International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Breast Cancer Detection Using Machine Learning

Shirin Raut[1], Gauri Kalanke[2], Shreyash Borode[3], Rutuja Awankar[4], Prof. J. C. Bambal[5]

[1]Shirin Raut, P.R. Pote Patil College of Engineering and Management, Amravati
[2]Gauri Kalanke, P.R. Pote Patil College of Engineering and Management, Amravati
[3]Shreyash Borode, P.R. Pote Patil College of Engineering and Management, Amravati
[4]Rutuja Awankar, P.R. Pote Patil College of Engineering and Management, Amravati
[5]Prof. J. C. Bambal, P.R. Pote Patil College of Engineering and Management, Amravati

***ABSTRACT*—** Each year number of deaths is increasing extremely because of breast cancer. It is the most frequent type of all cancers and the major cause of death in women worldwide. Any development for prediction and diagnosis of cancer disease is capital important for a healthy life. Consequently, high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients. Machine learning techniques can bring a large contribute on the process of prediction and early diagnosis of breast cancer, became a research hotspot and has been proved as a strong technique.

In this study, we applied machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors' (KNN) on the Breast Cancer Wisconsin Diagnostic dataset, after obtaining the results, a performance evaluation and comparison is carried out between these different classifiers. The main objective of this research paper is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective whit respect to confusion matrix, accuracy and precision. It is observed that Support vector Machine outperformed all other classifiers and achieved the highest accuracy (97.2%). All the work is done in the Anaconda environment based on python programming language and Skit-learn library.

I.INTRODUCTION

Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide, according to statistics released by the International Agency for Research on Cancer (IARC) in December 2020. In the past two decades, the overall number of people diagnosed with cancer nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020. Today, one in 5 people worldwide will develop cancer during their lifetime. Projections suggest that the number of people being diagnosed with cancer will increase still further in the coming years, and will be nearly 50% higher in 2040 than in 2020.

More than one in six deaths is due to cancer. This reinforces the need to invest in both the fight against cancer and cancer prevention.Our objective is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision and sensitivity.

II. LITERATURE REVIEW

Breast cancer detection using machine learning (ML) algorithms has been a topic of interest in medical research for several years. ML algorithms have been applied to various medical imaging modalities, including mammography, ultrasound, and magnetic resonance imaging (MRI), to improve the accuracy and efficiency of breast cancer detection. In 1994, the first research paper was published on the application of artificial neural networks (ANNs) for breast cancer detection using mammography. ANNs were found to be effective in detecting breast cancer from mammography images and had a higher accuracy than traditional methods.

A large number of machine learning algorithms are available for prediction and diagnosis of breast cancer. Some of the machine learning algorithm are Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbors (KNN Network) etc. A lot of researcher have realized research in breast cancer by using several dataset such as using SEER dataset, Mammogram images as dataset, Wisconsin Dataset and also dataset from various hospitals. By exploiting these dataset authors extract and select various features and complete their research.

Breast cancer detection using machine learning (ML) algorithms is an active area of research that has the potential to significantly improve the accuracy and efficiency of breast cancer diagnosis. The primary objective of this project is to develop an ML algorithm for breast cancer detection using mammography images. The project's scope includes developing a dataset of mammography images, preprocessing the images, selecting and training an appropriate ML algorithm

Some of the related work on this project includes a review of the state-of-the-art ML algorithms for breast cancer detection and diagnosis, including decision trees, support vector machines, random forests, deep learning, and transfer learning techniques.

The existing systems for breast cancer detection using ML algorithms have made significant progress in improving the accuracy and efficiency of breast cancer detection. However, there are still some limitations that need to be addressed. Some of the limitations of the existing system for breast cancer detection using ML algorithms include:

1. Limited dataset
2. Lack of diversity in datasets
3. Cost

III. SYSTEM ARCHITECTURE

The system architecture/design of a Breast Cancer Detection project using CNN involves developing a machine learning model..The model will be based on a CNN architecture, which has shown promising results in plant disease identification. The system architecture/design can be broken down into the following components:

Data Collection: The first component of the system architecture/design is data collection.

Data Preprocessing: The second component is data preprocessing. The collected dataset will undergo preprocessing, including resizing, normalization, and augmentation.

CNN Architecture: The third component is the CNN architecture.

Model Training: The fourth component is model training. The CNN model will be trained using transfer learning techniques, where a pre-trained model is fine-tuned for plant disease identification.

Model Validation: The fifth component is model validation. Once the model is trained, it will be validated using a separate test dataset.

Deployment: The final component is deployment. The model will be deployed as a web application, which can be used by farmers, plant pathologists, and other stakeholders in the agriculture industry

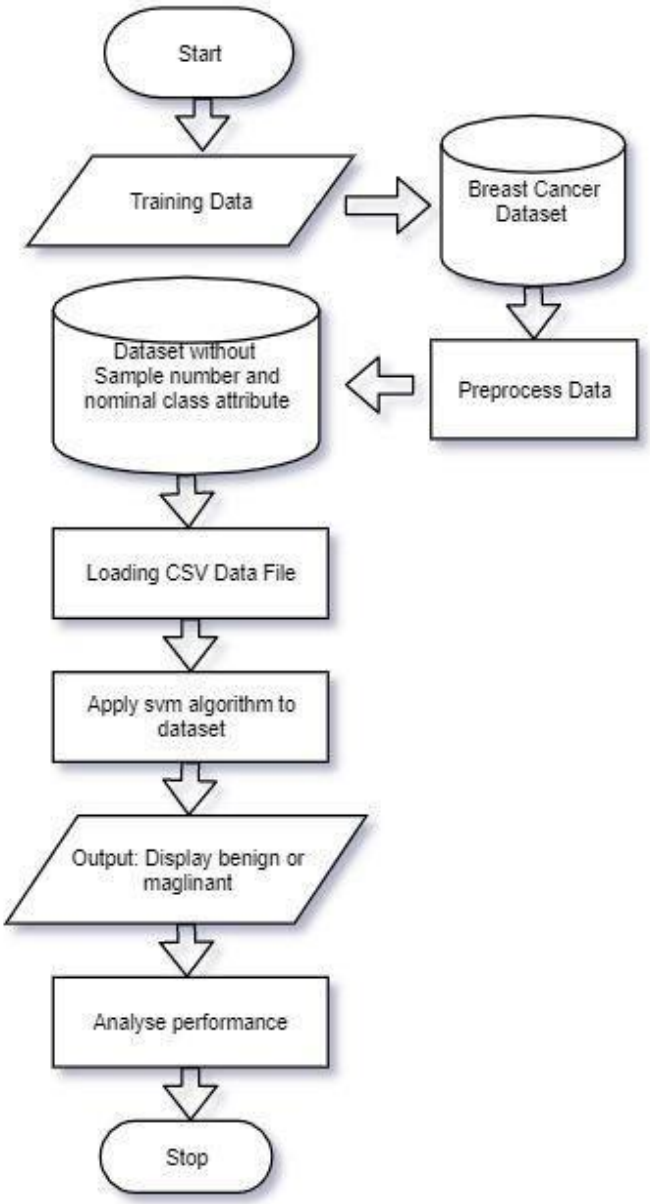

Fig 3.2: Flowchart of Proposed System

IV.OBJECTIVES

- The main objective of this project is to identify breast cancer at an early stage, allowing for more effective treatment to be used and reducing the risks of death from breast cancer.
- Since early detection of cancer is key to effective treatment of breast cancer, we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data.
- To predict Breast Cancer using Different ML Algorithms. Increase survival rate by early diagnosis.

V. ADVANTAGES & DISADVANTAGES

ADVANTAGES

- Reduced human error: Traditional methods of disease identification often rely on human expertise, which can lead to errors due to misinterpretation or misdiagnosis.
- Faster diagnosis: Automated disease identification systems can provide a faster diagnosis compared to traditional methods, which can take days or weeks to identify the disease.
- Reduce the mortality rate through appropriate therapeutic interventions at the right time.
- Scalability: The automated system can be scaled up to handle large datasets and can be deployed in various locations, making it a valuable tool.

DISADVANTAGES

- The accuracy of the model heavily depends on the quality and quantity of the training data.
- The algorithm model is a black box, meaning it is difficult to understand how the model is making its predictions.
- The model requires significant computational power to train, which can be expensive and time-consuming.

VI. CONCLUSION AND FUTURE WORKS

CONCLUSION

On the Wisconsin Breast Cancer Diagnostic dataset (WBCD we applied five main algorithms which are: SVM, Random Forests, Logistic Regression, Decision Tree, K-NN, calculate, compare and evaluate different results obtained based on confusion matrix, accuracy, sensitivity, precision, AUC to identify the best machine learning algorithm that are precise, reliable and find the higher accuracy. All algorithms have been programmed in Python using scikit-learn library in Anaconda environment.

It should be noted that all the results obtained are related just to the WBCD database, it can be considered as a limitation of our work, it is therefore necessary to reflect for future works to apply these same algorithms and methods on other databases to confirm the results obtained via this database, as well as, in our future works, we plan to apply our and other machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy

FUTURE SCOPE

The analysis of the results signifies that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables.

We are intending how to parametrize our classification techniques hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy.

According to a report published by National Cancer Registry Programme (NCRP), cancer cases are expected to increase from 13.9 lakh in 2020 to 15.7 lakh by 2025, assuming a 20 percent increase overall [3]. Common cancers can be prevented to be fatal if treated early. A breast cancer diagnosis made early can lead to effective treatment.

REFERENCES

[1] Wang, D. Zhang and Y. H. Huang "Breast Cancer Prediction Using Machine
Learning" (2018), Vol. 66, NO. 7.

[2] B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

[3] Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

[4] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", IJCSMC, Vol. 3, Issue. 1, January 2014, pg.10 – 22.

[5] Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

[6] R. K. Kavitha1, D. D. Rangasamy, "Breast Cancer Survivability Using Adaptive Voting EnsembleMachine Learning Algorithm Adaboost and CART Algorithm" Volume 3, Special Issue 1, February 2014

[7] P. Sinthia, R. Devi, S. Gayathri and R. Sivasankari, "Breast Cancer detection using PCPCET and ADEWNN", CIEEE' 17, p.63-65

[8] Vikas Chaurasia and S.Pal, "Using Machine Learning A lgorithms for Breast Cancer Risk Prediction and Diagnosis" (FAMS 2016) 83 (2016) 1064 – 1069