JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# BREAST CANCER DETECTION USING MACHINE LEARNING AND AWS(Amazon Web Service)

Pretty Pramod Kotian
*Department of Computer Engineering*
*Sinhgad Academy of Engineering, Pune*
kotianpretty@gmail.com

Snehal Babar
*Department of Computer Engineering*
*Sinhgad Academy of Engineering, Pune*
babarsnehal2000@gmail.com

Aishwarya Ningdali
*Department of Computer Engineering*
*Sinhgad Academy of Engineering, Pune*
ningdali172000@gmail.com

Muskan Altaf Shaikh
*Department of Computer Engineering*
*Sinhgad Academy of Engineering, Pune*
muskanshaikh00060@gmail.com

***Abstract :*** Breast cancer starts in the breast cell. It is a cancerous tumor where cancer cells grow and destroy nearby tissue. It was estimated in 2019 that 270,000 new breast cancer cases were diagnosed which is an alarming rise of cancer in women every year. With the advances of computer technology, we can save a life from cancer at an earlier stage. Hence, we have built the software with the help of machine learning to analyze breast cells before it gets fatal. This project aims to use machine learning algorithms and techniques to detect breast cancer and also do the prediction with Random Forest, KNN (k-Nearest-Neighbor) and Support Vector Machine algorithm. The Breast Cancer Wisconsin original dataset is used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision. We will perform data visualization in the form of graphs like histograms, boxplots and also study the correlation between each attribute. In the end, we will develop a classification report and confusion matrix to predict whether the dataset is benign or malignant breast cancer for every machine learning algorithm
.

Cloud Computing is a recently emerged model that is becoming popular among almost all enterprises.Cloud computing can also be described as the network that enables the distribution of processing, application, storage capabilities among many remote located computer systems. In cloud computing platforms plenty of IT resources are utilized and released as per the user requirement by using the internet.As the work tends to increase the awareness about the use of cloud computing in the medical field about storing of the data and the extent in using cloud computing in the medical field. This paper will provide the review of security research in the field of cloud usage. After research we have presented the working of AWS (Amazon Web Service) cloud computing to unite with Machine learning to produce amazing results in the field of medical. AWS is considered to be the most trusted provider of cloud computing by many users as they not only provide excellent cloud security but also provide excellent cloud services. Here we will summarize services provided by AWS that will help to choose the suitable features which will fulfill the long term requirements of the users.

***Keywords* -** Breast Cancer, Machine learning, Prediction, KNN and SVM, Cloud Computing, AWS(Amazon Web Service), Amazon EC2 instance.

# I. INTRODUCTION

Breast cancer is the second leading cause of female death (after lung cancer). Invasive breast cancer will be diagnosed in 246,660 women in the United States this year, with 40,450 women dying. Breast cancer is a cancer that originates in the breast and develops toward other areas of the body. When cells multiply uncontrollably, cancer develops. Breast cancer cells generate a lump that can be felt or seen on an x-ray. Breast cancer cells can spread to other parts of the body if they enter the blood or lymph system. Changes and mutations in DNA are among the causes of breast cancer. DCIS (ductal carcinoma in situ) and invasive carcinoma are both prominent kinds of breast cancer. Others are less prevalent, such as phyllodes tumors and angiosarcoma. There are a variety of algorithms for evaluating breast cancer outcomes. Fatigue, headaches, pain and numbness (peripheral neuropathy), bone loss, and osteoporosis are all side symptoms of breast cancer.

There are numerous algorithms for breast cancer classification and prediction. The performance of four classifiers is analyzed in this paper: SVM, Logistic Regression, Random Forest, and kNN, which are among the most popular data mining algorithms. Mammography or a portable cancer diagnostic equipment could be used to detect it early during a screening test. Cancerous breast tissues change as the disease advances, and this can be associated with cancer stage. The stage of breast cancer (I–IV) indicates how far the cancer has spread in a patient. Stages are determined using statistical indications like tumor size, lymph node metastasis, and distant metastases, among others. Patients must endure chemotherapy to prevent cancer from spreading. Patients must have breast cancer surgery, chemotherapy, radiation, and endocrine therapy to prevent cancer from spreading.

The study's objective is to identify and categorize malignant and benign individuals, as well as considering how to parametrize our classification. As a result, ways to achieve high precision have been developed. We're investigating a variety of datasets to see how Machine Learning may be applied to them. Breast cancer can be classified using machine learning techniques. We wish to lower the mistake rates by using maximum precision JUPYTER employs the 10-fold cross validation test, which is a machine learning technique. To assess and analyze information in terms of efficacy and efficiency

Breast cancer is the second most common type of cancer after lung cancer to cause death. Invasive breast cancer will be diagnosed in 246,660 women in the United States this year, with 40,450 women dying. Breast cancer is a cancer that originates in the breast and develops toward other areas of the body. When cells multiply uncontrollably, cancer develops. Breast cancer cells generate a lump that can be felt or seen on an x-ray. Breast cancer cells can spread to other parts of the body if they enter the blood or lymph system. Changes and mutations in DNA are among the causes of breast cancer.

DCIS (ductal carcinoma in situ) and invasive carcinoma are both prominent kinds of breast cancer. Others are less prevalent, such as phyllodes tumors and angiosarcoma. There are a variety of algorithms for evaluating breast cancer outcomes. Fatigue, headaches, pain and numbness (peripheral neuropathy), bone loss, and osteoporosis are all side symptoms of breast cancer. There are numerous algorithms for breast cancer classification and prediction.

The performance of four classifiers is analyzed in this paper: SVM, Logistic Regression, Random Forest, and kNN, which are among the most popular data mining algorithms. Mammography or a portable cancer diagnostic equipment could be used to detect it early during a screening test. Cancerous breast tissues change as the disease advances, and this can be associated with cancer stage. The stage of breast cancer (I–IV) indicates how far the cancer has spread in a patient. Stages are determined using statistical indications like tumor size, lymph node metastasis, and distant metastases, among others. Patients must endure chemotherapy to prevent cancer from spreading. Patients must have breast cancer surgery, chemotherapy, radiation, and endocrine therapy to prevent cancer from spreading. The study's objective is to Identifying and categorizing malignant and benign individuals, as well as considering how to parametrize our classification
As a result, ways to achieve high precision have been developed. We're investigating a variety of datasets to see how Machine Learning may be applied to them. Breast cancer can be classified using machine learning techniques. We wish to lower the mistake rates by using maximum precision JUPYTER employs the 10-old cross validation test, which is a machine learning technique. To assess and analyze information in terms of efficacy and efficiency.

## Cloud Computing:

Cloud computing is the supply of computing services over the internet, including servers, storage, databases, networking, software, analytics, intelligence, and more (Internet).

Cloud computing is a suitable substitute for on-premises data centers. We must manage everything with an on-premises data center, including obtaining and installing hardware, virtualization, installing the operating system and any other essential software, configuring the network, configuring the firewall, and configuring data storage. After we've completed all of the setup, we're in terms of maintaining it operating for the rest of the life.

However, if we choose Cloud Computing, a cloud vendor is responsible for ordering and managing the hardware. They also provide a variety of software and platform-as-a-service solutions. Any vital services can be purchased by us. Cloud computing services will be charged on a per-use basis.

The cloud environment provides an user-friendly online gateway that allows them to manage compute, storage, network, and application resources.

Benefits of cloud computing

- It reduces the cost by maintaining and managing the IT systems.
- It provides scalability and business continuity.
- It is collaboration efficient and provides flexibility for work practices.
- It provides access to automatic updates

Major Characteristics of Cloud Computing

- Resources Pooling.
- On-Demand Self-Service.

- Maintenance is easy.
- It is scalable and rapidly elastic.
- Cloud Computing proves to be economical in nature.
- Measured and Reporting Service.
- Cloud Computing provides Security.
- Cloud Computing provides automation.

Applications of Cloud Computing

- Online data storage
- Backup and Recovery
- Testing and development
- Cloud computing in medical fields
- Big Data Analysis

Limitations of Cloud Computing:

- Downtime is the biggest drawback of Cloud Computing.Cloud computing systems are internet- based, service outages are always an unfortunate possibility and can happen for any reason.
- Security and privacy Cloud service providers apply the better security norms and industry certificates, storing data and major files on external service providers always opens up threats.
- Vulnerability to attack In cloud computing, every element is online, which exposes possible vulnerabilities.
- Limited control and flexibility The service provider entirely possesses the cloud infrastructure, manages, and watches it and transfers minimum control over to the client.
- Cost interests taking up cloud solutions on a small scale and for short- term programs can be perceived as being costly. Nevertheless, the most significant cloud computing benefit is in terms of IT cost savings.

Security  Issues in Cloud Computing:

- Data Loss –

  This is also called Data Leakage. As we know that our sensitive data is in the hands of Somebody or anyone, and we don't have full control over our database. So, if the security of cloud service is to be breached by hackers also it may be possible that hackers will get access to our sensitive data or private files.
- User Account Hijacking –

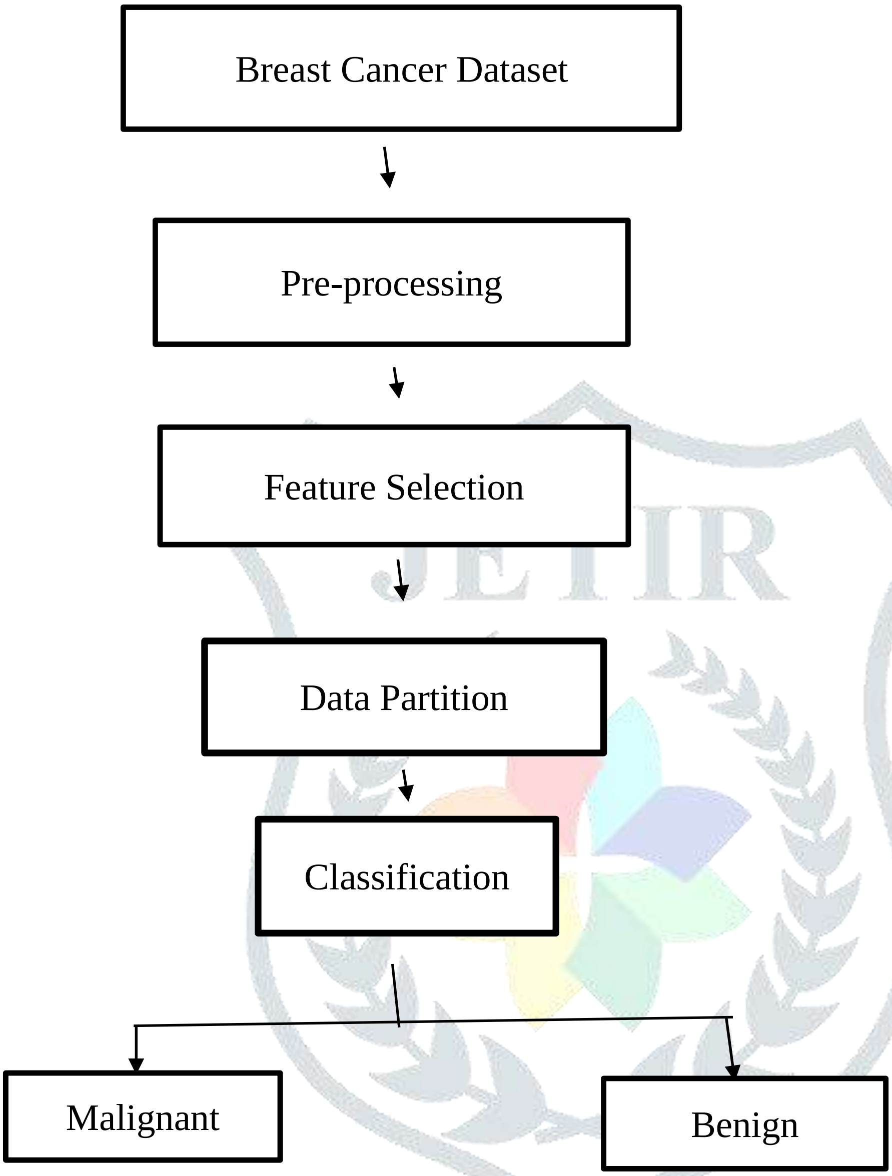
  If someway the Account of User or an Organization is hijacked by Hacker and they get full authority to perform Unauthorized Activities.
- Changing Service Provider –

  An organization wants to shift from AWS Cloud to Google Cloud Services also they face multiple problems like shifting of all data, also both cloud services have different ways and functions, so they also face problems regarding that. However, it may be possible that the charges of AWS are different from Google Cloud, etc.
- Denial of Service (DoS) attack –

  This type of attack occurs when the system receives too important business. Substantially DoS attacks occur in large associations similar to the banking sector, government sector, etc. When a DoS attack occurs, the data contained is lost .

## II. METHODOLOGY

### i. Proposed methodology:



Firstly we take the WBCD dataset i.e. Wisconsin Breast Cancer Dataset and then Pre-processing the data by using discretized filter and resampling it and making a Cleaned data for further process, After that there is feature selection process basically this process removes the non-informative data from the model. Next there is data partition process typically this process involves partitioning of the data into a training set and testing set. Then we use different classification techniques with machine learning and finally using this classification techniques, we predicts which type of cancer patient have are Malignant or Benign

### ii. Machine learning Algorithms Used:

k-Nearest Neighbor (kNN):-

k-Nearest Neighbour is a supervised learning method used to tackle problems mostly in the classification and regression fields. Its ease of use comes with a few disadvantages of its own; Depending on precision on data quality, sensitivity to large-scale data, and slow prediction, as well as the fact that its calculation is maintained permanently, makes it difficult.

As a result, it necessitates a large amount of memory, making it quite demanding. For most average datasets, a generic k-NN is frequently used to categorize the means of a particular cluster set.

We will discuss the following topics in this paper: Try to train it for pattern recognition and use it for the main goal of making predictions. The k-NN method is a non-parametric technique.

Because k-NN is instanced based learning, both scenarios are trained in a feature space. In output is given after classification. In classification, the output is determined by picking the most votes cast by neighboring clusters of a specified k, where k is an arbitrary number value or a predetermined value based on the goal Feature extraction is a way of extracting output from input data after transformation.

The following is a common approach for how the k-NN Algorithm works:

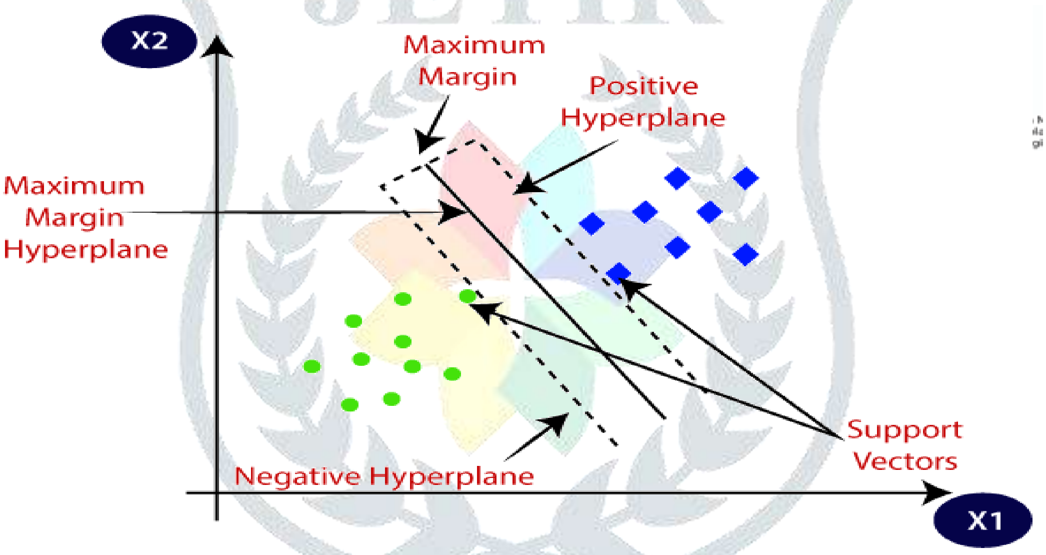
1)Assigning a value to k or assigning an arbitrary value to it.

2)On both the testing and training datasets, calculations are performed.

3)Classification based on the information provided.

4)Finishing the results received and converting them to output.

SVM (Support Vector Machine) :-



It is one of the most prominent and commonly used Supervised Learning Algorithms. The end goal and objective of this method is to produce a decision boundary that can set apart and isolate n-dimensional space into classes, putting them into current data points in the correct category, and this process is referred to as 'hyperplane'. Memory efficiency and high dimensional space are only a few of the advantages it has over others. It has the capacity to work with both linearly separable and non-separable data.

It is said that Kernel tricks, often known as generalized dot products, are a method of calculating the dot product of two vectors to see how much they affect each other. The possibilities of linearly non-separable data sets have higher probability in higher dimensions, according to Cover's Theorem.

### iii. Deployment of the Model in AWS(amazon Web Service) EC2 instances:

1. Built the Model
2. Exporting the Model we made using Pickle
3. Built a Flask website to serve this model
4. Deploy the website on AWS EC2.
    1) Create an AWS account
    2) Create an EC2 Instance
    3) Edit Security Group
    4) Download key generator(pem file)
    5) Download and install Putty and WinSCP
    6) Upload Flask website on EC2 using WinSCP
    7) Install packages on EC2 using Putty

By choosing AWS EC2 instance we have deployed our Machine Learning Model on ” Breast Cancer Detection” on by the following procedure.Firstly, we have made our model in Jupyter notebook and we have exported this model by using Pickle.Then we have built a website using flask to serve our model in Pycharm Application.

Before we proceed further we had already created an AWS account.In that AWS account we created the service EC2 instance and we deployed our website on that AWS EC2 instance.We have used Security Groups to create our own security code in it.We have downloaded and produced Key Generation (pem file) that we have used.

Lastly we needed a few packages more, that is Putty which is used to interact with the server directly and WinSCP which is used for file transfer to and from server.Then we have uploaded our Flask Website on Ec2 using WinSCP and packages are installed using Putty.

### iv. Proposed Methodology Result and Discussion:

All tests and experiments on the classifiers and algorithms described and laid out so far in this study were conducted using JUPYTER notebook libraries, Python 3, version 6.1.5, and scikit learning machine. We have divided and segregated our dataset in this study in a 70:30 ratio. Training accounts for 70% of the budget, while testing accounts for 30%. We chose JUPYTER because it includes a reputable collection of machine learning algorithms for pre-processing, clustering, classification, and regression that we could utilize on our dataset.

After training the model on our own dataset, we used the k-fold cross validation test to quantify the model's skill on new data or unexplored data sets.

# III. RELATED WORK

The main cause of breast cancer is when cells of the breast begin to grow abnormally. These cells divide rapidly than healthy cells then continue to accumulate, forming a lump or mass. Cells may spread through patients' breasts to their lymph nodes or to other parts of patients' bodies.

In Prateek P. Sengar, Mihir j. Gaikwad uses Logistic Regression and Decision Tree Classifier for accuracy Measures. The paper titled " Breast Cancer Detection Using Machine Learning Algorithms " used Random Forest, KNN(K-Nearest-Neighbor) and Naive Bayes algorithm and achieved accuracy 94% from each algorithm. Anusha Bharat and Pooja N & R Anisha Reddy(IEEE 2018) [4] compare the performance criteria of supervised classifiers such as KNN, Naïve Bayes, Logistic regression and SVM and achieve an accuracy of 99.1%.

In Kampreet S.Bhangu ,& Luxmi Sapra, International Conference on parallel, Distributed and Grid Computing (PDGC,2020) [4] here KNN ,Naïve Bayes , Logistic Regression, SVM and their accuracy came to be 98%. And last Nirdosh kumar, Gaurav Sharma [5] in that KNN, Logistic Regression, SVM, Naïve Bayes algorithms are used here we got logistic Regression=96.49%, SVM=98.24% KNN=97.20% Naïve Bayes=94.74% here we got better accuracy of SVM. We have used classification methods like KNN, SVM, Naive Bayes, Random Forest .

Literature survey For Machine Learning:

| Sr no | Literature Review | Author | Attributes | ML Algorithms | Accuracy Measures |
|---|---|---|---|---|---|
| 1 | Using Machine Learning algorithms for breast cancer risk prediction and diagnosis | Anusha Bharat, Pooja N & R Anishka Reddy (IEEE 2018) | Benign and Malignant | KNN, Naive Bayes, logistic regression and SVM (Support Vector Machine) | Accuracy of 99.1% for all algorithms. SVM using Gaussian kernel is the most suited technique for recurrence/non-recurrence prediction of breast cancer. |
| 2 | Breast Cancer Detection Using Machine Learning Algorithms | Shubham Sharma, Archit Aggarwal & Tanupriya Choudhury (IEEE 2018) | Diagnosis, Radius_mean, Texture_mean etc | Random Forest, KNN (kNearest-Neighbor) and Naïve Bayes algorithm | 94% is accuracy from each algorithm. KNN is the most effective in detection of breast cancer as it has the best accuracy, precision and F1 score over the other algorithms. |
| 3 | Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction | Prateek P. Sengar, Mihir J. Gaikwad & Prof. Ashlesha S. N agdive (ICSSIT 2020) | Perimeter, Smoothness, Compactness etc | Logistic regression , Decision Tree Classifier | a. Training Data: 75% b. Testing Data: 25% Here it can almost pinpoint accuracy using the Decision Tree Classifier algorithm. |
| 4 | Improving diagnostic accuracy for breast cancer using prediction-based approaches | Kamalpreet S. Bhangu, & Luxmi Sapra, International Conference on Parallel, Distributed and Grid Computing (PDGC,2020) | Concavity_mean, Concave points_mean, Symmetry_mean etc | KNN, Naive Bayes, logistic regression, SVM, XG Boost etc | Accuracy came to be 98%. XGBoost classifier with value of 0.99 performed better among all of them |
| 5 | The Machine Learning based Optimized Prediction Method for Breast Cancer Detection | Nirdosh Kumar, Gaurav Sharma & Lava Bhargava, (ICECA-2020) | Fractal dimension, Compactness etc | KNN, Naive Bayes, logistic regression, and SVM | Logistic Regression=96.49%, SVM=98.24 %, KNN=97.20 % and Naive Bayes=94.74 %. Here we have better accuracy of SVM |

Literature survey For Cloud Computing :

| Sr no. | Literature review | Author | Conclusion |
|---|---|---|---|
| 1. | Comparison and Analysis of Cloud Service Providers—AWS, Microsoft and Google | Dr. Manish Saraswat and Dr. R.C. Tripathi | Here the author made selections of vendors depending on business needs and technical requirements of particular companies i.e Aws, Azure, Gcp. Therefore AWS was found to be the best when we were looking for a platform with the broadest reliable and stable services when compared to Azure and GCP. |
| 2. | Cloud Computing Security: Amazon Web service | Saakshi Narula, Arushi Jain, Ms. Prachi | Cloud Computing provides many services but still they have problems that need to be solved to increase the market of world class technology. The concern is that cloud computing is Security around data protection. |
| 3. | A Survey on Health Care facilities by Cloud Computing | Puneet Saran, Durgi Rajesh, Hemant Pamnani. | Here the author stated the traditional ways are more prone to hacking because using the simple techniques as login and accessing the medical records by the patient or the doctor. They proposed an implementable plan with the help of the best security providing agency in the world . |

# IV. CONCLUSION AND FUTURE SCOPE

We proposed the implementation of breast cancer diagnosis model using two different machine learning algorithms, namely: SVM and KNN in Google Collab using Python Language.

These above-given algorithms gave satisfactory results.

In our predictions the accuracy came to be 98 percent a reasonable success of the experiment executed with the help of Machine Learning Algorithms.

Performance comparison of the machine learning algorithms techniques has been carried out using the Wisconsin Diagnosis Breast Cancer data set.

However, given the features according to that of this dataset, breast cancer can be predicted with almost pinpoint accuracy using our SVM and KNN algorithm.

Here we have also tried to store our data in cloud which will help with our implementation on a cloud platform for ease of usage.

Therefore AWS was found to be the best when we were looking for a platform with the broadest reliable and stable services and pricing is not a major challenge when the servers are running on windows.

The concern within cloud computing is Security around data, access and privacy protection. Cloud computing should be secure and robust and will mitigate the risks.

Amazon Web Service(AWS) has an outstanding performance in cloud computing because of the its excellent work in the area of Security of data. So building trust by providing security services was the main aim of choosing AWS in cloud computing.

The most important thing is that we have proposed an implementable plan with the help of the best security provided by AWS which almost solves the problem of the breach in accessing the medical files by the patient.

Furthermore, by taking more datasets there are higher chances of getting a better accuracy of the Breast Cancer Detection.

# V. REFERENCE

1. Anusha Bharat, Pooja N & R Anishka Reddy. On Using Machine Learning algorithms for breast cancer risk prediction and diagnosis. In 2018, IEEE Third International Conference on Circuits, Control, Communication and Computing.
2. Prateek P. Sengar, Mihir J. Gaikwad & Prof. Ashlesha S. Nagdive. On Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction. In the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020).
3. Shubham Sharma, Archit Aggarwal & Tanupriya Choudhury. On Breast Cancer Detection Using Machine Learning Algorithms. In the International Conference paper IEEE 2018.
4. Kamalpreet S. Bhangu, Jasminder K. Sandhu & Luxmi Sapra. On Improving diagnostic accuracy for breast cancer using prediction-based approaches. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC).
5. Nirdosh Kumar, Gaurav Sharma & Lava Bhargava. On The Machine Learning based Optimized Prediction Method for Breast Cancer Detection. In the Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020).
6. Maglogiannis, I., Zafiropoulos, E., & Anagnostopoulos - An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers.
7. M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan. On the Prediction of Breast Cancer using support vector machines and K-Nearest neighbors. In the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, 2017.
8. Sharma, Shubham, Archit Aggarwal, and Tanupriya Choudhury. On "Breast cancer detection using machine learning algorithms." In the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). IEEE, 2018.
9. Hussain, Lal, et al. "Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies." 2018 17th IEEE International Conference
10. Abien Fred M. Agarap. On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing(2018).
11. Dr. Manish Saraswat and Dr. R.C.Tripathi. On Cloud Computing: Comparison and Analysis of Cloud Service Providers - AWS, Microsoft and Google. In the 9th International Conference on System Modeling and Advancement in Research Trends (IEEE- Smart 2020).
12. Saakshi Narula, Arushi Jain and Ms.Prachi. On Cloud Computing Security: AMAZON WEB SERVICE. In 2015 Fifth International Conference on Advanced Computing and Communication Technologies.
13. Puneet Saran, Durgi Rajesh, Hemant Pamnani, Shikhar Kumar, T.G. Hemant Sai and Shridevi S. On A Survey on Health Care Facilities by Cloud Computing. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE).
14. Noman Islam and Aqueel ur Rahaman. On ”A Comparative Study ofMajor Cloud Service Providers for Cloud Computing ". In IEEE 4th International Conference on Computer and Communication Systems(ICCCS); 2019
15. Meiko Jensen, JorgSehwenk et al. “On Technical Security Issues in Cloud Computing”. In IEEE International Conference on Cloud Computing.
16. Deyan Chen, Hong Zhao. On the Data Security and Privacy Protection Issues in Cloud Computing. In 2012 International Conference on Computer Science and Electronics Engineering (2012 IEEE).