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

Selecting the Best Machine Learning Methods for Breast Cancer Risk Prediction

[1]Ambika L G, [2] Dr. T N Anitha, [3]Dr. Jayasudha K

[1]Assistent Professor, [2]Professor, [3]Associate Professor

[1]Information Science and Engineering,

[1]SJC Institute of Technology, Chickballapur, Karnataka.

Abstract : Every year, the number of people who die from breast cancer rises. It is the most frequent type of cancer in women and the top cause of death worldwide. For a healthy person, advancements in the identification and prediction of malignant illnesses are critical. To update patient treatment prospects and survival standards, high accuracy in predicting the growth of malignancies is required. ML approaches have been exhibited to significantly affect the most common way of screening bosom disease early finding and forecast. The Wisconsin Breast Cancer Diagnostic Dataset utilized six Machine learning calculations: Support Vector Machine (SVM), Random Forest, Logistic Regression, Naive Bayes (NB), Decision Tree (C4.5) and KNearest Neighbor (KNN). We performed a performance evaluation and comparison between these different classifiers after receiving the results. In the wake of obtain the outcomes, we played out an exhibition assessment and an examination between these various classifiers. Support vector machines have been found to outperform all other classifiers and achieve the highest accuracy of (97.3%).

Keywords: Breast cancer; Prediction; Diagnostic; SVM; NB; Logistic regression; C4.5; k-NN; Classification; Effectiveness; Accuracy; Precision.

I. INTRODUCTION

India is experiencing a bosom cancer epidemic, with an increasing number of young women being exposed to the disease. According to studies, India is expected to see around 1,70,000 additional instances of breast cancer by 2020. According to studies, one out of every 28 women will be impacted by the condition. While breast cancer affects almost mainly women, about 1-2 percent of men are also at risk. Breast cancer is a regular occurrence in both men and women. A breast tumour is an uncontrollable alteration in the cells of the breast.

Non-cancerous and malignant tumours are the two types of tumours [1]. Measurements by the International Agency for Research on Cancer (IARC) show that bosom expansion is the resulting cause of illness spreading at a rate of 198000 fatalities in the most urbanized areas. Breast most cancers is the maximum extreme malignancy that could strike a woman. A tumour takes place while cells with inside the bosom tissue cut up and multiply without the normal controls on mobileular dying and division. In this approach, most cancers of the bosom tissue is called bosom malignancy. Furthermore, approximately 10% of all ladies at numerous factors in their lifestyles discover this annoying [2].

Characteristics like age, family ancestry, and acquired risk are among the determinants. Notwithstanding the way that bosom malignant growth is the second most prominent reason for disease passing in ladies, the endurance rate is higher. Because of their early determination, 97 percent of women live for more than 5 years [3].Most of ladies have more than one realized risk factor for chest development, yet they won't ever get the infection. No matter how you look at it, being a woman and producing more experienced people is not merely the most.

In light of their extraordinary presentation in determining, diagnosing illnesses, reducing drug costs, and making ongoing decisions to save individuals' lives, data mining calculations utilized in the medical services industry are significant. The most common information mining displaying points are grouping and expectation, which utilize a scope of calculations for bosom malignant growth expectation. As indicated by the examination

local area, the most persuasive information mining calculations incorporate SVM, Logistic Regression, and Naive Bayes (NB), Decision tree (C4.5), RF and KNN [6].

Our goal is to predict and analyze malignant growth of the breast using different machine learning algorithms, and which classifications are based on how the confusion matrix, sensitivity and accuracy of each classifier are represented. It is to choose whether the vessel is the best. The rest of this document is organized as follows: Part 2 describes the related work of past breast cancer diagnostic studies. Part 3 describes the recommended method. Part 4 describes the results of the experiment. Work ends in Part 5.

II. Related Works

Many researchers have used different datasets to study breast cancer, including SEER datasets, mammography photographs as datasets, Wisconsin datasets, and facts from different hospitals. The author extracts and selects a large number of items from these datasets to complete the study. These are some important points. S Nayak [5] shows how to use 3D images to classify breast cancer using various supervised ML algorithms and concludes that SVM is ideal in view of in general execution. BM Gayathri [7] is dealing with a correlation concentrate on utilizing importance vector machines, which are less computationally costly than other AI calculations for recognizing bosom disease, and why RVM is better. Other AI strategies for diagnosing bosom disease with 97 %, in any event, when elements are limited, are made sense of. Hiba Asri [8] showed the best results in that the support vector machine (SVM) succeeded in predicting and diagnosing breast cancer with 97.13% accuracy and low accuracy and error rate. Y.khoudfi and M. Bahaj [9] as of late proposed a correlation of AI calculations that SVM is the best classifier with 97.9% exactness compared to ANN, RF, NB and is based on a 5-layer multilayer perceptron. Ahmed Hamza Osman [11] provided a diagnostic strategy for breast cancer in Wisconsin by combining a clustering approach with an effective probabilistic vector support machine (WBCD). The SVM method gave a 99.10% prediction. Our exploration centers on assessing calculations and ML strategies to decide the best techniques for distinguishing and anticipating bosom disease.

III. Methodology

Our review's fundamental objective is to track down a solid and prescient strategy for identifying bosom disease. On the Breast Cancer Wisconsin (Diagnostic) dataset, we utilized ML classifiers, for example, SVM, Random Forests, Logistic Regression, Naive Bayes, Decision tree (C4.5), and KNN and assessed the outcomes to figure out which model gives the best outcomes. Figure 1 depicts the suggested architecture.

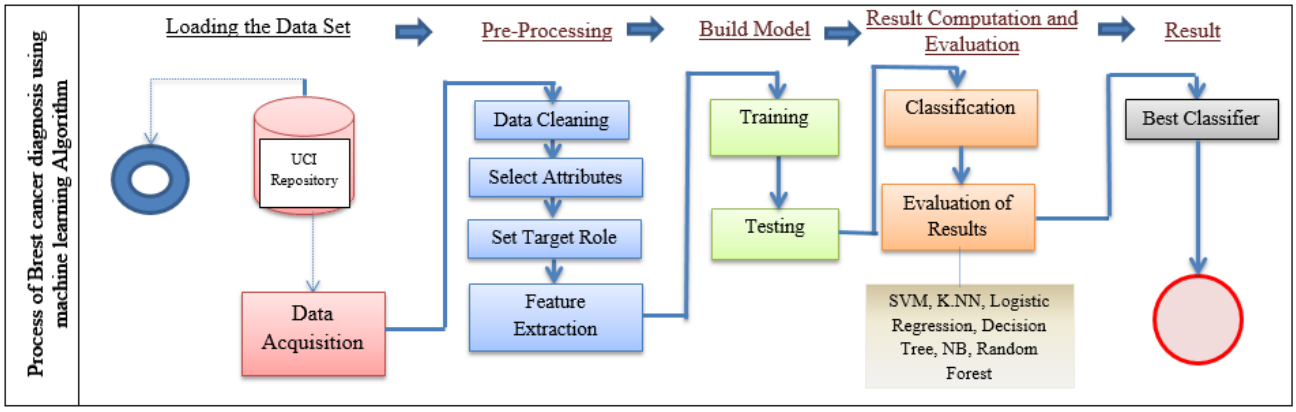

Fig. 1. Process Flow Diagram.

Following data accretion, The flow diagram consists of 5 steps: Loading the data set:UCI repository data, In preprocessing contain 4 steps: data cleansing, attribute selection, target Role selection, and feature extraction ,Build Model contained two steps: The 75% training data and 25% of testing data of the total data, is used to build our ML model. Following the testing of the models, we look at the outcomes and pick the computation that conveys the most elevated accuracy and perceive the most prescient calculation for bosom disease screening.

3.1. Dataset acquisition

The BC Wisconsin (Diagnostic) dataset was used in this investigation. Features are processed from a digitized picture of a fine needle suction (FNA) of a bosom mass. They portray qualities of the cell cores present in the picture. These characteristics determine the quality of the cell nuclei in the image. There are 569 occurrences (357 benign and 212 malignant), two classifications (63 percent benign and 37 percent malignant), and the attributes of Bosom Cancer Wisconsin Diagnostic.

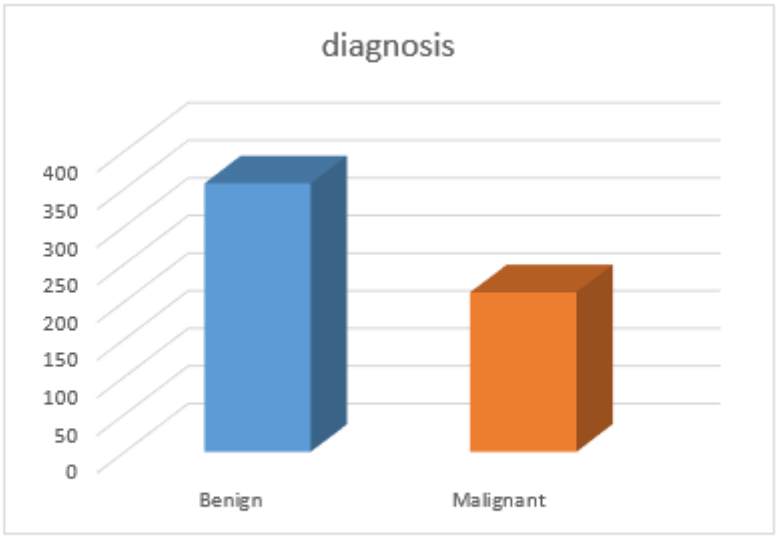

Fig. 2. Wisconsin Breast Cancer Diagnostic Datasets.

IV. Results and Discussion

AI Algorithms were utilized to investigate the Breast Cancer Wisconsin Diagnostic dataset. To survey and think about the models and recognize the best calculation for bosom disease expectation, we utilized Confusion Matrix, Accuracy, Precision, Sensitivity, Recall, and F1 Score as execution markers. The Confusion Matrix is a technique for assessing the exhibition of a grouping task with at least two sorts of result. A confusion matrix is a table having two dimensions: "Actual" and "Predicted," as well as "True Positives (TP), "True Negatives (TN)," "False Positives (FP)," and "False Negatives (FN)," and "False Negatives (FN)." The most often utilized execution metric is the accuracy of classification algorithms. The quantity of right reports returned by our AI model can be depicted as accuracy in record recoveries. The quantity of up-sides returned by your AI model is alluded to as responsiveness. We can utilize review to sort out the number of positive examples the ML that model accurately identified. The F1 score decides the symphonious mean of accuracy and responsiveness. Table 1 and Figure 2 showcase the accuracy rates for Wisconsin Breast Cancer Diagnostic datasets. Each of the classifiers in the preparation and testing sets have shifting exactness's, however SVM reliably outflanks different classifiers in the testing set (97.3 %).

Table 1. Exactness rate for breast cancer diagnostic dataset.

Algorithms	Accuracy Training Set (%)	Accuracy Testing Set (%)
SVM	98.2%	97.3%
Radom Forest	99.4%	96.8%
Logistic Regression	95.2%	95.7%
Naive Bayes	95.8%	95.9%
Decision Tree	98.9%	95.3%
K-NN	94.8%	93.9%

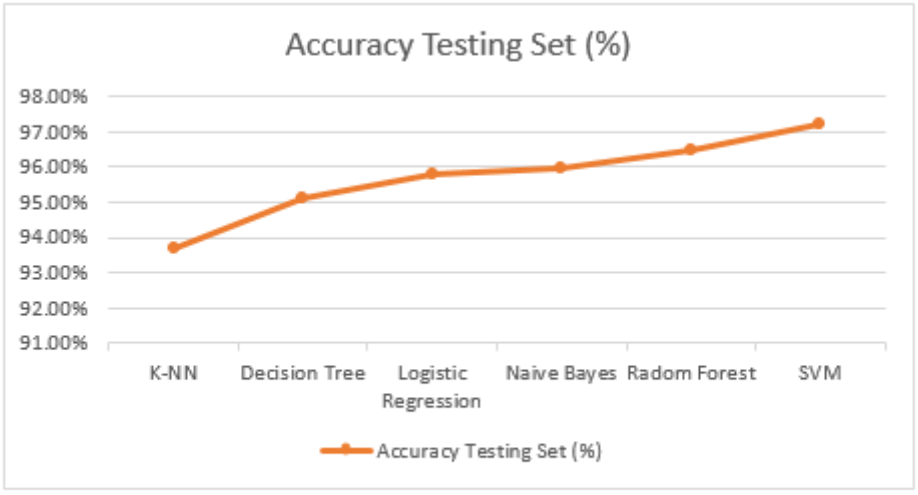

Fig. 3. Comparative graph of different classifiers

Figure 4 shows a heatmap demonstrating the correlation between the attributes of the WBCD dataset. The connection heatmap shows a two-layered relationship lattice with two discrete aspects, where the main aspect esteem addresses a line and the subsequent aspect esteem addresses a section. This heatmap utilizes shaded pixels on a monochrome scale to portray the subsequent relationship between the dataset's properties. The affiliation develops further as the variety power rises. The amount of measurements that match the dimensional values determines the colour value of the cells. The linearity between the two elements decides the layered worth (- 1 to +1). A positive relationship exists when the two factors change and move in a similar heading.

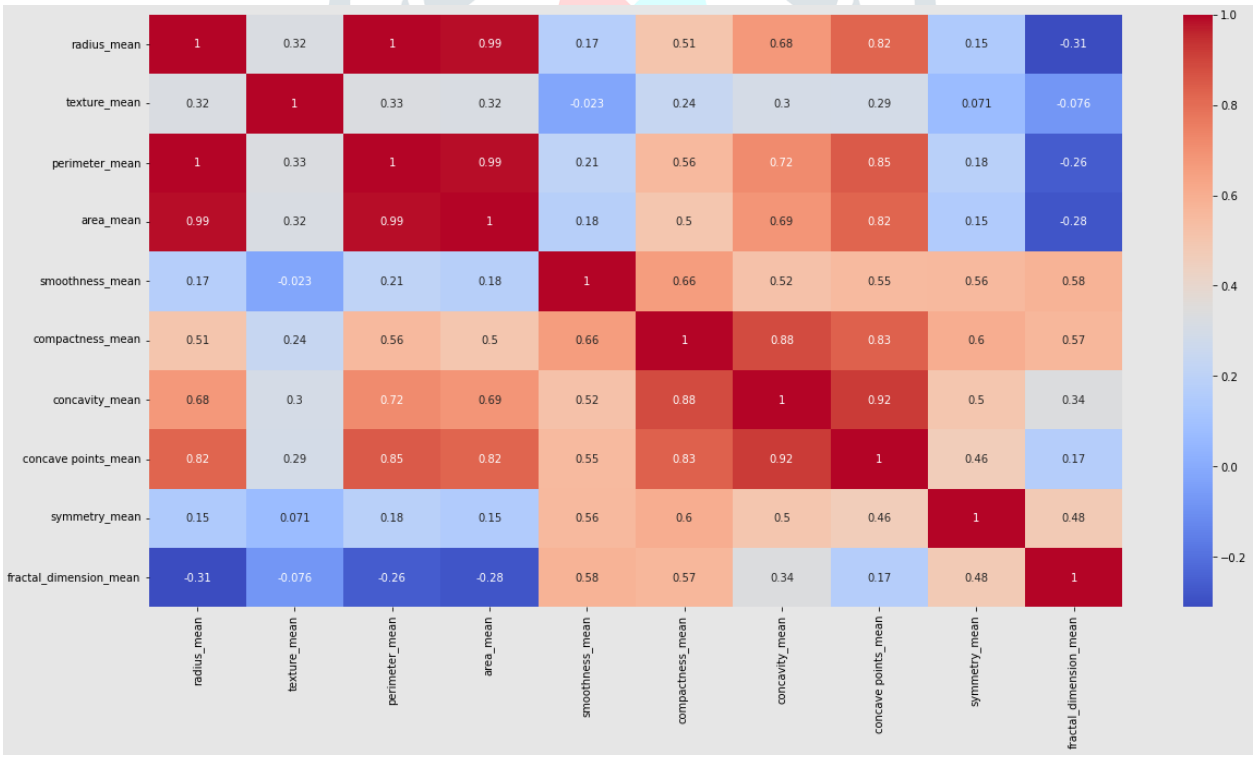

Fig. 4. Correlation between the different features

Table 2 represents the Confusion Matrix rates in each class, and Table 3 displays the obtained Classifiers performance measures for benign and malignant. As shown in Table 2, the SVM accurately predicts 556 cases out of 569, with 201 dangerous cases that are really harmful and 356 harmless cases that are truly harmless, and 11 cases that are inaccurately anticipated, with 11 threatening cases anticipated as harmless and 1 harmless case anticipated as dangerous. As a result, in terms of accuracy, Support Vector Machine beats other categorization techniques

Table 2. Confusion Matrix

	Malignant	Benign	
SVM	201	11	Malignant
	1	356	Benign
Radom Forest	196	16	Malignant
	7	350	Benign
Logistic Regression	201	11	Malignant
	5	352	Benign
Naive Bayes	197	15	Malignant
	7	350	Benign
Decision Tree	195	17	Malignant
	22	335	Benign
K-NN	201	11	Malignant
	7	350	Benign

Table 3. Classifiers performances

Algorithms	Precision	Sensitivity	Recall	F-Measure	Class
SVM	0.98	0.94	0.97	0.96	Benign
	0.97	0.99	0.96	0.98	Malignant
Radom Forest	0.96	0.94	0.95	0.95	Benign
	0.97	0.98	0.96	0.97	Malignant
Logistic regression	0.98	0.91	0.98	0.94	Benign
	0.95	0.99	0.97	0.97	Malignant
Naive Bayes	0.98	0.96	0.95	0.96	Benign
	0.98	0.97	0.97	0.94	Malignant
Decision Tree	0.94	0.92	0.95	0.93	Benign
	0.96	0.97	0.96	0.96	Malignant
K-NN	0.92	0.91	0.97	0.91	Benign
	0.95	0.96	0.91	0.95	Malignant

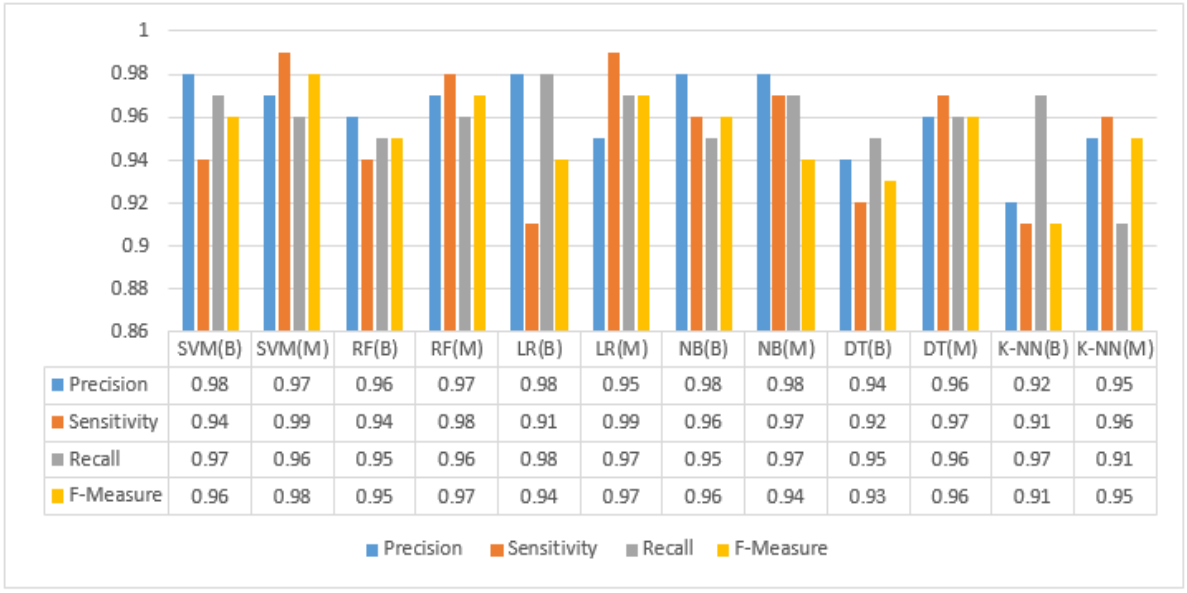

Fig. 5. Comparative diagram of machine learning algorithms with respect to Classifiers performances

The precision 0.98 percent, sensitivity 0.94 percent, recall 0.97 percent, and FMeasure 0.96 percent of SVM are greater than those of other classifiers, as shown in the table. In the bosom Wisconsin Diagnostic dataset disease, SVM reliably outperforms different classifiers as far as execution for the two classes harmful and harmless. A comparison of ML algorithms in terms of classification performance is shown in Figure 5.

V. CONCLUSION

On the Wisconsin Breast Cancer Diagnostic dataset (WBCD), we used six different algorithms to calculate, compare, and evaluate different results based on confusion matrix, accuracy, sensitivity, recall, and precision to find the best ML algorithm that is precise, reliable, and finds the highest accuracy. We determined that SVM exceeds all other methods with 97.3 % efficiency and 98 % precision after a thorough evaluation of our models. Finally, Support Vector Machine has been demonstrated to be powerful in the expectation and finding of bosom disease, with the best exactness and accuracy.

REFERENCES

[1] Lydia D Isaac1, Sureshkumar C2, "Diagnosis Prognosis and Prevention of Breast Cancer Based on Present Scenario of Human Life ", 2018 (ICCICT), Feb. 2-3, Mumbai, India, 2018.

[2] Sapiah binti sakri, Nuraini binti abdul rashid,"Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction" IEEE June 20, 2018. 2843443VOLUME 6, 2018,

[3] Daniel Duong, Morgan George, Brendan Abraham, "Using Patient-Generated Health Data to Facilitate Preoperative Decision Making for Breast Cancer Patients", 978-1-5386-1848- 6/17/$31.00 ©2017 IEEE ,2017.

[4] Varalatchoumy.M 1, Ravishankar.M 2," Comparative Study of Four Novel Approaches Developed for Early Detection of Breast Cancer and its Stages", Proceedings of the ICICI 2017, IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9,2017.

[5] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.

[6] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.

[7] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[8] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, 'Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[9] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 978-1-5386- 42252/18/$31.00 ©2018 IEEE.

[10] L. Latchoumi, T. P., & Parthiban, "Abnormality detection using weighed particle swarm optimization and smooth support vector machine," Biomed. Res., vol. 28, no. 11, pp. 4749–4751, 2017.

[11] A. H. Osman, "An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique," Int. J. Adv. Compute. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017.

[12] Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. Doi: 10.1038/nbt1206-1565.

[13] Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.

[14] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag; 2001.

[15]Quinlan JR. C4.5: Programs for Machine Learning. 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1.

[14] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set."

[15] 'WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[16] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.

[17] "www.indiancancersociety.org/breast-cancer/" breast cancer details.

[18] J. Sivapriya, A. Kumar, S. Siddarth Sai, and S. Sriram, "Breast cancer prediction using machine learning," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, 2019.

[19] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). IEEE, 2018, pp. 1-5.

[20] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, "Developing a web based system for breast cancer prediction using xgboost classifier," International Journal of Engineering Research Technology (IJERT), vol. 9, 2020