International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020
ISSN: 1475-7192

BREAST CANCER CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

[1]Ch. Kumudini Sreeja,[2] P.Venkata Sireesha, [3]R.Ramya Sri, [4]Praveen Tumuluru

ABSTRACT--*Cancer is the second reason behind death on the planet. Approximately eight million patients died because of cancer in 2019. The carcinoma is the leading reason for death amongst women. Several styles of studies are carried out on early detection of carcinoma to start remedy and growth the chance of survival. Most of the research concentrated on mammogram snapshots, MRI, and biopsy. However, mammograms, MRI, and biopsy photos have a risk of false detection that could endanger the patient's health. It is critical to hunt out alternative mechanisms that can be less complicated to implement and work with different information sets, which can be less expensive and safer, which may produce a greater dependable prediction. Classification, predictions are a form of the powerful processing strategies which are used to categorize and are expecting the records within the datasets, especially in a medical field, in which these strategies are widely utilized in prognosis and analysis to make decisions. The target of this paper is to match and perceive a correct model to predict the prevalence of carcinoma that supported various patient's medical records. The processing techniques make use of* ***the gadget*** *getting to know algorithms like a help vector device, naïve Bayes classifier, decision tree, Random Forest. It is anticipated that in actual application, physicians and patients can revel in the feature popularity outcome to prevent carcinoma, using these machines getting to know algorithms.*

Keywords- su*pport vector machine, Naïve Bayes , Decision tree, Random Forest*

I. INTRODUCTION

The challenge especially examined on finding accuracy of the dataset that we take. To perform this, we organized a huge survey which contains more often of the patients in a cancer hospital close to our locality. By this, we came to recognize that whether or not they have a malignant tumor or benign tumor basing on the facts this is provided by using the hospital. Further we have performed a confusion matrix consisting of some of the parameters like data accuracy with the help of machine learning algorithms like support vector machine, Naïve-Bayes [18], Decision tree, Random Forest.

1. Support vector machine:

[1] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Andhra Pradesh)

[2] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Andhra Pradesh)

[3] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Andhra Pradesh)

[4] Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Andhra Pradesh)

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

This is one of the algorithms in the grouping of large datasets. It is used for the classification of single-dimensional and multi-dimensional records. This mainly uses a multi-dimensional leveling to extremely acceptable dimension; information from these classifiers are usually apart through hyper-plane (a "selection barrier" keeping apart rows of one elegance from another). They are an awful lot low at risk of over-fitting. This vector provides a closeness characterization of the model.

A. Single-dimensional Support vector machine:

The fact components are separated with a hyper-plane and are treated as m-dimensional vector, and (m-1) dimensional particles. The several hyper-planes keeping apart the single-dimensional statistics, but we picked the big hyper-plane, which expands edges (zone among hyper-aircraft and the closest statistics components of any class).

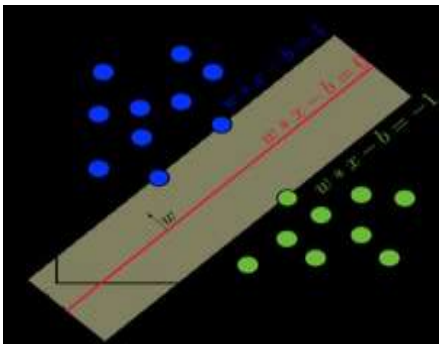


Figure 1(A): Single-dimensional Support Vector Machine

B. Multi-dimensional Support Vector Machine:

The information particles are multi-dimensional which are apart in m-dimensional zone. The hyper-planes are given by u kernel plots. Each kernel has a multi-dimensional function.

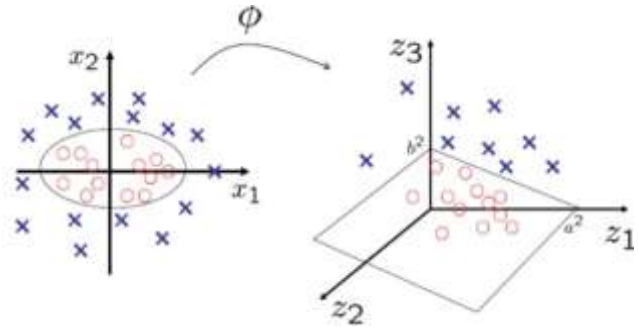


Figure 1.(B): Multi-dimensional Support Vector Machine

2. Naive-Bayes:

This is one of the algorithms in the grouping of large datasets. This works effectively for text classification. The working of naïve-Bayes is Bayes theorem

Bayes theorem:

This mainly uses the dependence possibility. The dependence possibility is chance of a fact as a way to happen, for the reason that a fact that previously occurred. The possibility of this event the usage of its previous information is given by

$$P\left(\frac{H}{E}\right) = \frac{P\left(\frac{E}{H}\right) * P(H)}{P(E)}$$

Where,

$P(H) \rightarrow$ Possibility of speculation H being absolute. (Previous possibility).

$P(E) \rightarrow$ Possibility of the proof.

$P(E|H) \rightarrow$ Opportunity of the evidence given that the speculation is true.

$P(H|E) \rightarrow$ Chance of the speculation for the reason that evidence is there.

3. Random Forest:

Random forest, like its name implies, consists of many individual selection bushes that perform as an ensemble, for classification, regression. Each character tree in the random woodland spits out a category prediction and the elegance with the maximum votes will become our model's prediction. Random forests can overcome the negative aspects of choice tree-like over becoming their training dataset.

4. Decision Tree:

It is a classifier Mechanism. It's simple and easy to implement. There is no requirement of domain know-how or parameter putting to handle excessive dimensional data. It produces results that can be clean to study and understand. The drill through a feature is handiest available in Decision Trees, which might be used to get entry to targeted patients' profiles. This is a tree-like structure, which consists of internal nodes, branches, and leaf nodes, and in which every branch represents a characteristic value; each internal node indicated a test on an attribute that is used for the tree.

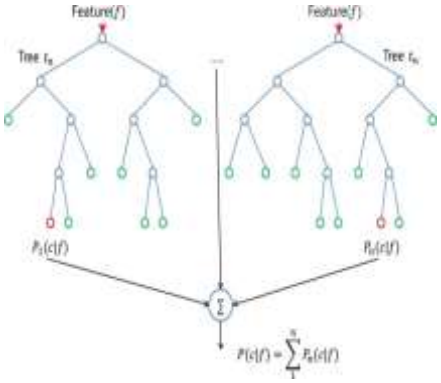


Figure 3: ***Random Forest***

Tree Pruning (Optimization)

Examine the branches which aren't beneficial to class as well as put off those in the selection tree.

- Pre-Pruning method
- Post Pruning method

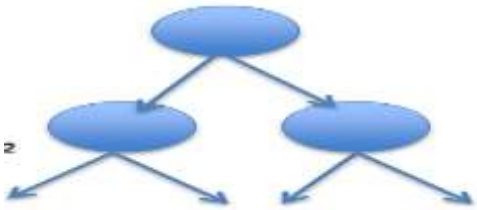


Figure 4: Sample Decision tree

II. PROCEDURE

The challenge is concentrated on the four most important phases. The first phase consisting of importing data which includes the data is collected from excel sheet .The second phase consists of data pre-processing which contains data cleaning, data transformation. The third phase consists of performance of algorithms .The fourth phase consists of comparison of accuracy between the algorithms.

Block diagram of the proposed system:

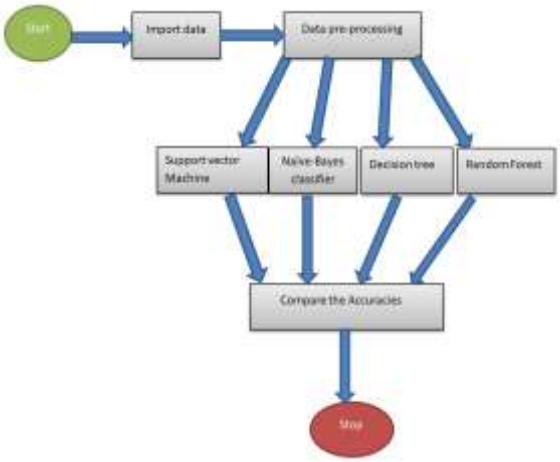


Figure 5 : Block Diagram

1 Collection of Data:

We created an internet survey the usage of Google paperwork for doctors to round out. This examine specially targeted the human's diseases from which they arise the sicknesses we concluded our survey along the easy query on perspectives as well as opinions on the presence of a tumor. After examination changed into end, it's able to get output of data in the shape of an Excel sheet. This sheet consists of questions on the x-axis, while each entry is on the y-axis [18]. After storing the excel sheet on computer, now proceed to the consecutive steps of performance and investigation.

The attributes from our statistics set are means, se, worst of radius etc.

2 Import Metadata:

The major steps in any process are gathering all the prerequisites. The major prerequisites are the software that we use and the facts with the essential libraries for the performance of facts.

I. Import Data:

As the data is in the excel format that we saved on computer is imported to the software that we use. Here, the software is 'R'.so we use setwd() to import the data. After the setwd() is done we read.csv() to load the data. Further the libraries are installed and performance of algorithm takes place.

II. Import Libraries:

The several types of equipment within 'R', has to import the essential Libraries. Some of the Libraries are 'e1071','caret'.

'rattle','stringr'.

The above are mainly vital as well as broadly used library functions in 'R ' allows to examine statistics several one-of-a-kind methods statistically.

3 Data pre-processing:

These records have few pleasant which appease the usage of essentials. The elements containing information agreeable, which include correctness or accuracy, completeness, consistency etc. In actual scenario, the statistics are grimy insufficient with flawed characteristic integrity, precise characteristics importance, as well as consisting of total data, and inadequate records are from, "Impossible" records cost it accrued, several concerns among the future whilst the facts accumulated and inspected.

I. Noisy:

Consists of failures as well as outliers in Noisy statistics (insufficient integrity) may be even invalid facts series devices, at the access of records, failures inside the transportation of facts.

e.g., id= "0"

II. Inconsistent:

Consists of discrepancy in labels, these records are from, several statistics resources, Functional dependency violation (e.g., changing a few joined statistics).

Inadequate information needs information purifying

e.g., Age="100" Birthday="03/07/1998"

III. Importance of Data Pre-Processing:

Nature of verdict is primarily on great records, Data-mining uses constant mixture of first-rate facts, extraction, purifying, and transportation of data containing the bulk paintings in constructing of facts in mining. e.g., inadequate or omitted data motive insufficient or maybe deceptive data

IV. Steps in Data Pre-processing:

The most important duties as well as strategies that we perform in information transformation is purifying i.e. removing the unwanted and omitted data.

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

V. Data cleaning:

Data Cleaning is a system figuring out as well as altering the information from a record as well as which group the inadequate, improper, imprecise statistics, and these statistics are utilized in substitution, correct, or exclude the unwanted records. This is done associated along facts tangle devices, or a cluster transform over setup. Later, facts should be fixed along different unique records units in classification.

VI. Omitted Data:

In Omitted Data, the facts aren't consistently usable

eg., Several facts records have no integrity cost for many characteristics, consisting of client profits on trade information, omitted records might be fixed.

Omitted data might be:

a. Gadget defect

b. Inequality statistics

c. .Accessing of data problem because of confusion

d. Convinced facts are not useful at the point of accessing

e. now unknown facts, conversely adjustments in information.

VII. Manipulation of Omitted Data:

Ignore the row with the elegance label is Omitted (Suppose work in grouping aren’t energetic alike whilst share of lacking integrity in keeping with characteristic varies.

a. Place omitted price manually: endless + impossible?

b. Place robotically along

i. An international consistent: e.g., "undiscovered”, brand new group?

ii. Characteristic suggest

iii. Characteristic implies for all specimen associate to the identical group: brilliant

iv. Better possible cost: interpretation-situation.

VIII. Noisy in Data:

I. Irregular blunders or variance in a regular attribute

II. Improper variables integrity can additionally be

1. Invalid statistics gathering devices

2. Access of information issues

3. Transportation of facts issues

4. Inequality in the identifying representation.

III. Alternative facts issues which lack information purifying

1. Replication of facts

2. Inadequate information

3. Inequality statistics

IX. Manipulation of Noisy Data:

• Binning

- Initially classify fact and separate them to (uniform-density) boxes.
- Next, even via box median, easy over way of box boundaries, etc.
- Regression

Even with the aid of meeting information to regression function

4. Performance of Algorithms:

Some of the algorithms that are used in this are system-gaining knowledge of techniques which are for categorization, regression, prediction, and anomaly-detection. The system gaining knowledge of techniques is used for both supervised and un-supervised.

I. *Performance of support vector machine*:

In this support vector machine used e1071 packages and some other packages for classification of type of tumor [17]. Using confusion Matrix with some parameters like accuracy, specificity, sensitivity, Pos pred values. neg pred values, and we predicted the output. In this the data is Train and tests are inside scaled in 70:30compositions. This model predicts nearly 98 percent in terms of accuracy.

II. Performance of Naïve-Bayes:

In this Naïve-Bayes we have used e1071 packages and some other packages for classification of type of tumor. Using confusion Matrix with some parameters like accuracy, specificity, sensitivity, Pos pred values. Neg pred values, and we predicted the output. In this the data is Train and tests are inside scaled in 70:30compositions.This version predicts nearly 93 percent in terms of accuracy.

III. Performance of Decision Tree:

In this Decision Tree we have used e1071 packages and some other packages for classification of type of tumor. Using confusion Matrix with some parameters like accuracy, specificity, sensitivity, Pos pred values. Neg pred values and we predicted the output. In this the data is Train and test are inside scaled in 70:30compositions.This version predicts almost 95 percent in terms of accuracy

IV. Performance of Random Forest:

In this Decision Tree we have used e1071 packages and some other packages for classification of type of tumor. Using confusion Matrix with some parameters like accuracy, specificity, sensitivity, Pos pred values. Neg pred values and we predicted the output. In this the data is Train and test are inside scaled in 70:30compositions.This model predicts nearly 95percent in terms of accuracy.

V. Comparison of Algorithms:

Later, the performance of all algorithms, we have compared all the algorithms in terms of accuracy, sensitivity, specificity, pos pred value and neg pred value, and they are represented in the form of both graphs and table as shown in results

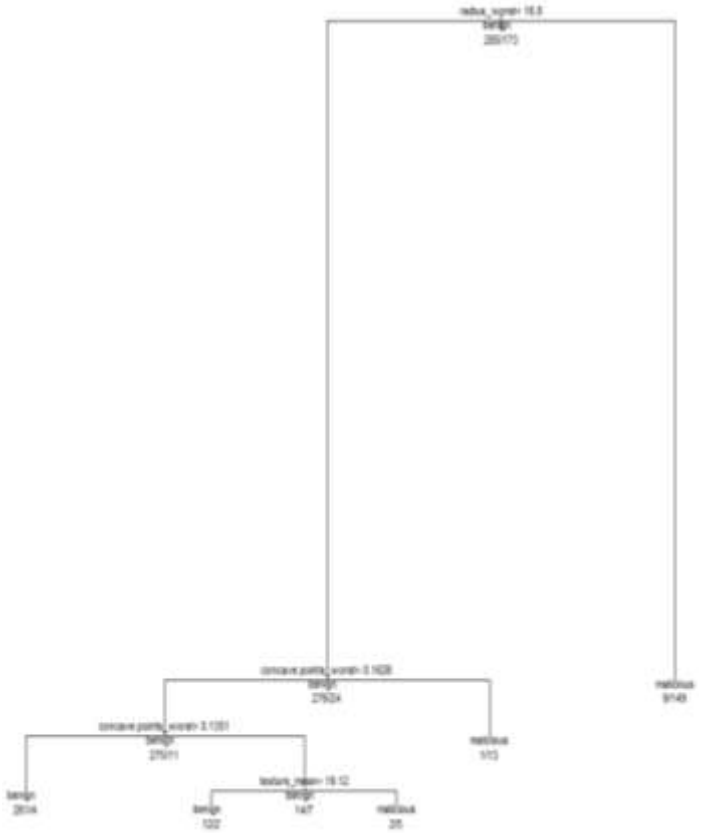
Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

III. RESULTS

Support vector machine:

```
Confusion Matrix and Statistics


pred_test   benign malicious
  benign        72         2
  malicious      0        37

               Accuracy : 0.982
                 95% CI : (0.9364, 0.9978)
    No Information Rate : 0.6486
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.96

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9487
         Pos Pred Value : 0.9730
         Neg Pred Value : 1.0000
             Prevalence : 0.6486
         Detection Rate : 0.6486
   Detection Prevalence : 0.6667
      Balanced Accuracy : 0.9744

       'Positive' Class : benign
```

Decision Tree:



Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

```
Confusion Matrix and Statistics


tree_pred_full benign malicious
     benign       340         7
     malicious     17       205

               Accuracy : 0.9578
                 95% CI : (0.9379, 0.9728)
    No Information Rate : 0.6274
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9106

 Mcnemar's Test P-Value : 0.06619

            Sensitivity : 0.9524
            Specificity : 0.9670
         Pos Pred Value : 0.9798
         Neg Pred Value : 0.9234
             Prevalence : 0.6274
         Detection Rate : 0.5975
   Detection Prevalence : 0.6098
      Balanced Accuracy : 0.9597

       'Positive' Class : benign
```

Naive-Bayes:

```
  Confusion Matrix and Statistics


  preds_naive  B  M
            B 92  6
            M  5 67

               Accuracy : 0.9353
                 95% CI : (0.8872, 0.9673)
    No Information Rate : 0.5706
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8677

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9485
            Specificity : 0.9178
         Pos Pred Value : 0.9388
         Neg Pred Value : 0.9306
             Prevalence : 0.5706
         Detection Rate : 0.5412
   Detection Prevalence : 0.5765
      Balanced Accuracy : 0.9331

       'Positive' Class : B
```

Random Forest:

```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 105   5
         M   2  58

               Accuracy : 0.9588
                 95% CI : (0.917, 0.9833)
    No Information Rate : 0.6294
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9109

 Mcnemar's Test P-Value : 0.4497

            Sensitivity : 0.9206
            Specificity : 0.9813
         Pos Pred Value : 0.9667
         Neg Pred Value : 0.9545
             Prevalence : 0.3706
         Detection Rate : 0.3412
   Detection Prevalence : 0.3529
      Balanced Accuracy : 0.9510

       'Positive' Class : M
```

GRAPHS**:**

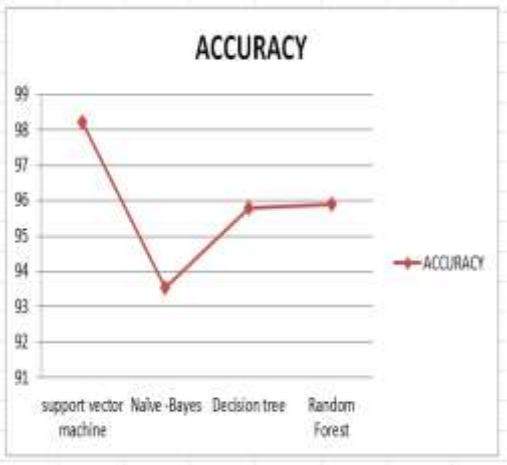


Figure 6: Accuracy of all algorithms

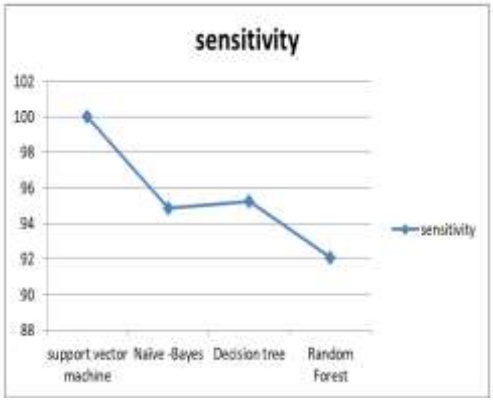


Figure 7: Sensitivity of all algorithms

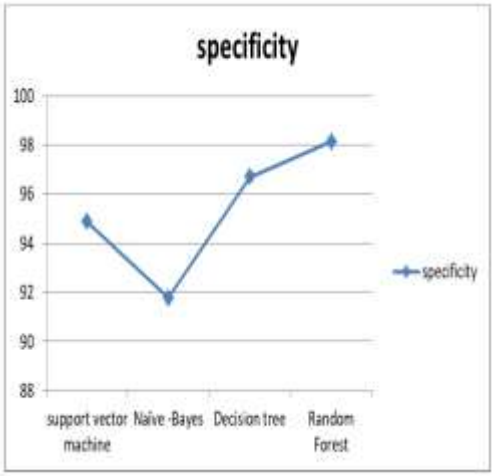


Figure 8: specificity of all algorithms

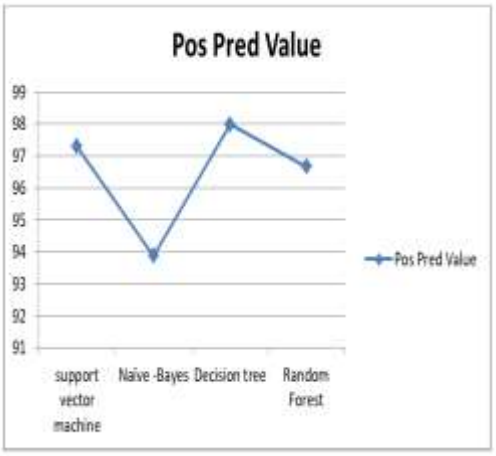


Figure 9: Pos Pred value of all algorithms

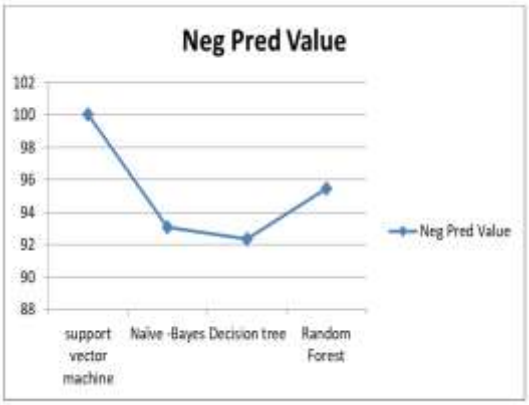


Figure 10: **N**eg Pred value of all algorithms

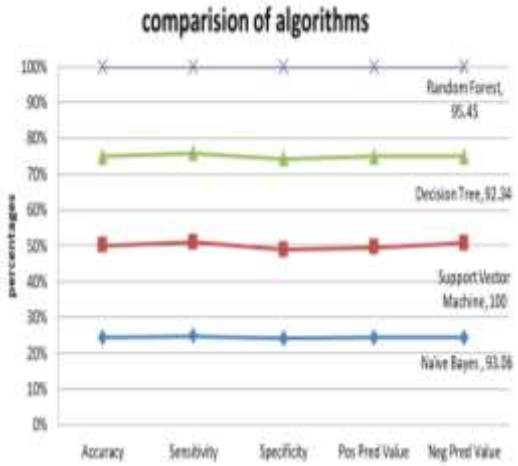


Figure 11: Comparison of all algorithms

Table 1: Comparison table

Values	Support vector machine	Naïve Bayes	Decision tree	Random Forest
Accuracy	98.2	93.53	95.78	95.88
Sensitivity	100	94.85	95.24	92.06
Specificity	94.87	91.78	96.7	98.13
Pos Pred Value	97.3	93.88	97.98	96.67
Neg Pred Value	100	93.06	92.34	95.45

IV. CONCLUSION

Breast Cancer is mostly seen in women at the early stages of their lives. As the prognosis it takes a lot of time with normal mechanisms of equipment in systems. So, we have used some of the machine learning algorithms like support vector machine, Naïve-Bayes classifier, random forest, a decision tree for the classification of the type of tumor-like benign and malignant. With these machine learning algorithms, we have trained and tested the data for the type of tumor and chose the best algorithm in terms of accuracy, and this algorithm can be used for equipment in systems during the prognosis and get the results very fast. So, we can control the mortality rate among women in the world.

REFERENCES

1. R.Preetha, S. Vinila Jinny-A Research on Breast Cancer Prediction Using Data Mining Techniques.
2. Ch. Shravya, K. Pravalika, Shaik Subhani-Prediction of Breast Cancer Using Supervised Machine Learning Techniques.
3. Yi-Sheng Sun, Zhao hao, Han-Ping-Zhu, "Risk factors and Preventions of Breast Cancer" International Journal of Biological Sciences.
4. Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.
5. VikasChaurasia, BB Tiwari and Saurabh Pal – "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms and Computational Technology.
6. Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction Using Data Mining Method, IEEE Conference paper.
7. D.Dubey, S.Kharya, S.Soni and –"Predictive Machine Learning techniques for Breast Cancer Detection", International Journal of Computer Science and Information Technologies, Vol.4(6),2013.
8. Nidhi Mishra, NareshKhuriwal.- "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm", 2018 IEEMA Engineer Infinite Conference (eTechNxT),2018.

9. Chao-Ying, Joanne, PengKukLida Lee, Gary M. Ingersoll –"An Introduction to Logistic Regression Analysis and Reporting ", September/October 2002 [Vol. 96(No.1)] Logistic Regression for Machine Learning.
10.
11. Mohammad Bolandraftar and Sadegh Bafandeh Imandoust - "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background"- International Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013
12. EbrahimEdrissEbrahim Ali1, Wu Zhi Feng2- "Breast Cancer Classification using Support Vector Machine and Neural Network"– InternationalJournalofScienceandResearch(IJSR) Volume 5 Issue 3, March 2016
13. Padmaja P and B. Lakshmi Ramani. "Adaptive Fuzzy System with Robust GSCA-based Fuzzy Rule Extraction for Data Classification," JARDCS, Vol. 10, 01, 2018.
14. Tumuluru, P. and Ravi, B. "GOA-based DBN: Grasshopper Optimization Algorithm-based Deep Belief Neural Networks for Cancer Classification". International Journal of Applied Engineering Research 12 (24) (2017).
15. Tumuluru, P. and Ravi, B. "Chronological Grasshopper Optimization Algorithm- based Gene Selection and Cancer Classification. Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, No. 3, 2018.
16. Praveen Tumuluru, Bhramaramba R, "A Framework for Identifying of Gene to Gene Mutation causing Lung Cancer using SPI - Network", International Journal of Computer Applications, vol. 152, no. 10, Oct 2016.
17. Praveen T, et al. "Credentials of Lung-Cancer Associated Genes Using Protein-Protein Interaction Network", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, No. 3, March 2016.
18. Praveen Tumuluru, Bhramaramba Ravi "Dijkstra's based Identification of Lung Cancer Related Genes using PPI Networks", IJCA, Vol. 163, No. 10, 04-2017.
19. Praveen T, Bhramaramba Ravi "A Survey on Gene Expression Classification Systems", International Journal of Scientific Research and Review ISSN NO: 2279-543X, Volume 6, Issue 12, 2017.
20. Praveen Tumuluru, Burra Lakshmi Ramani et al. "OpenCV Algorithms for facial recognition", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.
21. B. Lakshmi Ramani, Praveen T et al. "Deep Learning and Fuzzy Rule-Based Hybrid Fusion Model for Data Classification" IJRTE, ISSN: 2277-3878, Volume-8 Issue-2, July 2019.
22. Praveen Tumuluru, Radha Manohar Jonnalagadda et al. "Extreme Learning Model Based Phishing Classifier" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.
23. B. Lakshmi Ramani, Dr. Padmaja Poosapati "Adaptive Lion Fuzzy System to Generate the Classification Rules using Membership Functions based on Uniform Distribution" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14421-14433.
24. Tumuluru, P., Lakshmi, C.P., Sahaja, T., Prazna, R. "A Review of Machine Learning Techniques for Breast Cancer Diagnosis in Medical Applications "Proceedings of the 3rd International Conference on

I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019.

25. Nalajala, S., Akhil, K., Sai, V., Shekhar, D.C., Tumuluru, P. "Light Weight Secure Data Sharing Scheme for Mobile Cloud Computing" Proceedings of the 3rd International Conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019.