Breast Cancer Diagnosis using Supervised Regression Methodologies

T Susmitha
Student, Dept. of Computer science and engineering,
Vasireddy Venkatadri Institute of Technology, Guntur, India

P S L P Greeshma
Student, Dept. of Computer science and engineering,
Vasireddy Venkatadri Institute of Technology, Guntur, India

S R C Hemanth
Student, Dept. of Computer science and engineering,
Vasireddy Venkatadri Institute of Technology, Guntur, India

T Revathi
Student, Dept. of Computer science and engineering,
Vasireddy Venkatadri Institute of Technology, Guntur, India

Dr. SK Khaja Mohiddin
Asso.Prof, Dept. of Computer science and engineering,
Vasireddy Venkatadri Institute of Technology, Guntur, India

Abstract- Breast cancer is the second leading cause of cancer death in women. Early diagnosis can increase the recovery rate, reducing the death rate in patients. There are two kinds of breast lumps, Benign (non-cancerous) and Malignant (cancerous). We used Breast Cancer Wisconsin (Diagnostic) medical data sets from the UCI machine learning repository. Here, we use Logistic regression (Binary classification) and Softmax regression (Multi-class classification) techniques to diagnose breast cancer. Simulation and result proved that the proposed approach gives better results in terms of different parameters. The prediction results obtained by the proposed approach were very promising (99.60% true accuracy). The Aim of the paper is to implement Classification techniques to diagnose and understand the sensitivity and specificity of cases. It is also comparable with the existing machine learning and soft computing approaches present in the related literature.

Keywords – Area under Curve (AUC), Benign, Breast cancer, Malignant, Logistic regression, Receiver Operating Characteristics (ROC) Curve, Softmax regression.

I. INTRODUCTION

Breast cancer is the most common type of cancer in women after skin cancer. According to WHO there were 2.3 million women diagnosed with breast cancer and 6,85,000 deaths globally in the year 2020 [1]. Early screening, correct detection and diagnosis of Breast Cancer are very important to improve the survival rates significantly and to increase chances of recovery. Computer-aided intelligent and automated diagnosis systems, developed by machine learning approaches, are important means in the analysis of breast cancer, and it supports medical experts (oncologists) in the medical decision-making process. Breast cancer is not just limited to women; men can also suffer with this. But the percentage is very less compared to women. There are certain factors to increase the risk of breast cancer like increasing age, obesity, alcohol, family history of breast cancer, history of radiation exposure etc.

1.1 Symptoms of breast cancer

The common symptoms of breast cancer are

- a breast lump or thickening in breast
- alteration in size, shape or appearance of breast

- abnormal nipple discharge
- unusual pain in any part of breast

1.2 Types of Breast Lumps

Breast lumps can be classified into two types. They are Benign and Malignant

1.2.1. Benign Tumors

Benign tumors are non-harmful growths within the body. Unlike cancerous tumors, they don't spread to other parts of the body.

Benign tumors could form at any spot in the body. If you discover a lump or mass in your body which could be felt from the skin, you'd possibly immediately assume it's cancerous. As an example, women who find lumps in their breasts during self-examinations are often alarmed. However, most breast growths are benign. As a matter of fact, many growths throughout the body are benign.

Benign grows are the most common ones, with 9 out of 10 women Trusted Source showing benign breast tissue changes.

1.2.2. Malignant Tumors

Malignant tumors have cells that grow uncontrollably and spread locally and/ to distant sites. Malignant tumors are cancerous as they invade other healthy organs. They spread to distant sites via the bloodstream or the systemalymphaticum. This spread is termed metastasis. Metastasis can occur anywhere within the body and most typically is found within the liver, lungs, brain, breasts and bone.

Malignant tumors can spread rapidly and need treatment to avoid spread. If they're caught early, treatment is probably going to be surgery with possible chemotherapy or radiotherapy. If the cancer has spread, the treatment is probably going to be systemic, like chemotherapy or immunotherapy.

Breast cancer diagnosis is a prime example of Classification problem. Here, positive indicates the lumps are of malignant type, negative indicates that lumps are of benign type.

1.3 Classification in Machine learning

Classification is a technique of mapping the inputs into two class labels, either Benign or Malignant. Other examples of classification are predicting and email is spam or not, sentimental analysis, dog breed detection etc. Classification can be divided into 4 parts, they are

- Binary Classification
- Multi-class Classification
- Multi-label Classification
- Imbalanced Classification

In Binary Classification we have only 2 labels, Ex: email spam or not. They are more than two classes in Multi-class Classification. In Multi-label Classification we have one or more class in every input. Ex: A picture containing various elements like a man, dog, car etc. In Imbalance Classification, numbers of inputs are not equally spread. Our Breast Cancer falls under Binary Classification.

II. RELATED WORK

Several Machine learning (ML) and soft computing approaches have been applied in the analysis and classification of the data acquired from a digitized image of a fine needle aspirate (FNA) of a breast mass. (E.g. Breast Cancer Wisconsin (Diagnostic) Data Set). These approaches include SVMs Iranpour, M, Almassi, S, and Analoui, M [2], breast cancer detection from FNA using SVM and RBF classifier. Kadam, V.J., Jadhav, S.M. and Vijayakumar, K. Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Auto encoders and Softmax Regression got 98.60% true accuracy [3]. N. Khuriwal and N.Mishra had developed a model to diagnose breast cancer using Convolutional Neural Network in which they have used Wisconsin data set from UCI Machine learning

repository. They gained 99.67% accuracy using this model [4]. Mohamad Mahmoud Al Rahhal had used Convolution Neural Network's algorithm for Classification over histogram images and got 86.6% accuracy [5]. Teresa Anayjio proposed deep learning approach to classify haematoxylin and eosin stained breast images using Neural Networks and gained 83.3% accuracy [6]. R.D.Ghongade proposed the RF random & RF.ELM model for classification of breast lumps, for this they used digital mammography and stored 98% accuracy [7].

III. REVIEW OF LITERATURE

Machine learning (ML) is that the study of computer algorithms that improve automatically through experience by executing different tasks. ML is that the subset of computer science. The algorithms of machine learning build a mathematical model supported data provided, referred to as "training data", so as to create predictions or decisions without being explicitly programmed, to try to so ML algorithms are employed in a kind of applications, like email filtering and computer vision, where it's difficult or infeasible to develop conventional algorithms to perform the needed universe tasks.

Machine learning is expounded to computational statistics, which focuses on making predictions using machines (computers). The mathematical optimization study gives methods, theory and application domains to the sector of machine learning. Data processing may be a field of study associated with ml, specializing in exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is called as predictive analytics.

3.1 Regression Analysis

Regression analysis may be a predictive modelling technique that analyzes the relation between the target or variable and variable quantity in an exceedingly dataset. The various sorts of multivariate analysis techniques get used when the target and independent variables show a linear or non-linear relationship between one another, and also the target variable contains continuous values. The regression technique gets used mainly to work out the predictor strength, forecast trend, statistic, and just in case of cause & effect relation.

Regression analysis is that the primary technique to resolve the regression problems in machine learning using data modelling. It involves determining the most effective fit line, which may be a line that passes through all the information points in such some way that distance of the road from information is minimized.

There are different types of regression analysis techniques. We use different types of techniques depending upon the number of factors. Let us briefly know about some of the regression analysis techniques.

3.1.1. Linear Regression

Linear regression is one in every of the foremost basic styles of regression in machine learning. The linear regression model consists of a variable and a variable quantity related linearly to every other. Just in case the info involves over one experimental variable, then linear regression is named multiple simple regression models.

3.1.2. Logistic Regression

Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete. Example: 0 or 1, true or false, etc. This means the target variable can have only two values, and a sigmoid curve denotes the relation between the target variable and the independent variable.

3.1.3 Polynomial Regression

Polynomial Regression is another one in all the kinds of multivariate analysis techniques in machine learning, which is that the same as Multiple rectilinear regression with a touch modification. In Polynomial Regression, the connection between independent and dependent variables, that's X and Y, is denoted by the n-th degree.

It is a linear model as an estimator. Least Mean Squared Method is employed in Polynomial Regression also.

3.1.4 Bayesian Linear Regression

Bayesian Regression is one of the types of regression in machine learning that uses the Bayes theorem to find out the value of regression coefficients. In this method of regression, the posterior distribution of the features is determined instead of finding the least-squares. Bayesian Linear Regression is like both Linear Regression and Ridge Regression but is more stable than the simple Linear Regression.

3.1.5. Softmax Regression

The Softmax regression could be a type of logistic regression that normalizes an input value into a vector of values that follows a probability distribution whose total sums up to 1. The output values are between the range [0, 1] which is good because we are able to avoid binary classification and accommodate as many classes or dimensions in our neural network model. This can be why softmax is usually observed as a multinomial logistic regression.

IV. PROBLEM STATEMENT

It is a simple classification problem which predicts the tumor is Benign or Malignant, and we can also change the prediction according to the perspective of doctor or patient by tweaking the threshold values in the implementation.

4.1. Data set description

Here we have used Breast Cancer Wisconsin (Diagnostic) Dataset [8]. It can be downloaded from the UCI Machine Learning repository. This dataset consists of 569 different instances, each of 32 attributes or features. Among the 32 attributes, one is ID and one is classification. These remaining 30 attributes are features extracted from the cancer cells. Hence, we have used 30 parameters for training and testing. This is how the Wisconsin diagnostic dataset is prepared. Medical college of Wisconsin, Milwaukee recorded the data from mammography examinations conducted from April 5, 1999, to February 9, 2004. This database includes 48,744 examinations among which only 477 are malignant and the remaining are non-malignant. The examinations are conducted over 18,720 patients of various age groups. Various observations made from these are collected and organized as the dataset.

Of all the 569 total instances each of them can be classified either to Benign or Malignant. In dataset Benign (Non-cancerous) is represented with 0 and Malignant (Cancerous) is represented with 1. In this dataset 357 observations are of benign type and 212 are of malignant type.

Mean value	**Variance**	**Maximum value**
mean radius	radius error	worst radius
mean texture	texture error	worst texture
mean perimeter	perimeter error	worst perimeter
mean area	area error	worst area
mean smoothness	smoothness error	worst error
mean compactness	compactness error	worst compactness

mean concavity	concavity error	worst concavity
mean concave points	concave points error	worst concave error
mean symmetry	symmetry error	worst symmetry
mean fractal fractal	Fractal dimension error	worst fractal dimension

Table 1: Attributes of the Wisconsin Breast Cancer Data Set.

4.2. Sensitivity vs Specificity

In order to understand sensitivity vs specificity, we must understand the four important terms. They are true positives, true negatives, false positives, and false negatives.

True positive means something we predicted as true is actually happened to be true. Here in our problem

Actual: Malignant

Predicted: Malignant

True negative means something we predicted as true is actually happened to be false. Here in our problem

Actual: Benign

Predicted: Benign

False positive means something you incorrectly predicted as true but it is not. Here in our problem

Actual: Benign

Predicted: Malignant

False negative means something you incorrectly predicted something as false but it is true. Here in our problem

Actual: Benign

Predicted: Malignant

So, we can understand that our prediction can be one among the above four cases, which can also be seen in following figure.

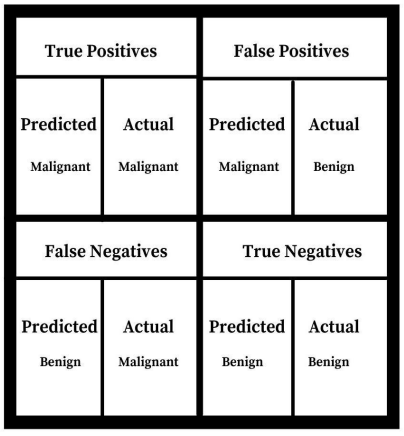

Fig 1: All possible predictions of model

Now, we can see two views in this problem based on the increase and the decrease of threshold values.

4.2.1. Patient view (sensitivity)

A patient always wants the correct information regarding his/her health. Hence, the need to

- Maximize the true positives
- Minimize the false positives

Hence, we need to decrease the threshold here

4.2.2. Hospital view (specificity)

In the other case, hospital wants the positive rate to their patient. Such that they will go for further costly treatment. Hence, the need to

- Minimize the true positives
- Maximize the false positives

Hence, we need to increase the threshold here.

This conflict between patient and hospital is *sensitivity vs specificity*.

V. FLOW DIAGRAM OF PROPOSED WORK

The fig 2 shows the operation of flow diagram or proposed work. Download the Wisconsin breast cancer (diagnostic) dataset, which is publicly available. The second step is input the data and dividing the dataset for training and testing the data. Scaling data, followed by splitting data into training and testing data. Finally, applying regression algorithms to plot ROC curve and observing AUC values for both training and testing data.

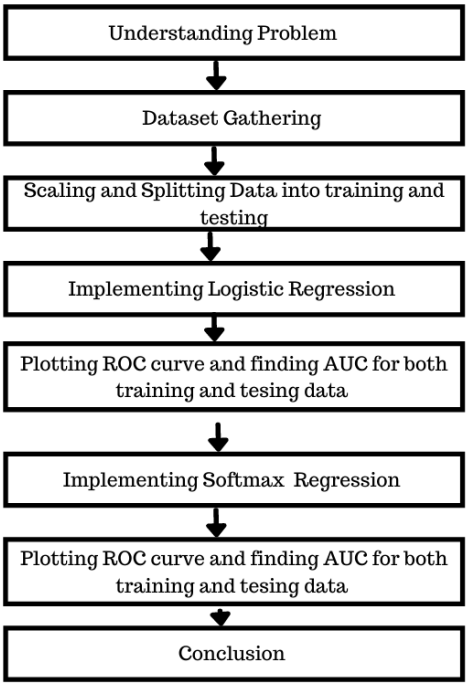

Fig 2: Flow diagram of proposed model

VI. IMPLEMENTATION

The Implementation sequence is shown in fig 2. The dataset downloaded is converted into required format after removing the unnecessary parameters. Then scaling data is followed by splitting data in two parts in 80% and 20% for training and testing respectively. Then we apply regression algorithms, i.e. Logistic and Softmax for training and testing data. So plot ROC curve and observe AUC.

6.1 Anaconda

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. Anaconda distribution comes with over 250 packages automatically installed, and over 7,500 additional open-source packages will be installed from PyPI additionally because the conda package and virtual environment manager. It also includes a GUI, Anaconda Navigator, as a graphical alternative to the command interface (CLI)

6.2 Python

Python could be a easy artificial language for beginners to find out that mixes the features of C and Java. It provides a sublime way of developing programs like C. Python offers classes and objects like Java. Python is open source software and available to download at no cost. Python features a dynamic type system and automatic memory management as its main features. Python supports multiple programming paradigms, including object oriented, functional and procedural, and contains a large and comprehensive standard library. Python interpreters are available for several operating systems like windows, Linux etc. In python Object Oriented programming and structured programming are fully supported, and lots of of its features support functional programming and aspect-oriented programming. Many other paradigms are supported through extensions, including design by contract and logic programming. For memory management, Python uses dynamic typing, and a mixture of reference counting and a cycle-detecting garbage collector. Python also has dynamic name resolution, which binds method and variable names during program execution. The standard library has two modules that implement functional tools taken from Haskell and Standard ML.

To run a python script, use python program_name.py

Libraries used to implement this model are:

- Keras is a deep learning API written in python, running on top of the machine learning platform tensor flow.
- Pandas are the data analysis tool used to do the manipulations on data ex: reading, writing, etc into files.
- Import pandas as Pd
- Tensorflow is the API consisting of many packages used for deep learning concepts.
- Scikit-learn could be a free software machine learning library for the Python programing language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is intended to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

6.3 ROC curve

ROC stands for Receiver Operating Characteristics. ROC curve is used for visual comparison of classification models, which shows the relationship between the true positive rate and the false positive rate. The area under curve (AUC) of the ROC curve is the measure of the accuracy of the model.

The Receiver operating curve (ROC) plots the possible thresholds of a trained model. ROC curve is plotted by taking false positive rates on X-axis and True positive rates on Y-axis.

Here True positive rate is the correctly predicted positives divided by all actual positives, similarly false positive rate is the incorrectly predicted positives divided by all actual negatives.

The standard form of the ROC curve is shown in the following figure 3.

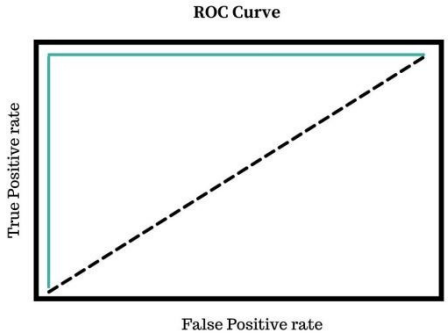

Fig 3: Standard form of ROC Curve

6.4. AUC

AUC is the complete area which is under the curve provides a measure of accuracy across all possible threshold value. The AUC value can be in the range of 0 and 1. In which 0 stands for Complete incorrect predictions and 1 stands for all correct predictions.

6.5. Logistic Regression

Logistic Regression is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome.

The outcome is measured with a dichotomous variable, meaning it will have two possible outcomes.

- Activation function = sigmoid
- Loss = Binary cross entropy

6.5.1. ROC curve of training data

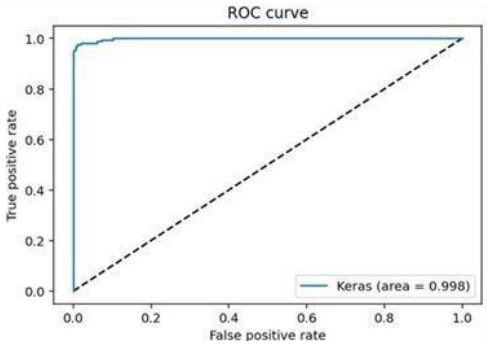

Fig 4: ROC curve of training data in logistic regression

AUC of training data = 0.998

6.5.2. ROC curve of testing data

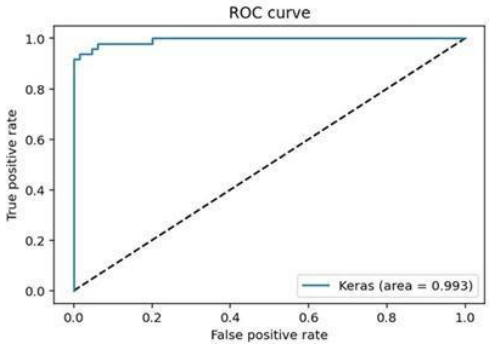

Fig 5: ROC curve of testing data in logistic regression

AUC of testing data = 0.993

6.6. Softmax Regression

Softmax Regression is a machine learning algorithm used for classification unlike Logistic regression, Softmax regression can be used for multi class classification. Here

- Activation function = Softmax
- Loss = Categorical cross entropy

6.6.1. ROC curve of training data

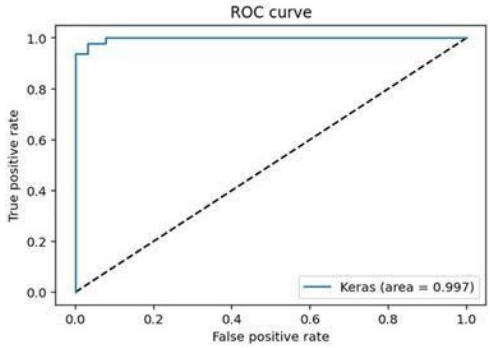

Fig 6: ROC curve of training data in softmax regression

AUC of training data = 0.997

6.6.2. ROC curve of testing data

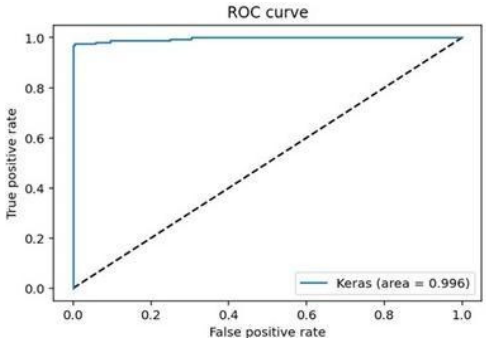

Fig 7: ROC curve of testing data in softmax regression

AUC of testing data = 0.996

VII.CONCLUSION

With the observations that we made from the ROC curve and AUC of training and testing data under both models, we want to conclude that for this specific problem both the algorithms work in the same way. The reason why we use Softmax regression is, the major drawback of logistic regression is it can be applied only to binary classification. In case in future if the new class is arising, then we can implement Softmax regression.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets /detail/breast-cancer

[2] Iran pour, M., Almassi, S., and Analoui, M., Breast cancer detection from FNA using SVM and RBF classifier. In: First Joint Congress on Fuzzy and Intelligent Systems, Ferdowsi University of Mashhad, Iran, and 29–31 Aug 2007.

[3] Kadam, V.J., Jadhav, S.M. & Vijayakumar, K. Breast Cancer Diagnosis Using Feature Ensemble Learning Based on Stacked Sparse Autoencoders and Softmax Regression. J Med Syst 43, 263 (2019).

[4] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 98-103, doi: 10.1109/ICACCCN.2018.8748777.

[5] Mohamad Mahmoud Al Rahhal, "Breast Cancer Classification in Histopathological Images using Convolutional Neural Network" International Journal of Advanced Computer Science and Applications (IJACSA), 9(3), 2018.

[6] Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. (2017) Classification of breast cancer histology images using Convolutional Neural Networks. PLoS ONE 12(6): e0177544.

[7] R. D. Ghongade and D. G. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM algorithm," 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), Kolkata, 2017, pp. 1-6.

[8] https://archive.ics.uci.edu/ml/datasets/Breast+ Cancer+Wisconsin+%28Diagnostic%29