

### UGC-CARE List

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	2278-6904
E-ISSN	NA
Discipline	Social Science
Subject	Social Sciences (all)
Focus Subject	General Social Sciences
UGC-CARE coverage years	from September-2019 to Present

शिक्षण आणि समाज



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**

UGC  
University Grants Commission

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



**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**

- 1 BIG DATA ANALYTICS : APPLICATIONS, TOOLS , CHALLENGES AND FUTURE INNOVATIONS
- 2 ROLE AND CHALLENGES OF SKILL DEVELOPMENT FOR APPRENTICESHIP IN MANUFACTURING INDUSTRIES W.R.T. PUNE CITY
- 3 STEGANOGRAPHY: APPLICATIONS, TECHNIQUES AND SECURITY CHALLENGES AND OPPORTUNITIES
- 4 A GUIDE ON SAVING AND INVESTMENTS FOR YOUTH OF INDIA
- 5 INDIA'S SUSTAINABLE BANKING SYSTEM - A CASE STUDY OF THE STATE BANK OF INDIA
- 6 COMPARATIVE ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS USED IN BREAST CANCER PREDICTION
- 7 A STUDY ON THE SOCIO- ECONOMIC CONDITIONS OF THE MATCH INDUSTRY WORKERS WITH SPECIAL REFERENCE TO SIVAKASI TALUK
- 8 GROUND REALITIES ABOUT SCHEDULED CASTES IN THE STUDY AREA REGARDING LAND REFORMS POLICIES
- 9 AN OVERVIEW OF SUSTAINABLE ECONOMIC DEVELOPMENT OF POULTRY PRODUCTION AND ITS EFFECT ON HUMAN HEALTH
- 10 CONFERENCES THEME TRACK-4 SUSTAINABLE / GREEN INVESTING AIR QUALITY INDEX DURING COVID -19 IN SALEM DISTRICT OF TAMIL NADU
- 11 DEVELOPMENT OF DIGITAL PAYMENT PLATFORM AFTER DEMONETISATION IN INDIA
- 12 A CASE STUDY OF CORPORATE SOCIAL RESPONSIBILITY (CSR) IN TATA GROUP OF INDUSTRIES
- 13 IMPACT OF MONETARY INCENTIVES ON EMPLOYEES' PERFORMANCE IN EDUCATIONAL INSTITUTIONS OF NEW DELHI
- 14 A STUDY ON SUPPLY CHAIN MANAGEMENT IN THE CONTEXT OF FOOD AND DRINK IN A STAR HOTEL IN WEST BENGAL
- 15 SUSTAINABILITY PRACTICES ADOPTED BY COMMUNITY HELPERS IN DEALING WITH COVID-19 IN INDIA
- 16 ✓ ICT IS AN EMERGING TECHNOLOGY IN THE HIGHER EDUCATIONAL INSTITUTE
- 17 SUSTAINABILITY MEASURES FOR THE MANGO POST HARVEST MANAGEMENT
- 18 SUSTAINABLE DEVELOPMENT: PROBLEMS AND PROSPECTS

- 19 IMPACT OF 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 CAPITALIZATION COMPANIES IN NSE



SAVITRIBAI PHULE PUNE UNIVERSITY

सवित्रीबाई फुले पुणे विद्यापीठ

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

**Mr. Prashant Wadkar**, Research Scholar, IIMS Chinchwad

**Dr. Shivaji Mundhe**, Research Guide, IIMS Chinchwad

**Dr. Sachin Misal**, Assoc. Professor, IIMS Chinchwad

[pnwadkar@gmail.com](mailto:pnwadkar@gmail.com); [drshivaji.mundhe@gmail.com](mailto:drshivaji.mundhe@gmail.com); [missal.sachin@gmail.com](mailto: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 ML model 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 breast cells grow fast which cannot be controlled. Breast cancer is the most common type of cancer in women and has the highest mortality rate. If a suspicious lump is found in an x-ray, the doctor normally conducts a diagnosis to determine whether it is cancerous and, if so, whether it has spread to other parts of the body or not, and at what stage it is. For the present study the breast cancer dataset has been obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg, USA. At the beginning of the study the attempt has been made to do the exploratory analysis of the dataset, also checked for whether 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 accuracy of the each algorithm has been checked and tried to find out the best ML model Algorithm. The Machine learning algorithms which has been used for creating the model were Logistic 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 by the 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 the research 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 six columns namely mean\_radius, mean\_texture, mean\_perimeter, mean\_area, mean\_smoothness and diagnosis as shown in Figure. 1.

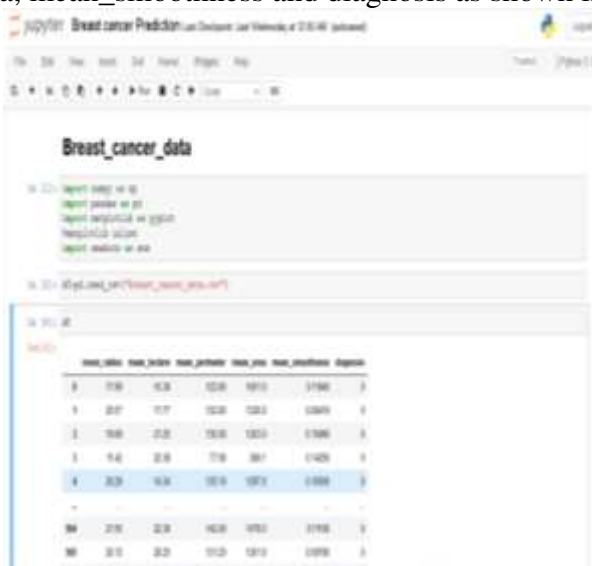


Figure 1 . Breast Cancer Dataset shown in Dataframe.

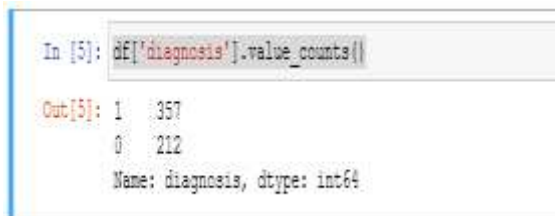
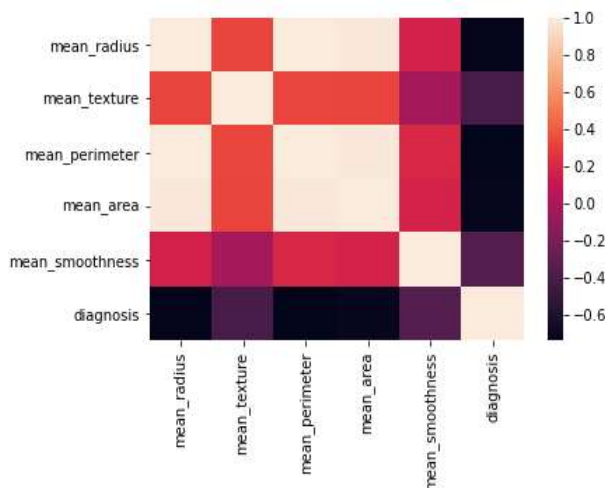


Figure 2. The dependent/target column 'diagnosis' showing the two categories and count of each.

Out of 569 there are 357 cancer positive patients (categorized as '1') and 212 as cancer negative patients (categorized as '0'). The dependent column and independent columns from the dataset has been identified and assumed that the 'diagnosis' as the dependent (target) column and rest columns as independent column.

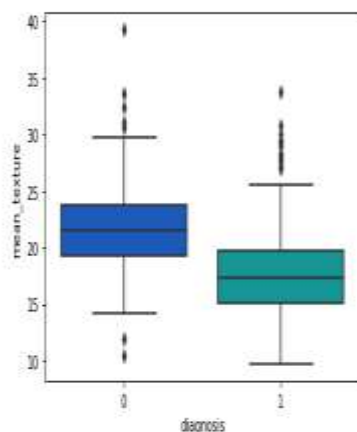




Graph 1. Correlation between parameters of the dataset

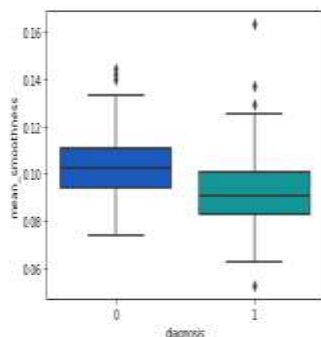
Graph 1. Shows the strong correlation between the dependent column 'diagnosis' and mean\_texture, mean\_smoothness, so researcher further proceed for boxplot to do further exploratory analysis to study their impact on cancer patients and noncancerous patients.

```
[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 [14]: sns.boxplot(x='diagnosis', y='mean_smoothness', data=df, palette='winter')
Out[14]: <AxesSubplot: xlabel='diagnosis', ylabel='mean_smoothness'>
```



Graph 3. Box plot: diagnosis v/s mean\_smoothness

Graph 2 and Graph 3 have been drawn assuming that the mean\_texture and mean\_smoothness are playing an important role and affecting the values of the dependent column as 'diagnosis'. We can say "lower the value of mean\_texture and/or lower the mean\_smoothness tends 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

```
In [17]: from sklearn.model_selection import train_test_split

In [24]: x_train,x_test,y_train,y_test=train_test_split(df.drop('diagnosis',axis=1),df['d

In [25]: from sklearn.linear_model import LogisticRegression

In [26]: logmodel=LogisticRegression()
logmodel.fit(x_train,y_train)

Out[26]: LogisticRegression()
```

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

```
In [32]: from sklearn.metrics import confusion_matrix

In [33]: confusion_matrix(y_test,predictions)

Out[33]: array([[55, 11],
               [10, 95]], dtype=int64)
```

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

		True Class	
		Positive	Negative
Predicted Class	Positive	TP <b>55</b>	FP <b>11</b>
	Negative	FN <b>10</b>	TN <b>95</b>

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)

$$\text{Accuracy} = \frac{TP+TN}{\text{TOTAL}}$$

$$\text{Accuracy} = \frac{55+95}{171} = 87 \%$$

$$\text{Inaccuracy} = \frac{FP+FN}{\text{TOTAL}}$$

$$\text{Inaccuracy} = \frac{11+10}{171} = 13 \%$$

**Logistic Regression Showed Accuracy: 87%**

## 2.Support Vector Machine Algorithm (SVM)

```
In [7]: from sklearn.model_selection import train_test_split

In [9]: x_train, x_test, y_train, y_test = train_test_split(data['diagnosis'], data['diagnosis'], test_size = 0.3)

In [10]: from sklearn.svm import SVC

In [11]: classifier = SVC()

In [12]: classifier.fit(x_train, y_train)

Out[12]: SVC()
```

Figure 6 :Applied SVM and created Model

```
In [16]: from sklearn.metrics import confusion_matrix

In [17]: confusion_matrix(y_test, ypred)

Out[17]: array([[ 59,  11],
               [  1, 100]], dtype=int64)
```

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)

```
Train Test Split
Use train_test_split to split your data into a training set and a testing set

In [26]: from sklearn.model_selection import train_test_split

In [27]: x_train, x_test, y_train, y_test = train_test_split(data['diagnosis'],
                                                            data['diagnosis'],
                                                            test_size=0.30)

Using KNN
Import KNeighborsClassifier from sklearn

In [28]: from sklearn.neighbors import KNeighborsClassifier

Create a KNN model instance with k_neighbors=1

In [29]: knn = KNeighborsClassifier(n_neighbors=1)

Fit the KNN model to the training data

In [30]: knn.fit(x_train, y_train)

Out[30]: KNeighborsClassifier(n_neighbors=1)

In [31]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                             metric_params=None, n_jobs=-1, n_neighbors=1, p=1,
                             weights='uniform')

Out[31]: KNeighborsClassifier(n_jobs=1, n_neighbors=1)
```

Figure:8 Applied KNN and created Model

```

Return your model with the best K value up to you to decide what you want, and to do the classification report and the confusion matrix

In [40]: # KNN k=5
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
pred = knn.predict(X_test)

print("KNN k=5")
print("\n")
print(confusion_matrix(X_test, pred))
print("\n")
print(classification_report(X_test, pred))

KNN k=5

[[ 50  15]
 [ 15  50]]

      precision    recall  f1-score   support

 0       0.77     0.80     0.78       65
 1       0.80     0.77     0.78       65

 accuracy         0.78     0.80     0.78      130
 macro avg         0.78     0.78     0.78      130
 weighted avg         0.78     0.78     0.78      130

```

Figure 9. Accuracy of KNN

### Choosing a K Value

Let's go ahead and use the elbow method to pick a good K value!

\* Create a for loop that trains various KNN models with different K values, then keep track of the error\_rate for each of these models with a list. Refer to the lecture if you are confused on this step \*

```

In [37]: error_rate = []

# Will store some time
for k in range(1, 41):

    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    pred_k = knn.predict(X_test)
    error_rate.append(np.mean(pred_k != y_test))

Now create the following plot using the information from your for loop

```

Figure10. Choosing K Value for KNN

```

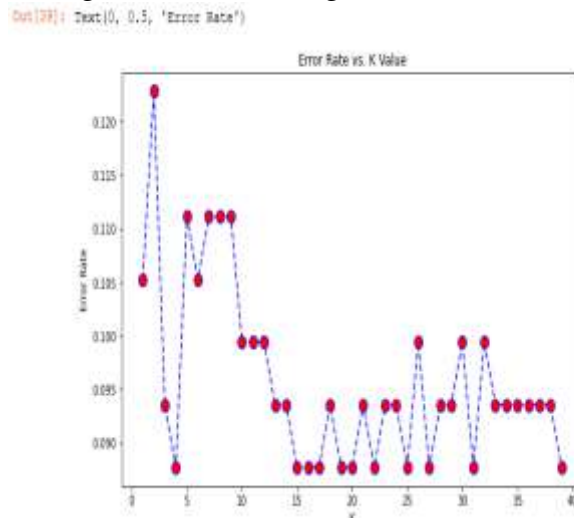
In [38]: import matplotlib.pyplot as plt
         @matplotlib inline

In [39]: plt.figure(figsize=(10, 6))
         plt.plot(range(1, 41), 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[38]: Text(0, 0.5, 'Error Rate')

```

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



Graph 4: K value v/s Error Rate

```

Retrain with new K Value
Retrain your model with the best K value (up to you to decide what you want) and re-do the classification report and the confusion matrix.

In [43]: # SVM KNN n=35
knn = KNeighborsClassifier(n_neighbors=35)

knn.fit(X_train, y_train)
pred = knn.predict(X_test)

print("KNN n=35")
print('k')
print(confusion_matrix(y_test, pred))
print('k')
print(classification_report(y_test, pred))

NOTE: n=35

[[ 51  11]
 [  6 102]]
    
```

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 has found 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](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. Sathy, 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](http://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](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](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>