Breast cancer prediction using machine learning

Bala Sujith Reddy Basani
700742601
dept.Computer Science
University of Central Missouri
bxb26010@ucmo.edu

Ruchitha Ekkaladevi Srinivas
700742174
dept.Computer Science
University of Central Missouri
rxe21740@ucmo.edu

Ranjith Rao Jupalli
700733541
dept.Computer Science
University of Central Missouri
rxj35410@ucmo.edu

Sharmila Spoorthy Pothi Reddy
700741830
dept.Computer Science
University of Central Missouri
Sxp18300@ucmo.edu

***Abstract*—Breast cancer is a condition where the growth of the cell is abnormal. Most breast cancers are seen in the lobules part of the breast. Breast cancer can spread to the outer parts of the breast.Once the patient starts to see the symptoms like a lump in the breast and blood discharge from the breasts the next step is to screen the breasts. There are multiple screening methods available to find breast cancer like mammogram, breast ultrasound, breast examination and Fine Needle Aspiration(FNA).Out of all the methods FNA is considered to be the definitive method to rule out the cancerous cells in the breast. The process of FNA includes collecting breast tissue from the suspected area with the help of the needle and this is studied under the microscope.Data to predict the cancer comes in multiple dimentions.In our dataset we have 32 real-valued features computed for each cell nucleus.These features include radius, texture and concavity parameters.To analyze the huge number of parameters and looking for patterns is difficult for traditional methods. Machine learning methods solve the problems of multi dimensional data easily with pattern recognition as the core responsibility.Recent studies have shown that we can gain 95 percent accurate results on the Wisconsin Breast Cancer dataset with machine learning methods. Since the manual methods consume more time, automating the task makes the process easy. Methods to arrive at the solution to this problem are PCA and LDA. Principal Component Analysis and Linear Discriminant Analysis methods are used to reduce the dimensions of data. Principal Component Analysis (PCA) applied to these features and calculates the principal components, or directions in the feature space that account for the most variance in the data.In the end a comparative study of both methods are conducted with the supervised machine learning methods include Support Vector Machine and Logistic regression.**

***Index Terms*—Dimentionality reduction, FNA(Fine Needle Aspiration), LDA(Linear Discriminant Analysis), PCA(Principal Component Analysis), Screening methods**

I. Introduction

Breast cancer is an abnormal growth of the cells. There are different kinds of breast cancers the type breast cancer depends on where the abnormal growth of cell starts on which part the growth of the cell goes out of control. Looking at the anatomy of the breasts contain three main parts. Those parts are lobules, connective tissues and ducts these 3 do different taks. Lobules are the glands, ducts are the tube shaped parts. the purpose of lobules is producing the milk, where as the function of ducts is to carry the milk to the nipple and the function of connective tissues is to keep everything together. Breast cancer is called metastatized when the beast cancer speads outside of the breast. The transfer of the disease starts at the breast and can be spread to other parts of the body through blood vessels.

There are different kinds of breast cancers in nature. But the most common breast cancer types are: Invasive ductal carcinoma and Invasive lobular carcinoma. The first category of the cancer starts in the ducts and starts to spread out of the ducts. This Invasive ductal carcinoma is spread to the other parts of the body or metastatised. The second type of breast cancer starts in the lobule part of the breast and can spread to the tissue part of the breast and it can be spread to other parts of the breast.In additiona to the most common breast cancer there are few types of breast cancers those are less common in nature. One of them is Paget's disease. Paget's disease is the rare breast cancer type which occur in the breast part of areola which is the black circle part of breast. This cancer can not exist alone this comes with the pre existing tumors like lobule or ducts tumor. By the time we find pagets disease the patient will have at least one or two tumor in lobule or ducts.Another rare type of breast cancer is Inflamatory breast cancer. Inflamatory breast cancer starts at the lyph vessels and starts to spread to the other parts of the body. Most common type of inflamatory breast cancer is Invasive duct cancer it blocks the milk carrying vessels. This cancer is called is inflamatory cancer because it makes the breast swollen and red.This breast cancer type is rare because it appears only 1 to 5 percent of the people.

Patient starts to see various changes in the breast when diagnosed with breast cancer. Most common symptoms include:

- One of the most common symptom to diagnose the disease is locating the tumor in the breast that is lump in the breast area
- Abnormal changes in the shape or size of the breast
- Experiencing dimpling of skin around the breast
- Having a new inverted nipple on the breast
- Peeling and scaling of the skin around areola and inflammation

The risk factors of the breast cancer disease decide the progress of the disease. If the risk factors grow your disease progress also grows. But few risk factors like hereditary and age we can not control them. We can control few risk factors. Having too many risk factors does not mean that we have the disease growing. Governable risk factors are life style changes:

- Drinking alcohol is one of the main risk factor that we can have control over. If the woman takes 1 drink per day that increases the risk of having cancer by 7 - 10 percent compared to the people who do not drink. This statistics change for the woman who takes two drinks per day. If the woman drinks 2 drinks per day the risk of having cancer increases by 20 percent from this we can observe that drinking alcohol has relation with the increasing trend of getting the disease.
- Another risk factor that governable is being obese, studies say that obese women have high blood pressure and high insulin levels. Having high insulin levels have the direct impact on having the cancer to certain percentage. And the increasing risk of the disease depends on the obese before and after menopause. People who became obese after the menopause are more prone to the disease than the people who are obese before the menopause.
- Another most important risk factor that we can control is being physically active. Generally we can avoid most of the diseases with this factor. But this is the most ignored factors due to different reasons. Though we do not have the evidence how being physically active reduces the intensity of the disease. But it reduces the disease progression. Spending maximum of 150 minutes and minimum of 300 minutes on moderate exercise or 75 - 150 minutes on intense activities will reduce the progression the cancer drastically.
- The other factor of reducing the disease progression is breastfeeding. Most studies show that breastfeeding is a key to reduce the breast cancer disease. Breast feeding is two way benefit it helps the child in so many ways like high immunity, lower risk of Infant Death Syndrome and lower long and short term diseases. At the same time mother can also benefit from breastfeeding the bay i.e. lower risk of diabetes levels and breast cancer.

There are factors that we can not control them. Those are age, being born as female, family history, personal history of breast cancer. Being born as female: This is one of the risk factors that we can not control. This disease is occurs in men also this is much more common in women. This factor can not be controlled. The second risk factor is family history can not be controlled because of its nature. The last important risk factor that can not be controlled is personal history of breast cancer. These factors can not be controlled with any external conditions. So these can not be considered to control the affects.

After finding the symptoms in the women body next step is to go to screening and with the health professionals. First they will check the options for the screening method after the shared discussion both the patient and doctor will decide the screening methods which is a best option for the patient's health. The screening method can not help to cure the disease it can help to diagnose the disease which is a important step in breast cancer prediction. The screening methods help to diagnose the disease in early stages before it becomes the severe. There are different types of breast cancer screening methods:

- Mammogram: Mammogram is an X-Ray of the breast where it shows the lumps or any abnormal changes of the breast. This is one of the easy method of screening for any age groups. Having regular mammogram check ups help to find the disease early and reduces the chances of dying.
- Breast MRI(Breast Magnetic Resonance Imaging): The type of screening method is selected based on the risk level also. People who are at low risk level will be suggested few types of screening methods and the people who are at the high risk will be suggested different screening methods the reason is the cost of the screening test and accurate results of the tests. In the breast MRI method breast image is taken and studied the chances of getting breast cancer. Breast MRI is not suggested for low risk category because it shows abnormal things even when the disease is normal. So Breast MRI is studied in combination with mammograms for better results. For this reason this screening method is suggested for high risk patients.
- Other than screening methods we have clinical exams also to find the disease. One of the clinical exam is Clinical Breast Exam this is a non invasive method it can be used as a basic step to find the lumps in the breast. In this test nurse or doctor uses his/her hand to feel the lumps in the breast.
- Self Awareness is the key to diagnose the disease early and reducing the chances of dying. regular checking of breasts for changes in breast color and size is saves the lives of the peoples from breast cancer. Cancer is the second leading cause of death. In that breast cancer is a common type of cancer. This disease is much more common in women than men. Women with this disease may die it is fatal. With the advancement of treatments and AI diagnosing procedure became simpler and much more easy. With the advancement of AI/Machine learning we can analyse the images and data find the patterns in the data and we can predict the disease. Machine learning is the most sought after technology for finding the patterns and disease prediction. Artificial Intelligence is more advanced technology than machine learning which uses sophisticated methods to predict, classify and detect the diseases. Machine learning is implemented in various languages Python, Java script are the most common languages. Implementing the machine learning algorithms with Python gives the advantages of abundant libraries. We have multiple libraries that can be used to implement

the algorithms. Pytorch, Tensorflow, keras and sklearn libraries help to implement the algorithms otherwise these are implemented from scratch. Implementing the algorithms gives the flexibility of just reusing without much effort. Our project is a binary classification problem. In this project we are using sklearn library to implement algorithms.

In addition to implementing the algorithms we have other libraries also like numpy, pandas and matplolib for mathematical analysis, data analysis and data visualization. In addition to matplotlib we have other sophisticated data visualization tools seaborn and plotly. The difference between matplolib and seaborn is the aesthetic appearance of the visualizations. Plotly and seaborn gives the dynamic visualizations.

Always the clinical data comes with high dimensions of data. In machine learning we have multiple methods to reduce the dimensions of the data. One of them are PCA(Principal Component Analysis) and LDA(Linear Discriminant Analysis). These are two different methods of dimensional reduction methods. PCA isa non linear dimentionality reduction method where as LDSA is a linear dimentionality reduction method. In our project we have impleented PCA and LDA both the dimentionality reduction methods with the machine learning algorithms. After implementing the algorithms we have compared the performance of the algorithms for a comparative analysis. Both PCA and LDA are implemented from sklearn library. For PCA we have constructed the covariance matrix. Later we have calculated the eigen values and eigen vectors. From eigen values and eigen vectors explained variance is calculated to know the maximum variance of the data.

II. Motivation

Breast cancer is a condition of abnormal growth of the cells in the breasts. The anatomy of the breast focuses on 3 main parts: lobules, connecting tissues and ducts.Lobules produce the milk, ducts carry the milk to the nipples and connecting tissues bind it together.There are different kinds of breast cancers in nature.The type of the cancer depend on where in which it started.Most breast cancers are seen in the lobules. Breast cancer can spread to the outer parts of the breast.Once the patient starts to see the symptoms like a lump in the breast and blood discharge from the breasts the next step is to screen the breasts.There are multiple screening methods available to find breast cancer like mammogram, breast ultrasound, breast examination and Fine Needle Aspiration(FNA).Out of all the methods FNA is considered to be the definitive method to rule out the cancerous cells in the breast. The process of FNA includes collecting breast tissue from the suspected area with the help of the needle and this is studied under the microscope. After analysing the samples under the microscope we have used machine learning algorithms to find the patterns of the data to predict the type of the cancer.

III. Objectives

Main objectives of the project are:

- Reduce the dimensionality using PCA and LDA
- predict or classify the type of the cancer benign or malignant using machine learning classifiers
- If the predicted class is benign that means the lump is non cancerous this could be due to hormonal changes.Medical examination is suggested for further changes.If the disease is malignant it says the presence cancerous cell in the breast and the further steps are finding the stage of the cancer for appropriate treatment.
- Conducting comparative analysis on the results achieved
- Publishing the results in the web application

IV. Related work

Breast cancer is the most common cancer. This can be seen in 1 in 100 women. Occurrence of this disease depends on the age factor it is mostly seen in the elder women whose age is more than 40 years.But the current research methods show that the age bracket is now decreasing. Researching the risk factors of the breast cancer in younger women is the less research topic. So in this topic we have explored the data of breast cancer data of the women whose age below 40 years. We have implemented the c5.0 algorithm on the data. Experimental results show that the accuracy of the algorithm shows excellent performance than the machine learning algorithms [20].

Breast cancer is the diagnosis of the malignant tumors in the breast of a woman or a man. This disease takes a toll on the health of the patient even causing death sometimes. In our paper we are implementing the combination of Random Forest and Ada Boost classifier algorithms which is an ensemble method. Ensemble methods show highest accuracy on the pattern finding. In this paper we are classifying the given input into benign or malignant tumors.After performing ensembles algorithms we compared the results of the ensembled model results with the single Random Forest and SVM models. The experimental results show that the accuracy of the ensemble model is improved by 4.8 in the lower limit and 9 percent as the upper limit [19].

After diagnosing the cancer disease next step is to get the chemotherapy done. But the process of chemotherapy is very painful. Advance methods like Neoadjuvant Chemotherapy(NAC) is the main treatment option when it comes to breast cancer. But before taking this therapy we need to take pCR test to find the number of series of steps to treatment. For this machine learning used and the accuracy of the model is evaluated before applying the in real time application [6].

With the increasing trend of breast cancer in women WHO(World Health Organization) released the report of the patients who died with the disease in the year of 2018 and number is 620k which is huge mortality it is close to 15 percent. Solution to this problem is diagnosing the disease in early stages reduces the risk of fatality. In the screening method image of the breasts predicts better than the other methods. So in this project we are implementing the image classification models for predicting the breast cancer [1].

Medically before going to any screening tests there is a much advanced topic of finding the breast cancer that is Estrogen Receptors(ER) which decides the breast cancer advancement. In this paper we have conducted study on breast cancer prediction using the status of ER whether it is positive or negative. Breast cancer samples are collected from cohort of breast in total they have collected the 278 samples. From these we have found the patterns using ANN(Artificial Neural Network) and the results are compared using comparative analysis [5].

Machine learning is growing popular for disease prediction especially cancer. With popularity of machine learning in future the advancement may even lead to detecting the cancer at our fingertips even without going to the hospital. In this paper we have designed a model to predict 3 types of caners lung cancer, breast cancer and prostrate cancer. We have given the model different attributes of the cancer type while training the model. and we have used different models for different kinds of cancer. We have used SVM for prediction of breast cancer and for lung and prostrate cancer we have used Random Forest algorithm. The attributes of lung cancer include yellow fingers, smoking, anxiety and pressure.For breast cancer radius, texture and area are used [16].

Breast cancer is growth of the cells going out of control. In this paper we have analysed the risk factors of the breast cancer like family history, physical activity, increase in breast size and psychological stress are considered these are not the exhaustive list of the parameters in total 12 features are considered and the instances are 275. To classify the data we have used ensemble models Random Forest and Extreme Gradient Boosting (XGBoost) is used. In the result analysis we have got 74 percent accuracy for Random Forest and 5 percent for XGBoost algorithm [8].

Breast cancer is the one of the main cause of death in women. Machine learning is the novel approach that is used in the many medical applications for detecting the type of cancer cells whether it is benign or malignant. There are many algorithms for making classifications SV(Support Vector Machine), NB(Naive Bayes), Decision Tree(CART), k Nearest Neighbors(k NN). These algorithms are implemented on wisconsin dataset. As a final result we have compared the accuracies of the each algorithms [4].

In today's world the task machine learning is pivotal when predicting the diseases. One of the concerning cancer diseases is breast cancer. In this paper we have used various classification algorithms and clustering algorithms. Classification algorithms are supervised machine learning techniques and clustering algorithms are unsupervised learning algorithms. In conclusion we have found that supervised classification algorithms are far superior than the clustering algorithms [7].

One of the cancers that are fatal in many women after skin cancer is breast cancer. The data is homogeneous so to predict the data we need superior networks to find the patterns. After machine learning deep learning techniques play an important role in finding the patterns of the data. In this paper we are using ANN(Artificial Neural Network) to predict the results of the data. Using ANN we have classified data into benign or malignant. When we compare the performance of the traditional algorithms it showed superior results [14].

Breast cancer can be analysed using various machine learning techniques. After uisng AI tecniques on Wisconsin dataset we have concluded that SVM(Support Vector Machine) is the best algorithm to find the patterns on the data. SVM performs well on the small datasets since the breast cancer disease dataset is a small dataset we have implemented the SVM algorithm on the dataset [9].

Breast cancer is one the challenges in the medical industry. To tackle this problem machine learning and data mining techniques provide simple and efficient methods to solve the problem. Once breast cancer occur and cured by certain methods there is a chance that it can occur at any time. Most common time to re occurrence of the disease is 3 - 5 years. In this paper we have used associating rule of data mining for investigating the occurrence of the disease using SEER dataset(Surveillance, Epidemology and End Results) [17].

Breast cancer or in general any disease contains multiple causes or factors that affect the person. To find the most important features is challenging task. If we find the most important factors that affect we can primarily focus on those factors. In general machine learning or data mining primary goal is to find the patterns and filter the most important features based on filter techniques. In this paper we are using the most efficient data mining technique that is attribute filtering for finding the important features. From the Wisconsin dataset we have filtered the important features after filtering the features we have used classification algorithms like Naive Bayes, Decision Tree and KNN algorithms [15].

The screening methods are the primary steps in the diagnosis of the disease. Patients are called for the screening tests depending on the severity of the disease. The number of the visits of the patient increases with the severity of the disease. But in general the number of visits and attendance plays an important role. In this paper we are proposing a novel approach to classify the attendance of the patient. We also filter the percentage of

the non attendees. In the experimental analysis we have observed that we not only focused on attended patients we have performed well on the non attendees also. To perform classification we have used AI-ATT algorithm(AI-Attendance) and we have scored 76 percent accuracy [2].
Breast cancer can be caused by many parameters.In this paper we are proposing a hybrid algorithm to classify the cancer type. In the hybrid approach we have used the MAD normalizing technique to normalize the data since the data is diverse. After normalizing the data we have used K-means clustering for weighing the data to create the cluster groups. After clustering the data we have trained AdaBoostM1 classifier to classify the given samples. For a baseline algorithm we have used Normalised data (MAD) and AdaBoostM1 classifier. And in the second stage we have implemented the hybrid approach and the results are 75 percent and 92 percent. In this we have used a data set of 116 patients records where the dataset is divided into 52 and 64 data group here 52 is the healthy group and 64 is the patient group. And the features are age of the patient, BMI of the patient, Glucose levels of the patient, insulin levels etc., [11]
With the increasing trend of cases of breast cancer in women. Our goal of this paper is to classify or analyse the features of the malignant tumor of the patients. We have gathered the features of the growth of the malignant tumor in the patients and separated the data of the people whose disease stage is normal i.e. the growth of the tumor is not progressive. To classify or categorize the features from the malignant tumor to benign tumor for the classification purpose we have deployed the Decision Tree algorithm. To analyse the algorithm performance we have used accuracy(AUC), time complexity, sensitivity and AUC value [12].
One of the main reasons of the fatality of the breast cancer is it is recurrent it can cause many times. So the precautionary steps taken by the patient is he/she should be careful of the disease recurrence. With self awareness and regular check ups we can avoid the problem. In this paper we have proposed the algorithm to classify the data from recurrent to non recurrent data samples. We have collected the data from Wisconsin breast cancer dataset. To train the data we have used Holo Entropy enabled Decision Tree(HDT).After training the algorithm we calculated the various performance measurement parameters and analysed based on them [13]. Since breast cancer is most progressive disorder we need best predictive models to predict and track the progress of the disease. The purpose of this paper is to classify the cancer prediction using GA(Genetic Algorithm), Fuzzy logic neural networks one is RFNN(Recurrent Fuzzylogic Neural Network) and AFNIS network to classify the data.For this task data is collected from UCI machine learning repository and the data has 116 instances in that we have used 82 samples for training purpose and 34 instances for testing purpose. In this data we have considered the records of both the healthy person and the patient records. To measure the performance of the algorithms we have used precision, accuracy and specificity are used and a result summary is created using the comparative analysis [18].
One of the countries that has the high rate of deaths recorded due to cancer like cervical and breast cancer is Indonesia. In the year of 2013 the number is 623k which is huge number of deaths in women alone.As per the recent studies early detection of any cancer can have multiple advantages like increasing the chances of survival, cost of treatment and cure of the disease. With this step we have collected the mobile perosoanl health record data from Hospital Surabaya and conducted a classification task on breast cancer.We have used Logistic Regression, Naive Bayes and other algorithms [3]. Apart from the traditional symptoms we can diagnose the disease using appearance level of the patient. In this paper we have implemented the Find-s and elimination algorithm to classify the data. As a baseline algorithm we have implemented the Naive Bayes algorithm and compared the results of the both algorithms [10].

V. Data Description

Characterstics of the data:

- Breast Cancer Wisconsin (Diagnostic) Data Set is collected from UCI machine learning repository.Dataset includes different features collected from breast biopsy i.e. Fine Needle Aspiration (FNA) of the breast
- Dataset format is CSV
- Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast biopsy.
- Features in the dataset describe characteristics of the cell nuclei present in the image
- All feature values are recorded with four precision
- Dataset contains 561 instances with 32 features.Missing attribute values: none
- Target Class distribution: 357 benign, 212 malignant

S.No.	Column name	Description
1.	Radius	mean of distances from center to points on the perimeter
2.	perimeter	Cell nuclei perimeter
3.	area	Cell nuclei area
4.	smoothness	local variation in radius lengths
5.	compactness	$\text{perimeter}^2/area - 1.0$
6.	concavity	severity of concave portions of the contour
7.	concave points	number of concave portions of the contour
8.	symmetry	Cell nuclei symmetry
9.	fractal dimension	"coastline approximation" - 1
10.	Diagnosis	Diagnosis ('M': Malignant 'B': Benign)
11.	texture	standard deviation of gray-scale values

Table I
Dataset feature description

VI. Proposed Framework

Project workflow is broken down into 5 stages. The first stage is preprocessing stage.Preprocessing or data cleaning is the primary step in any machine learning

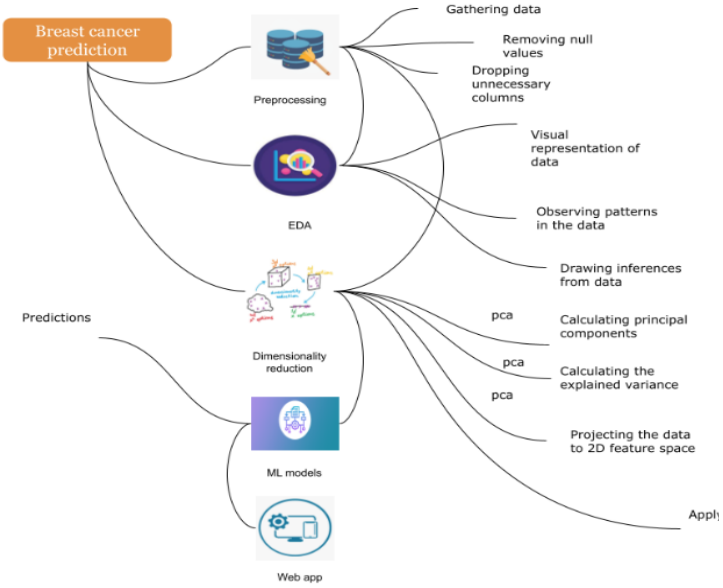

Figure 1. Workflow

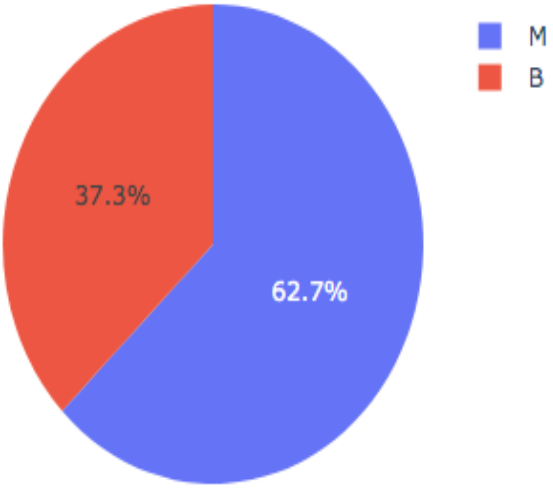

Figure 2. targets distribution

pipeline.In this step all the unnecessary data is removed and the goal of this step is to clean the data and make it ready to feed the machine learning model.In the second step Data is explored to find the patterns and distributions.The goal of this step is to find the outliers and draw observations.The above steps are very common steps for any project.The third step depends on the nature of the data.And from this problem specific methods applications starts.In the third step we have chosen dimensionality reduction methods PCA(Principal Component Analysis) and LDA(Linear Discriminant Analysis).After applying these methods we can be able to project the data into lower dimensional space.This methods makes the model interpretability easy.These methods are different from feature selection methods where they eliminate they features.Here, It will not eliminate the features instead they create new features with existing ones.These new features are the linear combinations of the existing features.After reducing the dimensions data is feed to the machine learning models for predicting the results. In the final step a user interface is created to show the predictions using python and Flask framework.

A. Exploratory Data Analysis

Exploratory data analysis is the primary step in finding the patterns of the data. In this we have implemented the various plots using data visualization techniques. We have chose plotly data visualization for dynamic representation of data. For example correlation of the data, outlier finding and data distribution.

B. PCA

Principal component analysis is an unsupervised learning technique to reduce the dimensions of the data.PCA is

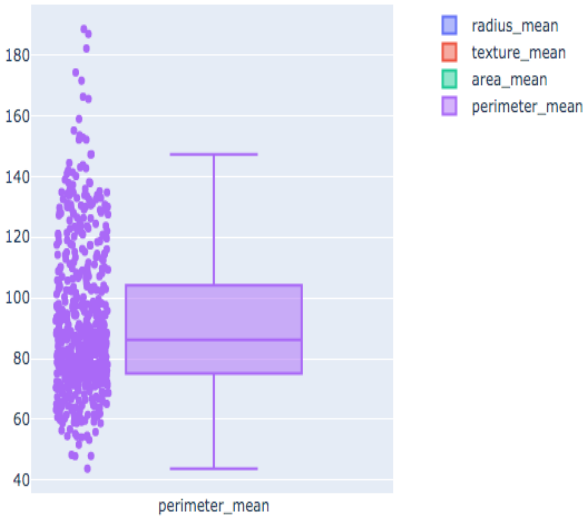

Figure 3. boxplot of perimeter mean

implemented using scikit-learn library.Data is split into training and testing.Train and test data is normalized before applying PCA. Calculated explained variance and cumulative variance and visualized using plotly data visualization.From the explained variance plot optimum number of components are selected to train the data. Train and test data is fit to the PCA object. In our project we have achieved 72 percent of the variance with the first 3 principal components

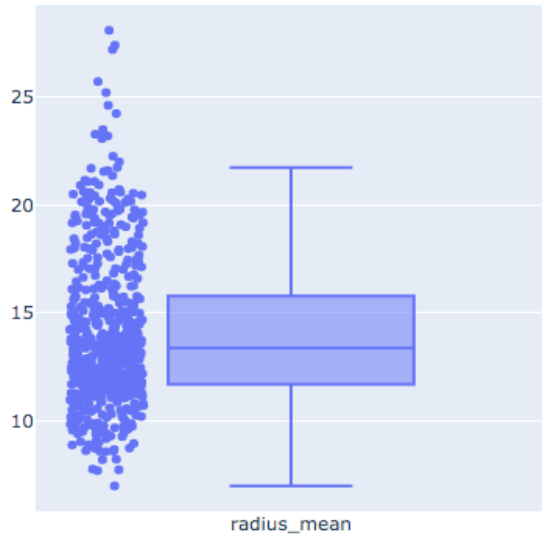

Figure 4. boxplot of radius mean

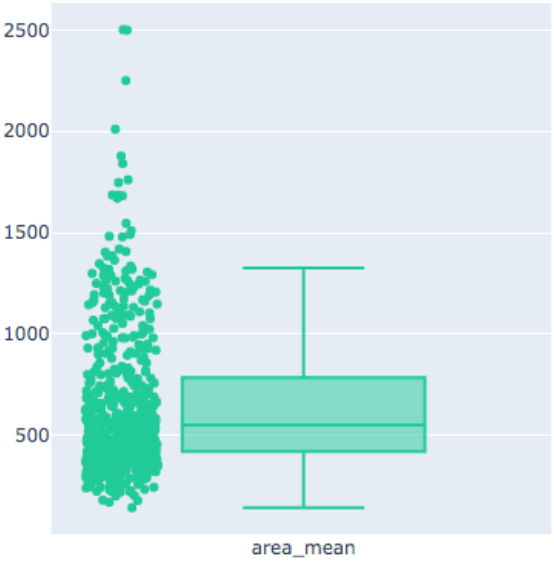

Figure 5. boxplot of area mean

C. LDA(Linear Discriminant Analysis)

LDA is implemented using scikit-learn library LDA takes the input solver and n components solver:Singular value decomposition (default). Does not compute the covariance matrix, lsqr: Least squares solution. eigen: Eigenvalue decomposition. To choose amongst the solvers we have performed GridSearchCV to select the best solver and the result is svd solver which uses the Singular Value Decomposition method.It is useful for a large number of features. N components is set to 1 which is (number of classes-1) this parameter affects the transformation.

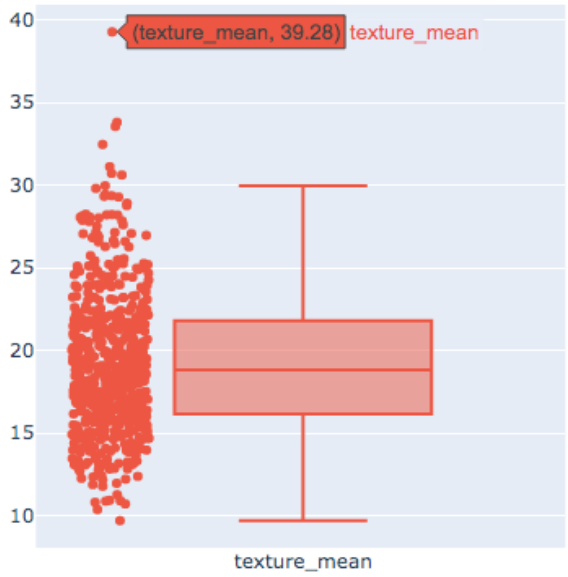

Figure 6. boxplot of texture mean

D. 3.SVM and Logistic Regression:

After reducing the dimensions data is fed to a classification models.In our project we have implemented SVM and Logistic Regression algorithms.To implement the algorithms scikit-learn library is used.Since the class is imbalance we are considering the precision,F1 score and recall parameters as metrics to evaluate the model performance.Best performing model is saved in pickle format and deployed into web interface

VII. Results Analysis

A. PCA Results

Principal Component Analysis (PCA) applied to this features and com (principal components, or directions in the feature space) that account for the most variance in the data. Here we plot the different samples on the 2 first principal components. For the 3 principal components we have achieved 72 percent explained variance.

B. Machine learning models

- **Logistic Regression:** Logistic regression is the simple classification algorithm implemented using scikit-learn library.
- Logistic regression achieved 94 percent accuracy on test data
- To understand the performance of the algorithm we have implemented classification report and confusion matrix.

From the classification report we can see the test accuracy of the model is 95 percent. The purpose of the classification report is finding the precision, recall and F1 score. Malignant category has highest precision that is 97 percent and benign category has highest recall

and F1 score that is 99 and 96 percent.The confusion matrix shows the classification rates for true and false positive and negative rates.Malignant category has the lowest incorrect classification samples that is 1 sample. Benign category has the highest incorrect classification samples that is 5.Total correctly classified samples are 108 samples.

	precision	recall	f1-score	support
B	0.93	0.99	0.96	72
M	0.97	0.88	0.93	42
accuracy			0.95	114
macro avg	0.95	0.93	0.94	114
weighted avg	0.95	0.95	0.95	114

Figure 7. Logistic regression classification report

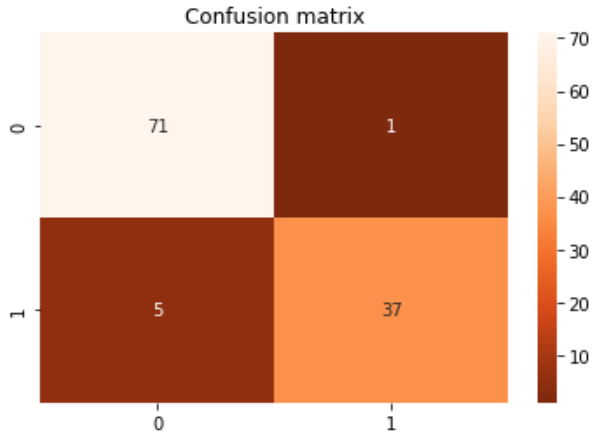

Figure 8. Logistic regression confusion matrix

– Support Vector Machine(SVM):

The second algorithm we have used for binary classification is SVM(Support Vector Machine). In the similar manner we have constructed the classification report and confusion matrix. From the confusion matrix we can observe that incorrect classification samples is high for benign class that is 6 and for malignant class it is 1. From the classification report we can observe that the test accuracy of the model is 94 percent. Precision of the model high for malignant class and recall and F1 score is high for benign class.

	precision	recall	f1-score	support
B	0.92	0.99	0.95	72
M	0.97	0.86	0.91	42
accuracy			0.94	114
macro avg	0.95	0.92	0.93	114
weighted avg	0.94	0.94	0.94	114

Figure 9. Classification report SVM

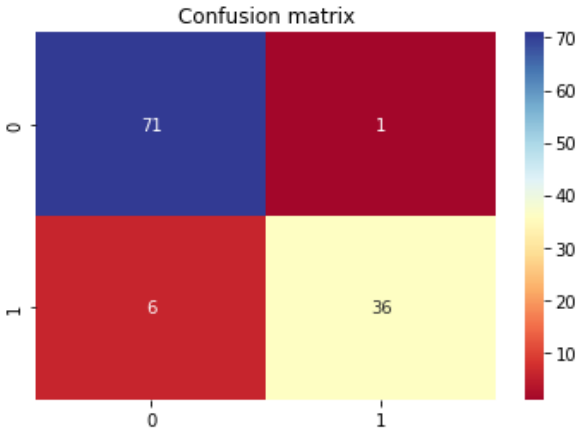

Figure 10. Confusion matrix SVM

C. LDA results

ROC-AUC curves are visualized for better understanding the of the accuracy of the models for Logistic regression and SVM have performed the PCA and LDA on Logistic Regression and SVM. Both the algorithms performed the same with the accuracy 95 percent.When compared to the SVM logistice regression performed well on the other parameters like precision, recall and F1 score.

	precision	recall	f1-score	support
0	0.95	0.97	0.96	72
1	0.95	0.90	0.93	42
accuracy			0.95	114
macro avg	0.95	0.94	0.94	114
weighted avg	0.95	0.95	0.95	114

Figure 11. Logistic regression classification report

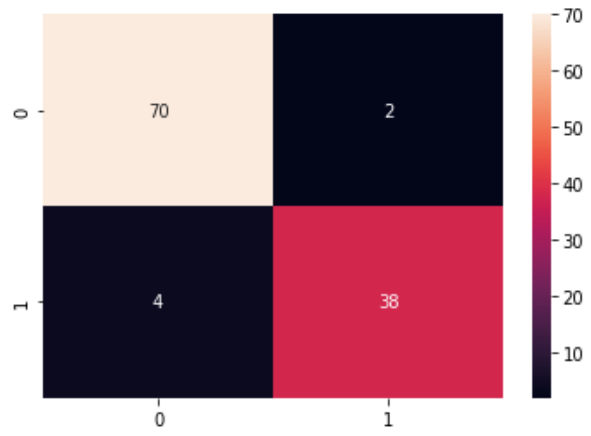

Figure 12. confusion matrix of logistic regression

	precision	recall	f1-score	support
0	0.92	1.00	0.96	72
1	1.00	0.86	0.92	42
accuracy			0.95	114
macro avg	0.96	0.93	0.94	114
weighted avg	0.95	0.95	0.95	114

Figure 13. Classification report of SVM

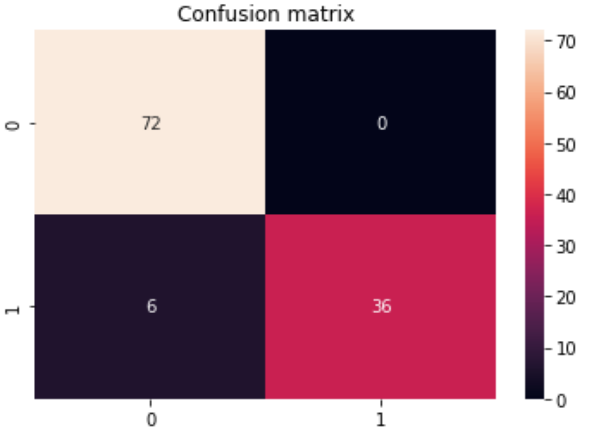

Figure 14. Confusion matrix of SVM

VIII. Results Summary

ROC-AUC curves are visualized for better understanding the of the accuracy of the models for Logistic regression and SVM have performed the PCA and LDA on Logistic Regression and SVM. Both the algorithms performed the same with the accuracy 95 percent.When compared to the SVM logistice regression performed well on the other parameters like precision, recall and F1 score. Logistic regression is best algorithm.

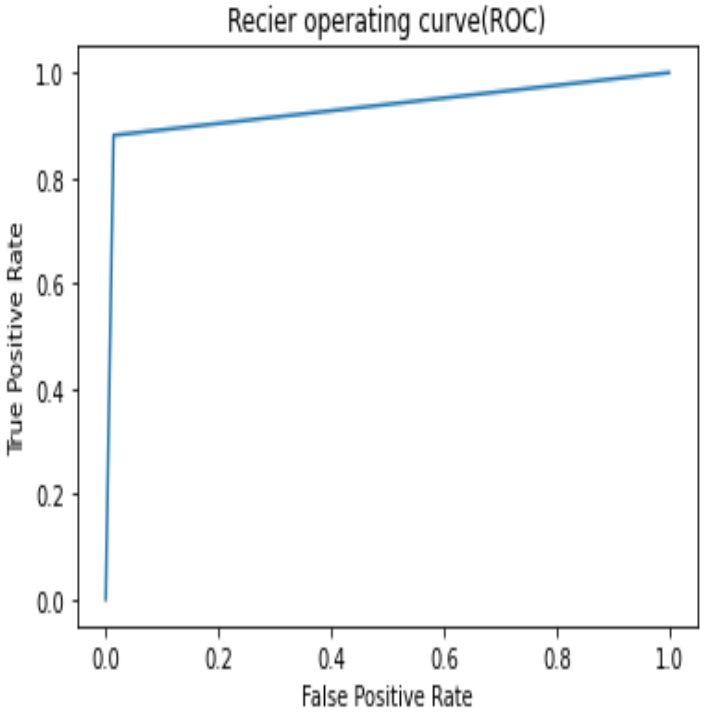

Figure 15. ROC curve

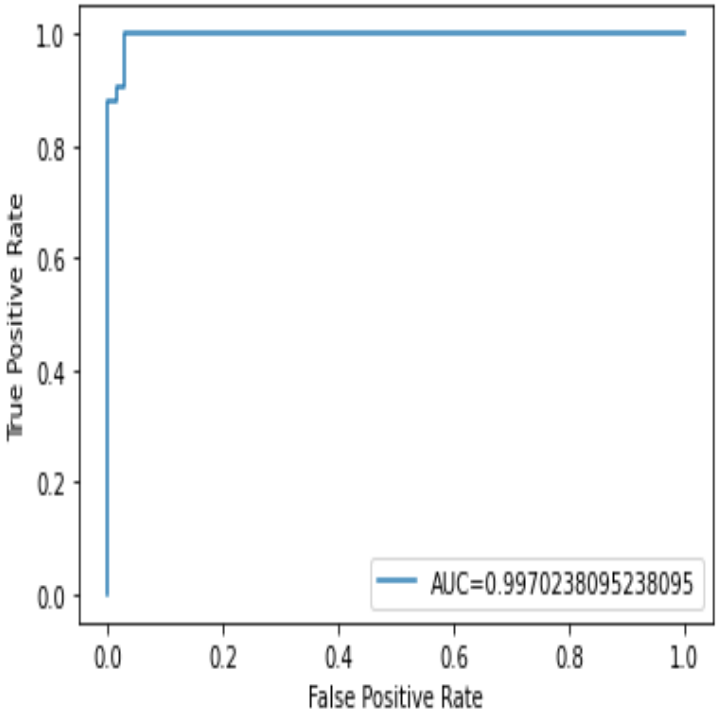

Figure 16. AUC curve

REFERENCES

[1] Yassine Amkrane, Mohammed El Adoui, and Mohammed Benjelloun. Towards breast cancer response prediction using artificial intelligence and radiomics. In *2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, pages 1–5. IEEE, 2020.

[2] Hisbel E Arochena, Raouf NG Naguib, Alison G Todman, Margot E Wheaton, and MG Wallis. Blind study for the prediction of attendance to a uk breast cancer screening unit. In *4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, 2003.*, pages 219–221. IEEE, 2003.

[3] Tessy Badriyah, Rimawanti Fauzyah, Iwan Syarif, and Prima Kristalina. Mobile personal health record (mphr) for breast cancer using prediction modeling. In *2017 Second International Conference on Informatics and Computing (ICIC)*, pages 1–4. IEEE, 2017.

[4] Anusha Bharat, N Pooja, and R Anishka Reddy. Using machine learning algorithms for breast cancer risk prediction and diagnosis. In *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, pages 1–4. IEEE, 2018.

[5] Gopal K Dhondalay, Dong L Tong, and Graham R Ball. Estrogen receptor status prediction for breast cancer using artificial neural network. In *2011 international conference on machine learning and cybernetics*, volume 2, pages 727–731. IEEE, 2011.

[6] Xin Feng, Lelian Song, Shaofei Wang, Haoqiu Song, Hang Chen, Yuxuan Liu, Chenwei Lou, Jian Zhao, Quewang Liu, Yang Liu, et al. Accurate prediction of neoadjuvant chemotherapy pathological complete remission (pcr) for the four sub-types of breast cancer. *IEEE Access*, 7:134697–134706, 2019.

[7] Dona Sara Jacob, Rakhi Viswan, V Manju, L PadmaSuresh, and Shine Raj. A survey on breast cancer prediction using data miningtechniques. In *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pages 256–258. IEEE, 2018.

[8] Sajib Kabiraj, M Raihan, Nasif Alvi, Marina Afrin, Laboni Akter, Shawmi Akhter Sohagi, and Etu Podder. Breast cancer risk prediction using xgboost and random forest algorithm. In *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–4. IEEE, 2020.

[9] Anuj Mangal and Vinod Jain. Prediction of breast cancer using machine learning algorithms. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 464–466. IEEE, 2021.

[10] CL Nithya, Sunanda Dixit, and BI Khodhanpur. Prediction of breast cancer using find-s and candidate elimination algorithm. In *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, volume 4, pages 1–4. IEEE, 2019.

[11] Kemal Polat and Umit Sentürk. A novel ml approach to prediction of breast cancer: combining of mad normalization, kmc based feature weighting and adaboostm1 classifier. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–4. Ieee, 2018.

[12] S Saranya and S Sasikala. Diagnosis using data mining algorithms for malignant breast cancer cell detection. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1062–1067. IEEE, 2020.

[13] Shabina Sayed, Shoeb Ahmed, and Rakesh Poonia. Holo entropy enabled decision tree classifier for breast cancer diagnosis using wisconsin (prognostic) data set. In *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 172–176. IEEE, 2017.

[14] Parag Singhal and Saurav Pareek. Artificial neural network for prediction of breast cancer. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on*, pages 464–468. IEEE, 2018.

[15] Ankita Sinha, Bhaswati Sahoo, Siddharth Swarup Rautaray, and Manjusha Pandey. Improved framework for breast cancer prediction using frequent itemsets mining for attributes filtering. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 979–982. IEEE, 2019.

[16] Ganta Sruthi, Chokkakula Likitha Ram, Malegam Koushik Sai, Bhanu Pratap Singh, Nikhil Majhotra, and Neha Sharma. Cancer prediction using machine learning. In *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, volume 2, pages 217–221. IEEE, 2022.

[17] DR Umesh and Bharathkumar Ramachandra. Association rule mining based predicting breast cancer recurrence on seer breast cancer data. In *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, pages 376–380. IEEE, 2015.

[18] Kaan Uyar, Umit Ilhan, Ahmet Ilhan, and Erkut Inan Iseri. Breast cancer prediction using neuro-fuzzy systems. In *2020 7th International Conference on Electrical and Electronics Engineering (ICEEE)*, pages 328–332. IEEE, 2020.

[19] Duan Yifan, Lu Jialin, and Feng Boxi. Forecast model of breast cancer diagnosis based on rf-adaboost. In *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 716–719. IEEE, 2021.

[20] Xia Zhang and Yingming Sun. Breast cancer risk prediction model based on c5. 0 algorithm for postmenopausal women. In *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 321–325. IEEE, 2018.