International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 08 | Aug 2022 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF BREAST CANCER,COMPARATIVE REVIEW OF MACHINE LEARNING TECHNIQUES AND THEIR ANALYSIS

Venkata Krishna Pokala[1], Geetha Mohana Priya Puli[2], Mounika Ravipati[3] ,Sriram Pokala[4], S.V.Rama Krishna[5]

[1,2,3] B.Tech Students, Vignan's University, Vadlamudi, Guntur, Andhra Pradesh, India
[4] B.Sc Student, Sri Baba Gurudev Degree And PG College, Sattenapalli, Andhra Pradesh, India
[5] Asst. Professor, Vignan's University, Vadlamudi, Guntur, Andhra Pradesh, India

---

Abstract - In the pool of emerging technologies, Machine learning has gained much popularity in medical field due to its high performance and accuracy. In these days it is very much essential to use machine learning models in every aspect for higher accuracy ,specifically in medical field since health is being given more importance to survive. Breast cancer is one of the most dangerous cancer disease among all the cancer types known till date. Not only early detection is not the solution but also curing the disease is the most important issue to be considered in the emerging world. As the population is growing rapidly , deaths due to breast cancer has increased exponentially. Here we are going to be more focused on detection as earlier the detection, higher the chance to cure. In this study, some of the machine learning algorithms have been employed to detect the disease such as SVM,KNN etc on WBCD which is publickly available and used dataset for most of the applications. The main motto of this research work is significant comparision and analysis on the applied algorithms in terms of accuracy, precision, recall, f-score. These studies demonstrate that modern machine learning methods could increase the accuracy of early cancer tumour prediction. The wbcd comprises of 569 instances and 32 attributes with no missing values which helps us to identify the target either malignant or benign.

***Key Words*: Breast cancer prediction, Classifier algorithms, SVM, KNN.**

1. INTRODUCTION

Breast Cancer is one of the most significant issue to be considered seriously in medicine or hospitality now a days since deaths due to Breast cancer are increasing exponentially.According to the new reports and reviews 53% of the indian women among all the reported cases has died due to Breast cancer(87090/162468).To dive into deep, some of the reasons for causing breast cancer is harmones, radiation therapy ,obesity etc. The intresting fact which I came across while I'm writing this paper is that Men can get the breast cancer too. However lessthan 1% men can face breast cancer which is negligible. But Building the model which can figure out that 1% men too is a challenging task. For early detection of breast cancer there exists some techniques such as mammography, computer aided detection(CAD) etc,.In this paper we will come to know the influence of the Machine learning algorithms.Recent year study's has proven that ML models has been gained a percentage of 30 in their predicting power. Breast cancer is one of the most common ailments/illnesses in India, causing many deaths in the present day. The shape of most malignancies cases in women is changing day by day as a result of changes in food and lifestyle. It is the second most common cause of women's lack of lifestyles in the world.To begin with, the paper is about the dataset and the some of the important insights of the data has mentioned for clear understanding about the dataset. The proprocessing techniques if neccessary since it is highly recommended to preprocess the data for improved performance and accuracy. The results of this work is represented in the form of table in a comparitive way which includes the accuracy, precision and recall etc. Although there available many algorithms , Logistic Regression is identified as the best for WBCD dataset as it is giving high accuarcy.

This uses mind of Machine analyzing (ML) to assume breast most cancers based truely certainly in reality totally on the statistics received. The several ranges of breast maximum cancers are diagnosed thru right remedy and detailing. If we do not provide proper remedy to our patients, it's going to bring about their loss of lifestyles. Earlier strategies for classifying statistics have been used, but their lower accuracy, because of the truth they might be used for correct categorization and prediction. Deep analyzing algorithms and numerical dataset system studying techniques are used to extract skills and hidden skills. This traditional approach, it is based mostly on regression, detects the lifestyles of maximum cancers, at the same time as new ML techniques and algorithms are constructed on model introduction. In its training and finding out degrees, the model is supposed to forecast unknown facts and offers a pleasing anticipated very last effects. These techniques used to differentiate amongst benign and malignant tumors.

2. PROBLEM STATEMENT

As you can see in the below imahe there are six different varieties of cancers that leads to the death. The ultimate and

very primary goal of this paper is to implement different ML algorithms that figures out which ML model best suits for predicting the breast cancer. To determine the number of patients with non-cencerous and cancerous tumours, as well as the type of tumour.

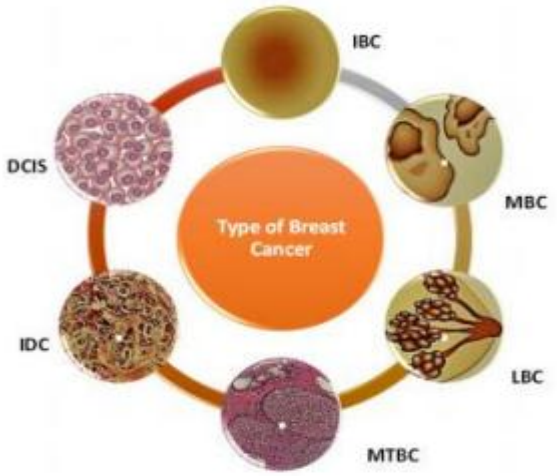

Fig -1: Various Kinds of Breast Cancer

3. EXISTING SYSTEM

On the basis of historical data and the existed situations, several researchers have already developed variety of methodologies for risk analysis and prediction of breast cancer. The analysis of breast cancer data for risk identification was the main focus of their study. However, a system that forecasts risk based on historical data and current data is a more crucial requirement. The clinical oncologists can utilize their model to help them make decisions. However, different types of persons who are initially exposed to the danger must also be taken into account. To more accurately anticipate the dangers, a rule-based system that can recognize the symptoms sooner and do a temporal analysis on their data will be helpful.

China is the world's most populous country. According to a recent report by the organization (GLOBOCAN-2018), the male-to-female breast cancer ratio is 8.6% for males and 19.2% for females. This disease claims the lives of over 1.3 million people each year. According to statistics, approximately 400 men and 41,900 women are predicted to die as a result of this disease.

According to a survey from the United States, 3.8 million women are alive yet have breast cancer. In 2019, 59,838 cases of Ductal Carcinoma in Situ (DCIS) breast cancer were detected in the United States. The total number of breast cancer deaths is 458,000. Breast cancer was the leading cause of mortality in China in 2012.Cancer accounted for 48% of all deaths in 2012, whereas the global death rate was 52%. In 2015, the statistics of 1,517 women were evaluated to determine the breast cancer survival and recurrence rate.

4. LITERATURE REVIEW

[1]The second major cause of cancer-related mortality in women is breast cancer. Breast cancer development is a multi-step process linked to a few different mobile types, and worldwide prevention is still challenging. One of the best ways to prevent this illness is by early detection of breast cancer. Because to early detection and treatment, the 5-350-day relative survival rate of breast cancer patients is above 80% in several developed worldwide locations. Significant progress has been achieved in both the creation of prevention measures and the statistics of breast cancer in general during the past ten years. The identification of breast cancer stem cells is a valuable resource for understanding the aetiology and processes behind tumor treatment resistance, and several breast cancer-related genes have been identified. Humans now have more medication choices for chemo prevention of breast cancer, and natural prevention has improved recently to improve the quality of life for cancer patients. We may highlight significant research on the pathophysiology, related genes, risk factors, and preventative measures of breast cancer in the recent years in this evaluation..

[2] The author conveyed that Breast cancers is one of the most not unusual cancers among women in the worldwide, accounting for the general public of recent maximum cancers times and maximum cancers-associated deaths steady with worldwide records, making it a notable public health hassle in contemporary day society. In this paper, we're capable of gift an outline of the evolution of huge information inside the health tool, and check 4 studying algorithms to a breast maximum cancers information set. In this study the author tried his best to explain how dangerous this cancer is.Author implemented variety of algorithms and finally achieved a good percentage of accuracy which is perfectly fine to explain someone the seriousness of the cancer. The experimental consequences display that SVM offers the brilliant accuracy ninety seven. Nine Percent. The locating will assist to choose out the terrific kind device-reading set of tips for breast most cancers prediction.

[3] A malignant growth called a breast tumor develops inside the glandular epithelium of the breast. It is regarded as one of the malignancies that affects women the most frequently in the globe. However, there isn't always a very effective method of treating breast cancer. The early diagnosis and assessment of breast tumours, however, is a crucial component in lowering the risk of mortality. Assessment of medical pictures from many modalities is typically required for an accurate appraisal of breast malignancies. There is a great demand for an automated equipment that could properly examine the photographs. In this paper, we introduce some commonly used scientific imaging techniques for analysis of breast most cancers, and based totally on them we take a look at some presently proposed techniques for breast most cancers detection with laptop

vision and device reading techniques. Finally, we've got a take a look at and feature a look at the detection not unusual standard everyday typical overall performance of numerous techniques on histological images and mammograph pix respectively Breast most cancers is a malignant tumor that takes region within the glandular epithelium of the breast. Sometimes, the technique of mobile increase goes incorrect. New cells form even the frame doesn't need them and antique or damaged cells do not die as they want to. When this takes region, a boom of cells frequently workplace work a mass of tissue referred to as a lump, boom, or tumor. Its onset is frequently associated with heredity, and the incidence of breast maximum cancers is higher among ladies maximum of the a while of forty and 60 or at a few degree inside the menopause.

4. PROPOSED SYSTEM

We obtained the Wisconsin Breast Cancer Diagnosis dataset which is publicly available on the inetrnet and we've utilised Google Colab as the development platform. Support Vector Classifier Machine Learning algorithm, KNN, Random Forest, Adaboost, and Xgboost Classifier are among the supervised learning algorithms and classification techniques used in our methodology, along with the K-fold cross validation technique.

Mainly there are exists three methods which exactly defines our proposed model. They mainly concentrate on the data as well as on the algorithm that is being used which means model that is built to predict the output. Primarily, Feature extraction has a pivotal role in predicting the output for any Machine Learning Model that is built. The accuracy of any ML model completely depends on how the data is being extracted and arranged in such a way that it learns and tests. Secondly, Model building is a process where the Machine learning algorithm is employed to predict the accuracy. However the output is predicted based on the data patterns. First of all the model learns from the data and predicts based on test data. Finally, Model evaluation is the final step which accurately gives the output of model that is being developed.

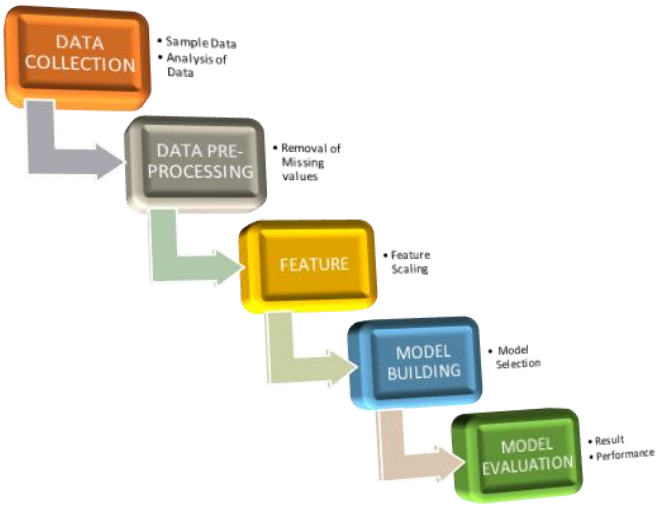

Fig -2:Work Flow

4.1 DATA VISUALIZATION

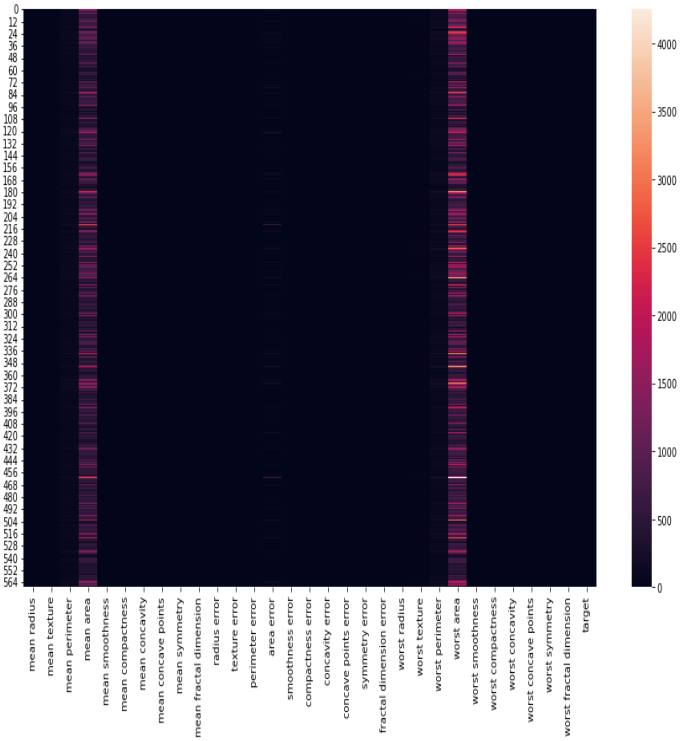

Fig -3:Heat map

Above heatmap consists of variety of different feature's value. As far as the heatmap is concerned ,it is used for showing the correlation among the features. As shown in Fig,.3 "Thicker the colour shows the more correlation than the lighter colour" which means that As the strength of colour increases the correlation between the feature gets moved towards positive correlation.

4. EXPERIMENTAL RESULTS

As per our earlier discussion we have used WBCD dataset which is available on the internet.Our subsequent step is to apply numerous magnificence algorithms to categorise our result in to 2 instructions in particular: - Benign and Malignant. Overall our dataset consists of 579 values in which 357 are Benign i.,e non cancerous cells and 212 are malignan I.,e cancerous cells. The below mentioned tables shows the accuracy gained by different machine learning algorithms used in this project. Each Algorithm has its own advantages and disadvantages as we all are well aware that a coin has both head and tail, right? So, XGboost is a gradient boosted tree algorithm that gains maximum accuracy and also got 0% Type-II error which is pretty good and indicates that our model has gained perfect accuracy without any error or misclassification.

Table -1: Accuracy of proposed system

S.NO	Algorithm	ACCURACY
1.	KNN	93.85%
2.	SVM	96.49%
3.	RANDOM FOREST CLASSIFIER	97.36%
4.	LOGISTIC REGRESSION	95.61%
5.	NAIVE BAYES	94.73%
6.	DECISION TREE	94.73%
7.	ADA BOOST	94.73%
8.	XG BOOST	98.24%

Table -2: Confusion Matrix

	0	1
0	46	2
1	0	66

As per the above table it is being proven that we have got 0% Type-II error which is perfectly Fine. It means that our model predicts that there are no such values called False Negative furthurmore it represents there exists no incorrect data prediction.

5. CONCLUSION

Research in recent years has shown that machine learning models are gaining much popularity due to high accuracy and prediction power. In this paper is is being proven that XG BOOST model has got 98.24% accuracy. Currently these are many number of machine learning techniques exists to analyse medical data. Building precise and computationally effective classifiers for medical applications is a challenging problem in the era of digital technologies. In order to discover the optimum classification accuracy, we used machine learning algorithms on the Wisconsin Breast Cancer (WBCD) dataset in this paper. The XGBOOST classifier provided the highest level of classification accuracy. Early diagnosis is therefore crucial, and invasive techniques' detection makes mass forecasts much simpler. The results of the examination are proven to be quite accurate in predicting breast cancer. The suggested device can quickly ascertain the severity of the sickness and forecast whether the patient will survive the illness or if it will develop to malignancy.

REFERENCES

[1] "Y.-S. Sun Et Al., "Risk Factors And Preventions Of Breast Cancer," International Journal Of Biological Sciences, Vol. Thirteen, No. Eleven, P. 1387, 2017"

[2] "Y. Khourdifi And M. Bahaj, "Applying Best Machine Learning Algorithms For Breast Cancer Prediction And Classification," In 2018 International Conference On Electronics, Control, Optimization And Computer Science (Icecocs), Pp. 1–5, Ieee."

[3] "Y. Lu, J. Y. Li, Y. T. Su, And A. A. Liu, "A Review Of Breast Cancer Detection In Medical Images," In 2018 Ieee Visual Communications And Image Processing (Vcip), Pp. 1–Four, Ieee"

[4] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; DOI 26:1-24

[5] 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

[6] 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

[7] 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.

[8] 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).

[9] Priyanka Jain & Santosh Kr. Vishwakarma (2016). Collaborative Analysis of Cancer Patient Data using Rapid Miner. International Journal of Computer Applications, 145, 8-13.

[10] 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

[11] S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007

BIOGRAPHIES

"I am a constant learner in programming and computer science field and also has a keen interest in ML and Data science "

"Ms.Geetha is an enthusiast of Machine learning and it's applications in the contemporary world,She carried out several projects in the same "

"I am very enthusiastic about computers and programming. As I am an adaptive person, I always learn new technologies in this field."

" A Passionate coder interested in computer science and related field.Addictied to learn new technologies in related field."