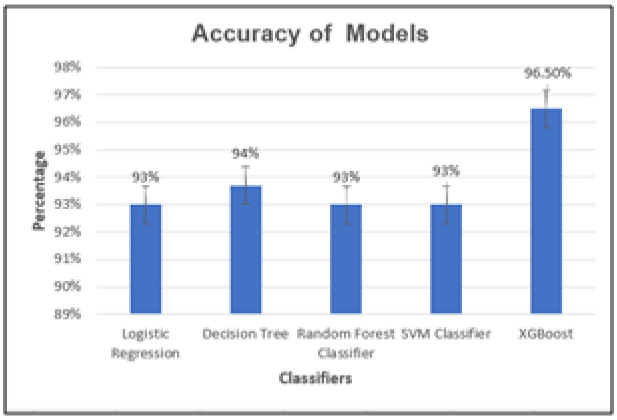


Fig.6 Graph showing accuracy of models

VI. CONCLUSION

We can observe that the accuracy achieved by the model are when they are not optimized and tuned. We can see when models are given every single feature they are predicting well and when we are giving them just 6 features, they are performing low. It is evident here that instead of training model with few features, limiting their depth in case of random forest and hyper tunning the model to reach near 98% accuracy is useless. Instead using all the features, getting98% accuracy and then optimizing our models to increase accuracy to 99 % or 100 % is beneficial.

So, early detection is very important and detection by invasive techniques makes predictions easier. There are many algorithms apart from them, in our case Random Forest exhibited the highest accuracy. Thus, the most accurate classifying model can be used to detect the cancer in early stages.

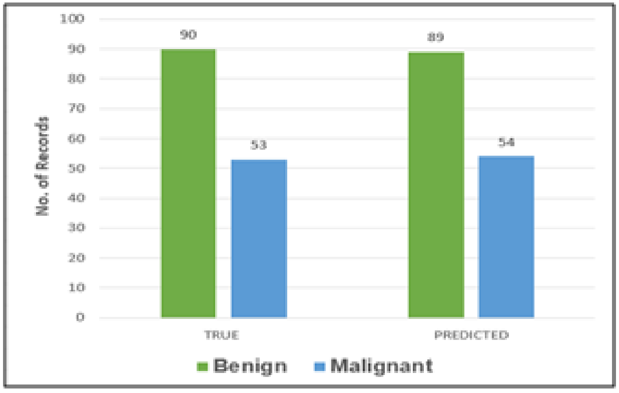


Fig. 7. Graph showing true and predicted numbers of Benign and Malignant.

You can see from graph that our model predicted 1 less value for benign and 1 value more of malignant and that is the only error in our Random Forest model.

REFERENCES

1. Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; DOI 26:1-24
2. B M Gayathri and C P Sumathi. An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer. International Journal of Computer Applications, 2016.DOI 10.5120/ijca2016911146
3. Houston, Andrea L. and Chen, et. al. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review 1999; DOI 13:437-466
4. Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques 2006 DOI 10.1186
5. K. Balachandran and R. Anitha, “Ensemble based optimal classification model for pre-diagnosis of lung cancer”,2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE (2013), DOI 10.1109/ICCCNT.2013.6726467.
6. M. Kumar, S. S. Tomar and B.Gaur, “Mining based Optimization for Breast Cancer Analysis: A Review”,International Journal of Computer Applications, vol. 19, no. 13,(2015).
7. Priyanka Jain & Santosh Kr. Vishwakarma (2016). Collaborative Analysis of Cancer Patient Data using Rapid Miner. International Journal of Computer Applications, 145, 8-13.
8. Priyanka Gupta & Prof. Shalini L(2018): Analysis of Machine Learning Techniques for Breast Cancer Prediction. International Journal Of Engineering And Computer Science *7*(05),ISSN:2319-7242
9. S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007.

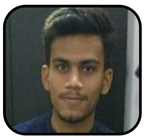
AUTHORS PROFILE

Dr Shakkeera L is working as a Senior Assistant Professor at VIT Bhopal University, Madhya Pradesh, India She has a teaching experience of 15 years and has published more than 35 research publications in refereed International/National Journals and International/National Conferences.
shakkeera.l@vitbhopal.ac.in

Rahul Raj Pandey is a second-year student in Computer Science and Engineering, at VIT Bhopal University, Madhya Pradesh, India - 466114
pandeyrahulraj99@gmail.com

Sidhya Virya Singhis a second-year student in Computer Science and Engineering, at VIT Bhopal University, Madhya Pradesh, India - 466114
sidhya.v.singh@gmail.com

Siddhartha S. Mukherjeeis a second-year student in Computer Science and Engineering, at VIT Bhopal University, Madhya Pradesh, India - 466114
siddharthmukherjee556@gmail.com



Rahul Bhardwajis a second-year student in Computer Science and Engineering, at VIT Bhopal University, Madhya Pradesh, India - 466114
rahulbhardwaj301201@gmail.com