IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Breast Cancer Prediction

Gargi Gupta, Nilesh Anand

School of Computer Science & Engineering,
Galgotias University, Greater Noida, Uttar Pradesh, India

Abstract: Breast cancer is causing an alarming increase in the number of deaths each year. It is the most common type of cancer and the leading cause of death in women around the world. Any advancement in cancer illness prediction and detection is critical to living a healthy life. As a result, high accuracy in cancer prognosis is critical for updating therapy aspects and patient survivability standards. Machine learning approaches, which have been shown to have a significant impact on the process of breast cancer prediction and early diagnosis, have become a research hotspot and have been proven to be a powerful technique. On the Breast Cancer Wisconsin dataset, we used five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbours (KNN). The objective of this project is to train machine learning models to predict whether a breast cancer cell is Benign or Malignant. Data will be transformed and its dimension reduced to reveal patterns in the dataset and create a more robust analysis. The optimal model will be selected following the resulting accuracy, sensitivity, and f1 score, amongst other factors. We will later define these metrics. We can use machine learning methods to extract the features of cancer cell nuclei and classify them. It would be helpful to determine whether a given sample appears to be Benign ("B") or Malignant ("M"). The machine learning models that we will apply in this report try to create a classifier that provides a high accuracy level combined with a low rate of false negatives (high sensitivity). This project will make a performance comparison between different machine learning algorithms in order to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity, and specificity, in order to find the best diagnosis. Diagnosis in an early stage is essential to facilitate the subsequent clinical management of patients and increase the survival rate of breast cancer patients. The major models used and tested will be supervised learning models (algorithms that learn from labelled data), which are most used in these kinds of data analysis. The utilization of machine learning approaches in medical fields proves to be prolific as such approaches may be considered of great assistance in the decision-making process of medical practitioners

Keywords — Breast Cancer, machine learning, feature selection, classification, prediction.

I. INTRODUCTION

Breast cancer is a disease in which cells in the breast grow out of control. This kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer is the second cause of death among women. After skin cancer, breast cancer is the most common cancer diagnosed in women.

In today's world, one out of every five people will develop cancer at some point in their lives. According to projections, the number of persons diagnosed with cancer would rise even more in the following years, reaching about 50% higher in 2040 than in 2020. Cancer fatalities have risen as well, from 6.2 million in 2000 to 10 million by 2020. Cancer is responsible for more than one-sixth of all deaths. This emphasizes the importance of investing in both cancer treatment and cancer prevention. The successful implementation of information and communication technology (ICT) in medical practice is critical to the health-care system's transformation, particularly in cancer care. Big data has transformed the size of data as well as the creation of value from it. By evaluating vast amounts of unstructured, heterogeneous, non-standard, and incomplete healthcare data, big data has revolutionized business intelligence. It not only forecasts but also assists in decision-making, and it is increasingly being recognized as a breakthrough in continual advancement with the goal of improving patient care quality while lowering healthcare costs. Data mining algorithms used in the healthcare business play an important role because of their excellent performance in disease prediction, diagnosis, and cost reduction, as well as making real-time decisions to save people's lives. Classification and prediction are the most typical data mining modelling aims, and numerous methods are used to predict breast cancer. Machine learning in histopathology has developed an interest over the decade due to its improvements in classification and localization tasks. Breast cancer is a prominent cause of death in women. Computer-Aided Pathology is essential to analyze microscopic histopathology images for diagnosis with an increasing number of breast cancer patients. The convolutional neural network, a deep learning algorithm, provides significant results in classification among cancer and non-cancer tissue images but lacks in providing interpretation. The image frames a classification problem as weakly supervised multiple instances learning problems and use attention on instances to localize the tumor and normal regions in an image. Attention-based multiple instance learning (A-MIL) is applied on BreakHis and BACH datasets. A method used in this paper produces better localization results without compromising classification accuracy.

I. RESEARCH METHODOLOGY

The main goal of our study is to find the most effective and predictive algorithm for breast cancer detection, so we used machine learning classifiers like Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN) on the Breast Cancer Wisconsin Diagnostic dataset and compared the results to see which model has the best accuracy.

1.1Method

Our method starts with data collection and then moves on to pre-processing, which consists of four steps: data cleaning, attribute selection, target Role selection, and feature extraction. Machine learning algorithms are built using the prepared data to forecast breast cancer for a new set of measures. We show the model new data for which we have labels to evaluate the algorithms' performance. This is commonly accomplished by using the Train test split method to split the labelled data we've acquired into two pieces. The training data, also known as the training set, accounts for 75% of the data utilised to develop our machine learning model. Test data, or test set, is 25% of the data that will be used to see how well the model works. After testing the models we compare the obtained results to select the algorithm that provides the high accuracy and identify the most predictive algorithm for the detection of breast cancer.

1.2 *Algorithms for Machine Learning:*

Machine learning is an application of artificial intelligence (AI) that allows computers to learn and develop without having to be programmed manually. Machine learning is based on the occurrence of computer programmes that examine the data and use it to learn for themselves. The learning process begins with facts or datasets, examples, experiences, or instructions, from which they might deduce a pattern and, if necessary, enhance it in the near future.
The predictive analysis of machine learning algorithms is achieved in our study.

1.3 *Machine Learning Techniques*:

Machine learning is an application of artificial intelligence (AI) that allows computers to learn and develop without having to be programmed manually. Machine learning is based on the occurrence of computer programmes that examine the data and use it to learn for themselves. The learning process begins with facts or datasets, examples, experiences, or instructions, from which they might deduce a pattern and, if necessary, enhance it in the near future.
The predictive analysis of machine learning algorithms is achieved in our study. The following are the machine learning techniques used in our project:

• Random forests, also known as random decision forests, are an ensemble method for classification, regression, and other tasks that works by training a large number of decision trees and then outputing the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random choice forests correct the tendency of decision trees to overfit their training set. Random forests, also known as random decision forests, are a type of tree that produces results using ensemble learning approaches for classification and regression. The properties it employs to build such trees are bagging and feature randomness. In comparison to the decision tree, the random forest has the advantage of not overfitting the data.

• Artificial neural networks (ANNs), also known as neural network systems, are computer systems that simulate the operation of the human brain. The algorithm's main goal is to give a faster and more accurate output than an older or traditional approach. If the algorithm is provided data or an image regarding a specific thing, it will be able to swiftly detect or categorise photographs that do not contain the object.

• Support Vector Machine (SVM) is a classifier that separates datasets into classes in order to find a maximum marginal hyper plane (MMH) using the closest data points. Support vector machines, often known as supervised models in machine learning, are supervised models. When classifying objects, a support vector machine builds a hyperplane. A hyperplane is a line that divides the two classes on a plane. An SVM training algorithm builds a model that assigns new examples to at least one of two categories, making it a non-probabilistic binary linear classifier (though methods like Platt scaling exist to use SVM in a probabilistic classification setting). Given a set of coaching examples, each marked as belonging to at least one of two categories, an SVM training algorithm builds a model that assigns new examples to at least one of two categories, making it a non-probabilistic binary linear classifier New examples are then mapped into that same space and assigned to a category based on which side of the gap they land on.

Given a knowledge set, DTs struggle to group and label observations that are similar between them, and search for the simplest rules that partition the observations that are not similar until they reach a particular degree of similarity. They employ a layered splitting technique, in which they try to split the data into two or more groups at each layer, so that data belonging to the same group is as similar as possible (homogeneity), and groupings are as distinct as feasible.

II. LITERATURE REVIEW

For the prediction and detection of breast cancer, a significant number of machine learning algorithms are available. Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN Network) are examples of machine learning algorithms. Many researchers have conducted breast cancer research using a variety of datasets, including the SEER dataset, mammogram pictures as a dataset, Wisconsin Dataset, and data from other hospitals.

There are some important findings. Ultrasound characterisation of breast masses by S. These cells divide more briskly and disperse faster than healthy cells do and continue to accumulate, forming a lump or mass that the may start causing pain. Cells may spread rapidly through your breast to your lymph nodes or to other parts of your body. Some women can be at a higher risk for breast cancer because of their family history, lifestyle, obesity, radiation, and reproductive factors. In the case of cancer, if the diagnosis is made promptly, the patient can be saved because cancer treatment has advanced. The Naive Bayesian Classifier, k-Nearest Neighbour, Support Vector Machine, Artificial Neural Network, and random forest are the four machine learning classifiers used in this study. Image resolution and lesion characterization have been found to improve with harmonic imaging and real-time compounding. Recently, USG elastography has appeared to be fairly promising. The specificity and positive predictive value of USG in the characterization of breast masses have improved, according to preliminary findings. The relative difference in density and acoustic resistance of the lesion as opposed to the surrounding breast tissue is the reason why any lesion is apparent on mammography or USG. Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach proposed a system in which Breast cancer prediction is an open topic of research. Different machine learning methods are employed in this paper to detect Breast Cancer Prediction. Prediction methods include decision trees, random forests, support vector machines, neural networks, linear models, adabost, and naive bayes.

Methodology

The main goal of our study is to find the most effective and predictive algorithm for breast cancer detection, so we used machine learning classifiers like Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision Tree (C4.5), and K-Nearest Neighbors (KNN) on the Breast Cancer Wisconsin Diagnostic dataset and compared the results to see which model has the best accuracy. Our method starts with data collection and then moves on to pre-processing, which consists of four steps: data cleaning, attribute selection, target Role selection, and feature extraction. Machine learning algorithms are built using the prepared data to forecast breast cancer for a new set of measures. We show the model new data for which we have labels to evaluate the algorithms' performance. This is commonly accomplished by using the Train test split method to split the labelled data we've acquired into two pieces. The training data, also known as the training set, accounts for 75% of the data utilised to develop our machine learning model. Test data, or test set, is 25% of the data that will be used to see how well the model works. After testing the models we compare the obtained results to select the
algorithm that provides the high accuracy and identify the most predictive algorithm for the detection of breast cancer.

III. PROJECT DESIGN

Machine learning is an application of artificial intelligence (AI) that allows computers to learn and develop without having to be This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input

We have also gained an understanding of the reasons why our model is able to predict this well. Our earlier analysis showed the difference in morphology between cell metrics for malignant v benign tumours, which could be seen visually in the images and were expressed in different distributions of values for particular feature measurements of the cells that we observed.

Our use of higher level analytical tools such as TDA also allowed us to gain a much better understanding of the dataset, in particular a better idea of the range of feature values that were typical for malignant and benign tumours as 'groups' within the dataset.

The dataset is instantiated directly in code in the form of a dictionary. It can be found in the load_data.py script.

The mlp.py script contains the instantiated model and can be further tuned according to personal discretion. Function calls can also be modified to produce graphs and figures (off by default).

By running mlp.py, the classifier is trained and validated with Monte Carlo Cross Validation, as per the paper by Patricio et al., 2018. Within the mccv_keras.py script, which is called by mlp.py, we implement our cross validation mechanism exactly as per the svm.py file. This is important because by running svm.py we reproduce the original experiment from Patricio et al., 2018 with very similar figures, thus proving that the
reimplemented validation mechanism works exactly the same as the original, seeing as
the original implementation was carried out in the R programming language.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

It is hoped that the data derived from these images of
the cells is able to capture the differences between these types of cells.

In this project, we aim to develop a machine learning model that will aim to predict Malignant tumours with the highest accuracy.

Python Package	Version
Keras	2.2.4
Tensorflow-gpu	1.12.0
Scikit-learn	0.20.3
Scipy	1.2.1
Numpy	1.16.2

3.1 System Proposed

This system compares the following machine learning (ML) algorithms: Support machine vector (SVM), Decision Tree (DT), Random Forest (RT), Artificial Neural Networks (ANN), Naive Bayes (NB), and Nearest Neighbor (NN) search. The Wisconsin datasets were utilized to create the data collection. The dataset was divided into training and testing sets in order to implement the machine learning methods. There will be a comparison of all six algorithms. The website will be given a model of the algorithm that produces the best results. The website will be built using the flask python framework. The database will be hosted on Xampp, Firebase, or the native Python and flask libraries. The UCI Machine Learning Repository has this data set available.
It is made up of 32 multivariate real-world properties.
The total number of cases in this data collection is 569, and there are no missing values.
The proposed system's procedure is as follows:
1. The patient uses our website to schedule an appointment.
2. After that, the patient will meet with the doctor in person for the appointment.
3. The doctor will manually examine the patient before doing a breast mammography or an ultrasound. This ultrasound will produ ce a picture of the breast that will show whether or not there are any lumps.
4. If the lumps are detected, a biopsy will be performed. The digitised image of the Fine Needle Aspirate (FNA) is what forms the features of the dataset.
5. Those numbers will be provided to the system by the doctor and the model will detect if it's a benign or a malignant cancer.
6. The report will be then forwarded to the patient on their respective account

3.2 Formulation of Problem

- ID number
- Diagnosis (M = malignant, B = benign)
- Features 3-32 are ten real-valued features are computed for each cell nucleus:
  a) radius (mean of distances from 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)
- The mean, standard error (se) and "worst" or largest (mean of the

three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

3.3 Tool Database and Technology Used:

The various database tools and technology that we used in our app are:
● Jupiter
●Python
• [Wisconsin Breast Cancer Diagnostic Dataset]
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/version/2
The dataset's features describe characteristics of the cell nuclei on the image.
The features information are specified below:
-Attribute Information:
1.ID number
2. Diagnosis (M = malignant, B = benign)

IV. RESULTS AND DISCUSSION

Breast cancer, if detected early, can save the lives of thousands of women and even men.
These programmes assist patients and clinicians in gathering as much information as possible in the real world.
The data for the idea we offered was gathered through research on nine papers. We will be able to classify and forecast whether a cancer is benign or malignant using machine learning algorithms. The accuracy rate of our project is 94%.
Machine learning algorithms can be utilised in medical research because they improve the system, minimise human errors, and reduce manual errors. The most frequently used ML methods were decision trees (19 studies, 61.3%), artificial neural networks (18 studies, 58.1%), support vector machines (16 studies, 51.6%), and ensemble learning (10 studies, 32.3%). The median sample size was 37256 (range 200 to 659820) patients, and the median predictor was 16 (range 3 to 625).
Further optimization of the performance of the proposed model is also needed in the future, which requires more standardization and subsequent validation.
In conclusion, in this proposed system earlier prognosis of breast cancer to improve the survivability of patients is emphasized. Data collected from Wisconsin dataset along with biosensor output are stored in database and processed to produce Intelligent Data. Real-time output from a biosensor device improves the accuracy of disease prediction

REFERENCES

[1] Fabian Pedregosa and all (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.

[2] 'WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020)

[3] "Analysis of Machine Learning Techniques for Breast Cancer Prediction" by the Priyanka Gupta and Prof. Shalini L of VIT university, vellore, 5 May 2018.