CHILUKURI, N.V.S.G.S.S., BANO, S., THOLETI, G.S.R., KAMMA, S.P. and NIHARIKA, G.L. 2022. An analytical prediction of breast cancer using machine learning. In *Kumar, A., Senatore, S. and Gunjan, V.K. (eds.) Proceedings of the 2nd International conference on data science, machine learning and applications (ICDSMLA 2020)*, 21-22 November 2020, Pune, India. Lecture notes in electrical engineering, 783. Singapore: Springer [online], pages 185-202. Available from: https://doi.org/10.1007/978-981-16-3690-5_17

An analytical prediction of breast cancer using machine learning.

CHILUKURI, N.V.S.G.S.S., BANO, S., THOLETI, G.S.R., KAMMA, S.P. and NIHARIKA, G.L.

2022

An Analytical Prediction Of Breast Cancer Using Machine Learning

N.V.S Guru Sai Sarma Chilukuri [1], Shahana Bano [2], Guru Sree Ram Tholeti [3],
Sai Pavan Kamma [4], Gorsa Lakshmi Niharika [5]
Department of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
[1] gurusai21102000@gmail.com
[2] shahanabano@icloud.com
[3] usertgsr@gmail.com
[4] saipavank2000@gmail.com
[5] niharikagorsa2000@gmail.com

Abstract: Breast Cancer is one of the most occurring cancer among women affecting about 2 million people. There is 98% percent chance of 5 years survival rate if detected at early stage. The data about Breast cancer used in this paper is the Wisconsin dataset which is taken from Kaggle. This is a classification problem, there are two classes (0 representing a non-malignant tumor, 1 representing malignancy). Min Max scalar is used for preprocessing of data to limit data within certain range (known as scaling). The algorithms used for classification are Support Vector Classifier, Random Forest, Naïve Bayes, Decision Tree, K-Nearest Neighbours. Suport Vector Clasifier and Random forest gave the highest accuracy, Evaluation metrics such are Area Under Curve-Rectfied Operational Charecterstics curve, confusion matrix, Recall score, accuracy. To avoid overfitting cross validation is used where k fold value is 3.

Keywords: Artificial Neural Network, Random Forest, Decision Tree, Cross Validation.

1. Introduction

Being the most frequently occurring cancer in women, breast cancer affects around 10% of women at some point in their life. It is the second leading contributor to women’s death after lung cancer. 25% of all cancers in women including 12% of all new cases are caused by breast cancer [3]. Topics like medial science rise rapidly when certain approaches like data mining is applied due to better possibility of prediction of diseases, reducing medicine costs, improving health of patient by revamping the quality of healthcare along with value by saving people’s lives through real time decisions. The paper provides you with an analysis of performance and comparison of accuracy in classification between the algorithms such as: Logistic Regression, Suport Vector Machine, Random Forest and Naïve Bayes, being the major influential algorithms of data mining used in the research community [9].

KNN: Assumes that similar things exit with in proximity.

ALGORITHM:

1. Load the data.
2. Initialize K value to choose number of neighbours.
3. For each sample in the data
    a. Calculate the distance between the query and the current sample from the data.
    b. Add the distance and the index of the example to an ordered collection.
4. Sort the ordered collection of distances and indices in ascending order by the distances
5. Pick the first K entries from the collection

In general distance between the samples is calculated using Euclidean distance (Eq.1).

$$d(x , y) = \sqrt{\sum_{i=0}(x_i - y_i)^2} \text{ --- Eq.1}$$

Different types of distances are

1.Manhattan distance: $D (x, y) = \sum_{i=0}^{n}|x_i - y_i|$

2.Chebyshev distance: $D (x, y) = \max(|x_i - y_i|)$

Advantages:

KNN needs less training period because of instance base leaning. New data can be added seamlessly which will not impact the accuracy because KNN requires no training for classification. KNN is easy to implement. Less number of parameters

Disadvantages:

Does not work well with larger datasets because the cost of calculating the distance between the data points is very high. Does not work well with higher dimensions. Sensitive to noisy data, missing values, and outliers.

Naïve Bayes

Navie Bayes (Eq.2) is probabilistic classification algorithm whose crux is based on Bayes theorem.

$$P(h|D) = \frac{P(D|h)*P(h)}{P(D)} \text{ --- Eq.2}$$

P (h | D) is posterior probability

P (D | h) is likelihood

P (h) is prior probability

h is hypothesis

D is Data

Assumptıons:

Predictors/features are in independent(Eq.3) (i.e. one feature does not affect another feature).

$$P(A \cap B) = P(A) * P(B) \text{ --- Eq.3}$$

Algorithm:

- Calculate the probabilities of membership of each class label (i.e. probability of data points associated to a class).
- The class having the highest probability is the most suitable class.
- The above statement refers to calculating MAP (Maximum A Posteriori).
- MAP for hypothesis is:
    - MAP (h) = $max\ P(h|D)$
    - MAP (h) = max $\frac{P(D|h)*P(h)}{P(D)}$
    - MAP (h) = max $P(D|h) * P(h)$

Gaussian Naïve Bayes classifier (Eq.4):

$$P(x_i \mid y) = \frac{1}{\sqrt{2*\pi*\sigma_y^2}} * \exp\left(-\frac{(x_i-\mu_y)^2}{2*\sigma_y^2}\right) \text{ --- Eq.4}$$

Advantages:

The convergence is quicker given that Naïve Bayes condition of Independence holds. As Naïve Bayes is generative model it easy to deal with missing values.Needs less training data.

Dısadvantages:

Naïve Bayes assumes that the features are independent. There is zero frequency problem where Naïve Bayes algorithm assigns zero to class that model has never seen during training. To solve zero frequency, we need to use smoothing techniques.

Random Forest:

Random forest is voting based supervised discriminative classification algorithm.

Algorithm:

- Randomly select “Q” features from total “N” features where Q<<N
- Build the decision trees associated with “Q” features from selected data points
- Find the best tree within the forest
- Take the test features and use the randomly created decision trees for classification and store the predicted outcomes
- Calculate the votes for each predicted target
- Highly voted predicted target would be the final answer

Advantages:

It can be used for both classification and regression problem There is no problem of overfitting given that there are a greater number of trees present in the forest. Can handle missing values.

Disadvantages:

Needs a lot of computational power and memory storage to calculate a greater number of decision trees**.** Predictions are slow. Needs longer training period

Logistic Regression:

Logistic Regression is supervised discriminative classification algorithm mostly used when target feature is dichotomous**.** Logistic regression fits a S shaped curve(Eq.5) known as sigmoid curve to the target variable

Sigmoid Function:

$$s(x) = \frac{1}{1+e^{-x}} \text{ --- Eq.5}$$

Output of sigmoid curve will always be within range of [0,1]

Assumptions:

Observations to be independent of each other**,** assumption of linearity.

Advantages:

Easy to train, Less number of computations, Easy to Implement and Interpret and Features need not to be scaled.

Disadvantages:

For Multinomial classification we need to use one versus all or one versus one Classification, Prediction is not possible and Target variable should be discrete.

2. Methodology

2.1 K - Nearest Neighbour

K-NN is very simple in the implementation. K- Nerest Neighbour is higly efficent regarding the search space; non linear separability can be achived with K-Nerest Neighbour. Few parameters to tune distance metric and k-value.

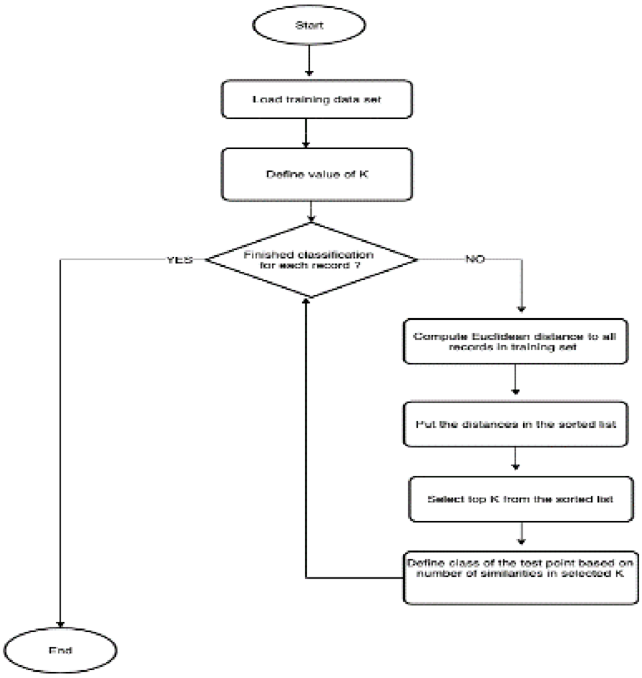

Start
Load training data set
Define value of K
Finished classification for each record ?
YES
NO
Compute Euclidean distance to all records in training set
Put the distances in the sorted list
Select top K from the sorted list
Define class of the test point based on number of similarities in selected K
End

Fig:1 KNN Flow chart

2.2 Random Forest

Random Forest is ensemble classification algorithm in which features selected are done at random. Among all the available classification methods, random forests provide the highest accuracy. The random forest algorithm can also handle big data with huge number of variables running into thousands. When the data is imbalenced it automatically balance data sets. Random forest also handles variables fast, making it suitable for complicated tasks.

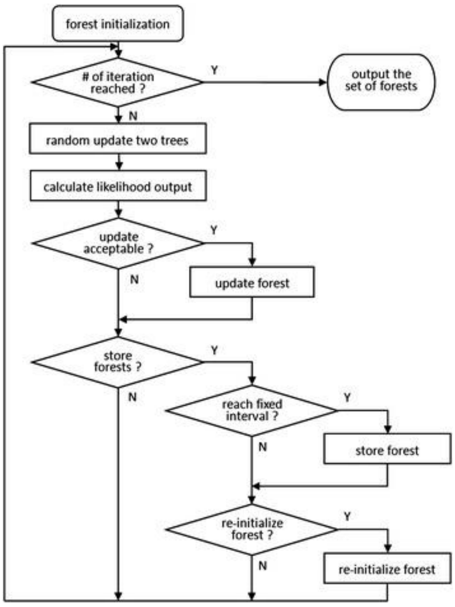


Fig:2 Random forest Flow chart

2.3 Artificial Neural Network

Artificial Neural Networks will identify the patterns in the and has a capability to learn the hidden patterns by them selfes. Input of neural network is stored in the networks instead of a database; hence there is no loss of data and does not affect its working. Learning in neural network is change in weights of neural by backpropagation which uses optimiztion algorithms such as gradient descent, adam etc.. Neural networks can be implemented in parallel by using multiple cores of a processor without affecting the performance of the system.

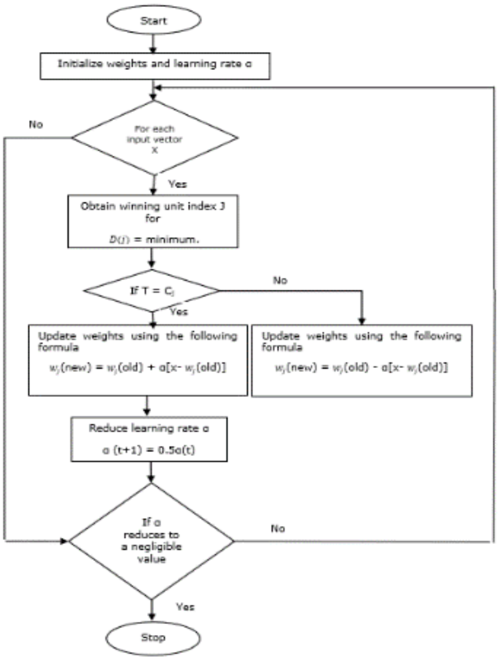

Start
Initialize weights and learning rate α
No
For each input vector X
Yes
Obtain winning unit index J for
D(J) = minimum.
If T = C_j
No
Yes
Update weights using the following formula
w_j(new) = w_j(old) + α[x- w_j(old)]
Update weights using the following formula
w_j(new) = w_j(old) - α[x- w_j(old)]
Reduce learning rate α
α (t+1) = 0.5α(t)
If α reduces to a negligible value
No
Yes
Stop

Fig:3 ANN Flow chart

3. Procedure

3.1 Data Set and Pre-Processing:

Breast cancer Wisconsin Data set is taken from Kaggle website which is source for several datasets. There are 32 parameters and 570 rows present in dataset. Some of parameters present in dataset: Diagnosis of Breast cancer**,** Radius Mean, Texture Mean, Perimeter Mean, Area Mean. Id number of patients is dropped from data set. Diagnosis is the target variable which consists of two classes (Malignant = 1 or Benign = 0). Malignant means critical, Benign means not harmful. 63% of the diagnosis feature is Benign, 37% is Malignant. The data set has no missing values. Diagnosis feature is categorical in nature. As target feature (Diagnosis) is dichotomous (Malignant and Benign classes) we use Point Biserial Correlation is used to measure the strength of association between the Independent features and Dependent variable.

- ○ $r_{pb} = \frac{M_0 - M_1}{s_y} \sqrt{\frac{n_0}{n} * \frac{n_1}{n}}$ --- Eq.6

- $M_0 = Mean\ of\ data\ group\ 1$
- $M_1\ = Mean\ of\ data\ group\ 2$

- Sy = Standard deviation of continuous data
- n0, n1 = number of items in respective groups
- n = total number of elements in two groups

Diagnosis feature is bimodal distribution. Which means distribution consists of two peaks.Min Max scalar(Eq.7) is applied to Independent features to limit the values between 0 to 1 (i.e. Scaling of independent features is done using Min Max Scalar).

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \text{ --- Eq.7}$$

80% of data is used for training and 20% of data is used for testing.To stop overfitting of data during training cross validation is used with K-fold value as 3.

3.2 Decision Tree:

Decision Tree is constructed based on Information Gain and Gini Index. We need to calculate the entropy to obtain Information Gain. Entropy tells us about appropriate measure of the randomness of a system. Lower the value of Entropy (Eq.8) higher information is obtained by the model.

$$E(T) = \sum_{i=1}^{c} -\mathrm{p}_i * log_2(\mathrm{p}_i) \text{ --- Eq.8}$$

$$E(T, X) = \sum_{x \in c} P(C) * E(C) \text{ --- Eq.9}$$

- T refers to Target variable
- X refers to individual attribute

Target variable in my data set is Diagnosis which consists of two classes Malignant (1) and Benign (0)

E(Diagnosis) = E (Malignant, Benign)

= E (357, 212)

= E (0.62, 0.38)

$$= -(0.62) * log_2(0.62) - (0.38) * log_2(0.38)$$

E(T) = E(Diagnosis) = 0.96

Information Gain (I):

$$I(T, X) = E(T) - E(T, X) \text{ ---Eq.10}$$

Max I (T, X) (Eq.10) is selected as root node and attribute values are taken as branches to that node [13].

3.3 Gaussian Naïve Bayes:

Gaussian NB (Eq.10) is a variant of Naïve Bayes algorithm. Gaussian NB is generative model which means that the model will learn the join probability distribution [13].

$$p(x_i|y) = \frac{1}{\sqrt{2*\pi*\sigma_y^2}} * e^{-\frac{(x_i-\mu_y)^2}{2*\sigma_y^2}} \quad \text{--- Eq.10}$$

- $\sigma\ is\ standard\ deviation$
- $\mu\ is\ mean$

From the above formula we could calculate the Gaussian distribution of data to obtain likely hood (L).

Assumptıons:

Variance is independent of y and x**.** All features are independent . As dependent feature consists of two classes M and B

Prior probability of cancer being Malignant (M):

P (cancer = M) = $\frac{357}{(357+212)}$ = 0.62

Prior probability of cancer being Benign (B):

P (cancer = B) = $\frac{212}{(357+212)}$ = 0.38

Given a record of a patient we could determine the type of cancer based on the features by calculating the score for M and B.

data = (radius-mean = 17.99, texture-mean = 10.38, perimeter-mean = 122.8, area-mean = 1001.0, …, fractal-dimension-worst = 0.1189).

Probability for calculating the given record is Malignant:

P (cancer = M | data) = p(cancer = M) *
L (radius-mean = 17.99 | cancer = M)
* L (texture-mean = 10.38 | cancer = M) *
L(perimeter-mean = 122.8 | cancer = M) * …
* L(fractal-dimension-worst = 0.1189 | cancer = M)
P (cancer = M | data) = 0.78

Probability for calculating the given record is Benign:

P (cancer = B | data) = P(cancer = B)
* L (radius-mean = 17.99 | cancer = B) *
L(texture-mean = 10.38 | cancer = B) *
L(perimeter-mean = 122.8 | cancer = B) * …
* L(fractal-dimension-worst = 0.1189 | cancer = B)
P (cancer = B | data) = 0.22

$$max\ (P(cancer = M\ |data), P(cancer = B|data)) = \text{max (0.78,0.22)}$$

So, given record of a patient’s cancer is Malignant.

3.4 Artificial Neural Network:

ANN is a supervised learning algorithm used for finding certain patterns in the given features. It is also used for classification, regression.

$$\text{y} = \sum_{i=1}^{30} w_i * x_i + b \text{ --- Eq. 11}$$

- w is weights
- x is inputs
- b is bias

$$y_{out} = activation(y) \text{ – Eq.12}$$

Randomly initialize 30 weights
Consider the following record:
record = (radius-mean = 17.99, texture-mean = 10.38, perimeter-mean = 122.8, area-mean = 1001.0, …, fractal-dimension-worst = 0.1189).

$$\boldsymbol{y = w_1 * \textbf{radius} - \textbf{mean} + w_2 * \textbf{perimeter} - \textbf{mean} + \cdots w_{30} * \textbf{fractal} - \textbf{dimension} - \textbf{worst}}$$

$$y = 0.125 * 17.99 + 0.256 * 10.38 + \cdots + 0.584 * 0.1189$$

$$y = 10.8755$$

$$y_{out} = \frac{1}{1 + e^{-y}}$$

$$y_{out} = \frac{1}{1 + e^{-10.8755}}$$

$$y_{out} = 0.3215$$

Now by applying backpropagation we update the weights

$$Error(e) = \sum 0.5 * (target - y)^2$$

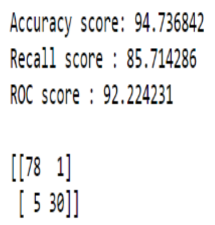
$$\frac{d(Error)}{dw} = -2(target - y)$$

4. Results

Algorithm	Accuracy	AUC	Recall	TP out of 79 positives	TN out of 35 Negatives
K-NN	94.73	0.98	85.71	78	30
LR	95.61	0.99	92.30	73	36
Naïve Bayes	93.85	0.99	92.30	71	36
SVC	92.98	0.99	82.05	74	32
Random Forest	95.61	0.99	92.30	73	36

TABLE: 1 Results

4.1 K-Nearest Neighbor Evaluation



```
Accuracy score: 94.736842
Recall score : 85.714286
ROC score : 92.224231

[[78  1]
 [ 5 30]]
```

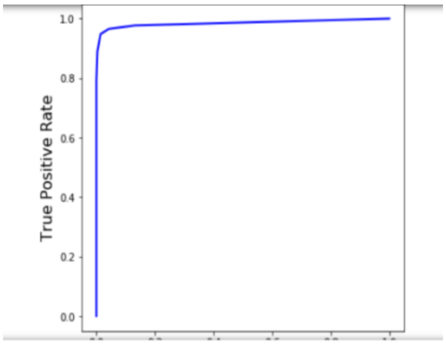
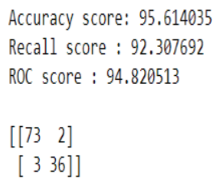


Fig.4: Evaluation of KNN Fig.5: KNN ROC curve

Area under the curve (AUC score): 0.98

4.2 Logistic Regression Evaluation



```
Accuracy score: 95.614035
Recall score : 92.307692
ROC score : 94.820513

[[73  2]
 [ 3 36]]
```

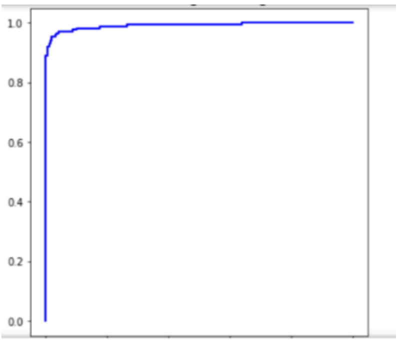
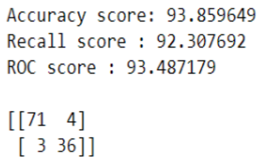


Fig.6: Evaluation of Logistic Regression Fig.7: LR ROC CURVE

AUC score: 0.99

4.3 Naïve Bayes Evaluation:



```
Accuracy score: 93.859649
Recall score : 92.307692
ROC score : 93.487179

[[71  4]
 [ 3 36]]
```

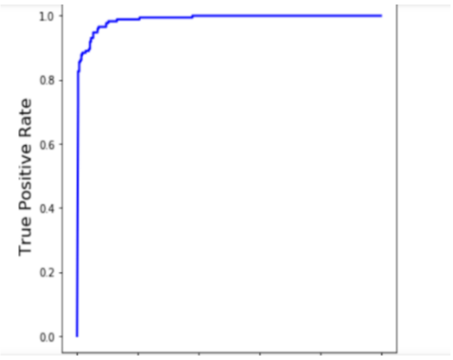
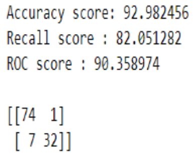


Fig.8: Evaluation of Naïve Bayes

Fig.9: Naïve Bayes ROC Curve

AUC score: 0.99

4.4 Svc Evaluation:



```
Accuracy score: 92.982456
Recall score : 82.051282
ROC score : 90.358974

[[74  1]
 [ 7 32]]
```

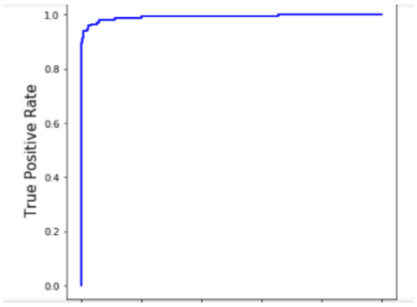
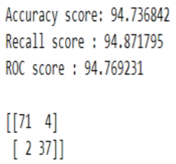


Fig.10: Evaluation of SVC

Fig.11: SVC ROC curve

AUC Score: 0.99

4.5 Decision Tree Evaluation:



```
Accuracy score: 94.736842
Recall score : 94.871795
ROC score : 94.769231

[[71  4]
 [ 2 37]]
```

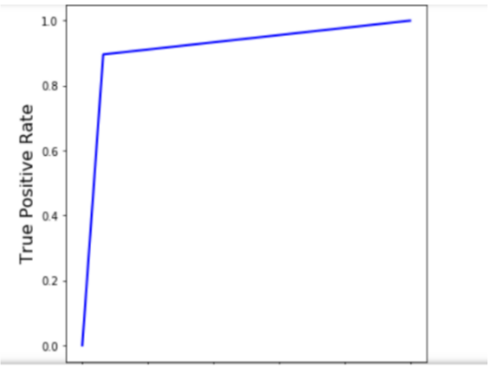
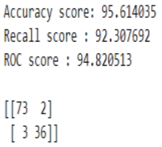


Fig.12: Evaluation of Decision Tree Fig.13: Decision Tree ROC curve

AUC score: 0.92

4.6 Random Forest Evaluatıon



```
Accuracy score: 95.614035
Recall score : 92.307692
ROC score : 94.820513

[[73  2]
 [ 3 36]]
```

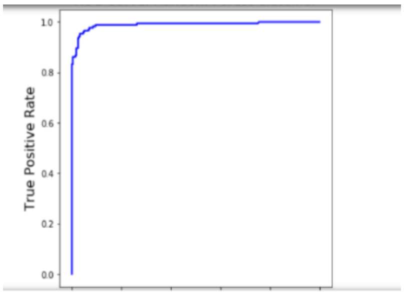


Fig.14: Evaluation of Random Forest Fig.15: Random Forest ROC curve

AUC SCORE: 0.99

5. Conclusion

K-NN, S.V.M, logistic regression, Naïve Bayes, and Decision Tree were used in the project. Random forest and Logistic regression have the highest accuracy, recall score, and AUC. Logistic regression and Random forest have 95% accuracy (as mentioned in Fig:6 and Fig:15) and has a smaller number of false positives and false negative compared to remaining classification algorithms.

6. References

[1] Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2), 105-

115.

[2] Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR, Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence 2013 Health & Medical Informatics.

[3] Cancer Statistics, 2016. CA: A Cancer Journal for Clinicians

[4] Asri, Hiba & Mousannif, Hajar & Al Moatassime, Hassan & Noël, Thomas. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science.

[5] Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF (2017) SVM and SVM Ensembles in Breast Cancer Prediction. PLOS ONE 12(1): e0161501.

[6] Longo, G. A., Zilio, C., Ortombina, L., & Zigliotto, M. (2017). Application of Artificial Neural Network (ANN) for modeling oxide-based nanofluids dynamic viscosity. International Communications in Heat and Mass Transfer, 83, 8-14.

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

[8] M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391453.

[9] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301.

[10] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S, “Breast Cancer Prediction using Machine Learning”, 2019 International Journal of Recent Technology and Engineering (IJRTE).

[11] Jadhav, Mamta & Thakkar, Zeel & Chawan, Pramila. (2019). Breast Cancer Prediction using Supervised Machine Learning Algorithms.

[12] Ch. Shravya, K. Pravalika, Shaik Subhani , Prediction of Breast Cancer Using Supervised Machine Learning Techniques, 2019 International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[13] T. Roshini, P. V. Sireesha, D. Parasa and Shahana Bano, "Social Media Survey using Decision Tree and Naive Bayes Classification," 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT).

[14] R. RamyaSri, S. IshaSanjida, D. Parasa and Shahana Bano, "Food Survey using Exploratory Data Analysis," 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT).

[15] Jean Sunny, Nikita Rane, Rucha Kanade, Sulochana Devi, 2020, Breast Cancer Classification and Prediction using Machine Learning, 2020 INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT).

[16] M L Varsha, M H P Kashyap, E Bodhith, M S R Prasad, “Prediction Of Heart Diesease Using Machine Learning Techniques”, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 02, FEBRUARY 2020, pp no. 4389 to 4392.

[17] Gurram, D., & Narasinga Rao, M. R. (2017). A comparative study of support vector machine and logistic regression for the diagnosis of thyroid dysfunction. International Journal of Engineering and Technology(UAE), 7(1.1), 326-328.

[18] Siva Kumar, P., Sarvani, V., Prudhvi Raj, P., Suma, K., & Nandu, D. (2017). Prediction of heart disease using multiple regression analysis and support vector

machines. Journal of Advanced Research in Dynamical and Control Systems, 9(18 Special Issue), 675-682.

[19] Banchhor, C., & Srinivasu, N. (2016). CNB-MRF: Adapting correlative naive bayes classifier and MapReduce framework for big data classification. International Review on Computers and Software, 11(11), 1007-1015. doi:10.15866/irecos.v11i11.10116

[20] Rachapudi, V., Venkata Suryanarayana, S., & Subha Mastan Rao, T. (2019). Auto-encoder based K-means clustering algorithm. International Journal of Innovative Technology and Exploring Engineering, 8(5), 1223-1226.

[21] Srinivas, V., Aditya, K., Prasanth, G., Babukarthik, R. G., Satheeshkumar, S., & Sambasivam, G. (2018). A novel approach for prediction of heart disease: Machine learning techniques. International Journal of Engineering and Technology(UAE), 7(2.32 Special Issue 32), 108-110.

[22] Chelladurai, R., Selvakumar, R., & Poonguzhali, S. (2018). Automatic segmentation of multiple lesions in ultrasound breast image. International Journal of Engineering and Technology(UAE), 7, 665-670.

[23] Anisha, P. R., & Vijaya Babu, B. (2018). EBPS: Effective method for early breast cancer prediction using wisconsin breast cancer dataset. International Journal of Innovative Technology and Exploring Engineering, 8(2S), 205-211.