INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1978)
Available online at: https://www.ijariit.com

Breast Cancer Tumor Detection

Siddagangu S.
siddagangus.is17@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Nadhiem Latief
nadhiemlatief.is17@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Rekha B. S.
rekhabs@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Smitha G. R.
smithagr@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

Priya Bansal
priyabansal.is16@rvce.edu.in
RV College of Engineering, Bengaluru, Karnataka

ABSTRACT

Women in India confront serious fatalities such as respiratory difficulties, but some also confront serious diseases in the form of breast cancer, which differs for each lady; This dangerous illness has been detected in more than half of middle-aged women. Breast cancer is found by looking for lumps in women's breasts that look like tumors. These anomalies' cells can be treated right away if they're found. However, benign tumors, which are fully non-cancerous, can form lumps with cells of the same size and no structural changes between them. Machine Learning(ML) techniques are frequently utilized to create tools for physicians and diagnosis of carcinoma, which can dramatically improve patient survival rates. The goal of this project is to create a machine learning system that can predict if a tumor is benign or malignant, as well as visualize the properties of both cancers will next be illustrated via graph plotting. Support Vector Classifier(SVC), Logistic Regression, Decision Tree Classifier, and KNN are the used Machine Learning(ML) Algorithms for Breast Cancer Tumor Detection. To compare and assess the accuracy and ROC plotting performance of Machine Learning Classifiers.

Keywords: *Support Vector Classifier, Decision Tree, Knn, Roc, Machine Learning*

1. INTRODUCTION

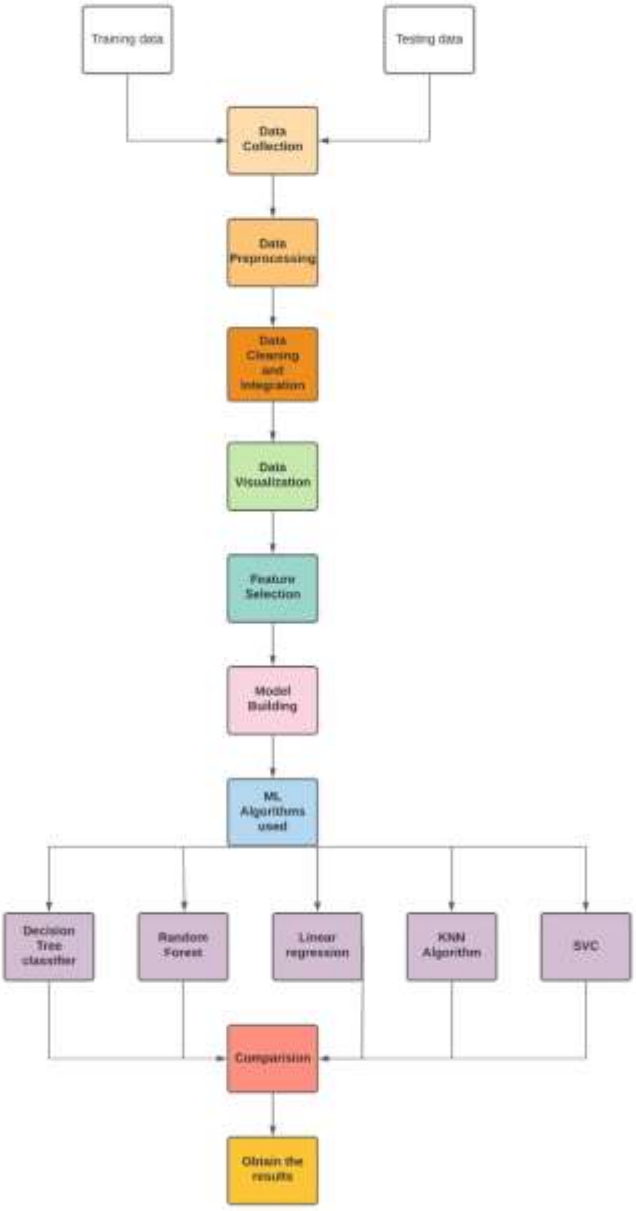
Breast cancer is a dangerous breast cell improvement. The condition spreads to many parts of the body if it is not addressed. Malignant growth, except skin cancer, is the most well-known type of malignancy in women, accounting for one out of every three malignancy diagnoses. Sound cells in the breast alter and outgrow their control, forming a lump or sheet of cells known as a tumor. We applied various machine learning approaches to detect breast cancer tumors in this article. To diagnose cancerous tumors with a high accuracy score, we used Support Vector Classifier(SVC), Logistic Regression, Random Forest(RF), Decision Tree Classifier, Ensemble Learning approach, and K Nearest Neighbors. Our goal is to apply Machine Learning Techniques to detect Breast Cancer tumors with the highest accuracy possible.

2. LITERATURE SURVEY

The standards of [1] are in reality altogether different from the late deals with how to battle dangerous malignant growth, it takes an extraordinary course out, The examination of anticipation is centered around three estimates which include: 1) foreseeing all the danger evaluations that may prompt the infection 2) anticipating what are the odds of the tumor to develop again and 3) odds of the patient getting by from the illness. Be that as it may, passing by the current situation of the most recent innovation [3] tracks down a profound learning technique by applying CNN for isolating X-beams which overwhelmingly performed superior to any past models, The outcomes were introduced on digitized film where just one model accomplishes a picture AUC score of 0.88 while improving the model to four averaging models gave the AUC score of 0.91. [4] gave an understanding of Artificial Knowledge strategies that are equipped for outperforming human specialists in bosom malignancy forecast. [2] introduced a fairly or fundamental procedure that thinks about calculations like Random forest, KNN (k-Nearest-Neighbor), and Naive Bayes. The results acquired were very fascinating and have all in all a decent measurement for additional medicines and discovery. [5] and [6] consolidated the procedure for a half-breed model of a picture preparing procedure with CNN that delivered high goal pictures into the highlights that were the very pinnacle of significance in grouping the tumors. [7] and [8] zeroed in on the significance of Region of Interest as opposed to zeroing in on the whole tumor cells and needed to go computational speed as contrasted with different models. [9] forced a novel technique for hereditary programming and AI calculations that precisely characterizes the right outcomes.

3. PROPOSED METHOD

The proposed method of Breast Cancer Tumor Detection follows which is shown in Fig.1.



Data Collection

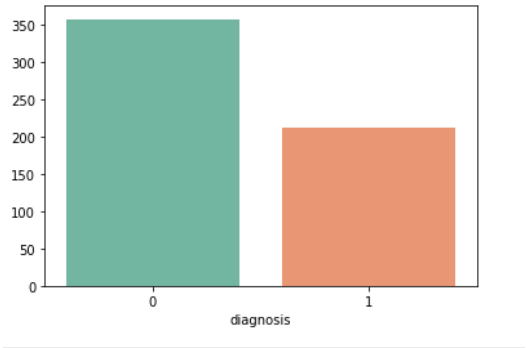
The dataset for Wisconsin Breast Cancer Diagnosis(WBCD) was gathered over the web. For analysis, the table has 568 rows and 64 columns. Comma-separated values(CSV) format is used to store the obtained data.

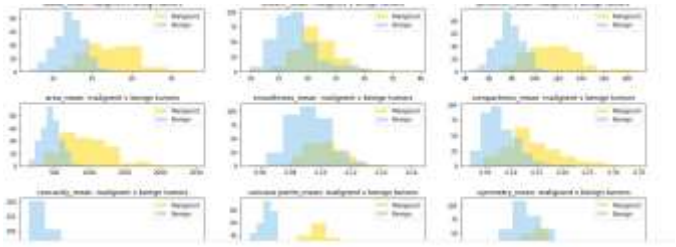
Data Preprocessing

Data preparation is the process of converting raw data into a usable format. In which unneeded columns were deleted

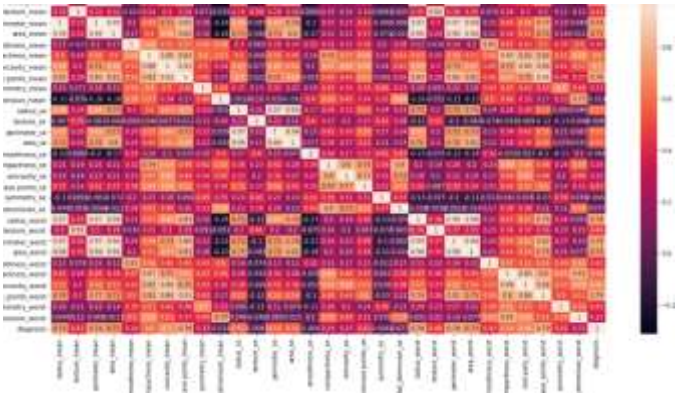
Data Visualization

We divided the dataset into two categories: malignant and benign, and then created a heatmap to visualize the features.





Heat map



Model Building

The process of selecting various Machine Learning(ML) algorithms is known as model building. There are a variety of ML approaches that can be applied. In this paper, the dataset has two types of dependent variables itself has two sets. Thus different algorithms are used.

```
from sklearn.svm import SVC
#To find the best parameters for the model
from sklearn.model_selection import GridSearchCV

param_grid = {'C': [0.01, 0.1, 0.5, 1, 10, 100],
              'gamma': [1, 0.75, 0.5, 0.25, 0.1, 0.01, 0.001],
              'kernel': ['linear', 'poly', 'rbf']}

grid = GridSearchCV(SVC(), param_grid, refit=True, verbose=1, cv=5)
grid.fit(X_train, y_train)

best_params = grid.best_params_

svc = SVC(**best_params)
svc.fit(X_train, y_train)
predict_svc = svc.predict(X_test)
```

```
#Applying logistic regression algorithm
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(random_state=0)
lr.fit(X_train,y_train)
predict_lr = lr.predict(X_test)
```

```
from sklearn.tree import DecisionTreeClassifier

dt = DecisionTreeClassifier()
params = {'criterion':['gini', 'entropy'],
          'random_state':[0]}
dt1 = GridSearchCV(dt, param_grid=params)
dt1.fit(X_train,y_train)
predict_dt = dt1.predict(X_test)
```

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier()
rf.fit(X_train,y_train)
predict_rf = rf.predict(X_test)
```

```
from sklearn.ensemble import VotingClassifier

model = VotingClassifier(estimators=[('lr',lr), ('dt', dt1), ('svm',svc), ('rf', rf)], voting='hard')
model.fit(X_train,y_train)
predict_en = model.predict(X_test)
```

Machine learning techniques

In ML(Machine Learning techniques, the learning interaction can be divided into two categories: Supervised learning, a set of data examples is used to train the machine and is to train the machine to produce the desired result. There are no pre-determined informational collections of the expected outcome, implying that the goal is more difficult to achieve. One of the

most well-known strategies in directed learning is characterization. It builds a model using recorded tagged data, which is then used to estimate the future. In the clinical field, hospitals and clinics maintain massive data banks with patient histories, diagnoses, and outcomes. As a result, analysts use this data to develop arrangement models that rely on verifiable cases for induction. The clinical deduction has thus become a considerably more straightforward task, thanks to machine-based assistance and the vast amount of clinical data now available. It's worth noting that this paper's full set of techniques is classified as a classification model.

Support vector classifier

An SVC's (Support Vector Classifier) purpose is to fit the data we provide, returning a "best fit" hyperplane that partitions or categorizes our data. Following that, after obtaining the hyperplane, we would be able to provide specific highlights to our classifier for it to determine what the "expected" class is. As a result, this computation appears to be realistic. SVC is taken in a voting mechanism.

Decision tree classifier

In machine learning algorithms, the decision tree algorithm is classified as supervised learning. In the tree structure, the decision tree is represented. The decision tree receives input based on specific criteria, and the output is shown as true or false. The node's values are determined by comparing each attribute.

Logistic Regression

The supervised learning classification of logistic regression is used to predict the likelihood of an objective variable. To sort the data into two categories, we used logistic regression. whether it's cancerous or benign.

Random forest classifier

RF (Random Forest) ensembles a forest of trees by combining multiple decision trees. The argument is that a single decision tree can offer either a simple or a highly specific model. The random forest provides more stability. This suggests that RF(Random forest) is unaffected by the input data set's noise. One The capacity of RF(Random forest) to manage data minorities is one of the key reasons for its use in cancer detection.

Ensemble learning method

Ensemble techniques are meta-calculations that combine several AI techniques into a single predictive model to reduce fluctuation, bias or increase predictions. For the model, we employed Majority Voting Classification, using Logistic Regression, Decision Tree Classifier, Support Vector Classifier, and Random Forest as estimators. Fig.7.Shows that Majority voting classification. For each test case, each of these model classification outcomes is computed, and the final output is projected based on the majority of the results.

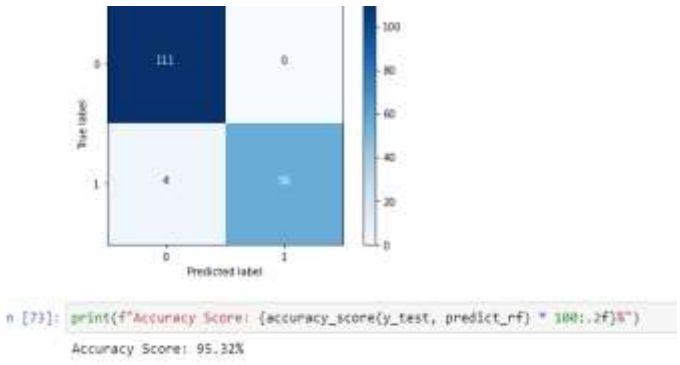


Fig. 7: Ensemble Learning Method.

K Nearest Neighbors

K-NN (K-Nearest Neighbors) is a learning method based on examples. The majority of an object's neighbors decide its classification in KNN. Which is not taken in the voting mechanism. K-NN can give better accuracy than other ML(Machine Learning) algorithms.

4. METHODOLOGY

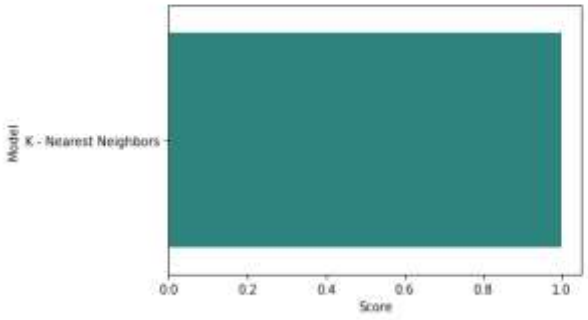
The classifiers SVC(support vector classifier), Decision tree classifier, and RF(random forest), Linear regression, ensemble learning, and KNN are all examples of linear regression (K-Nearest Neighbor). The model created from the trained dataset is applied to the test data set, and the results are obtained. The acquired outcomes are assessed using metrics such as the Accuracy score. Here, the ROC curve area is used, and a comparison is drawn, leading to the best approach being chosen. We used open-source Python software in this paper.

Confusion matrix

The actual class labels and the expected class labels based on the class labels by the classifiers are compared. Varied Classifiers such as SVC (Support Vector Classifier), Decision Tree Classifier, Random Forest, Linear Regression, Ensemble learning technique, and KNN yield the highest accuracy scores in this study. In this paper, KNN gives more accuracy than other ML(Machine Learning) algorithms.

```
Classification_Models(x_train,y_train,x_test,y_test)
```

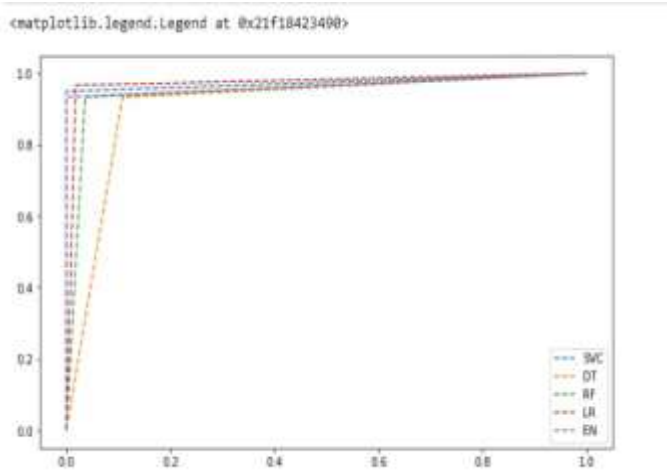
	Model	Score
0	K - Nearest Neighbors	1.0



5. RESULT AND DISCUSSION

Algorithms	Precision	Recall	F1-score	Support
Support Vector Classifier	0.97	1.00	0.99	111
Logistic Regression	0.98	0.98	0.98	111
Decision Tree	0.96	0.89	0.93	111
Random Forest	0.96	0.96	0.96	111
Ensemble Learning Method	0.97	1.00	0.98	111

The result of each model is compared with the training and testing set. Almost all the models worked well and gave above 90% accuracy in the prediction. By Considering the Confusion matrix, It is observed that KNN(K Nearest Neighbors) has more accurate predictions than others. The decision tree classifier applied in the data set gave an accuracy of 90.64% whereas the accuracy of SVC and logistic regression came as high as 98.25% and 97.66% respectively. Both the random forest and the ensemble learning model gave us an accuracy of 95.32%. Table 1 gives a clear comparison of the accuracies provided by different models used with respective F1 scores.



6. CONCLUSION

Breast malignancy offers an unprecedented setting for clinical end by similarly considering the patient's condition and their response to treatment. Machine Algorithms have shown a huge improvement in bosom disease development discoveries. Notwithstanding, challenges stay for the specific end and checking of bosom tumors regardless of improved advancements. There is a need to consolidate the natural social and fragment stream of information to improve the perceptive models. We have proposed a Logistic relapse to assess the exhibition, Support vector classifier, Decision tree classifier, Random Forest classifier, and ensemble learning strategy. We have proposed the presentation of these classifiers utilizing diverse execution measures i.e exactness, accuracy, review, F1 score. and support. We used different weighted mechanisms including macro average.

7. REFERENCES

[1] Mohammed S.A., Darrab S., Noaman S.A., Saake G. (2020) Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. In: Tan Y., Shi Y., Tuba M. (eds) Data Mining and Big Data. DMBD 2020. Communications in Computer and Information Science, vol 1234. Springer, Singapore.

[2] Siham A. Salah A. Noaman, Sadeq Darrab, "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques, "An international conference on Data mining and big data 2020:pp 108-117,11 july 2020.

[3] Ajit Kumar, Raj Kumar Patra, Anupam Ghosh, "Model Selection for Predicting Breast Cancer using Supervised Machine Learning Algorithms, " 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), DOI: 10.1109/ICCE50343,2020.

[4] Noreen Fatima, Sha hong and Haroon ahamed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," Received July 30, 2020, accepted August 9, 2020, date of publication August 14, 2020, date of current version August 26, 2020. Digital Object Identifier 10.1109/ACCESS.2020.3016715

[5] Raed Shubair, "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis, " Conference: The 2016 IEEE 5th International Conference on Electronic Devices, Systems, and Applications (ICEDSA'2016),Dec 2016.

[6] Puneet Yadav,Rajat Varshney,Vishan kumar Gupta, "Diagnosis of Breast Cancer using Decision Tree Models and SVM, "International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 03 , Mar-2018

[7] Nusaibah kh. AI Salihy,Turgay, "Classifying breast cancer by using decision tree algorithms," Proceedings of the 6th International Conference on Software and Computer Applications(ICSCA),2017.

[8] Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms", Journal of Healthcare Engineering, vol. 2019, Article ID 4253641, 11 pages, 2019.

[9] Hiba Asri, Hassan Al Moatassime, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis",Computer Science 83:1064-1069,December 2016.

[10] Moh'd Rasoul Al-hadidi, Abdulsalam Al Arabiyyah, Mohannad Alhanahnah, "Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm, "9th International Conference on Developments in eSystems Engineering,2016.

[11] Shubham Sharma, Archit Azarwal, Tanupriya Choudhury, "Breast Cancer Detection Using Machine Learning Algorithm, 2010 meramona Unlerenon Computational Techniques, Electronics and MechanicsSystems (CTEMS), 2012.

[12] Minghao Piao, "Discovery of Significant Classification Rules from Incrementally Indwara Decision Tree Ensemble for Diagnosis of Disease, lecture Notes in Computer Science, 2009.

[13] DI T Choudhury, V Kumar, D Nigam, B Mandal "intelligent classification of lung & oral cancer through diverses mining algorithms", International Conference on Micro Electronics and Telecommunication Engineering 2015.

