

Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer

Shishir Roy ^[1], Rutuja Gawande ^[2], Aman Nawghare ^[3], Shivam Mistry ^[4]

Department of Information Technology, Government College of Engineering,
Amravati, Maharashtra - India

ABSTRACT

This paper aims to present a comparative study of different machine learning algorithms for the detection of breast cancer. These algorithms include Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes Classifier, Decision Tree Classifier and Random Forest Classifier. This performance comparison is done on the Breast Cancer Wisconsin (Diagnostic) Dataset created by Dr. William H. Wolberg, a physician of the University Of Wisconsin Hospital at Madison, Wisconsin, USA. The performance metrics used for the comparison are accuracy, sensitivity, specificity, false discovery rate, false omission rate, Matthew's correlation coefficient and area under curve.

Keywords: - Machine Learning, Breast Cancer Wisconsin (Diagnostic) Dataset, Logistic Regression, K-Nearest Neighbour, Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, Naive Bayes Classifier.

I. INTRODUCTION

Breast Cancer is the most common cancer in India. Breast cancer is one of the top three leading causes of death among women worldwide [1]. In 2018, 27.7% of all new cancers detected in women were breast cancer, 1,62,468 women were newly detected with breast cancer and 87,090 women died of breast cancer in India [2]. India incorporates an overriding young population and hence the numbers of women being diagnosed with breast cancer is simply about to increase. Breast cancer cannot be prevented. If it has to happen, it will happen. However, the deaths due to breast cancer can definitely be reduced by early detection.

With an increasing count of breast cancer cases, comes huge data which is of significant use in further medical research, and there comes the application of data science and machine learning. Prior studies [10]-[17] have been proposed that use data mining and machine learning algorithms for the detection of breast cancer using various datasets with varying degrees of success.

The remainder of this paper is organized as follows - Section 2 presents the literature review, Section 3 describes the research methodology which includes introduction of machine learning libraries used, description of dataset, experimentations on data pre-processing, steps in data visualization, machine learning algorithms and performance metrics, Section 4 lists the experimental results and Section 5 shows the conclusion made.

II. RELATED WORK

There are numerous modern techniques that have been evolved with the evolution of technology for the prediction of breast cancer. The work related to this field is outlined shortly as follows -

The authors in [11] demonstrated a system for identification of breast cancer which is trained and tested for Breast Cancer Wisconsin (Diagnostic) Dataset using Support

Vector Machines, Random Forest and Bayesian Networks. The performance metrics such as Accuracy, Recall and Precision are calculated. The outcome illustrated an accuracy of 97%.

The authors in [12] demonstrated a system for identification of breast cancer which is trained and tested for Breast Cancer Wisconsin (Original) Dataset using Support Vector Machines and K-Nearest Neighbor. The performance metrics such as Accuracy, Sensitivity and Specificity are calculated. The outcome illustrated an accuracy of 98.57% and 97.14% respectively for SVM and K-NN.

The authors in [13] proposed an algorithm based on support value on deep neural networks for identification of breast cancer which is trained and tested on data obtained from M.G. Cancer Hospital and Research Institute, Visakhapatnam, India. The performance metrics such as Accuracy, Recall and Precision are calculated. The outcome illustrated an accuracy of 97.21%.

The authors in [14] demonstrated a system for identification of breast cancer which is trained and tested for Breast Cancer Wisconsin (Diagnostic) Dataset using Support Vector Machines, Random Forest and Naive Bayes. The performance metrics such as Accuracy, Recall, f1-score and Precision are calculated. The outcome illustrated an accuracy of 94.74%, 95.90% and 94.47% for RF, K-NN and NB respectively.

The authors in [15] demonstrated a system for identification of breast cancer which is trained and tested for Breast Cancer Wisconsin (Diagnostic) Dataset using Logistic Regression and Decision Tree. The performance metrics such as Accuracy, Recall and Precision are calculated. The outcome illustrated an accuracy of 94.40% and 95.10% for LR and DT respectively.

The authors in [17] demonstrated a system for identification of breast cancer which is trained and tested for Breast Cancer Wisconsin (Diagnostic) Dataset using GRU-SVM, Linear Regression, MLP, L1-NN, L2-NN, SoftMax

Regression and SVM with accuracy of 93.75%, 96.09%, 99.03%, 93.56%, 94.73% 97.65% and 96.09% respectively.

III. METHODOLOGY

A. Machine Learning Library

Jupyter Notebook [9] is used to implement the machine learning algorithms in this project with the help of other scientific computing libraries - scikit-learn [3], numpy [4], matplotlib [5], pandas [6], imblearn [7] and seaborn [8].

B. Dataset

The dataset used in this project is commonly known as Breast Cancer Wisconsin (Diagnostic) Dataset. Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA created this dataset.

Dr. Wolberg used Xcyt, a computer program, to perform the cytological analysis on the digitized images of the fine needle aspirate of the solid breast mass samples from patients that describe the characteristics of cell nuclei. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, then it calculates the (1) mean value, (2) standard error and (3) extreme value (worst i.e. mean of the three largest values) of each feature for the image, returning a 30 real-valued vector [10].

The ten features include - (1) radius (mean of distances from center to points on the perimeter), (2) texture (standard deviation of gray-scale values), (3) perimeter, (4) area, (5) smoothness (local variation in radius lengths), (6) compactness (perimeter² / area - 1.0), (7) concavity (severity of concave portions of the contour), (8) concave points (number of concave portions of the contour), (9) symmetry and (10) fractal dimension (“coastline approximation” - 1) [10].

The dataset dimensions are 569 x 33 which is 569 data entries (rows) and 33 attributes (features) out of which 357 are Benign (non-cancerous) and 212 are Malignant (cancerous).

C. Data Pre-processing

The original dataset has 33 features out of which the first (id) and last (unnamed) contributes nothing to our result (whether Benign or Malignant?), hence these two columns are dropped (removed) from the dataset.

One of the features of the dataset taken is ‘diagnosis’ which is of the type character and the general rule of thumb is not to feed any data which not of numeric type to Machine Learning (ML) model, hence this column is converted into numeric type using LabelEncoder().fit_transform() of scikit-learn.

To avoid inappropriate assignment of relevance, the dataset was standardized using the following equation -

$$z = (x - u) / s$$

where u is the mean of the feature to be standardized and s is the standard deviation of those features. The standardization was implemented using StandardScaler().fit_transform() for training data and StandardScaler().transform() for testing data. The standardization is done after splitting the train-test data in order to avoid data leaks from testing data to training data.

To avoid majority bias in our model, the dataset was resampled using the imblearn library. The resampling was implemented using SMOTE, KMeansSMOTE and SMOTETomek. The resampling is done after splitting the train-test data in order to avoid data leaks from testing data to training data.

D. Feature Selection

1) **Benign vs Malignant cells:** The dataset contains 569 records out of which 357 (62.7%) are Benign (non-cancerous) and the remaining 212 (37.3%) are Malignant (cancerous). The following figure shows the comparison of Benign and Malignant cells in our data.

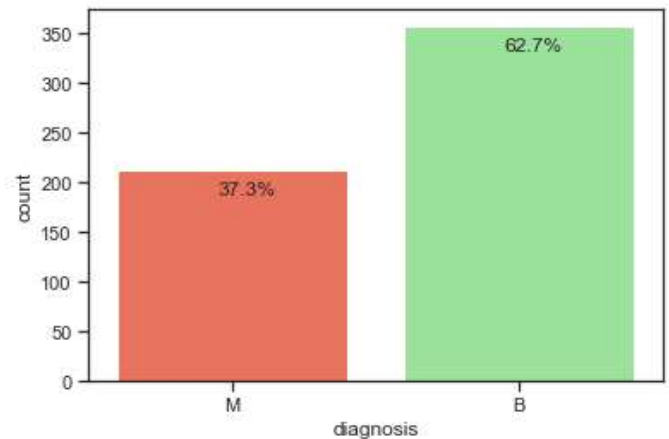
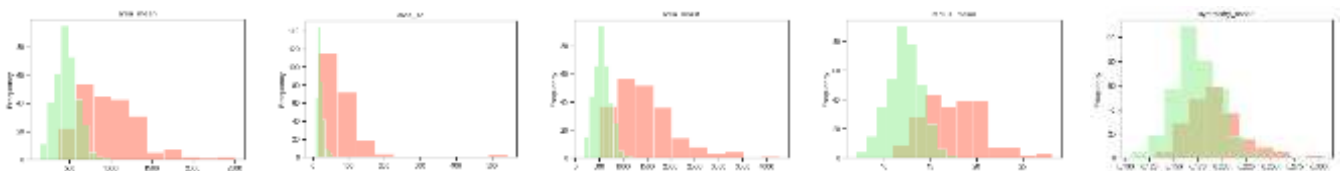


Fig. 1 Benign vs Malignant cells

2) **Histograms:** Histograms provide a great way to explore and study single variables at a time. It represents the distribution of a continuous variable over a given interval or period of time. It is used to inspect the underlying frequency distribution (eg. Normal distribution), outliers, skewness, etc.



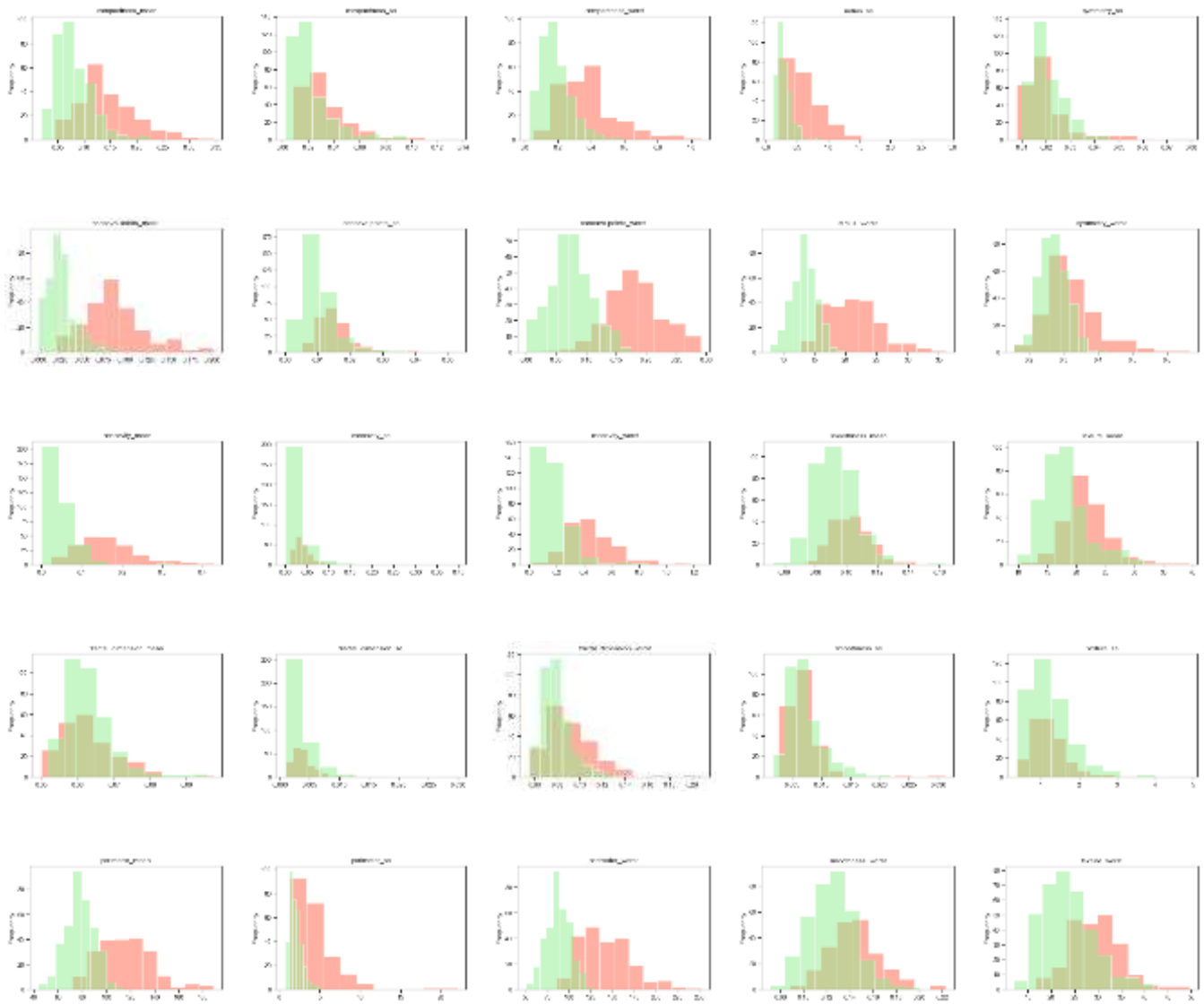


Fig. 2 Histograms for all features of dataset

3) **Pair Plot:** A pairs plot (also called a scatterplot matrix) allows us to see both distribution of single variables (kernel density estimate) and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis.

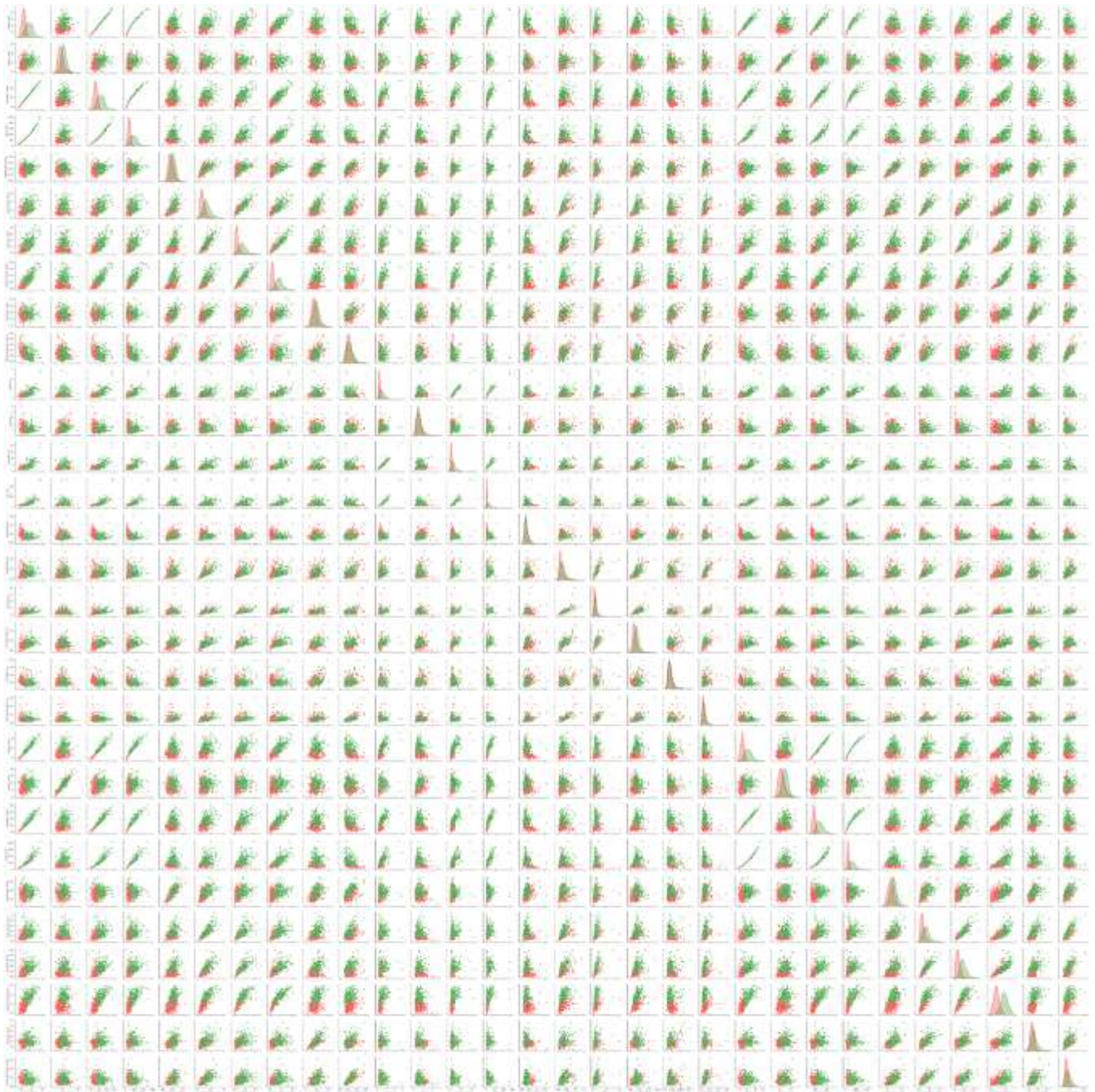


Fig. 3 Pair Plot for all features of dataset

4) **Correlation Heatmap:** Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. Higher the correlation value between a dependent variable and an independent variable, higher the significance of that independent variable in determining the dependent variable. The correlation value can

be positive, negative and zero depending on the direction of the change.

A high correlation between dependent and independent variables is desired whereas the high correlation between two independent variables is undesired. Highly correlated pairs of independent variables are considered as redundant, wasting time and space.

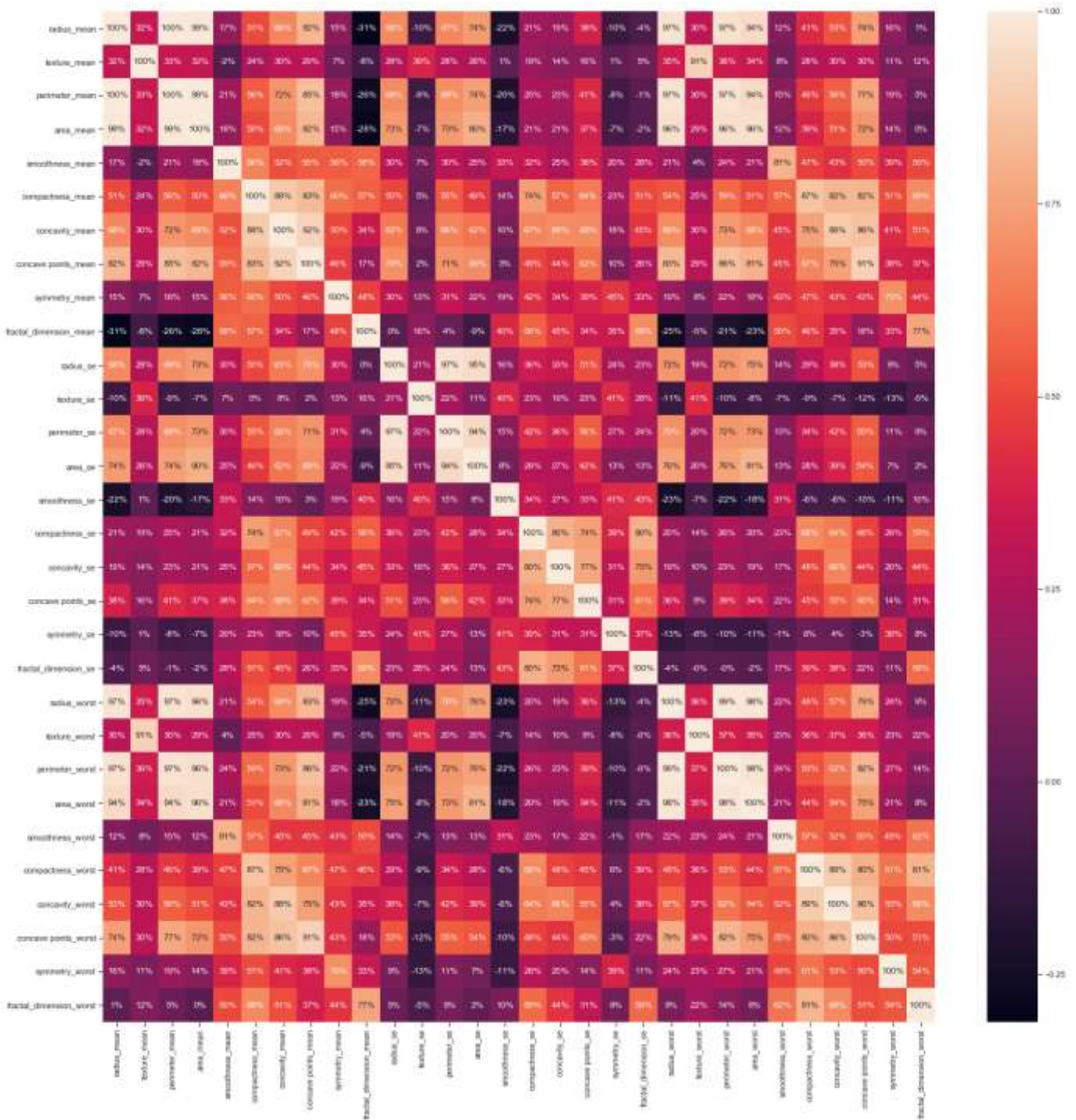


Fig. 4 Heatmap representing correlation between all dependent and independent variable

E. Machine Learning Algorithms

1) **Logistic Regression:** Logistic regression is a supervised learning algorithm which is used for classification. This algorithm predicts the probability of a dichotomous target (dependent) variable. It gives probabilistic values which lie between 0 and 1. Based on these probabilistic values the algorithm does the

classification. Logistic Regression is very similar to the Linear Regression, the only difference is the hypothesis function where the former uses a sigmoid function ($0 \leq \text{output} \leq 1$) while the later uses a regression function ($-\infty \leq \text{output} \leq \infty$); thereby different uses where the former is used for classification and the later is used for regression. Fig.5 is the Confusion Matrix generated on the results obtained from the Logistic Regression Model.

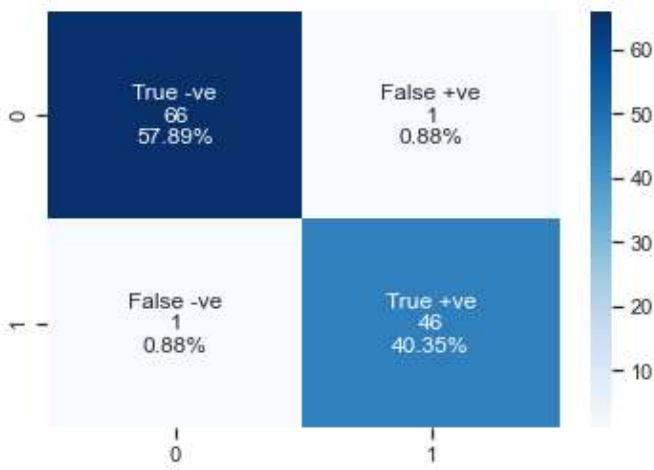


Fig. 5 Confusion Matrix on results from Logistic Regression

2) **K-Nearest Neighbor:** K-Nearest Neighbor is a supervised machine learning algorithm which is used for classification. This algorithm stores all the training data and classifies new data points on the basis of similarity of the nearest (with least Euclidean distance) k training data points. It is a non-parametric algorithm i.e. it does not make any assumption on underlying data. The algorithm does not immediately start learning from the training data provided rather it sits on training data and when some new data is provided for classification, then it performs required actions on the dataset; thereby it is also called a lazy learning algorithm. Fig.6 is the Confusion Matrix generated on the results obtained from the K-Nearest Neighbor Model.

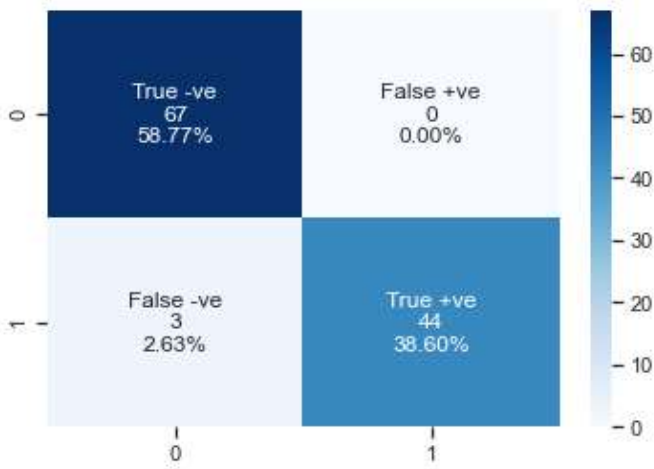


Fig. 6 Confusion Matrix on results from K-Nearest Neighbor

3) **Support Vector Machine:** Support Vector Machine is another supervised machine learning algorithm which is used for classification. This algorithm divides the dataset into different classes using a hyperplane such that there is maximum margin for better accuracy. Where a hyperplane can be described as a function which best separates the classes and margin can be described as the distance between the two closest data points of different classes to the

hyperplane. Fig.7 is the Confusion Matrix generated on the results obtained from the Support Vector Machine Model.

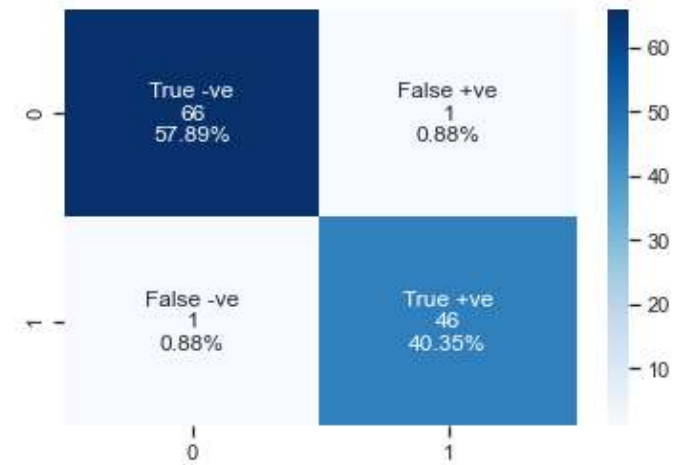


Fig. 7 Confusion Matrix on results from Support Vector Machine

4) **Naive Bayes Classifier:** Naive Bayes Classifier is a supervised machine learning algorithm which is used for classification. This algorithm is based on Bayes Theorem which is used for calculating the probabilities of the objects, thus it is also a probabilistic classifier like Logistic Regression. The algorithm is called Naive Bayes Classifier because the algorithm assumes that the presence of any feature in a class is independent (unrelated) to the presence of any other feature of that class. Fig.8 is the Confusion Matrix generated on the results obtained from the Naive Bayes Classifier Model.

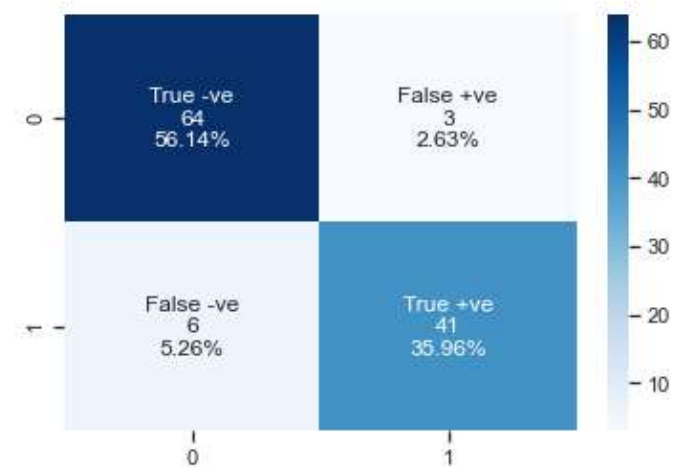


Fig. 8 Confusion Matrix on results from Naive Bayes Classifier

5) **Decision Tree Classifier:** Decision Tree Classifier is another supervised machine learning algorithm which is used for classification. This algorithm creates a tree structure for classifying different classes called a decision tree, where internal (decision) nodes represent the features of a dataset used to make any decision, branches represent the decision rules and each leaf node represents the outcome (class). A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees. Fig.9

is the Confusion Matrix generated on the results obtained from the Decision Tree Classifier Model.

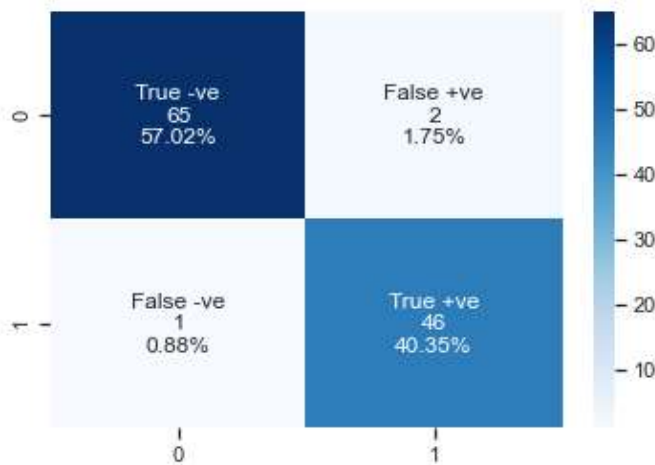


Fig. 9 Confusion Matrix on results from Decision Tree Classifier

6) Random Forest Classifier: Random Forest Classifier is another supervised machine learning algorithm which is used for classification. This algorithm, as the name suggests, combines a number of Decision Tree Classifiers on different subsets of the provided dataset and takes the average of their outcomes to predict a single outcome which improves the predictive accuracy of the algorithm. The greater the number of trees, the greater is the accuracy and lesser is the chance of overfitting. This concept is called ensemble learning. Fig.10 is the Confusion Matrix generated on the results obtained from the Random Forest Classifier Model.

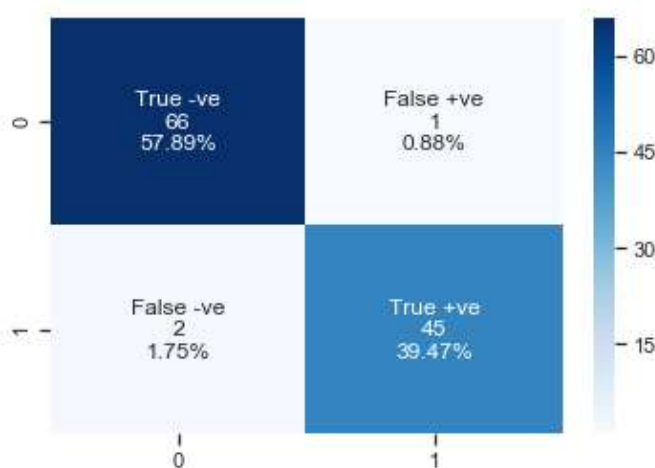


Fig. 10 Confusion Matrix on results from Random Forest Classifier

F. Performance Metrics

1) Confusion Matrix: Confusion Matrix is a 2x2 matrix which contains TN, FP, FN and TP at positions 11, 12, 21 and 22 respectively. It categorizes the outputs predicted by the classification model into 4 categories viz. TN = True Negative (Correctly Rejected), FP = False Positive (Incorrectly Accepted), FN = False Negative

(Incorrectly Rejected) and TP = True Positive (Correctly Accepted).

2) Accuracy: Accuracy is defined as the proportion of all the predictions that are accurately (correctly) predicted.

$$\text{Accuracy (Acc)} = (TP+TN) / (TP+TN+FP+FN)$$

3) Sensitivity: Sensitivity is defined as the proportion of actual positive cases that are predicted as positive. It is also called ‘Recall’.

$$\text{Sensitivity (Recall)} = (TP) / (TP+FN)$$

4) Specificity: Specificity is defined as the proportion of actual negative cases that are predicted as negative.

$$\text{Specificity} = (TN) / (TN+FP)$$

5) False Discovery Rate: False Discovery Rate is defined as the proportion of actual negative cases that are predicted as positive. It is also called ‘False Positive Rate’.

$$\text{False Discovery Rate (False Positive Rate)} = (FP) / (TN+FP).$$

The sum of Specificity and False Discovery Rate is 1.

6) False Omission Rate: False Omission Rate is defined as the proportion of actual positive cases that are predicted as negative. It is also called ‘False Negative Rate’.

$$\text{False Omission Rate (False Negative Rate)} = (FN) / (TP+FN)$$

The sum of Sensitivity and False Omission Rate is 1.

7) Matthews Correlation Coefficient: Matthews Correlation Coefficient is actually the correlation coefficient of test data and predicted data. The higher the value of Matthews Correlation Coefficient, the better the prediction. MCC is also perfectly symmetric in nature.

$$\text{MCC} = (TP \times TN - FP \times FN) / \{(TP+FP)(TP+FN)(TN+FP)(TN+FN)\}^{1/2}$$

The value of MCC is always between -1 and 1, where 1 is perfect classifier, -1 always misclassifies and 0 is pure random classification.

8) AUC-ROC Curve: The Receiver Operator Characteristic (ROC) curve is the plot of TPR vs FPR at various threshold values. Area Under the Curve (AUC) is the measure of separability which depicts how well a model can classify classes successfully. Higher the AUC, the more classes are correctly predicted.

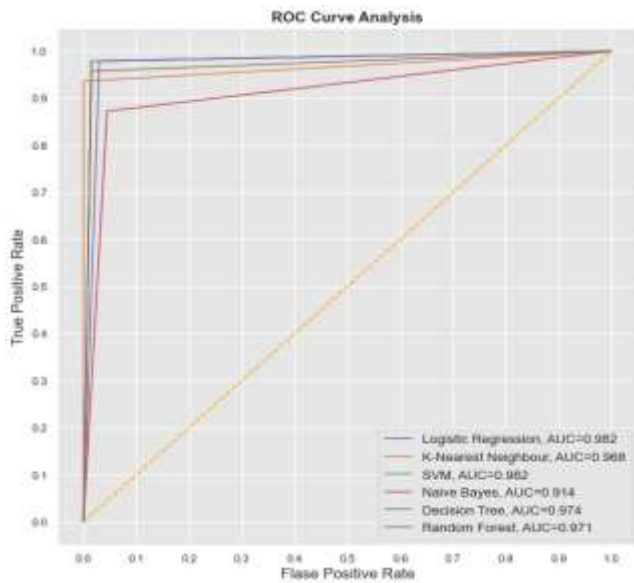


Fig. 11 ROC Curve Analysis

IV. RESULTS

This project is conducted on a laptop with Intel Core i7-8750H CPU, 16GB DDR4 Ram, and Nvidia GTX 1050Ti GPU. Table.1 summarizes the result.

TABLE I

PERFORMANCE METRICS OF DIFFERENT MACHINE LEARNING ALGORITHMS

Model	Accuracy	Sensitivity	Specificity	False Omission Rate	False Discovery Rate	MCC	AUC
Logistic Regression	98.245%	97.872%	98.507%	2.127%	1.492%	96.379%	98.2%
K-Nearest Neighbor	97.368%	93.617%	100%	6.382%	0%	91.189%	96.8%
Support Vector Machine	98.245%	97.872%	98.507%	2.127%	1.492%	96.379%	98.2%
Naive Bayes	92.105%	87.234%	95.522%	12.765%	4.477%	83.678%	91.4%
Decision Tree	97.368%	97.872%	97.014%	2.127%	2.985%	94.602%	97.4%
Random Forest	97.368%	95.744%	98.507%	4.255%	1.492%	94.567%	97.1%

V. CONCLUSIONS

The paper presents a comparative study of different machine learning algorithms for the detection of breast cancer. This performance comparison is done on the Breast Cancer Wisconsin (Diagnostic) Dataset created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA [10]. The algorithms studied are Logistic Regression, K-Nearest Neighbor, Support Vector

Machine, Naive Bayes Classifier, Decision Tree Classifier and Random Forest Classifier with an average accuracy of 96.783%. It is observed that the Support Vector Machine and Logistic Regression are the most efficient algorithms when it comes to accuracy, sensitivity, specificity, false omission rate, false discovery rate, Matthew’s correlation coefficient and area under curve over other algorithms studied here.

ACKNOWLEDGMENT

Prof. A.W. Bhade, Head of Department of Information Technology, Government College of Engineering, Amravati, has been a tremendous source of support and guidance to the authors throughout this project.

REFERENCES

- [1] "Latest Statistics of Breast Cancer in India." <https://www.breastcancerindia.net/statistics/trends.html> (accessed Feb. 2, 2021).
- [2] "BREAST CANCER INDIA." <https://www.breastcancerindia.net/> (accessed Feb. 2, 2021).
- [3] "Scikit-Learn User Guide." https://scikit-learn.org/stable/user_guide.html (accessed Feb. 1, 2021).
- [4] "NumPy User Guide." <https://numpy.org/doc/stable/user/index.html> (accessed Feb. 1, 2021).
- [5] "Matplotlib Overview." <https://matplotlib.org/stable/contents.html> (accessed Feb. 1, 2021).
- [6] "Pandas User Guide." https://pandas.pydata.org/docs/user_guide/index.html#user-guide (accessed Feb. 1, 2021).
- [7] "Imblearn User Guide." https://imbalanced-learn.org/stable/user_guide.html#user-guide (accessed Feb. 1, 2021).
- [8] "Seaborn User Guide and Tutorial." <https://seaborn.pydata.org/tutorial.html> (accessed Feb. 1, 2021).
- [9] "Documentation." <https://jupyter.org/documentation> (accessed Feb. 1, 2021).
- [10] William H Wolberg, W Nick Street, and Olvi L Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Dataset". 1992. Distributed by Kaggle. <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>
- [11] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, United Arab Emirates, 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818560.
- [12] M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, Bangladesh, 2017, pp. 226-229, doi: 10.1109/R10-HTC.2017.8288944.
- [13] Anji Reddy Vaka, Badal Soni, Sudheer Reddy K., "Breast cancer detection by leveraging Machine Learning", ICT Express, Volume 6, Issue 4, 2020, Pages 320-324, ISSN 2405-9595, <https://doi.org/10.1016/j.ict.2020.04.009> (<https://www.sciencedirect.com/science/article/pii/S2405959520300801>)
- [14] S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.
- [15] P. P. Sengar, M. J. Gaikwad and A. S. Nagdive, "Comparative Study of Machine Learning Algorithms for Breast Cancer Prediction," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 796-801, doi: 10.1109/ICSSIT48917.2020.9214267.
- [16] 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. https://doi.org/10.1007/978-981-15-7205-0_10
- [17] Abien Fred M. Agarap. 2018. 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 (ICMLSC '18). Association for Computing Machinery, New York, NY, USA, 5–9. DOI: <https://doi.org/10.1145/3184066.31>