IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

Breast Cancer and Cervical Cancer Detection Using Machine Learning

P. Bhuvaneswari[1], D. Divya Dharshini[2], M. Srimathi[3], G. Sneka[4], P. Dinesh Kumar[5]
Assistant Professor, Department of Information Technology[1]
Students, Department of Information Technology[2,3,4,5]
Hindusthan Institute of Technology, Coimbatore, India

Abstract: *Women are seriously threatened by breast cancer and cervical cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy. Four algorithms SVM, Logistic Regression, Random Forest and KNN which predict the breast cancer outcome have been compared in the paper using different datasets. All experiments are executed within a simulation environment and conducted in JUPYTER platform. Aim of research categorises in three domains. First domain is prediction of cancer before diagnosis, second domain is prediction of diagnosis and treatment and third domain focuses on outcome during treatment. The proposed work can be used to predict the outcome of different technique and suitable technique can be used depending upon requirement. This research is carried out to predict the accuracy. The future research can be carried out to predict the other different parameters and breast cancer and cervical cancer research can be categorises on basis of other parameters.*

Keywords: Breast Cancer, Cervical cancer, machine learning, feature selection, classification, prediction, KNN, Random Forest, ROC, etc.

I. Introduction

Breast Cancer is one of the leading cancer developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue. For the execution of the ML calculations, the dataset was parceled into the preparation set and testing set. A correlation between every one of the six calculations will be made. The calculation that gives the best outcomes will be provided as a model to the site. The site will be produced using a python framework, called flask.

Cancer is a significant health problem, especially as it is one of the most common causes of death in many countries around the world. Breast, cervical and thyroid cancer are the most common types of cancer among women. In the Kingdom of Saudi Arabia (KSA), cancer statistics are significantly increasing. The total number of cancer cases among women registered in the Saudi Cancer Registry (SCR) is 8,565 and cancer in females accounts for 52.8% of all cancer cases in the KSA. Cervical cancer was the fourth most common cancer among Saudi females in 2015, with 403 cases, representing 6.1 % of all cancer cases diagnosed among Saudi women. In 2010, there were 220 cervical cancer cases among Saudi women, representing 4.1 % of all cancer cases, which indicates an annual increase of 9 % in the number of cervical cancer cases. Since then, the number of cases increased even further, to 1073 by the end of 2018, according to a report by the World Health Organization.

II. Literature Review

Traditional diagnosis involves trained physicians to visually examine the medical images of breast for any signs of tumor development in the region. However due to the large scale of the medical image data, this manual diagnosis is often laborious and can be highly subjective due to inter-observer variability. Inspired by the advanced computing technology which is capable of performing complex image processing and machine learning, researches had been carried out in the past few decades to develop computer aided diagnosis (CAD) systems to assist clinicians detecting breast cancer. [2]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

The King Hussein Cancer Center, the only comprehensive cancer center in Jordan, has changed the practice of oncology in the country via implementation of a multidisciplinary approach to treatment, monitoring of treatment outcomes, and investments in ongoing cancer research. However, there remains no national system for ensuring provision of high-quality cancer care nationwide. Here, we review the epidemiology of breast cancer and the current status of breast cancer care in Jordan, we compare our treatment outcomes with international ones, and we highlight challenges and improvement opportunities.[3]

We propose a machine learning technique that allows a joint and fully supervised optimization of dimensionality reduction and classification models. We also build a model able to highlight relevant properties in the low dimensional space, to ease the classification of patients. We instantiated the proposed approach with deep learning architectures, and achieved accurate prediction results (top area under the curve AUC = 0.6875) which outperform previously developed methods, such as denoising autoencoders. Additionally, we explored some clinical findings from the embedding spaces, and we validated them through the medical literature, making them reliable for physicians and biomedical researchers.[4]

Cancer is a pestilent disease. One of the most important cancer kinds, cervical cancer is a malignant tumor which threats women's life. In this study, the importance of test variables for cervical cancer disease is investigated by utilizing Stability Selection method. Also, Random Under-Sampling and Random Over-Sampling methods are implemented on the dataset. In this context, the learning model is designed by using Random Forest algorithm. The experimental results show that Stability Selection, Random Over-Sampling and Random Forest based model are more successful, approximately 98% accuracy.[5]

It is very necessary to diagnose the benign tumor tissues in early stages with their exact location to save the life of human being. This paper includes a survey on different segmentation techniques to diagnose the benign tumor cells in MRI images and it also includes the proposed for the same using fuzzy c-means algorithm to find out the benign tumor tissues.[6]

These implications relate the degradation of patients to the cervical stages. A healthcare monitoring system could utilize the vicious impacts of Sthese implications to provide a customized care to patients [7]

III. Existing System

Breast Cancer is one of the leading cancer developed in many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue. 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.The cervical cancer risk factor dataset was used to construct the classification model through a voting method that combines three classifiers: Decision Tree, K-N Neighbor and Random Forest.

Disadvantages:

- Accuracy is Low.
- Basic Algorithms are used with minimum efficiency.

IV. Proposed Methodology

This proposed system presents a comparison of machine learning (ML) algorithms: Support machine vector(SVM), Decision Tree(DT), Random Forest(RT), Artificial Neural Networks(ANN), Naive Bayes (NB), Nearest Neighbour (NN) search. The data-set used is obtained from the Wisconsin datasets. For the implementation of the ML algorithms, the dataset was partitioned into the training set and testing set. A comparison between all the six algorithms will be made. The algorithm that gives the best results will be supplied as a model to the website. The website will be made from a python framework, called flask. And it will host the database on Xampp or Firebase or inbuilt Python and flask libraries. This data set is available on the UCI Machine Learning Repository. It consists of 32 real world attributes which are multivariate. The total number of instances is 569 and there are no missing values in this data set. The project aims to implement a self-learning protocol such that the past inputs of the disease outcomes determine the future possibilities of the cervical disease to a particular user. The proposed model makes use of strong preprocessing tools so that the classification and prediction do not show any errors

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

relating to the dataset. A huge number of training sets will be used to make the prediction more and more accurate. Not only does the datasets but also the attributes to be used are selected taking into consideration the various important parameters and attributes.

4.1 Advantage

- Accuracy is High.
- Advanced Algorithms are used with Maximum Accuracy.
- Deployed the model as a Web page.

IV. DESCRIPTION OF MODULES

1. Dataset Collection
2. Hypothesis Definition
3. Data Exploration
4. Data Cleaning
5. Data Modelling
6. Feature Engineering

4.1 Dataset Collection

A data set is a collection of data. Departmental store data has been used as the dataset for the proposed work. Sales data has Item Identifier, Item Fat, Item Visibility, Item Type, Outlet Type, Item MRP, Outlet Identifier, Item Weight, Outlet Size, Outlet Establishment Year, Outlet Location Type, and Item Outlet Sales.

4.2 Hypothesis Definition

This is a very important step to analyse any problem. The first and foremost step is to understand the problem statement. The idea is to find out the factors of a product that creates an impact on the sales of a product. A null hypothesis is a type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations. An alternative hypothesis is one that states there is a statistically significant relationship between two variables.

4.3 Data Exploration

Data exploration is an informative search used by data consumers to form true analysis from the information gathered. Data exploration is used to analyse the data and information from the data to form true analysis. After having a look at the dataset, certain information about the data was explored. Here the dataset is not unique while collecting the dataset. In this module, the uniqueness of the dataset can be created.

4.4 Data Cleaning

In data cleaning module, is used to detect and correct the inaccurate dataset. It is used to remove the duplication of attributes. Data cleaning is used to correct the dirty data which contains incomplete or outdated data, and the improper parsing of record fields from disparate systems. It plays a significant part in building a model.

4.5 Data Modelling

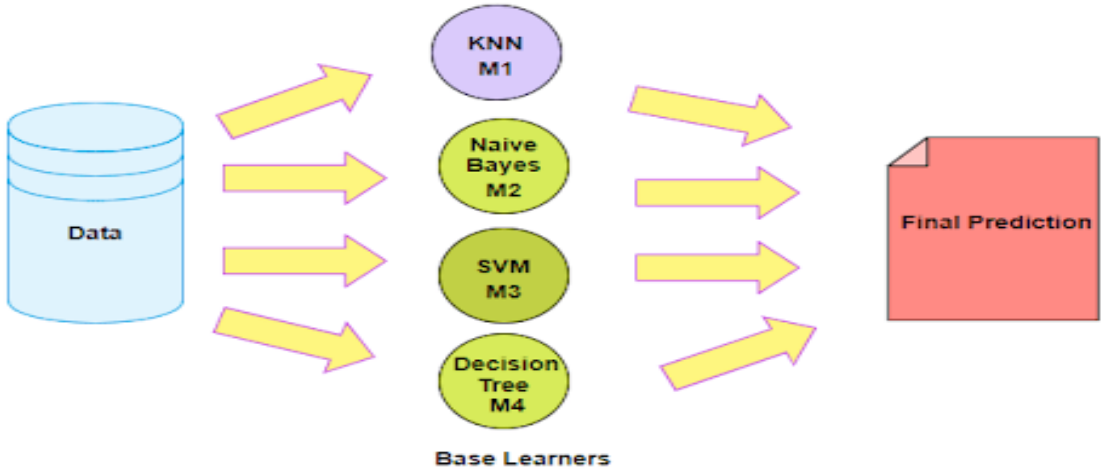
In data modelling module, the machine learning algorithms were used to predict the Wave Direction. Linear regression and K-means algorithm were used to predict various kinds of waves. The user provides the ML algorithm with a dataset that includes desired inputs and outputs, and the algorithm finds a method to determine how to arrive at that results. Linear regression algorithm is a supervised learning algorithm. It implements a statistical model when relationships between the independent variables and the dependent variable are almost linear, shows optimal results. This algorithm is used to show the direction of waves and its height prediction with increased accuracy rate.

K-means algorithm is an unsupervised learning algorithm. It deals with the correlations and relationships by analysing available data. This algorithm clusters the data and predicts the value of the dataset point. The train dataset is taken and are clustered using the algorithm. The visualization of the clusters is plotted in the graph.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

4.6 Feature Engineering

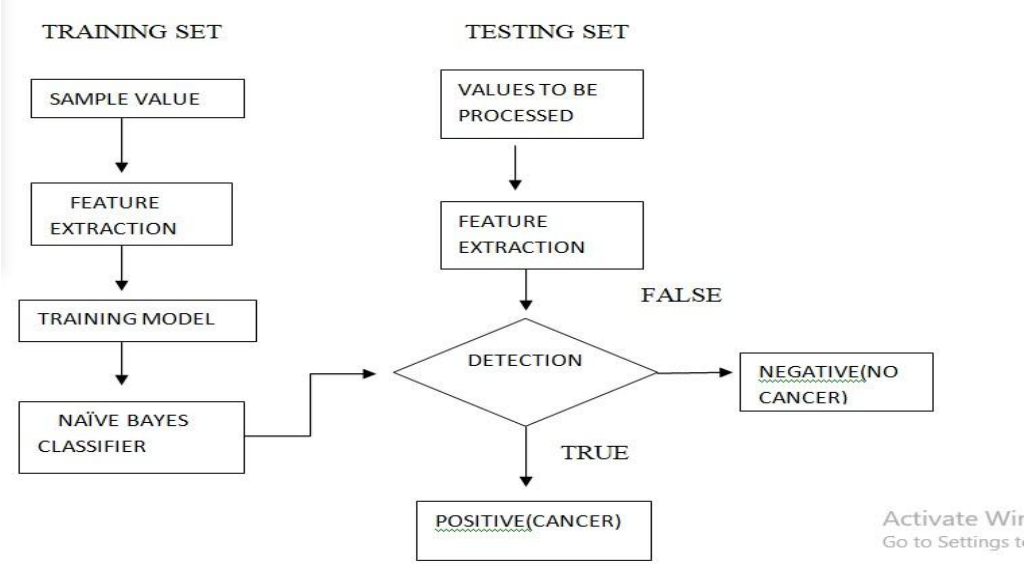
In the feature engineering module, the process of using the import data into machine learning algorithms to predict the accurate directions. A feature is an attribute or property shared by all the independent products on which the prediction is to be done. Any attribute could be a feature; it is useful to the model.



V. Machine Learning Algorithms Used

1. Support Vector Machines (SVM)
2. Logistic Regression
3. K Neighbors Classifier
4. Naive Bayers
5. Gradient Booster Classifier
6. Random Forest Classifier
7. Decision Tree Classifier

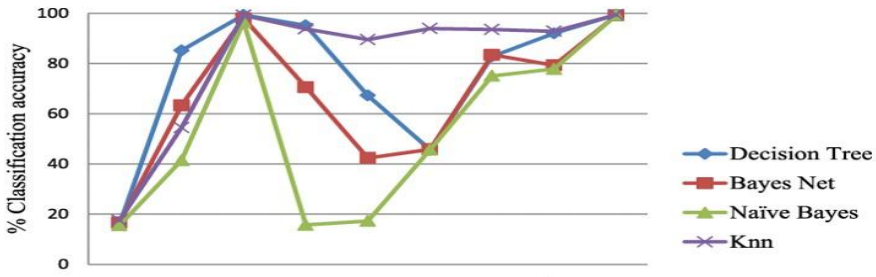
5.1 Data Flow Diagram



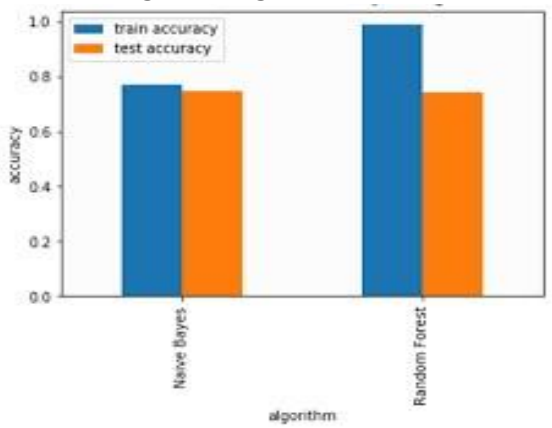
Activate Wir
Go to Settings t

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

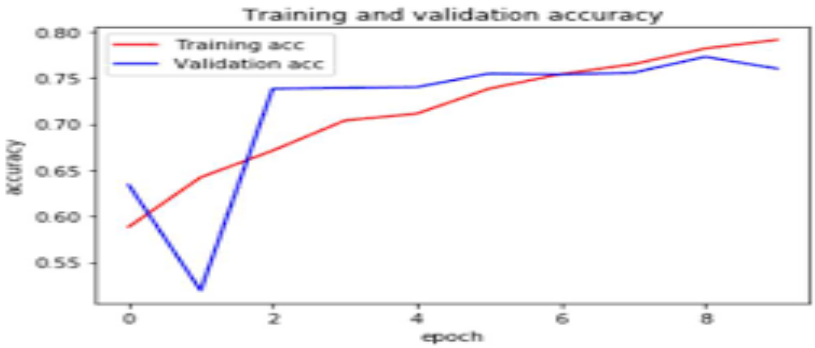
5.2 Implementation



5.3 Graph for Breast Cancer Prediction using These Algorithms



5.4 Graph for Cervical Cancer Prediction using These Algorithms



VI. Conclusion

Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, April 2022
Impact Factor: 6.252

REFERENCES

[1]. Shwetha K, Chaithra D , Sindhu .S "breast cancer and cervical cancer detection using deep learning technique",2018.
[2]. H. Abdel-Razeq, A. Mansour, and D. Jaddan, "Breast Cancer Care in Jordan," JCO global oncology, vol. 6, pp. 260-268, 2020.
[3]. P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal,Supervised deep learning embeddings for the prediction of cervical cancer diagnosis,2015 pp.1-8
[4]. K. Akyol, "A Study on Test Variable Selection and Balanced Data for Cervical Cancer Disease," Inf. Eng. Electron. Bus., vol. 5, pp. 1–7,2018.
[5]. F. S. Ahadi, M. R. Desai, C. Lei, Y. Li, and R. Jia, "Feature-Based classification and diagnosis of breast cancer using fuzzy inference system," in 2017 IEEE International Conference on Information and Automation (ICIA), 2017, pp. 517-522.
[6]. Gogate U. et al., 2018 implemented healthcare monitoring system for illness of cardiac patients [3]. On the cervical cancer, Snijders (P. J. F et al., 2006)

BIOGRAPHY

P. Bhuvaneswari is Assistant Professor in the Department of Information Technology at Hindusthan Institute of Technology at Hindusthan Institute Of Technology, Coimbatore. She Completed master degree Software engineering Sri Krishna College Of Engineering and Technology Coimbatore. She Completed her B.Tech IT Vivekananda Institute Of Engineering and Technology for Women,Nammakal in (2011).

Divya Dharshini D is a Final year student in the Department of Information Technology at Hindusthan Institute of Technolog.Her area of interest is in Machine Learning.

Srimathi M is a Final year student in the Department of Information Technology at Hindusthan Institute of Technolog.Her area of interest is in Block chain.

Sneka G is a Final year student in the Department of Information Technology at Hindusthan Institute of Technolog.Her area of interest is in Machine learning.

Dinesh Kumar P is a Final year student in the Department of Information Technology at Hindusthan Institute of Technolog. His area of interest is in Machine learning.