

#### A PRELIMINARY REPORT ON

**“BreastCancerDetectionUsingK*-*NearestNeighbours*,* LogisticRegressionandEnsembleLearning”**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY , PUNE IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

OF

## BACHELOR OF ENGINEERING (COMPUTER ENGINEERING)

#### SUBMITTED BY

PALASH DANDGE PRN: 72010297K

## DEPARTMENT OF COMPUTER ENGINEERING

#### R.M.D. SINHGAD SCHOOL OF ENGINEERING WARJE, PUNE 411058

**SAVITRIBAI PHULE PUNE UNIVERSITY**

**2019 - 2022**

**CERTIFICATE**

This is to certify that the project report entitled

#### Breast Cancer Detection using K-Nearest Neighbours , Logistic Regression and

#### Ensemble Learning

submitted by

#### PALASH DANDGE PRN : 72010297K

This is to certify that Ms. Palash Dandge from Third Year Computer Engineering has successfully completed her Seminar work titled “Breast Cancer Detection Using K – Nearest Neighbours , Logistic Regression and Ensemble Learning” at RMD Sinhgad School of Engineering, Warje, Pune in the partial fulfillment of the Bachelors Degree in Engineer- ing.

**Ms. Mrunalinee Patole External Examiner** Project Guide Department of Computer Engg. Department of Computer Engg.

#### Mrs. V.M.Lomte Dr. V.V.Dixit

Head Of Department Principal

Department of Computer Engg. RMDSSOE Warje,Pune-58

Place: Pune Date:

### Acknowledgments

*It gives us great pleasure in presenting the preliminary project report on* ***‘“Breast Cancer Detection Using K – Nearest Neighbours , Logistic Regression And Ensemble Learning ”.’****.*

*I would like to take this opportunity to thank my internal guide* ***Prof. Mrunalinee Patole*** *for giving me all the help and guidance I needed. I am really grateful to them for her kind support. Her valuable suggestions and directions were very helpful.*

*I am thankful* ***Prof. Grishma Bobhate*** *for providing us detailed instructions related to this study. Information about how this study should be done helped us alot.*

*I am also grateful to* ***Prof. Vina Lomte****, Head of Computer Engineering De- partment, RMD Sinhgad School of Engineering for her indispensable support.*

*Finally I wish to thank and appreciate to all our teachers and friends for their constructive comments, suggestions and guidance and to all those who directly or indirectly helped me in completing this seminar.*

Palash Dandge (PRN: 72010297K)

(B.E. Computer Engg.)

**Abstract**

Breast Cancer is one of the most severe disease s that is faced by women leading nowhere other than increased death rates in society and it is considered to be one of the most intense disease in the history of medical science. Looking at the number of deaths caused by Breast cancer, it is considered to be a major threat but today's advancement in medical science has the capability to cure such threat completely if detected at its early stages without causing any harm to the patient. The major challenge arise during the detection of cancer and differentiating between the diagnosis that affirms whether the patient has a benign or malignant type of cancer. Machine Learning Algorithms like K-Nearest Neighbors, Support Vector Machine (S VM) and Artificial Neural Network (ANN) helps us solve this problem by achieving results with high precision and accuracy. The following paper helps in diagnosis of breast cancer using Logistic Regression (LR), K-Nearest Neighbors (KNN) and Ensemble Learning with Principal Component Analysis (PCA) and a comparative study is also made with other papers on the basis of accuracy.

(Keywords— Breast Cancer Diagnosis, Logistic Regression (LR), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Principal Component Analysis (PCA), Ensemble Learning, Machine learning.)

**TABLE OF CONTENTS**

**Certificate**

**Acknowledgements**

**Abstract**

1. **Introduction**
   1. **Objectives**
   2. **Problem Statement**
2. **Literature Survey**
3. **Architecture of the System**
4. **Technology used in the System with models**
5. **Analytical Study**
   1. **Logistic Regression**
   2. **K-Nearest Neighbour**
   3. **Data Pre – Processing**
   4. **Performance Measure Indince**
6. **Visualization**
   1. **Head Map For Correlation of Attributes**
   2. **Variance Graph of Dataset**
7. **Case Study**
8. **Conclusion**
9. **Future scope**
10. **Bibliography : References**

**1 Introduction**

Cancer is considered to the most dangerous health problems leading to severe health conditions , cancer is the result of abnormal excess growth of cells in body which can cause severe health issues and can cause death. Breast cancer is a cancer which is a result of excess growth of breast tissue. Breast cancer is life threatening and is observed to be the second highest cause of deaths after lung cancer in women. About 1 in 8 women suffers from breast cancer once in their lifetime. Breast cancer alone lead to 627,000 deaths in 2018 with a frequency of 2.1 million in 2018 worldwide and more than 1 million cases are observed every year in India [1]. Breast cancer can be identified by observing presence lumps in breast.

* 1. **OBJECTIVES**

Cancer is considered to the most dangerous health problemsleading to severe health conditions , cancer is the result of abnormal excess growth of cells in body which can cause severe health issues and can cause death Breast Cancer is one of the most severe diseases that is faced by women leading nowhere other than increased death rates in society and it is considered to be one of the most intense disease in the history of medical science

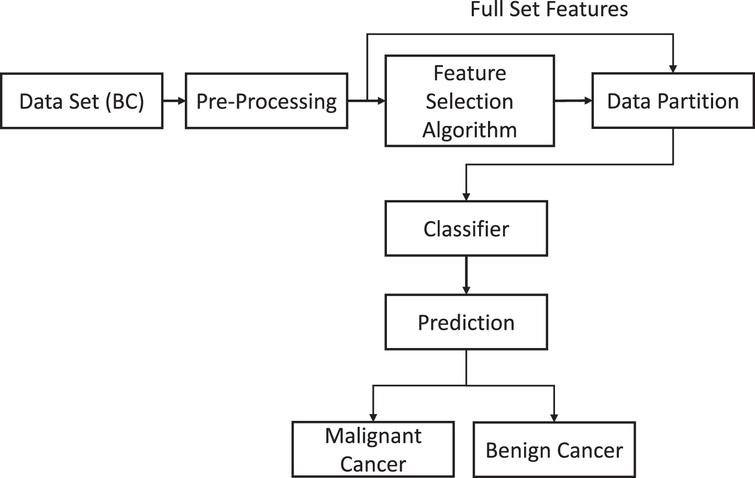
* 1. **Problem Statement**

To Develop a Machine Learning Model To Detect Weather the Women is Having Breast Cancer or Not . Radiologists can predict if the mammography images have cancer or not. we propose a new method to detect the breast cancer with high accuracy .

**2. Literature Survey**

The authors aimed to update a literature review on the effectiveness of breast cancer screening and diagnosis conducted in 1996. Their main objective was to assess the early diagnosis of breast cancer by primary care health professionals. The authors searched for English language studies published between January 1996 and October 1998. Searches were conducted on MEDLINE, EMBASE, HealthSTAR, Current Contents, DARE, NHS EED, the Cochrane Library and CINAHL; the search terms were listed in detail in the report. The reference lists of identified papers and documents published by members of the International Network of Agencies for Health Technology Assessment were also searched.. Systematic reviews with meta-analysis, randomised trials, cohort studies, cost-effectiveness studies, case-control studies, before-and-after studies, and cross-sectional descriptive and ecological studies were eligible for inclusion. The studies actually included in the review were randomised trials, cohort studies, case-control studies and other cross-sectional studies. Studies were excluded if they were published before January 1996 or after October 1998, had a participation rate lower than 50%, had a sample size less than 25, or did not clearly describe their methods or results. Studies were eligible for inclusion in the review if they assessed breast cancer screening and diagnosis techniques in primary care. Studies were included on the triple diagnostic test for breast cancer and its individual components (clinical examination, fine-needle aspiration and mammography), ultrasound, core biopsy and population-based breast cancer screening. The authors did not specify any inclusion criteria relating to the diagnostic reference standard. Some of the studies included in the review used surgery and long-term follow-up to assess the presence of breast cancer and survival ratesThe authors did not specify any inclusion criteria relating to the participants. Details of each study sample were tabulated in the report. The authors did not specify any outcomes that the studies had to assess in order to be eligible for inclusion. The main outcomes included in the review were sensitivity, specificity, and overall and disease-free survival. The authors rated the articles according to study design, based on a schedule developed by the New Zealand Guidelines Group of the National Health Committee (see Other Publications of Related Interest). The level of evidence was graded using an adaptation of the U.S. Preventive Services Task Force protocol (1989; full citation not provided). The articles were appraised using the validity schedules described. The authors did not state how many reviewers performed the validity assessment. The authors did not state how the data were extracted for the review, or how many reviewers performed the data extraction. Data were extracted on the level of evidence, country, sample size, participant characteristics, outcomes, main results and study limitations. he authors included 232 studies, of which 33 focused on evaluating diagnostic tools for identifying breast cancer (5 on fine-needle aspiration, 6 on the triple test, 6 on core biopsy and 7 on scintigraphic imaging). mammography is more sensitive than clinical examination for diagnosing breast cancer in women of all ages. Fine-needle aspiration: the sensitivity is generally high (over 90%), but the specificity varies (78 to 95%). Both the sensitivity and specificity vary depending on the placement of the needle. In lesions larger than 1 cm, core biopsy may be an alternative to fine-needle aspiration.Other tests: ultrasound is recommended as the first radiological investigation for women younger than 35 years. There was no strong evidence that breast self examination improves survival. There was insufficient evidence about the role of scintimammography, colour Doppler and magnetic resonance mammography in new breast lesions.

**3. Architecture of the System**

****

**4. TECHNOLOGY USED IN SYSTEM WITH MODEL**

**Python** is an interpreted high-level general-purpose programming language.Python's design philosophy emphasizes code readability with its notable use of significant Indentation.

**Python Packages :**

* **Pandas** – Highly Used for  data manipulation and analysis
* **Logistic Regression** - used for solving binary classification problem
* **K – Nearest Neighbour** - simplest forms of classification technique

**5. ANALYTICAL STUDY**

**5.1 LOGISTIC REGRESSION**

Logistic Regression is statistical model used for modelling binary classification problem using logistic function and many more complex extensions are exist for logistic regression.

Equation of Linear Regression is as follows: -

…………(1)

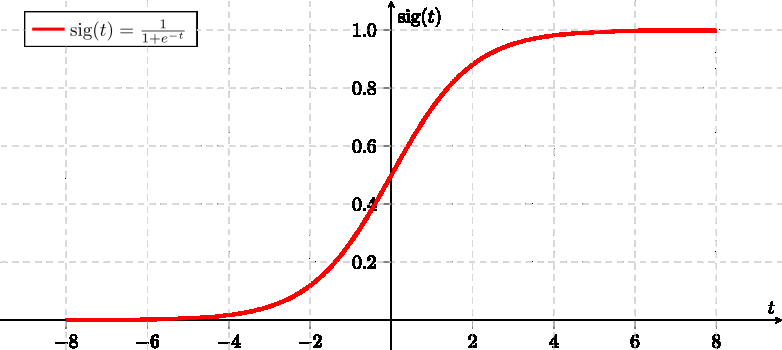
Where Y is the dependent variable and X1,X2…Xn are explanatory variables

**Sigmoid Function:-**

………… (2)

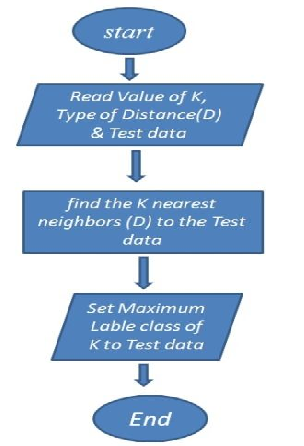
After applying sigmoid function in equation 2 on equation 1,

****



**5.2 K-Nearest Neighbour**

KNN (K-Nearest Neighbor) is one of the simplest forms of classification technique. In this technique, k training samples, whose attributes are relatively similar (closest) to the test samples.



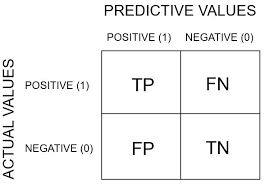
* 1. **Data Pre-Processing**

Data is pre-processed using various techniques. To get rid of defects and redundancy present in the data, pre-processing on the dataset has been done. It is an important to pre-process the dataset and remove the redundancy before training and testing the model. For our model, the defects in the WDBC dataset have been pre-processed and corrected.

The ‘diagnosis’ column give information about the class label. Out of the 32 columns in the dataset, the column named ‘id’ is the patient id for each patient in the dataset which do not serve any purpose in our classification process, therefore ‘id’ column is dropped from the dataset .

**5.4 Performance Measure Indices**

The performance of the model a confusion matrix is first computed



Performance is later measured using the formulas stated below-



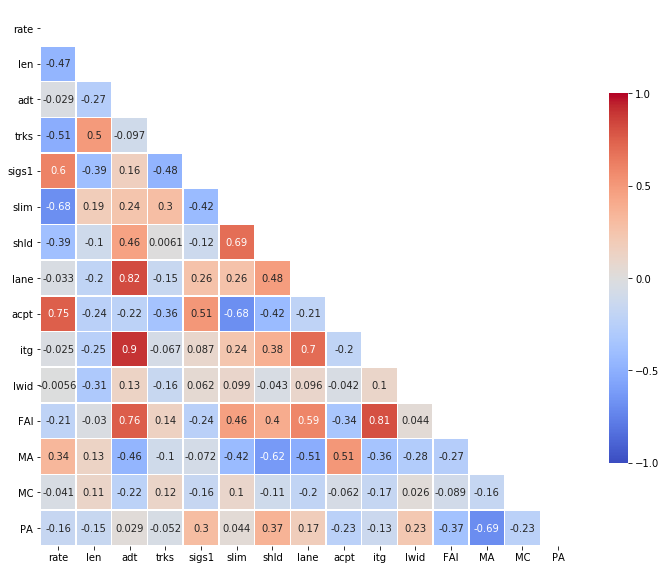


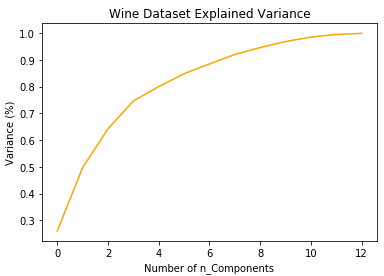




1. **Visualization**

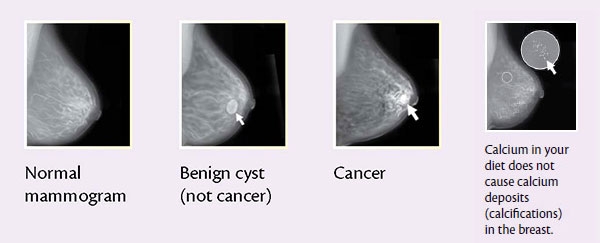
**6.1 Heatmap For Correlation of Attributes**



**6.2 Variance Graph of Dataset**

**7 Case Study**

Premenopausal woman, had a screening mammogram which revealed an abnormality in the right breast. It had no palpable masses on breast exam. A mammographically localized surgical biopsy was done and revealed a small (0.9 cm) grade III infiltrating ductal carcinoma with some associated ductal carcinoma-in-situ (DCIS). The surgical margins were not clear . Estrogen and progesterone receptors are negative. So, I used KNN & LR Analysis method in our proposed system

****

**8. CONCLUSION**

The aim is to classify the types of tumour , i.e. Benign tumour and Malignant tumour, various machine learning techniques allows us to find the most suitable model which is capable enough classify the tumour with high accuracy. Wisconsin breast cancer diagnosis (WDBC) data set is taken from UCI machine learning repository. Initially the data is pre-processed followed by applying Principal Component Analysis (PCA) with 17 components on the dataset. After which, different machine learning techniques like K-Nearest Neighbor, Logistic Regression and Ensemble Learning are applied and their results are evaluated using confusion matrix.

**9. FUTURE SCOPE**

The dataset consist of 569 instances and in future more data would be added to the database which would increase help in better training of machine learning models and would work more accurately, which will also brief us about the relationship among various attributes. After analysing the results using confusion matrix, Logistic Regression gave 97.90% accuracy and confusion matrix of Logistic Regression is shown in Figure 6. K-Nearest Neighbor gave 98.60% accuracy and its confusion matrix is shown in Figure 7. Ensemble Learning technique of 5 machine learning algorithms (Logistic Regression, K-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Classifier and Random Forest Classifier) gave the accuracy of 99.30%. Figure 8 shows the confusion matrix of Ensemble Learning.

**10. Bibliography : References**

1. WHO (World Health Organisation) Fact Sheet https://[www.who.int/en/news-room/fact-sheets/detail/cancer](http://www.who.int/en/news-room/fact-sheets/detail/cancer)
2. C. Chen, “ Curriculum Assessment Using Artificial Neural Network and Support Vector Machine Modeling Approaches: A Case Study,” Jan, 2010.
3. Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan, *"*Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors *"*, 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dec 2017.
4. Jinan Alhajjaj, Reem Alyami, Abdullah Alqahtani, Batool Alnajrani, Ilham Elaalami, Taoreed O. Owolabi , Nahier Aldhafferi, and Sunday

O. Olatunji, "Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines*"*, IEEE, 2017

1. Dan C. CireSan, Alessandro Giusti, Luca M. Gambardella, Jurgen Schmidhuber," Mitosis Detection in Breast Cancer Histology Images and Deep Neural Network", International Conference on Medical Image Computing and Computer-Assisted Intervention 2013, pp 411- 418
2. B. Santhi, R. Nithya, *"* Classification of Normal and Abnormal Patterns in Digital Mammograms for Diagnosis of Breast Cancer *"*,

Aug. 2011

1. G. Manikandan, B. Karthikeyan , P. Rajendiran , R. Harish , T. Prathyusha , V. Sethu, " Breast Cancer Prediction Using Ensemble Techniques", Indian Journal of Public Health Research & Development, July 2019, Vol.10, No. 7
2. S. k. Mandal, A. Gupta and Animesh Hazra, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", IJCA (0975 – 8887) Volume 145 – No.2, July 2016
3. UCI Machine Leaning Repository, Center of Machine Learning and Intelligent Systems,

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Dia

gnostic)

1. Ram Murti Rawat, Shivam Panchal, Vivek Kumar Singh, Yash Panchal, "Breast Cancer Detection Using Support Vector Machine with Principal Component Analysis", IJCRT, March 2020, Volume 8, Issue 3, ISSN: 2320-2882.
2. Ms. Steffi Thomas, Mr. Atharva Joshi, Ms. Rutu Kalhapure, Mr. Deepraj Bhosale, Prof. Deepali Sonawane, 2020, Bionic ARM for Prosthetist, International Journal Of Engineering Research & Technology (IJERT) ICSITS – 2020 (Volume 8 – Issue 05)