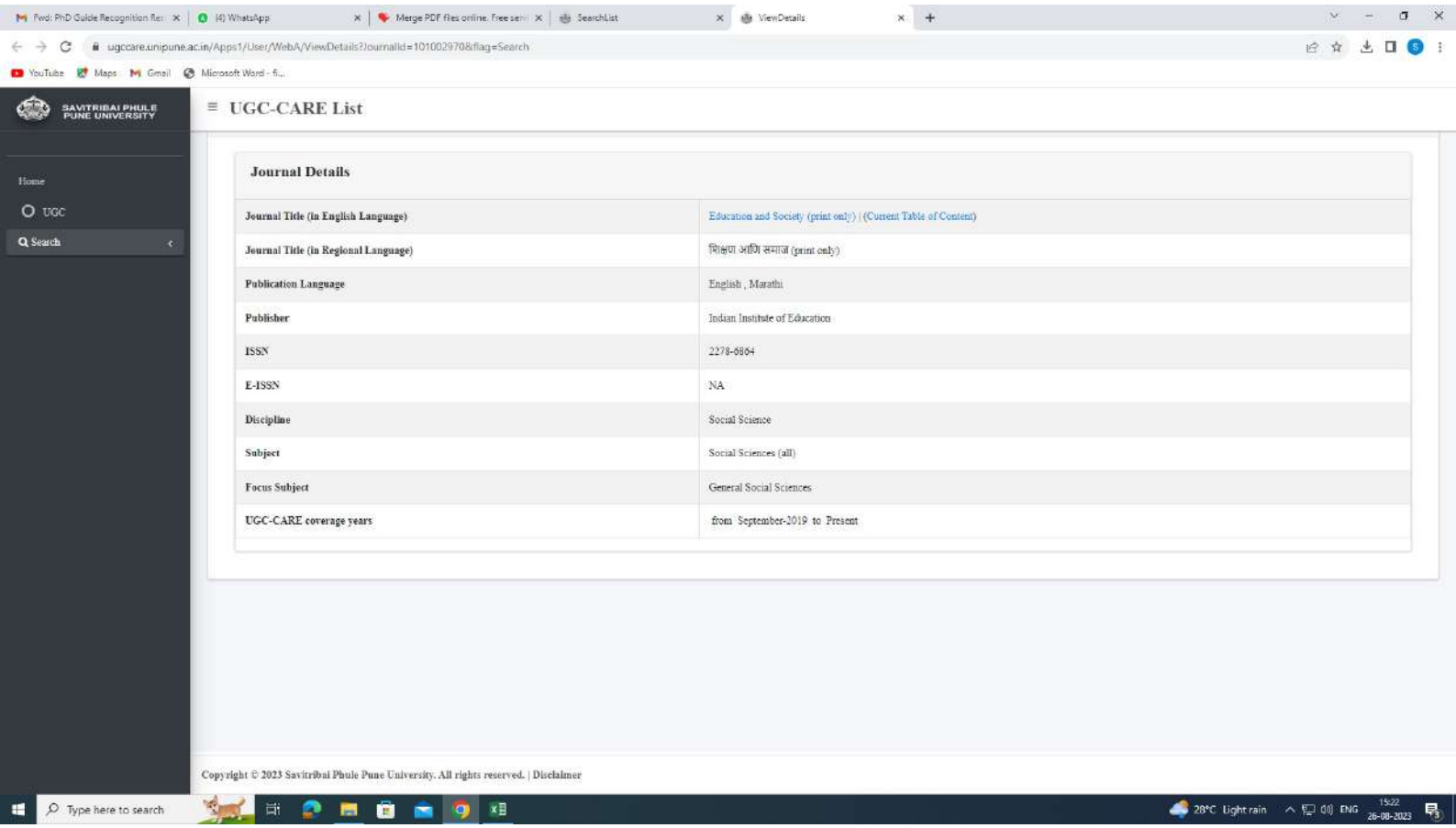

Fwd: PhD Guide Recognition Re
(4) WhatsApp
Merge PDF files online. Free ser
SearchList
ViewDetails
ugccare.unipune.ac.in/Apps1/User/WebA/ViewDetails?journalId=101002970&flag=Search
YouTube
Maps
Gmail
Microsoft Word - fi...
SAVITRIBAI PHULE PUNE UNIVERSITY
UGC-CARE List
Home
UGC
Search
Journal Details
Journal Title (in English Language)
Education and Society (print only) | (Current Table of Content)
Journal Title (in Regional Language)
शिक्षण आणि समाज (print only)
Publication Language
English , Marathi
Publisher
Indian Institute of Education
ISSN
2278-6864
E-ISSN
NA
Discipline
Social Science
Subject
Social Sciences (all)
Focus Subject
General Social Sciences
UGC-CARE coverage years
from September-2019 to Present
Copyright © 2023 Savitribai Phule Pune University. All rights reserved. | Disclaimer
Type here to search
28°C Light rain
ENG
15:22
26-08-2023

शिक्षण आणि समाज

SAVITRIBAI PHULE PUNE UNIVERSITY
सावित्रीबाई फुले पुणे विद्यापीठ
॥ यः क्रियावान् स पण्डितः॥

Education and Society

CERTIFICATE OF PUBLICATION

This is to certify that the article entitled

COMPARATIVE ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS USED IN BREAST CANCER PREDICTION

Authored By

Dr. Sachin Misal, Assoc. Professor, IIMS Chinchwad

Published in
Education and Society (**शिक्षण आणि समाज**) : ISSN **2278-6864** with IF=6.718
Vol. 47, Issue 01, No. 10, January - March : 2023
UGC CARE Approved, Group I, Peer Reviewed,
Bilingual, Multi-disciplinary Referred Journal

ज्ञान-विज्ञान विमुक्तये
UGC
University Grants Commission

शिक्षण आणि समाज
Education and Society

Editor
Shikshan Ani Samaj

UGC CARE LISTED PERIODICAL
ISSN : 2278 - 6864

Education and Society

Since 1977

Vol-47, Issue-1, No.-10, Jan - Mar : 2023

Indian Institute of Education

INDEX

19 IMPACT 0F CLIMATE CHANGE ON AGRICULTURE IN TAMIL NADU

20 INTRODUCTION TO ANIMATION

21 ANALYSIS OF CASH FLOW AT BHARATH EARTH MOVERS LIMITED, KGF - A STUDY

22 INFLUENCE OF SOCIAL MEDIA FOR ACADEMIC BENEFITS: A STUDY AMONGPOST-GRADUATE STUDENTS AND RESEARCH SCHOLARS OF THE UNIVERSITY OF SCIENCE TECHNOLOGY, MEGHALAYA

23 CAPITAL MARKET DEVELOPMENTS AND ITS REFORMS IN INDIA

24 IMPACT OF ECONOMIC FACTORS ON THE STOCK PRICE BEHAVIOR OF SELECTED LARGE MARKET CAPITALIZTION COMPANIES IN NSE

SAVITRIBAI PHULE PUNE UNIVERSITY

COMPARATIVE ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS USED IN BREAST CANCER PREDICTION

Mr. Prashant Wadkar, Research Scholar, IIMS Chinchwad
Dr.ShivajiMundhe, Research Guide, IIMS Chinchwad
Dr. Sachin Misal, Assoc. Professor, IIMS Chinchwad
pnwadkar@gmail.com; drshivaji.mundhe@gmail.com; missal.sachin@gmail.com

Abstract:
Machine learning has the concept of experiential learning which recognize patterns and make accurate predictions of future events. It has been also utilized in various sectors including the health sector. Machine learning is used to find the abnormalities at an early stage of various type of disease. The various Machine Learning algorithms have been given different types of accuracy for the model created for the same data set. For the present study the researcher has used the secondary data of breast cancer patients. The exploratory analysis of this dataset has been done and further the MLmodel has been created by the Researcher.In this research, the comparative analysis of different ML algorithms has been done, for this the machine has been trained and tested by three ML algorithms such as Logistic Regression, Support Vector Machine and K-Nearest Neighbour Algorithm. The accuracy of the each model has compared. The confusion matrix in this research has been used to check the accuracy of the model, and found Support Vector Machine (SVM) is the best with higher accuracy for the current data set. For the present study the Python Language and its various libraries has been used to create a model.

Key Words: *Machine Learning, Python, Breast Cancer, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbour Algorithms, sklearn, Confusion matrix.*

I. INTRODUCTION

In Breast cancer, the breastcells grows fast which cannot be control. The breast cancer is the most common typeof cancer in women and the highest mortality rate. If a suspicious lump isfound in x-ray, the doctor normally conducts a diagnosis todetermine whether it is cancerous and, if so, whetherit has spread to other parts of the body or not, and at what stage it is. For the present study the breastcancer dataset has been obtained from the University ofWisconsin Hospitals, Madison from Dr. William H.Wolberg, USA.At the beginning of the study the attempt has been made to do the exploratoryanalysis of the dataset, also checked for weather the dataset is clean. It has found that the dataset is clean, so no pre-processing is required. The 'Diagnosis' column found as the dependent column and rest all as independent columns. By commencing with Train, Test, the three ML models created by the researchers. Further the accuracyof the each algorithm has been checked and tried to find out the best ML model Algorithm. The Machine learningalgorithmswhich has been used for creating the model wereLogistic regression, SVM and KNN.

II. OBJECTIVE OF THE STUDY

1. To perform the exploratory data analysis of the breast cancer dataset.
2. To create the ML model by using Logistic Regression, Support Vector Machine and K-Nearest Neighbour Algorithm.
3. To do the comparative analysis of above model created.
4. To check the accuracy of the Machine learning model.
5. To find out best, accurate ML model for breast cancer diagnosis.

III TECHNOLOGY USED

The Python, which is dynamic programming language is used by the researcher. Also the different Python libraries has been used for data visualization, performing statistical operations, analysis, creating the model, such as pandas, numpy, seaborn, matplotlib, sklearn etc bythe researcher. The

Machine Learning Algorithms like Logistic Regression, Support Vector Machine and K-Nearest Neighbour Algorithm have been used for the present study for creating the model.

IV. RESEARCH METHODOLOGY ADOPTED

The Secondary data has been used for doing theresearch and it has been taken from the genuine and renowned website www.kaggle.com. Breast cancer dataset has been obtained from the University of Wisconsin Hospitals, Madison from Dr. William H.Wolberg, USA. The dataset has 569 rows and 6 columns.

V. DATA ANALYSIS AND INTERPRETATION

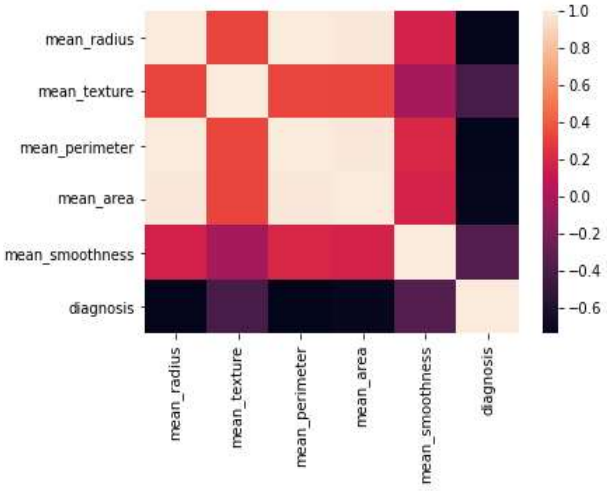
It has been seen that the cancer dataset has sixcolumns namely mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness and diagnosis as shown in Figure. 1.

Figure 1 . Breast Cancer Dataset shown in Dataframe.

```
In [5]: df['diagnosis'].value_counts()

Out[5]: 1    357
        0    212
        Name: diagnosis, dtype: int64
```

Figure 2. The dependent/target column ‘diagnosis’ showingthe two categories and count of each.

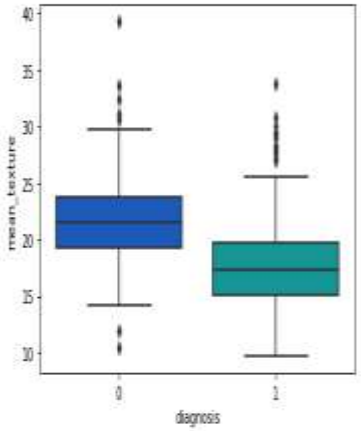
Out of 569 there are 357 cancerpositive patients (categorized as ‘1’) and 212 as cancer negativepatients (categorized as ‘0’). The dependent column and independent columnsfrom the dataset has been identified and assumed that the ‘diagnosis’ asthe dependent (target) column and rest columns asindependent column.



Graph 1. Correlation between parameters of the dataset

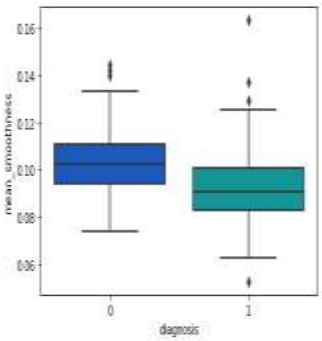
Graph 1. Shows the strong correlation between thedependent column 'diagnosis' and mean_texture,mean_smothness, so researcher further proceed forboxplot to do further exploratory analysis to studytheir impact on cancer patients and noncancerouspatients.

```
[5]: sns.boxplot(x='diagnosis',y='mean_texture',data=df, palette='winter')

[5]: <AxesSubplot:xlabel='diagnosis', ylabel='mean_texture'>
```



Graph 2. Box plot: diagnosis v/s mean_texture

```
In [16]: sns.boxplot(x='diagnosis',y='mean_smoothness',data=df, palette='winter')

Out[16]: <AxesSubplot:xlabel='diagnosis', ylabel='mean_smoothness'>
```



Graph 3. Box plot: diagnosis v/smean_smoothness

Graph 2 and Graph 3 have been drawn assuming that themean_texture and mean_smoothness are playing aimportant role and affecting the values of the dependentcolumn as 'diagnosis'. We can say "lower the value ofmean_texture and/or lower the mean_smoothnestends to be cancerous".

Later the researcher, proceed further for training the Machine Model and testing the Machine Learning Model by 70%, 30% of data respectively by making use of the below ML Algorithms.

Algorithms used to Train the Machine and to create the ML Model

1. Logistic Regression Algorithm
2. Support Vector Machine Algorithm(SVM)
3. K-Nearest Neighbor Algorithm (KNN)

1. Logistic Regression

Figure 3. Shows how the model is trained using Logistic Regression

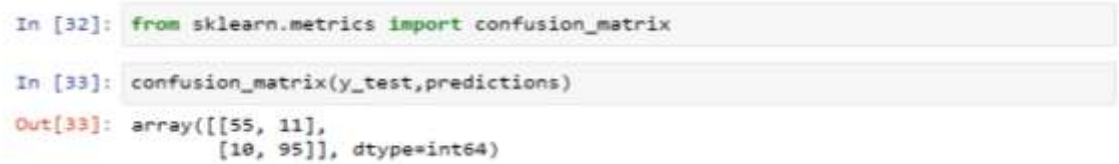


Figure 4. Confusion metrics applied after training and testing of the model by using Logistic Regression Algorithm

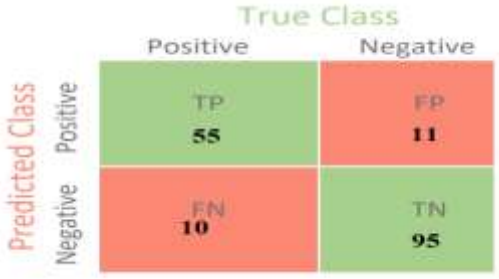


Figure 5. Confusion Matrix with their values

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The classifier made a total of 569 predictions. So the model has been checked for accuracy with the help of a confusion matrix. The Figure 5 shows the 55 and 95 as the TRUE Values (and 10 and 11 as FALSE values)

Accuracy = TP+TN/TOTAL
Accuracy = 55+95/171= 87 %

Inaccuracy =FP+FN/TOTAL
Inaccuracy = 11+10/171 = 13 %

Logistic Regression Showed Accuracy: 87%

2.Support Vector Machine Algorithm (SVM)

Figure 6 :Appllied SVM and created Model

Figure 7. Model trained using SVM

Accuracy = TP+TN/TOTAL
Accuracy = 59+100/171= 92.98 %

Inaccuracy =FP+FN/TOTAL
Inaccuracy = 1+11/171 = 7.02 %

Support Vector Machine Algorithm (SVM) Showed Accuracy: 92.98 %

3. K-Nearest Neighbour (KNN)

Figure:8 Applied KNN and created Model

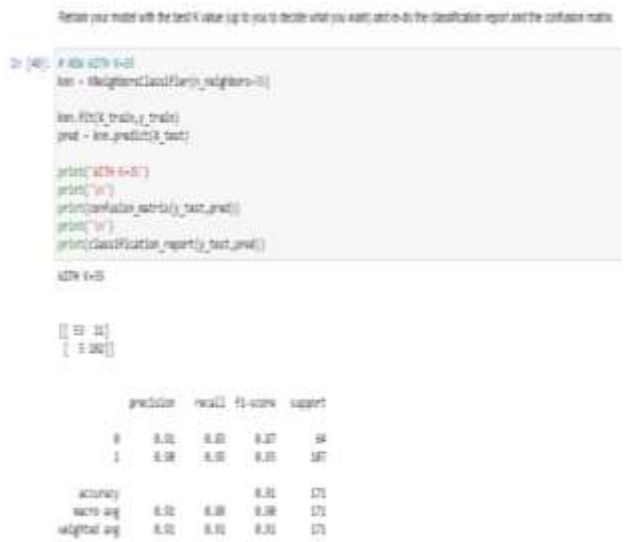


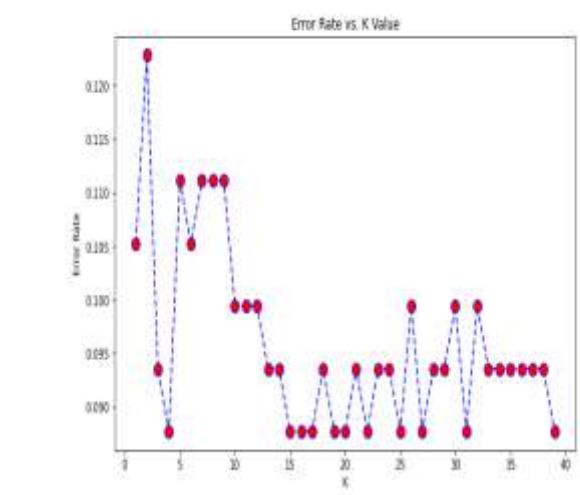
Figure 9. Accuracy of KNN

Figure10. Choosing K Value for KNN

```
In [38]: import matplotlib.pyplot as plt
         %matplotlib inline
```

```
In [39]: plt.figure(figsize=(10,6))
         plt.plot(range(1,40),error_rate,color='blue', linestyle='dashed', marker='o',
                 markerfacecolor='red', markersize=10)
         plt.title('Error Rate vs. K Value')
         plt.xlabel('K')
         plt.ylabel('Error Rate')

Out[39]: Text(0, 0.5, 'Error Rate')
```

Figure:11 For Plotting K value v/s Error Rate



Graph 4: K value v/s Error Rate

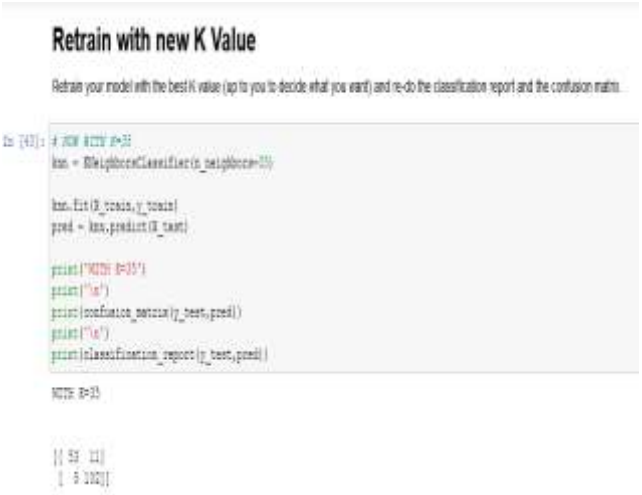


Figure 12. Model trained using KNN

K-Nearest Neighbour (KNN) Showed Accuracy: 91 %
Comparisons of ML Algorithms used with Accuracy

Table 1. Comparison of ML Algorithms used, with Accuracy

Logistic Regression	SVM	KNN
87%	**92.98 %**	91%

When researcher compared all three Model's accuracy by using confusion matrix, it hasfound that the SVM is proved with highest accuracy 92.98 %. The next KNN found thesecond best with 91% accuracy and the third Logistic Regression with 87% accuracy.

VI. FINDINGS

1. It has been found that the Machine learning has played vital role in prediction breast cancer in early stage.
2. The **Logistic Regression** proved with 87% accuracy.
3. The **SVM** proved with **92.98 %** accuracy.
4. The **KNN** proved with91%accuracy.
5. The SVM algorithm is proved best for this study due to highest accuracy among all three.
6. While doing this project it seems that python and its libraries played vital roles in EDA, Visualization and creating ML Model. And also proved for predicting Breast cancer in early stage.

It seems that for predicting the cancerouspatientmanually is very difficult and time consuming. Theabove research by using Machine Learningovercomes these problems due to speed andaccuracy.

VII.SUGGESTION AND RECOMMENDATIONS

It has been seen that the Machine Learning model seems to be best for predicting the early stage of cancer. So the researcher would suggest as further enhancement in present study with the implementation of best model (SVM) as live on website, so any doctor or patient can give the parameters of patients and check the early stage of cancer, and hence treatment can be commenced in advance to avoid death loss.

VIII. CONCLUSION

It seems that for predicting the cancerous patient manually is very difficult and time consuming. The above research by using Machine Learning overcomes these problems due to speed and accuracy. The above research has been proved how accurately the Machine learning model gives the accurate results. So this model can be used for early detection of cancerous patient and avoid the death of

patient and save the life of people. This study will help a lot to the doctors, patients and people for early diagnosis of cancer so it found to be beneficial to the society and hence it has been proved the significance of the study.

REFERENCES

Book reviews

1. Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python by ManoharSwamynathan
2. Machine Learning by Tom Mitchell

A. Journals, Articles and News.

1. Rawal, R. (2020). Breast cancer prediction using machine learning. Journal of Emerging Technologies and Innovative Research (JETIR), 13(24), 7.Wang, D. Zhang and Y. H. Huang "Breast Cancer Prediction Using Machine Learning" (2018), Vol. 66, NO.
2. Akbugday, B. (2019, October). Classification of breast cancer data using machine learning algorithms. In 2019 Medical technologies congress (TIPTEKNO) (pp. 1-4). IEEE.
3. Keleş, M. K. (2019). Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Tehničkivjesnik, 26(1), 149-155.
4. Chaurasia, V., & Pal, S. (2014). Data mining techniques: to predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 3(1), 10-22.
5. Anji Reddy, V., Soni, B. (2020). Breast Cancer Identification and Diagnosis Techniques. In: Rout, J., Rout, M., Das, H. (eds) Machine Learning for Intelligent Decision Science. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-3689-2_3
6. 'WHO | Breast cancer', *WHO*. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).
7. Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., &Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, 487-492.
8. Sethy, P. K., Pandey, C., Khan, M. R., Behera, S. K., Vijaykumar, K., &Panigrahi, S. (2021). A cost-effective computer-vision based breast cancer diagnosis. Journal of Intelligent & Fuzzy Systems, 41(5), 5253-5263.
9. Kanchana, M. Classification of breast cancer with mammogram images using various transformations and machine learning techniques.
10. Analysis Of Breast Cancer Dataset And It's Prediction Using Machine Learning Authored By Prof.Prashant N. Wadkar, Dr.Sachin Misal, SanketMundheicimit-22, Feb 2022, ISBN978-81-927230-0-10, International Institute Of Management Science. (IIMS).
11. Analysis Of Breast Cancer Dataset And It's Prediction Using Machine Learning authored by Prof.Prashant N. Wadkar, Dr. Sachin Misal, SanketMundhe, Yashomanthan (ISSN : 2347-8039),a peer reviewed multidisciplinary research journal2022, IMPACT FACTOR : 6.692, Vol | Issue | 2022, International Institute of Management Science. (IIMS).

C. Web References:

1. www.Kaggle.com
2. https://www.sciencedirect.com/science/article/pii/S2405959520300801
3. https://scholar.google.co.in/citations?view_op=view_citation&hl=en&user=MsqZt-4AAAAJ&citation_for_view=MsqZt-4AAAAJ:UeHWp8X0CEIC
4. https://reader.elsevier.com/reader/sd/pii/S1877050921014629?token=2400447BD4A06840160A8E13498693F1FE2E0A4231A00559F405A0FFE4B66773A7EF3D6928935BCCB0AE52A65464B155&originRegion=eu-west-1&originCreation=20220313083617
5. https://www.researchgate.net/publication/348019215_A_Cost-Effective_Computer-Vision_Based_Breast_Cancer_Diagnosis
6. http://hdl.handle.net/10603/253116
7. https://shodhganga.inflibnet.ac.in/handle/10603/253116
8. http://hdl.handle.net/10603/210351